Mining textual data for simplified reading

If students can find texts that are adapted to their reading abilities, the information becomes more accessible, and easier for them to transform to knowledge. Students with reading problems get easier texts whereas students with no reading problems get more advanced texts. Individually adapted simplifications of the texts will further support the students' learning, and stimulate their interest in reading, increase their actual reading efforts and facilitate source critique, as they understand the texts better.

In the project we will develop sophisticated measures for assessing a student's reading ability and a tool for the student and teacher to create a profile of this ability. 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 to a length that is adapted to the individual student and for automatic transformation to easy-to-read Swedish.

1 Introduction

A negative trend has been shown for Swedish students' reading ability, based on results from the last international investigation that Sweden has taken part in (Skolverket 2007, 2010). There is a decline in the proportion of strong readers among ten year old students. Among fifteen year old students this pattern is repeated, but in addition there is also a decline in reading ability among those who are the least strong readers. Thus, a shifting downwards is identified. At the same time 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). Even though the differences between students who do well and students who do less well are relatively small in Sweden in comparison to many other countries, it is possible to find everything from students who struggle with their reading to students who are very excellent readers in each school year and most classrooms. Consequently, it is a huge challenge for a teacher to find texts that are adequate and support learning for different groups of students in a classroom. The importance and teaching effect of finding the right level for each student, i.e. to find the 'zone of proximal development' (ZPD), is a well-known fact from both a theoretical and an empirical perspective (Vygotsky 1976).

The task of finding appropriate texts for different groups of students gets more demanding from grade 4 and onwards. To begin with they become longer and complex in their structure. Students who have a reading ability solely adjusted to more simple texts will get into problem. This phenomenon is often called the "4th grade slump". A little higher up in school around grade 7, the texts become more and more subject specific, which is especially visible in the vocabulary. Text structure and vocabulary are, thus, two very important aspects that can cause problems for students who are not so experienced readers. Their lack of experience can depend on that they are second language students at an early developmental language level and reading level in their second language. But it can also be first language students with restricted language and/or reading experiences (cf. "4th grade slump").

Both teachers and students would benefit tremendously if it were possible to test each student's language and reading ability and then suggest appropriate texts and, if needed, also shorten a text or simplify more advanced texts to a suitable level. Such testing needs tools if the teachers are to be able to assess each individual student.

2 Research objectives

The main aim of the project 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.

The project integrates three different research areas: studies of reading ability, measures of readability and development of techniques and tools for simplified reading.

Research objectives include:

- An understanding of how students reading motivations are affected by adapting texts to a student's individual reading abilities. Tools for assessing students' reading ability.
- New and more comprehensive measures of readability.
- A set of tools for text selection, automatic summarisation, and transformation to easy-to-read Swedish that are individually adapted to a student's reading abilities.

3 Selected theoretical background

The project's theoretical framework comprises three research areas: educational sciences, language technology and computer science.

3.1 Reading literacy – a broad theoretical perspective

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. Some studies of comprehension concern experiments where different aspects of the texts have been manipulated in order to understand the significance of these aspects.

In other studies interviews with individuals or group discussions are arranged in order to study how a text is perceived and responded to and how the reader moves within the text (e.g. Langer 2011; Liberg *et. al.* 2011). This last type of studies is very often based on a socio-interactionistic perspective. What is considered to be reading is, thus, extended to also include how you talk about a text when not being completely controlled by test items. In such a perspective Langer (2011 p. 22-23) has shown how students build their envisionments or mental text worlds when reading by *being out and stepping into an envisionment of the text content*, *being in and moving through such an envisionment*, *stepping out and rethinking what you know*, *stepping out and objectifying the experience*, and *leaving an envisionment and going beyond*. These so called stances are not linear and for a more developed reader they occur at various times in different patterns during the interaction between the reader and the text, i.e. the reader switches between reading e.g. "on the lines", "between the lines", and "outwards based on the lines".

In a socio-cultural perspective the focus is made even wider and reading is perceived as situated social practices. The term situated pinpoints that a person's reading ability varies in different situations and with different text types and topics. A

model of reading as social practice is proposed by Luke and Freebody (1999). They map four quite broad reading practices that they consider to be necessary and desirable for members of a Western contemporary society: *coding practices, text-meaning practices, pragmatic practices* and *critical practices*. The first two practices could be compared to what above is discussed as decoding, comprehension and reader-response. The last two practices on the other hand point to the consequences of the actual reading act, which at the same time is the raison d'etre of reading: there can be no reading without having a wider purpose than to read and comprehend. These practices concern on the one hand how to "use texts functionally" and on the other hand to "critically analyze and transform texts by acting on knowledge … that they represent particular points of views … and that their designs and discourses can be critiqued and redesigned in novel and hybrid ways" (ibid p 5-6). A person with a very developed reading ability embraces all these practices and can move between them without any problem. He/she is not only able to decode and comprehend the text but also able to use what has been generated from the text and to take a critical stance, all this in order to extend his/her knowledge sphere. All these perspectives taken together give both a very deep and a very broad understanding of the concept of reading. In order to mark this shift from a narrower to a much more widened concept the term, *reading literacy* is often preferred to reading (see e.g. OECD 2009 p. 23).

When assessing students reading ability the types of texts and reading practices tested have thus a much broader scope today than earlier. It facilitates a more delicate differentiation between levels of reading ability. Two well-tested and established studies of reading ability in the age span focused here are the international studies of ten year old students (PIRLS) and fifteen year old students (PISA). Both these studies are based on a broad theoretical view of reading, i.e. reading literacy. The frameworks of PIRLS and PISA concerning both the design of tests and the interpretation of results in reading ability levels will therefore be important sources and resources for constructing students' reading ability profile in this study (see e.g. Mullis *et. al.* 2009; OECD 2009).

3.2 Testing reading ability and creating profiles

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. In the construction of this test at least three degrees of linguistics difficulty will be used. Accordingly at least three prototypical texts will be chosen per school subject area. Items testing the following reading practices will be constructed for each of these three texts (cf. Mullis *et. al.* 2009 p. 23-29; OECD 2009 p. 34-44):

1. Retrieve explicitly stated information and make straightforward inferences (cf. Luke's and Freebody's text-meaning practices and Langer's first envisionment),
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. A multiple choice item format will be used in order to be able to score students' answers automatically. The reading practices will be scored on a three-point scale (cf. Liberg *et. al.* 2011). This test will be performed with students of different

ages within the age span 10 – 15 years old. The data regarding these students' reading ability will also be used to standardize the test.

As a result of this test student's reading ability profile will be generated automatically in form of a table showing his/her results concerning the three reading practices for texts of various degrees of linguistic difficulties in the chosen subject area. The profile will thus show with what type of reading practices a student is able to read a text of a certain difficulty degree. Based on the profile it should be able to choose texts that are more suitable for a student.

3.3 Texts and linguistic features

A wide range of readability measures exist, most of them originating from research in the period between 1920 and 1970. For Swedish, the standard readability index is LIX (Björnsson, 1968). It encodes the heuristic that the longer a word or a sentence is, the more difficult it is likely to be. Other readability metrics are OVIX (Hultman and Westman, 1977), which is an indicator of the lexical variation in a text, and the nominal ratio (NR), suggested to present the information density in a specific text passage (Melin and Lange, 2000). LIX and OVIX are easy to operationalize and the measures can be retrieved automatically on large text chunks. Computation of the NR requires a preprocessed text with part-of-speech annotation.

We will consider *global language measures* built upon lexical, morpho-syntactic and syntactic features of a given text. The complexity of lexical features can be captured by use of specific word lists, indicating frequency and dispersion of the specific items. One of the most robust findings in the word recognition literature is that frequency influences the efficiency with which units are processed. Lexical frequency profiles (LFP) were means designed by Laufer and Nation (1995) for measuring the active vocabulary of learners. The morpho-syntactic features regard the structure of specific word forms, and how they relate to the lexical base form – the lemma. Finally, the syntactic features contribute not only to the sentence complexity, but also to the overall text cohesion and coherence.

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 on the specific reader (Mühlenbock and Johansson Kokkinakis, 2009). The language properties of the *reader model* therefore also need to be considered. There are studies showing that a complex morphology seems to negatively influence the decoding process in "early" second language-learners. Persons with dyslexia are predominantly troubled by long words, while a text with complex syntax will be cumbersome for linguistically unpractised students. The text structure, e.g. breaks against cultural expectations, is a property that predominantly afflicts untrained readers. More generally, however, an individual's vocabulary knowledge seems to be a high predictor of reading comprehension.

Our measures of readability will consider linguistic features appearing at the surface in terms of raw text, but also at deeper language levels. For the latter task we are going to automatically process the text in four different steps: after pre-processing it will be annotated with part-of-speech and lemma information and finally it will be parsed with dependency annotation.

3.4 Techniques and tools for simplified reading

The work on techniques for simplified reading consists of several parts. The first task is to find the actual texts for a given situation. These texts should be adapted to each individual student's given background knowledge, reading ability profile, as measured above, and the situation at hand. Current work, see below, establishes a foundation

for experimenting with different automatic lexical metrics that can be used to classify texts into readability levels. When the texts have been found and a rough set of candidate texts have been selected, two important techniques for further simplification will be used, in order for the text to be further matched to the reader. These techniques include an interactive tool for automatic summarization of the texts from different genres and a set of rules for automatically transforming texts to easy-to-read Swedish.

3.4.1 Automatic text summarization

Automatic summarization can be done in various ways. A common distinction is extract versus abstract summaries. An extract summary is created by extracting the most important sentences from the original text. An abstract summary on the other hand is a summary where the text has been broken down and rebuilt as a complete rewrite to convey a general idea of the original text.

Furthermore, the summaries can be indicative (only providing keywords as central topics) or informative (content focused) (Firmin and Chrzanowski, 1999). The former might be more usable when a reader needs to decide whether or not the text is interesting to read and the latter when a reader more easily needs to get a grasp of the meaning of a text that is supposed to be read.

Extraction based summarizers are often based on the vector space model; a spatial representation of a word's meaning where every word in a given context occupies a specific point in the space and has a vector associated to it that can be used to define its meaning, e.g. HolSum (Hassel and Sjöbergh, 2007), SummaryStreet (Franzke *et. al.*, 2005) and CogSum (Smith and Jönsson, 2011a).

The vector space model is able to capture the meaning of words in terms of how similar their contexts are, that is, words that often occur with the same words are located close to each other in the vector space. This technique can for instance be used to tell if a certain sentence in a text "is about" the same thing as the document as a whole, that is, how the sentence is located in the vector space compared to the other sentences in the document. The closeness in space is measured as the cosine of the angle of their vectors. This way, the most important sentences can be extracted from a document, sentences that most resemble the document as a whole.

The summarizer used in the proposed project is vector space based using a space reduction technique called Random indexing. The summarizer, called CogSum, is written in Java and utilizes a Random Indexing toolkit available at Hassel (2011). Using Random Indexing the vectors are initialized in randomized directions (Sahlgren, 2005), but when more and more words are encountered, their vectors will be updated and the words will find their place in the space. The word representation in CogSum is incrementally created from a large training material that stabilizes the summariser and also provides better summaries (Smith and Jönsson, 2011b).

CogSum also uses the Weighted PageRank algorithm in conjunction to its Random Indexing-space to rank the sentences (Chatterjee and Mohan, 2007). A high similarity between vectors gives a low angle and tells us that the words have appeared in similar contexts and thus can be considered to have similar meaning.

Summaries can be used as a mean to make a text easier to read, as the text is shorter. It has also been shown that summaries provide texts that are easier to read based on common readability measures (Smith and Jönsson, 2011a). However, a summary also contains syntactic errors, such as missing antecedents to referring expressions (Kaspersson *et. al.* 2012), which needs to be handled in order to have the summary more readable.

3.4.2 Automatic transformation of text to easy-to-read Swedish

Today texts are manually transformed to easy-to-read Swedish; there are no techniques that automatically perform such transformations. Decker (2003) studied corpora of easy-to-read texts and normal texts. Her study resulted in 25 general transformation rules used to simplify a text syntactically. The rules can be grouped into two subsets of rules; 1) rules that remove or replace sub phrases and 2) rules that add new syntactical information to the text. An example of a rule from the first category is: np(det+ap+n) → np(n). This rule will replace any nominal phrase containing a determiner, an adjective phrase and a noun with a nominal phrase containing only the noun, which makes the text easier to read but at the same time also removes information from the text.

Other syntactic, or lexical, transformations are rather easy to do, for instance, to replace abbreviations with its extended form, which can be done, e.g., based on the list of abbreviations assembled by the Swedish Academy. Syntactic transformations need, however, to be applied with caution, otherwise there is a risk that too much meaning of a sentence is lost. In the proposed project we will use a tool for syntactic transformation called CogFlux (Rybing, Smith, and Silvervarg, 2010), which is based on Decker's (2003) transformation rules and allows for easy experimentation with various rules, and also for addition of new transformation rules when needed.

On the semantic level, the most important transformation is to change difficult words with their synonyms or hyponyms. This can be done, for instance, based on frequency, common words are assumed to be easier to understand, based on length, short words are assumed to be easier to understand, or a combination of the two where a difficult word is replaced if it is both long enough and has a high enough frequency. For the latter, the concept of "enough" needs to be quantified, and for length and frequency similar measures are needed. Replacing synonyms also needs to consider syntactic features, e.g. to ensure that synonyms have the same part of speech and inflection.

3.4.3 Combining techniques and individual adaptation

The techniques for automatic summarization and transformation to easy-to-read Swedish can be combined to achieve as easy-to-read texts as possible, e.g. transforming a summary to easy-to-read Swedish. Furthermore, transformations can be customized to fit the needs for specific reading ability, as discussed above.

The techniques lend themselves to individual adaptation thanks to high parameterisation. The vector space methodology is dependent on what kinds of texts that is used in training, how contexts are defined, what contexts to consider important and when, to name a few. The text simplification part is specified by what rewriting rules should be used and when. The text search is based on lexical measures that are subject to investigation. The metrics need, further, to be weighted and combined according to the reader's proficiency.

4 Work plan

The research in the project will be conducted in three parallel, but closely connected, parts:

1. Testing students' reading ability for a specific subject area. Selecting texts adapted to the student's reading ability, testing the student's vocabulary knowledge for the texts chosen in order to refine the search for an

appropriate text. This includes developing the actual tests, frameworks and the computer tool for creating readability profiles.
2. Develop new models for readability. This includes using the profiles and subject areas from 1) to assess the readability models.
3. Using the measures from 2) to find texts of suitable difficulty. Simplify the texts as necessary using a combination of summarization and transformations based on the students' readability profile, and finally evaluate this with students.

The language technology tools for selecting texts, automatic transformation to easy-to-read Swedish, and summarization will be developed iteratively and incrementally as the linguistic measures and studies of reading ability evolve. Each year of the project will include four phases:

1. Select representative texts for the subject area.
2. Conduct tests, using the texts from 1, to create readability profiles for students of a certain age.
3. 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.

Three subject areas will be investigated in the project, selected in order with increasing reading difficulty, based on language as well as content. The first year fiction texts will be processed. The content of such texts are often rather easy to understand and the vocabulary comprises few unknown words. There are also fewer alternative texts, which means that ample time can be given to the development of the language technology tools, including the tool for creating readability profiles. The second year will use texts from the social sciences, texts that resemble fiction with respect to content, but with a more difficult vocabulary. The second year allows us to refine the readability profiles and the language technology tools to cater for more varied texts. The third year will use texts from the natural sciences. Such texts are considered to be the most difficult with many previously unknown and also more abstract words and a content that often is difficult to understand. There are many alternative texts that need to be simplified carefully not to loose important details.

Throughout the project, user evaluations will be conducted to assure that:

1. the readability profiles can accurately predict a student's reading abilities,
2. the tool for creating readability profiles can be used by the teachers and students,
3. the models for readability are able to predict readability,
4. the texts selected based on the mapping between readability profiles and the model for readability are correct and adapted to the needs of individual students, and
5. the summarizations and automatic transformations provide texts that are considered useful with respect to readability, relevance and content.

Students from three different grades: 4, 6 and 8, will be enrolled in the project evaluations. The evaluations will be conducted at schools in the areas of Uppsala, Linköping, and Gothenburg. Evaluations focussed on the first three steps above will be done in order to identify reading profile standards for students in the three grades concerning the reading of texts of different degrees of difficulty in each specific subject area. Evaluations focussing steps 4 and 5 assess the language technology tools developed over the project period.

5 The research group

The research group comprise the necessary competence needed to carry out the proposed project, more specifically competence on reading ability, automatic measures of readability and development of techniques and tools for natural language processing.

The main applicant is Arne Jönsson, professor in computer science at the department of computer science, Linköping University. Professor Jönsson has been working with research on natural language interfaces; especially dialogue managers and empirical studies of human computer interaction including research on recommender systems and user modelling for more than twenty years. Lately his research activities include studies of collaboration using augmented reality, learning technology, especially design and evaluation of teachable agents, and the use of vector space models for language processing. He has been in charge of a number of research projects on dialogue systems development, multi-modal interaction and vector space techniques for language technology. Currently he runs two projects on simplified reading, one focussed on automatic summarization and transformation to easy-to-read Swedish and one on ranking webpages based on readability. In both projects PhD student Christian Smith, who is also proposed to participate in this project, is doing the development of the language processing techniques.

The co-applicant at Uppsala University, Caroline Liberg, is professor in educational sciences, specialized in reading and learning processes. During the past 15 years, she has headed projects and studies centered on students' encounters with reading and writing of texts in different subject areas in school. Senior lecturer Jenny W. Folkeryd who is going to be a co-worker in this project has also participated in several of these as well. Liberg has moreover been involved in similar projects in Norway (University College of Sör-Tröndelag, Trondheim). Some of these projects have been conducted in close collaboration with teachers, students and head masters and have been integrated in the schools' daily pedagogical practices. Liberg has furthermore been a scientific consultant at The National Agency of Education (Skolverket) for the last eight years concerning the Swedish part of the international study of ten year old students' reading ability (PIRLS 2006 and 2011) and Folkeryd has been the head of the scoring procedure in both 2006 and 2011. Folkeryd also takes part in the construction of the national test concerning reading and writing and Liberg is part of the reference group for these national tests. The research domain of the Uppsala research group, thus, concerns reading, writing and learning processes in different subject areas in school and language aspects of these subject areas. The Uppsala group is just now heading a national graduate school regarding the language of schooling in mathematical and science practices. The doctoral students at Uppsala University are studying the international tests PISA and TIMSS and the national tests in math and science.

The co-applicant Sofie Johansson Kokkinakis is senior researcher at the department of Swedish, Göteborg University. She has extensive experience on readability, second language learning and language use. She has been in charge of numerous projects in these areas, often also including development of methods and techniques for computer-based assessment. Katarina Mühlenbock is about to finalize her thesis entitled "Matching text to readers. Assessing readability for specific target groups". It is a corpus based, data mining approach, were the linguistic characteristics of easy-to-read texts are identified and implemented in a computerized text complexity classifier. Katarina is presently on temporary leave from a position as head of DART, which is a county affiliated center directed towards assistive

technology for persons with communicative disabilities. Katarina has a background as computational linguist and has been involved in a wide range of national and international research and development project in the field of language technology, assistive technology, readability and pedagogical software production.

Furthermore, some teachers and students will be engaged during the development of the learning environment in order to ensure its real world relevance. Clearly, the planned project requires a complex and well-composed team representing several different kinds of expertise. Fortunately, we have managed to compose such a team.

Finally, the proposed project will benefit from the experiences from other projects, previous and on going, and from the networks that they are associated with. Yet the research questions addressed in the proposed project stand-alone and are not explored in any other project.

6 References

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

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

Chatterjee, N. and Mohan, S. (2007). Extraction- based single-document summarization using random in- dexing. *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

Decker, A. (2003). Towards automatic grammatical simplification of Swedish text. Master's thesis, Stockholm's University.

Firmin, T. and Chrzanowski, M.J. (1999). *An Evaluation of Automatic Text Summarization Systems,* volume 6073, pages 325–336. SPIE.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005). Summary street®: Computer support for comprehension and writing. *Journal of Educational Computing Research,* 33(1):53–80.

Hassel, M. and Sjöbergh, J. (2007). Widening the holsum search scope. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa)*, Tartu, Estonia, May.

Hassel, M. (2011). Java random indexing toolkit, January 2011. http://www.csc.kth.se/xmartin/java/.

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

Kaspersson, T., Smith, C., Danielsson, H., and Jönsson, A. (2012). This also affects the context - Errors in extraction based summaries, *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC),* Istanbul, Turkey.

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

Laufer, B. and Nation, P. (1995). Vocabulary Size and Use: Lexical richness in L2 Written Production, *Applied Linguistics,* **16**, 307-322

Liberg, C. (2010). *Texters, textuppgifters och undervisningens betydelse för elevers läsförståelse. Fördjupad analys av PIRLS 2006*. Skolverket

Liberg, C., af Geijerstam, Å. and Folkeryd, J.W. (2011). Scientific Literacy and Students' Movability in Science Texts. In C. Linder, L. Östman, D.A. Roberts, P-O. Wickman, G. Erickson & A. MacKinnon (Eds.), *Exploring the Landscape of Scientific Literacy*. New York: Routledge (p. 74-89).

Luke, A. and Freebody, P (1999). Further Notes on the Four Resources Model. Reading Online. http://www.readingonline.org/research/lukefreebody.html [2012-02-13]

Melin, L. and Lange, S. (200). Att analysera text. Stilanalys med exempel, *Studentlitteratur.*

Mühlenbock, K. and Johansson Kokkinakis, S. (2009). LIX 68 revisited – an extended readability measure. In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Trong, K.L. and Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework*. Chestnut Hill, MA: Boston College.

OECD (2009). *PISA 2009 Assessment Framework. Key Competencies in Reading, Mathematics and Science.* Paris: OECD

Reichenberg, M. (2000). *Röst och kausalitet i lärobokstexter: en studie av elevers förståelse av olika textversioner*. Acta Universitatis Gothoburgensis, Diss. Göteborg : Univ.,Göteborg.

Rybing, J., Smith, C., and Silvervarg, A. (2010). Towards a Rule Based System for Automatic Simplification of Texts, *Swedish Language Technology Conference, SLTC*, Linköping, Sweden.

Sahlgren, M. (2005). An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, TKE 2005.

Skolverket (2007). *PIRLS 2006. Läsförmågan hos elever i årskurs 4 – i Sverige och i världen*. Rapport 305.

Skolverket (2010). *Rustad att möta framtiden? PISA 2009 om 15-åringars läsförståelse och kunskaper i matematik och naturvetenskap*. Rapport 352.

Smith, C., and Jönsson, A. (2011a). Automatic summarization as means of simplifying texts, an evaluation for Swedish. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010)*, Riga, Latvia.

Smith, C., and Jönsson, A. (2011b). Enhancing extraction based summarization with outside word space. *Proceedings of the 5th International Joint Conference on Natural Language Processing,* Chiang Mai, Thailand.

Vygotsky, L.S. (1976). *Tænkning og sprog. Volym I & II.* (2:a uppl). Köpenhamn: Hans Reitzel.