

Bachelor thesis



Keeping an Eye on the Context:
An Eye Tracking Study of Cohesion Errors in Automatic
Text Summarization

by

Evelina Rennes

LIU-IDA/KOGVET-G--13/028--SE

2013-07-02

Bachelor thesis



Keeping an Eye on the Context:
An Eye Tracking Study of Cohesion Errors in
Automatic Text Summarization

by

Evelina Rennes

LIU-IDA/KOGVET-G--13/028--SE

Supervisor : **Arne Jönsson**

Dept. of Computer and Information Science
at Linköping University

Examiner : **Carine Signoret**

Dept. of Behavioural Sciences and Learning
at Linköping University

Abstract

Automatic text summarization is a growing field due to the modern world's Internet based society, but to automatically create perfect summaries is not easy, and cohesion errors are common.

By the usage of an eye tracking camera, this thesis studies the nature of four different types of cohesion errors occurring in summaries. A total of 23 participants read and rated four different texts and marked the most difficult areas of each text.

Statistical analysis of the data revealed that absent cohesion or context and broken anaphoric reference (pronouns) caused some disturbance in reading, but that the impact is restricted to the effort to read rather than the comprehension of the text. Erroneous anaphoric reference (pronouns) was not detected by the participants which poses a problem for automatic text summarizers, and other potential disturbing factors were detected.

Finally, the question of the meaningfulness of keeping absent cohesion or context as a separate error type was raised.

Keywords : Automatic text summarization, cohesion errors, eye tracking, CogSum

Acknowledgements

I would like to thank everyone who, in different ways, made this work possible. A special thanks to my enthusiastic and motivational supervisor Arne Jönsson for the opportunity to contribute to this project. It has been truly inspirational. I would also like to thank Carine Signoret for all the feedback and support. Last but not least, I would like to thank my dad, for all the helpful comments on the text in its final stage.

Contents

List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Purpose of the study	2
2 Background	3
2.1 Automatic text summarization	3
2.1.1 The word space model	4
2.1.2 CogSum	5
2.2 Error types	5
2.3 Eye tracking	9
2.3.1 Fixation duration	9
2.3.2 Pupil size	10
3 Method	11
3.1 Participants	11
3.2 Equipment	11
3.3 Text selection	12
3.4 Creation of questionnaire	12
3.5 Experimental procedure	13
3.5.1 Pilot study	14
3.6 Data analysis	15

4	Results	17
4.1	Attitude to reading	17
4.2	Text rating	18
4.3	Error marking and subjective rating	20
4.4	Eye tracking results	21
5	Discussion	25
5.1	Text ratings	25
5.2	Eye tracking results	26
5.2.1	Number of fixations and fixation duration	26
5.2.2	Pupil size	27
5.3	Error markings and subjective rating	28
5.3.1	Marked areas	28
5.4	Procedure	30
5.4.1	Data analysis	31
6	Conclusion	33
6.1	Further research	35
	Bibliography	36

List of Tables

4.1	Mean and standard deviation of participants' self rated reading abilities and attitudes towards reading.	18
4.2	Mean and standard deviation (within brackets) of the text ratings.	18
4.3	Descriptives of the texts used in the test.	19
4.4	Distribution of cohesion errors and other categories that were marked by the participants.	20
4.5	Mean and standard deviation of the subjective rating for each error type.	21
4.6	Mean and standard deviation of the number of fixations, fixation duration and pupil size for each error type.	21
4.7	Pairwise comparisons from the Bonferroni post-hoc test. . .	23

List of Figures

2.1	Example of erroneous anaphoric reference.	6
2.2	Example of absent cohesion or context.	7
2.3	Example of broken anaphoric reference.	8
4.1	The number of fixations (mean) distributed over the different error types.	22

Chapter 1

Introduction

The task of automatic text summarization consists of reducing the length of a text, preserving most of its content. It is a growing research field due to the last few decades' development of an Internet based society, characterized by the constant need of easy access to textual information. Except for the obvious benefit of effective information mediation, the ability to summarize texts automatically might be of use to persons with poor reading skills, for example people with dyslexia, cognitive disabilities, aphasia, or the partially sighted. To manually abbreviate and simplify texts is very time consuming, and many documents remain inaccessible for poor readers.

There are various ways in which automatic summarization may be conducted, for example through *extraction* or *abstraction*. *Abstraction* paraphrases the text content by breaking out the key ideas in order to capture the general idea of the text (Hahn and Mani, 2000), whereas an extraction based summary extracts the most important sentences from a text, and produces a summary consisting of the most important words or sentences put together.

This thesis covers extraction based summarization, and especially summaries produced by the automatic text summarizer *CogSum* (Smith and Jönsson, 2011), and is specially concerned with the challenges this method poses. As the extraction based summarization does not consider the context of the text, errors tend to occur in the summaries. The higher the

level of summary is, the more errors are found (Kaspersson et al., 2012) and especially the anaphoric references are known to cause problems in automatic text summarization (Hassel, 2000).

Another problem is the lack of cohesion in the summaries, due to the limitation the extraction of complete phrases implies (Hahn and Mani, 2000). Kaspersson et al. (2012) identified different types of errors and classified them into three different categories (and sub-categories) including broken and erroneous anaphoric references and absent cohesion or context.

Kaspersson et al. (2012) reached the conclusion that errors vary depending on the level of summary, and that even though some errors belong to the same family of errors, they behave differently relative to degree of summary. However, this does not imply that a summary of lower summarization level is worse than a summary of a high level, since a short summary might result in a high information density, rather than making it easier (Keskiä, 2012).

1.1 Purpose of the study

The aim of this study was to investigate how different types of cohesion errors affect the reading of a text summarized by an extraction based automatic text summarizer. This was explored by tracking scan paths with an eye tracking camera and by letting participants rate and comment on the parts of the texts that were found to be difficult.

Earlier studies have shown that words that are used less frequently demands a longer lexical activation process (Just and Carpenter, 1980; Rayner, 1998). This motivates the choice to, in addition to the previously identified error types, also look for other factors that might affect the experience of reading automatically summarized texts, for example unusual of difficult words.

Thus, the main questions of the study are:

- To what degree is the reading disturbed by the errors?
- What errors result in the strongest disturbance?
- To what extent are other factors (for example difficult words) a reason to disturbance in reading?

Chapter 2

Background

This chapter provides a brief overview of the theoretical background of the fields of study that are concerned in this thesis. The properties of general automatic text summarization systems is explained, and more specifically the summarizer used in this thesis: CogSum. The chapter ends by accounting for the theoretical basis of the eye tracking method.

2.1 Automatic text summarization

The goal of an automatic text summarization system is to produce a summary of a document. A common distinction that is made is the one between abstraction based summaries and extraction based summaries, but both methods have one property in common: the produced summary is supposed to be shorter than the original document, conveying most of the relevant information, and without redundance.

Abstraction based summaries might be useful when the reader wants to understand the core of a text. This method uses semantic information to capture the general idea of a text (Hahn and Mani, 2000). The abstract summaries are often rewritten, in contrast to extraction base summarizers, that reproduces the sentences in their original forms.

The goal of extraction summaries is to extract the most important sen-

tences of a text, and might be of use for a reader to decide whether or not a text is interesting.

Early work on extraction summarization used statistical techniques focusing on features like word frequency, with the intention to extract the most important sentences by excluding the most common words, which often consisted of prepositions and and pronouns, and giving the rest of the words weights based on the frequency (Luhn, 1958). Edmundson (1969) developed the methods further by adding three extra components; cue words, title words and sentence location.

More recent extraction methods utilize methods developed within the area of artificial intelligence, but most methods are in some way based on graph or vector based models (Chatterjee and Mohan, 2007).

2.1.1 The word space model

A word space model consists of spatial representations of word meaning (Chatterjee and Mohan, 2007). Every word in the vocabulary of a given text is represented by *context vectors*, typically produced from data in a matrix where the rows contain the single words and the columns contain the context (Sahlgren, 2005). The words are given an specific place in a multidimensional word space, and from these relative positions it is then possible to compute the semantic similarities between the words (Sahlgren and Karlgren, 2005). This is one of the benefits of the method, it makes semantics computable.

However, there are some problems with the matrixes used in word space models. One issue is that they tend to get very large when faced to large data sets, and most of the cells of the matrix will have a zero value (Sahlgren, 2005), which has motivated the development of dimension reduction techniques. One such method is Random Indexing, which main function is to crowd the context vectors base on the occurrence of words in context.

The method of Random Indexing, based on the work on sparse distributed representations done by Kanerva (1988), is described as a two-step process where every unit of context is given a high-dimensional *index vector*. An index vector is a unique representation, which dimensions are described by a number of randomly set +1s and -1s, while the rest of the

vectors are set to a zero value. The text is then scanned through, and every time a certain word occurs in the text, the specific index vector of that word is added to the context vector of the same word, that has the same dimensionality as the index vector. This implies that the context vectors the words are represented by, are the sum of the index vectors of all the contexts in which a word occurs (Sahlgren, 2005). Instead of constructing a co-occurrence matrix before extracting context vectors, the Random Indexing method suggests a backwards-approach where the co-occurrence matrix is built after accumulating the context vectors.

2.1.2 CogSum

The summarizer used in this thesis, CogSum (Smith and Jönsson, 2011) is based on Random Indexing and a modified version of the Weighted Page Rank algorithm (Brin and Page, 1998), which is a method used for selecting what sentences are relevant in a text. The algorithm calculates a rank based on the Random Indexing vectors, which makes sentences that are similar in context support each other, and eventually result in a ranking of the sentences by their importance. The input is formed by the original text to be summarized and a list of stop words, consisting of common function words as prepositions. The summary is then conducted by applying the Random Indexing and Page Rank methods described above.

2.2 Error types

The different kinds of errors that are used in this thesis are derived from an earlier study within the Friendly Reader project. Kaspersson et al. (2012) categorized three error types and sub-types:

1. Erroneous anaphoric reference
 - (a) Noun-phrases
 - (b) Proper names
 - (c) Pronouns
2. Absent cohesion or context

3. Broken anaphoric reference

- (a) Noun-phrases
- (b) Proper names
- (c) Pronouns

Erroneous anaphoric references, as exemplified in Figure 2.1, describes errors that occur when the anaphoric references refer to an incorrect antecedent. This is often the case when the summary has not included the correct antecedent. There are three sub-types of erroneous anaphoric references: noun-phrases, proper names and pronouns.

The text in italic shows non-extracted sentences and the text in bold shows sentences that have been extracted. The underlined word highlights the anaphora that refers to an erroneous previously extracted antecedent.

On top of the crater wall there might have grown reefs of various kinds, and further down the crater rim other types of plants and animals were living.

As time goes by, the crater is filled with sediments which is beneficial to the flora and fauna.

The earth's crust is moving

An impact crater is almost always circular at the moment of the impact.

But the form might change if the earth's crust moves (picture).

By studying the form of the crater the researches are able to get an idea of the movements.

However, only the big craters are useful in this context.

Figure 2.1: Example of erroneous anaphoric reference.

Absent cohesion or context, as exemplified in Figure 2.2, describes the case when extracted sentences lack cohesion or context, which affects the comprehension of the summary.

The text in italic marks non-extracted sentences and the text in bold marks sentences that have been extracted. The last sentence is marked as error type 2: absent cohesion or context

It was founded in the 1960s by the National Bank on its 300th anniversary and is formally called "The Sveriges Riksbank prize in Economic Sciences in Memory of Alfred Nobel".

For 35 years Gary S Becker has been working on his research on economical analysis of questions that usually are considered to be covered by other academical disciplines.

His research was for a long time seen as controversial and is to date not fully accepted within the fields of study that he is considered infringing.

Becker's ideas also met objections for political purposes.

He has, for example, claimed that a high sick pay might result in people staying at home without being ill, and that it may be good for a family if the wife takes on the role as a housewife.

- Yes, he is often seen as cynical by people that think that the world is controlled by idealistic rhetoric, says Professor Assar Lindbeck, president of the Nobel Committee of the Royal Swedish Academy of Science.

Education, family, crime

Becker's most noted contribution treats the analysis of relationships between education, in Becker's words "investment in human capital", and salaries.

If the yield of the education is low, as it has been in Sweden in periods, people's willingness to get an education is reduced.

Figure 2.2: Example of absent cohesion or context.

Broken anaphoric references, as exemplified in Figure 2.3, are errors that occur when the summarizer does not extract the antecedent that is referred to in an anaphoric reference. There are three sub-types of broken anaphoric references: noun-phrases, proper names and pronouns.

The text in italic marks non-extracted sentences and the text in bold marks sentences that have been extracted. The underlined word represents the anaphora that refers to an antecedent that has not been extracted (*the rise in salary*).

If the yield of the education is low, as it has been in Sweden in periods, people's willingness to get an education is reduced.

Becker regards the family like a small factory.

It produces goods and services, for example meals and cleaning, by refining purchased goods.

The amount of time and money that is spent on housework is determined by the salaries of the family members.

If the salary is raised, the incentives to be at work increases and to stay at home decreases.

This has the effect that the efforts in the household are redistributed from time to goods, for example that one buys more prepared meals.

Figure 2.3: Example of broken anaphoric reference.

Kaspersson et al. (2012) concluded that errors depend on the level of summary, and the errors that displayed a statistical significance on level of summary were 1c (erroneous anaphoric reference - pronouns), 2 (absent cohesion or context), and 3a, c (broken anaphoric references - noun-phrases and pronouns). In another study it was claimed that references to people differ to other kinds of referring expressions, such as references to objects, in a summarized news text. The two differing properties of the referred person are the familiarity to the reader and the global salience in the news story (Siddharthan et al., 2011).

2.3 Eye tracking

Eye tracking is a method with many possible applications. The main idea of the method is that the eyes provide a kind of direct link to the cognitive processes and by studying the movement of the eye it is possible to gain insight into the cognitive state of a person executing a certain task. The eye's movement is a result of both goal driven and stimulus driven processes (Duchowski, 2007), and depend strongly on the type of cognitive task that is being performed.

The most common metrics used in eye tracking studies are *fixations* and *saccades*. A fixation is defined as the period of time where the eye is relatively still (about 200-300 ms) and a saccade is the rapid movement when the eye moves from one fixation to another (Holmqvist, 2011). During this movement the eyes are moving so quickly that the sensitivity to visual input is reduced, a phenomenon called *saccadic suppression* (Matin, 1974).

2.3.1 Fixation duration

Just and Carpenter (1980) formed a hypothesis that an object or a text is processed exactly as long as a fixation lasts, and therefor implies a relatively easy access to cognitive processing. However, this is not uncontroversial, and the hypothesis has been questioned, and there are several exceptions that should be considered.

The fixation duration indicates the effort needed for the cognitive processing, but the average fixation duration varies depending on the task and stimuli. The more complicated a text is, the longer are the average fixation durations, and factors like stress might result in shorter fixations (Holmqvist, 2011).

According to Rayner (1998), the average fixation duration is not an adequate measure since it underestimates the duration that the fixations last. The first fixation is often longer than the following fixations on the same word, and the mean duration is therefor in many cases slightly too low. Rayner (1998) claims that the first fixation duration is a better measure of cognitive processing.

In usability research, many short fixations imply that information that was expected to be found is missing (Ehmke and Wilson, 2007).

All words of a text are not fixated during reading. Long words are more likely to be fixated than short ones (Just and Carpenter, 1980), but other aspects as frequency and predictability from context are also proven to be a reason of shorter fixations or word skipping (Reichle et al., 1998).

2.3.2 Pupil size

Hess and Polt (1964) showed that the size of the pupil increases during problem solving, and claims that it seemed evident that the pupil response is a direct reflection of neurological activity. According to the study, the pupil size also seemed to increase correlated to the difficulty of the task, which implies that this could be used as a measure of cognitive activity. The diameter of the pupil can indeed be used to measure cognitive workload, though one has to be aware of the problems this method proposes. The pupil size is sensitive to various states of the participant. Except for cognitive workload, the pupil size increases as an effect of emotion, anticipation, pain or drug influence, and it might decrease due to factors like fatigue, diabetes or high age. It is important that the presented stimuli are of the same brightness and contrast and that the lighting of the room is kept constant, since the pupil size is light sensitive (Holmqvist, 2011).

Chapter 3

Method

This chapter describes the method used for designing and realizing the experiment and succeeding data analysis.

3.1 Participants

The study was conducted on current or former students of Linköping University. All participants were recruited by e-mail, and were offered a cinema ticket. A total of 27 participants completed the test, but four participants were excluded from all data analysis due to poor eye tracking data. Of the 23 remaining participants, 13 were men and 10 were women. They were all native Swedish speakers without any writing or reading disability and with normal or corrected-to-normal vision. The average age was 23.2 ($SD = 2.76$).

3.2 Equipment

The eye tracking equipment used for this study was SMI iView RED II 50 Hz Pupil/Corneal reflex camera mounted underneath a 19" computer monitor. The softwares used for recording and analyzing the eye tracking data were iView X, Experiment Center 3.0 and BeGaze 2.

A Sony PCM recording device was used to record the comments.

3.3 Text selection

The texts used in the tests were four texts from the Swedish popular scientific magazine *Forskning och Framsteg*. The texts were summarized by the automatic text summarizer CogSum to a summary level of 33%, meaning that 33% of the original text remained. The output of the summarizer was not in any way formatted, other than being divided into paragraphs in order to enhance readability. The texts were previously used and tagged for errors by Kaspersson et al. (2012).

The four texts varied in length from 11 to 14 rows and the number of tagged errors varied from 6 to 12 per text. In total there were 34 errors, and the error types that were present in the texts were:

- 1(c) Erroneous anaphoric reference - Pronouns
- 2. Absent cohesion or context
- 3(a) Broken anaphoric reference - Noun-phrases
- 3(c) Broken anaphoric reference - Pronouns

The remaining error types were not present in the texts.

The level of summary (33%) was chosen in order to get as many errors as possible in a text, while keeping it at a reasonable length that is still readable. Kaspersson et al. (2012) showed that the text summarized to 33% have the most errors per one hundred sentences (20.6 errors), and six texts that were considerably longer or shorter were removed. The order of the texts was not the same for all participants, in order to control possible boredom or order effects.

3.4 Creation of questionnaire

The questionnaire was created with the intention to catch the participants' reading strategies and prior attitudes to reading. The questions were an-

swered using an unipolar Likert scale varying from 1 to 5, where 1 corresponded to *do not agree* and 5 represented *agree completely*. The participants also filled in age, gender, profession or current education, and whether glasses or contact lenses were used during the experiment. The questionnaire was validated in three pilot tests in order to verify that the questions caught every desired aspect, and that no question was unclear or poorly formulated.

3.5 Experimental procedure

Before the experiment the participants were informed that the participation was completely voluntary, that they were going to be anonymous and that they were allowed to terminate the experiment if they did not want to continue.

The experiment consisted of four parts: answering a questionnaire, text reading, error marking and text rating.

Before positioning in front of the eye tracking equipment, the participants got to fill in the questionnaire described above, answering questions about reading strategies, and attitudes towards reading. They were then sat in front of a computer screen with an RED eye tracking camera positioned under the screen. Before the actual test a calibration of the eye tracking camera was performed. The participants were asked to find a comfortable position before starting the calibration, since it is important to keep the same position during the test. The calibration was repeated until a satisfying calibration value was achieved.

The reading part of the test consisted of the four texts presented one by one. The participants were not aware of that the texts were summarized. They were instructed to read the texts for as long as they wanted until they felt they understood it, and continue to the next text by pressing the spacebar. They were told that they were going to perform a task after finishing reading, but they were naive to what the task consisted of.

After reading, the participants were asked to mark the most problematic parts of each texts, using a highlighter pen on a printed copy of the texts. They were allowed to mark as many as they wanted, and then got to rank the marked areas on a scale 1-3 where 1 was the least difficult area and

3 the most difficult area. They were then allowed to comment on their markings, and the comments were recorded.

The participants also got to rate the texts regarding difficulty, how boring they were, how interesting they were and how exhausting they were. A Likert scale from 1 to 5 was used, where 1 represented *do not agree* and 5 represented *agree completely*.

After the test, the participants were asked whether they felt that the presence of the eye movement camera had any impact of their performance, and if their attitude towards the texts would be different if they knew in advance that they were summaries.

3.5.1 Pilot study

Three pilot studies were performed to test the experimental design. The first iteration detected a few weaknesses in the method. The initial idea was that the participants were to mark the problematic areas on the computer, in order to minimize the amount of different stimuli. The problems that emerged associated with this task were that the test as a whole became very heavy. One participant said that reading on a computer screen is very concentration demanding, and all participants said that it would be better to perform the final task on a printed copy of the texts. For these reasons the task was changed from marking up on the computer by clicking, to underlining problematic areas on printed copies of the texts.

All participants misunderstood the instructions given and marked more than three errors in each texts. However, this was not seen as a problem, and since no participant marked more than four errors per text, it was decided to change the instructions from *mark the three most difficult areas* to *mark the most difficult areas*.

To the second pilot study, an additional task was included in order to add another dimension to the test. After marking the errors, the participants got to rate the texts regarding difficulty, how boring they were, how interesting they were and how exhausting they were. A Likert scale from 1 to 5 was used, where 1 represented *do not agree* and 5 represented *agree completely*. After the whole test the participants were informed that the texts were automatically summarized and asked if their attitude towards the texts would be different if they would have known that the texts were

summarized by a computer. This question was asked to control if the result of the study would be different if the participants had prior knowledge of the nature of the texts.

After changing the task from marking the difficult areas on the computer screen to a printed copy of the texts, a problem emerged. The participants tended to mark larger areas, instead of marking the specific place where the problem occurred. For this reason the instructions were changed from *underline or circle the problematic areas* to *highlight the problematic areas*, and instead of a pencil, a highlighter pen was used for this task. This proved to be successful, since the participants were more inclined to specify the difficult areas in detail.

An additional question about the participant's profession or current education was added to the questionnaire. This was done in order to get an explanation to potential outliers.

3.6 Data analysis

To analyze the data recorded by the eye tracking equipment, areas of interest (AOIs) were defined. There were four different types of AOIs corresponding to the four error types and each area intended to define a specific type of error. In the case of error type 2, absent cohesion or context, it is often difficult to detect the specific place in a sentence where the error occurs, which motivated the decision to mark the whole sentence as an AOI of type 2. In some sentences there were more than one error, and all AOIs were placed so that they did not overlap, with the result that some type 2 errors lack data from an area corresponding to the area of the other error type in the same sentence.

The rating of each error was used in order to motivate definitions of further AOIs, that did not correspond to any of the pre-defined error types. If more than half of the participants marked an area as difficult, that same area were to be defined as an own AOI. However, this was not the case and no other area was considered in the analysis except for the already defined error types.

Since the AOIs of error type 2 were much larger than the other AOIs, the number of fixations was corrected by dividing with the size of the

AOIs. The corrected number was multiplied with 1000 in order to get a larger number, which had no impact on the statistical results. The same correction procedure was conducted for the area surrounding the AOIs in order to compare the number of fixations between areas of very different size.

The variables used for the statistical analysis of the eye tracking data were the number of fixations, fixation duration and pupil size.

Chapter 4

Results

This chapter presents the results from all parts of the experiment conducted for this thesis. First, the data collected by the questionnaire and text rating are presented, followed by the results of the error marking and subjective rating. Finally, the results of the eye tracking data from the reading sessions are explained.

4.1 Attitude to reading

The questionnaire that evaluated the participants' prior attitudes to reading gave the results presented in Table 4.1. The assertions were answered with a 1-5 Likert scale where 1 corresponds to *do not agree* and 5 to *agree completely*. The participants generally considered themselves be good readers.

Table 4.1: Mean and standard deviation of participants' self rated reading abilities and attitudes towards reading.

Assertion	Mean	Std.Dev.
I usually understand what I read	4.61	.58
I am a slow reader	2.52	.90
I find it easy to read	4.70	.70
I find it exhausting to read	1.65	.88
I am often pleased to get a rough idea of a text's content	3.70	1.02

4.2 Text rating

The texts used in this study were evaluated regarding three different criteria: whether they were easy to understand, boring or exhausting to read. The results are demonstrated in Table 4.2.

The texts differed slightly. According to the means, text 2 was considered the easiest, least boring and least exhausting text while text 1 was the most boring text and text 3 was the most exhausting text to read.

Table 4.2: Mean and standard deviation (within brackets) of the text ratings.

Assertion	Text 1	Text 2	Text 3	Text 4
The text was easy to understand	3.43 (.90)	3.96 (1.33)	2.91 (1.20)	3.78 (1.17)
The text was boring	3.52 (1.08)	1.78 (.85)	3.00 (1.08)	2.17 (.98)
The text was exhausting to read	2.96 (1.14)	2.00 (1.09)	3.48 (.99)	2.39 (1.1)

The statistical analysis (repeated measures ANOVA) revealed that for the criterium *easy*, texts differed significantly ($F(3, 66) = 4.02, p < .05$). A Bonferroni post-hoc test showed that there is a significant difference

between text 2 and text 3 ($p < .05$), implying that text 2 was significantly easier than text 3.

For the criterium *boring*, significant differences were found ($F(3, 66) = 15.28, p < .001$). Bonferroni post-hoc test showed that there are significant differences between text 1 and text 2 ($p < .001$), text 1 and text 4 ($p < .05$), and between text 2 and text 3 ($p < .05$). The results show that text 1 and text 3 were significantly more boring than text 2.

For the criterium *exhausting* significant differences were found ($F(3, 66) = 9.37, p < .001$). Bonferroni post-hoc test showed that there are significant differences between text 1 and text 2 ($p < .05$), text 2 and text 3 ($p < .001$), and between text 3 and text 4 ($p < .05$). Text 1 and text 3 were significantly more exhausting to read than text 2, and text 3 were significantly more exhausting to read than text 4.

From the results in Table 4.2, it was obvious that text 2 was considered the easier, least boring and least exhausting text. Table 4.3 shows the amount of tagged cohesion errors of each text and the number of rows of each text. The row labeled *Percentage* represents the ratio of the number of errors and the number of rows. Text 2 was the shortest text, with the least errors which resulted in a relatively low percentage of errors per row.

Text 3 and text 4 were of the same length (14 rows) but text 3 had a higher percentage of errors per row, and had the overall highest score of errors per row (85.71%).

Table 4.3: Descriptives of the texts used in the test.

	Text 1	Text 2	Text 3	Text 4
Number of errors	7	6	12	9
Number of rows	12	11	14	14
Percentage	58.33 %	54.55 %	85.71 %	64.29 %

4.3 Error marking and subjective rating

Other areas than the previously tagged errors were marked, and these results are presented in Table 4.4. The errors that had been identified in advance made up 38.3 % of the total amount of markings. The second most frequent reason of marking was different types of language related problems, for example long sentences or complicated word order (17.55 %). 11.7 % of the total amount of markings were *difficult words*. 11.17 % of the markings belonged to the *Other* category. General problems to understand the context were represented in 9.04 % of the markings and summarizer errors and numbers made up in 7.45 % and 4.79 % respectively of the total amount of marked areas.

Table 4.4: Distribution of cohesion errors and other categories that were marked by the participants.

Category	Percentage
Cohesion error	38.3%
Context	9.04%
Difficult words	11.7%
Summarizer errors	7.45%
Numbers	4.79%
Language	17.55%
Other	11.17%

The subjective rating refers to the number that the participants gave the marked errors. The mean of the subjective rating, between 1 (least difficult) and 3 (most difficult), of each error that the participants gave the marked errors is presented in Table 4.5.

All error types had similar scores ranging from 1.70 to 1.88. No statistical significance was found between the subjective ratings of each error type.

All participants reported that their attitude would be more lenient if they knew in advance that the texts used in the test were summaries.

Table 4.5: Mean and standard deviation of the subjective rating for each error type.

Error type	Mean	Std.Dev.
Error 1c	1.82	0.77
Error 2	1.85	0.81
Error 3a	1.70	0.86
Error 3c	1.88	0.78

4.4 Eye tracking results

This section presents the eye tracking results from the reading session. All participants reported that the presence of the eye movement camera did not have any significant impact on their reading performance.

The Table 4.6 presents the eye tracking data. The row labeled *None* corresponds to the area that has not been marked as an error type, and is thus seen as the rest of the text. The number of fixations depends on the size of the AOI, causing the number of fixation to be noticeably higher for larger areas. For this reason, the number of fixations was corrected for size according to the procedure reported in Chapter 3 before further analysis. The values corrected for the size of the AOI are within brackets.

Table 4.6: Mean and standard deviation of the number of fixations, fixation duration and pupil size for each error type.

Error	Number of fixations		Fixation duration		Pupil size	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
1c	13.61 (2.28)	6.22 (1.04)	291.88	76.14	10.82	1.20
2	210.30 (8.69)	51.91 (2.15)	280.25	41.63	10.76	1.21
3a	12.70 (1.75)	4.30 (.59)	269.17	52.83	10.73	1.20
3c	22.61 (6.14)	5.08 (1.38)	279.20	59.19	10.79	1.22
None	841.44 (1.25)	193.77 (.29)	273.33	41.13	10.82	1.21

A repeated measures ANOVA was used to test for differences between the four error types and the rest of the text. For fixation duration and pupil size no difference was found ($p > .05$). For the corrected number of fixations there was a significant difference $F(2.160, 47.522) = 251.86$, $p < .001$, Greenhouse-Geisser corrected.

The graph in Figure 4.1 presents the number of fixations (means), corrected for the size of the AOI, distributed over the error types. Error type 2 and error type 3c have the highest number of fixations.

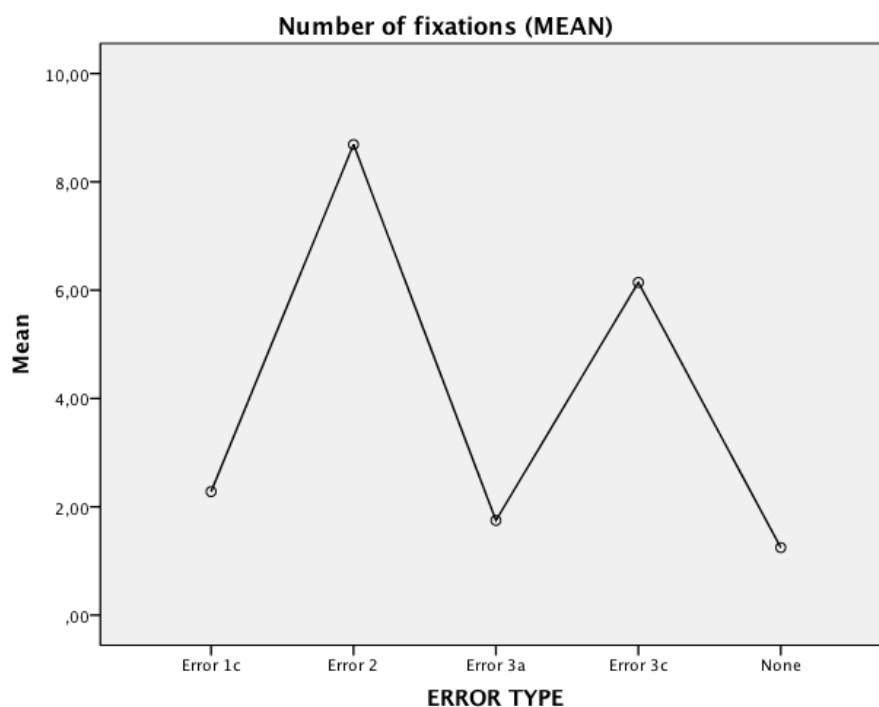


Figure 4.1: The number of fixations (mean) distributed over the different error types.

Bonferroni post-hoc tests revealed significant differences, as presented in Table 4.7. The rows labeled *None* represent the rest of the text. Statistically significant differences are marked in bold.

Table 4.7: Pairwise comparisons from the Bonferroni post-hoc test.

Pairwise	Comparisons	<i>M. Diff.</i>	Sig.
Error 1c	Error 2	-6.41	.000
	Error 3a	.53	.065
	Error 3c	-3.86	.000
	None	1.03	.000
Error 2	Error 1c	6.41	.000
	Error 3a	6.94	.000
	Error 3c	2.55	.000
	None	7.44	.000
Error 3a	Error 1c	-.53	.065
	Error 2	-6.94	.000
	Error 3c	-4.39	.000
	None	.50	.002
Error 3c	Error 1c	3.86	.000
	Error 2	-2.55	.000
	Error 3a	4.39	.000
	None	4.90	.000
None	Error 1c	-1.03	.000
	Error 2	-7.44	.000
	Error 3a	-.50	.002
	Error 3c	-4.90	.000

All error types were fixated significantly more than the rest of the text ($p < .05$).

Error type 2, absent cohesion or context, had significantly more fixations than all other error types ($p < .001$). The error type 3c, broken anaphoric reference (pronouns), had significantly more fixations than error type 1c and 3a ($p < .001$).

Significant differences were found between all error types except for 1c, erroneous anaphoric reference (pronouns), and 3a, broken anaphoric reference (noun-phrase) ($p = .065$). The marginal significance level suggests a tendency of slightly more fixations on error type 1c than on error type 3a.

Chapter 5

Discussion

This chapter presents a discussion about both the results and the method used in this thesis. First, the results of the different parts of the tests are discussed, followed by a discussion regarding the experimental procedure.

5.1 Text ratings

There were several differences between the texts. In the text rating, text 2 stood out, being considered the most easy, least boring and least exhausting text. This can be explained by the fact that it was the shortest of all texts, and that it had the lowest percentage of number of errors per row. Text 3 was considered the most exhausting text. It was the text that scored the highest on number of errors per row and was one of the longest texts. Text 4, which had the same length as text 3 but less errors, was considered less exhausting than text 3. No difference was found regarding the difficulties of these two texts. This suggests that the experience of the text is influenced by number of cohesion errors, rather than the text length.

Text 1 was considered the most boring text. Seen to the length and number of errors, it was similar to text 2, but differed in rating. The reason to why text 1 was considered more boring was probably because the topic was considered boring. Text 1 treated the Nobel Prize while text 2

treated polar bears, and it is possible that the second topic seemed more attractive to the reader.

No participant were aware of that the texts were summarized, which resulted in a critical attitude towards the texts. After finishing the test, the participants were asked whether their attitude towards the texts would be different if they knew in advance that the texts they had read were summaries. All participants claimed that they would be more lenient with the texts if they knew that they were automatically summarized, and this is probably an important factor when evaluating the automatic text summarizer. When used in real situations and when the user is aware of this fact, it is likely that the different errors are not seen as severe. It would be interesting to investigate whether the summaries are preferred over the original texts, despite of their errors.

5.2 Eye tracking results

The results from the eye tracking constitutes the main part of the analysis. In the previous chapter it was shown that there are significantly more fixations in the areas marked as 2, absent cohesion or context, and 3c, broken anaphoric reference (pronouns).

5.2.1 Number of fixations and fixation duration

The results of the statistical analysis of the eye tracking data suggests that error type 2 and error type 3c are the areas that cause the most reading disturbances. However, no difference could be observed for the duration of the fixations or the pupil size, which indicates that these areas are not more cognitively engaging than the rest of the text. These two claims are somewhat contradictory, and therefore interesting. The participants fixated significantly more on errors of type 2 and 3c, but the fixations were not significantly longer. According to the general hypothesis of fixation duration, long fixations means deep cognitive processing, which would imply that the errors did not cause any substantial effort. However, according to Ehmke and Wilson (2007), many short fixations might indicate confusion when expected information is missing. Although this claim is made within

the field of usability research and is applied on a web stimulus, it could be seen as a possible interpretation to the pattern of many but short fixations within the areas of error type 2 and 3c.

The sentences are often tagged as more than one error type. Of all sentences that are marked with error type 3c, all except for one is also marked with error type 2. It is therefore uncertain if the fixations within the two different AOIs can be said to be separated from each other.

Error type 1c, erroneous anaphoric reference, had significantly more fixations than the rest of the text, but less fixations compared to the other error types (except for error 3a where no statistical significance was found). The reason to why this error type is fixated less might be because it is difficult to identify, since the anaphoric expression refers to an existing (erroneous) antecedent. For the practical use of the automatic text summarizer, it is preferable that the errors are found. If the reader does not discover that the antecedent is erroneous, the comprehension of the text is inaccurate.

Error type 1c is indeed fixated more than the rest of the text, which would suggest that it is detected by the readers. However, one must consider the risk that the number of fixation is contaminated by the fixations of error type 2. Error type 1c always co-occurs with error type 2, which makes it possible that the statistically significant number of fixations might be due to error type 2. Thus, there is a risk that this error type has not been detected and might cause inaccurate comprehension of the text, an issue that motivates further work on the automatic text summarizer to investigate the impact of this error type.

5.2.2 Pupil size

The pupil diameter can, in certain circumstances, measure the cognitive workload, but since the size of the pupil might increase or decrease due to other factors, it is a metric that demands an experimental design that controls for all other potential factors. The experiment conducted for this thesis did not control for factors like fatigue or light variation, which might be a possible explanation to the little change in pupil size. It is also possible that the participants did not find the task cognitively involving. The given instructions were *read for as long as you want until you feel that you*

understand, and the participants were not informed about the task until after reading all texts. Since there was not a specific task to perform, the cognitive workload might have been lowered, and the problematic areas where only shown in number of fixations.

5.3 Error markings and subjective rating

This section discusses the results achieved in the participants' error marking task and the subjective rating of the marked areas.

5.3.1 Marked areas

As expected, the majority of areas marked by the participants (38.3 %) were marked due to the previously identified cohesion errors. However, there were other aspects of the texts that seemed to cause problems to the reader. As the instructions of the task were to mark the areas that the participants found difficult, other areas than the previously identified errors were found.

17.55 % of the markings were due to problems that arose from linguistic factors, such as long sentences, or phrases with a difficult word order. Some of these problems, for example non-intuitive word order, might emerge as a result of the automatic text summarizer extracting previous sentences, but the length of the sentences is not controlled by the summarizer. 11.7 % of the markings represented difficult words in the text. This highlights one important function of the automatic text summarizer: a synonym tool. By implementing the option to replace difficult words by a synonym, the total text comprehension is likely to be enhanced.

9.04 % of the markings were areas that the participants claimed were out of context, although not tagged by the absent cohesion or context error. The reason to this is probably that the error types affect other parts of the text as well, and that the error type is vague and hard to narrow down to apply for only one sentence.

7.45 % of the marked areas were due to poor formatting of the texts. The texts were deliberately kept as close to the actual output of the summarizer as possible, which resulted in that references to extracted headlines

and missing pictures were left in the text. This caused 7.45 % of the markings, which in hindsight was not beneficial to this study.

4.79 % of the marked areas consisted of numbers. It is possible that numbers are more confusing the more summarized a text is, since it is a factor that is also affected by absent cohesion.

It is probable that many of these errors emerge in the extraction process. Since the automatic text summarizer is extraction based, it is in its nature that information disappear from the original text, causing difficulties to understand the general context, or leaving sentences with a strange word order behind.

The fact that 9.04 % of the participants complained about missing context in other parts of the texts than the sentences tagged by error type 2 is curious. This suggests one of two things: either that the error type does not cover all cases of this error types satisfactory and should be expanded, or that the error type is vague and cannot possibly include all cases of absent context.

Generally, there is a strong relationship between the absent cohesion or context error and the other error types. In the texts used in this thesis, 74 % of the sentences were marked as broken or erroneous errors combined with absent cohesion or context, which suggests that there might be a link between the different error types. Kaspersson et al. (2012) briefly discusses the relationship between error type 2 and other cohesion types, and claims that it often is the reason for errors like erroneous or broken references. Since the error type 2 in the majority of cases co-occur with any type of missing anaphoric reference, it is reasonable to question the purpose of keeping it as a separate category. It is a vaguely defined type of error and there are only two cases where the error type 2 occurs independently of any erroneous or broken anaphoric reference

The subjective rating of the errors, rated with a 1, 2 or 3, depending on how difficult the participants experienced the areas to be, scored similarly and showed no significant difference, suggesting that no error type was considered more problematic.

No AOI was placed around the marked areas that did not belong to any cohesion type, which makes it difficult to compare them. Thus, the question of how severe these problems are is left open for further investigation on this field.

5.4 Procedure

The participants in this study were current or former students on Linköping University, which is a relatively narrow sample. The benefits of selecting students for this type of task is that it is probable that they have somewhat similar reading habits, which on the other hand also implies that the generalizability is reduced.

The stimuli was presented on a computer screen while the task was done on paper. The initial idea was to perform the task on the computer in order not to change stimuli, but the pilot tests showed that doing the tasks directly on the computer was too exhausting and concentration demanding. To change the stimuli during the experiment is not an optimal solution, but the benefits of avoiding a possible fatigue effect were considered more important than the disadvantage of changing stimuli. The three pilot studies that were conducted indicated that the participants were more eager to mark errors on the paper stimuli than the computer stimuli, but what impact this had on the actual results is not certain.

All metrics used for the eye tracking analysis are sensitive to stress, which might have affected the results in some ways. Some participants spontaneously said that they felt they needed to perform well on the test by, for example, reading fast. It is possible that this resulted in shallow reading and that the participants paid less attention to the meaning of the texts.

Since the participants were naive to what the task consisted of until finishing reading, some participants guessed what the task were going to be. Thus, it might have been better to add some kind of simple task, such as rating the texts or a few comprehension questions between each text, in order to make the reading seem more meaningful. Some participants said that they thought there were going to be a test of the memory capacity, and that they therefore tried to remember as many numbers and names as possible. To create a task that encourages reading in all its simplicity is very difficult, but it is clear that there should be some kind of task to give the participant a purpose of reading the texts.

The texts used for this study were of popular scientific type from the Swedish magazine *Forskning och Framsteg*. The four selected texts were chosen in order to have as many errors as possible, and to be of about

the same length. To this study, the unformatted texts were divided into paragraphs to enhance the readability and remove other factors that might cause reading disturbances. What could have been done to improve the readability further is to delete the few words that referred to pictures, such as *Picture (0)*. This was not done in order to keep the text as close to the original output as possible, but caused many participants unnecessary trouble. This type of errors made up 7.45% of the total amount of the markings. This is unfortunate since this kind of problem was not a part of the focus of this thesis. What impact this had on the study is not clear, but since these areas were easy to find, it is imaginable that it made the participants less sensitive to the more subtle problematic areas.

5.4.1 Data analysis

During the process of data analysis, several choices were made that affected the results. As mentioned in Chapter 3, the placement of the AOIs was an issue worth discussing.

The majority of the sentences had more than one error type represented, always an erroneous or broken anaphoric reference combined with error type 2, absent cohesion or context. The erroneous or broken anaphoric references are relatively easy to specify in a sentence, but the absent cohesion or context is a more vaguely defined error type, which made the placement of this AOI problematic. The solution was to place an AOI for error type 2 over the whole sentence, without overlapping the AOI for the other error type. Since no overlap was permitted, there might have been eye tracking data in the anaphoric reference area that belonged to error type 2.

Several options were considered in order to deal with this data loss, for example to compensate the missing area with an average of data from the rest of the area. However, since there were no satisfying way of doing this and error type 2 was seen as loosely defined and difficult to place, no compensation was made. This has obviously had an impact on the data analysis, but since the error type 2 showed a significant difference in number of fixations despite of presumed missing data, the effect was probably small.

As discussed, the different types of errors are of different range. The anaphoric reference errors can often be defined to consist of one word while the absent cohesion or context are more widely distributed over a sentence.

This resulted in AOIs that were of very different sizes, which was problematic as the error type 2 had significantly more of all data because of its larger area. To deal with this, the number of fixations was corrected by division with the size of the AOI. This was necessary in order to be able to compare the different areas.

Although the average fixation duration is not always considered to be an adequate measure since it tends to underestimate the duration that the fixations last, this variable was chosen for data analysis. This is motivated by the fact that the cohesion errors make the reader make regressions and return to previously read passages, which makes the first fixation duration an insufficient measure of the time spend on a certain word or sentence. Another reason to choose the average fixation duration over the first fixation duration was that it is then possible to compare the AOIs with the rest of the text, since the first fixation duration would give an erroneous value on the rest of the text.

Chapter 6

Conclusion

This study has investigated cohesion errors in texts summarized by an extracted based summarizer, *CogSum*. The main questions of the study were:

- To what degree is the reading disturbed by the errors?
- What errors result in the strongest disturbance?
- To what extent are other factors (for example difficult words) a reason to disturbance in reading?

The results of the experiment conducted for this thesis led to four conclusions:

- Cohesion errors are problematic, but not severely so:

It is clear that cohesion errors affect the experience of reading a summary negatively. The number of fixations was significantly higher in areas belonging to error type 2 (absent cohesion or context) and error type 3c (broken anaphoric reference, sub-type pronoun) which could suggest that the participants experienced difficulties when trying to read these error types in particular.

The more cohesion errors there are in a text, the more exhausting it is to read. Yet a text with a high amount of cohesion errors per row is not significantly more difficult, which suggests that the errors indeed cause problems during reading, but that the impact is restricted to the effort to read rather than the comprehension of the text. This result is supported by the average fixation duration and the pupil size, implying that the participants did not find the cohesion errors more cognitively involving than the rest of the text.

- **Erroneous antecedents - a possible problem:**
There is a risk that error type 1c, erroneous anaphoric reference (pronoun) was not detected by the readers, since it refers to an existing, but incorrect, antecedent. This results in an inaccurate text comprehension, and is obviously a major problem for the automatic text summarizer. Further investigation on this area is necessary in order to be able to study this phenomenon isolated from other error types.
- **Other factors matter:**
There are other factors except for cohesion errors that constitute a source of disturbance, for example linguistic factors, and many of them are a direct cause to the fact that extraction based summarizers eliminates information from a text.
- **Vague definition:**
Finally, the error type 2 has proven to be problematic in order to investigate cohesion errors. This thesis has for various reasons raised the question whether it is meaningful or not to make the absent cohesion or context an actual error type. It is in the nature of an extraction based summarizer to remove a certain amount of content and the particular error type is often caused by a broken or erroneous anaphoric reference. The subjective rating showed that the reader complains about missing context in other parts of the texts than the sentences actually tagged by this error type, which further advocates the removal of this error type as a separate category.

6.1 Further research

This is only one way to investigate the phenomenon of cohesion in automatically created summaries. There are many other potentially interesting point of views that could be of concern in further studies on this topic.

This study used participants that were unaware of the fact that the texts were summarized, and it is possible that the result would be different if the prior knowledge was different. All participants claimed that they would be more lenient with the texts if they knew they were automatically summarized, and an approach for a future investigation could be whether summaries are preferred over original texts, despite of their weaknesses.

The categories identified in the error marking in this study could be relevant in order to investigate further sources of disturbance caused by the summarizer, and there is obviously the issue of the erroneous anaphoric reference error that might cause major problems for this automatic text summarizer.

Bibliography

- Brin, Sergey and Page, Lawrence. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30 (1-7):107–117, 1998. ISSN 0169-7552. doi: [http://dx.doi.org/10.1016/S0169-7552\(98\)00110-X](http://dx.doi.org/10.1016/S0169-7552(98)00110-X).
- Chatterjee, Nilhadri and Mohan, Shiwali. Extraction-based single-document summarization using random indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455, 2007.
- Duchowski, Andrew T. *Eye tracking methodology: Theory and practice*. Springer-Verlag, London, 2007.
- Edmundson, HP. New methods in automatic extracting. *Journal of the ACM*, 1969.
- Ehmke, Claudia and Wilson, Stephanie. Identifying web usability problems from eye-tracking data. In *Proceedings of HCI 2007*, volume 1, 2007.
- Hahn, Udo and Mani, Inderjeet. The challenges of automatic summarization. *Computer*, 33(11):29–36, 2000.
- Hassel, Martin. Pronominal resolution in automatic text summarisation. Master’s thesis, Master thesis in Computer Science, Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden, 2000.

- Hess, Eckhard H. and Polt, James M. Pupil size in relation to mental activity during simple problem-solving. *Science*, 132(3611):1190–1192, 1964.
- Holmqvist, Kenneth. *Eye Tracking - A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.
- Just, M. A. and Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980.
- Kanerva, Pentti. *Sparse distributed memory*. Cambridge MA: The MIT Press, 1988.
- Kasperi, Thomas; Smith, Christian; Danielsson, Henrik, and Jönsson, Arne. This also affects the context - errors in extraction based summaries. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012*.
- Keskisärkkä, Robin. Automatic text simplification via synonym replacement. Master's thesis, Linköping University, Department of Computer and Information Science, 2012.
- Luhn, Hans P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(159-165), 1958.
- Martin, Ethel. Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12):899–917, 1974.
- Rayner, Keith. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.
- Reichle, Erik D.; Pollatsek, Alexander; Fisher, Donald L., and Rayner, Keith. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157, 1998.
- Sahlgren, Magnus. An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, 2005.

-
- Sahlgren, Magnus and Karlgren, Jussi. Counting lumps in word space: Density as a measure of corpus homogeneity. *Analysis*, pages 151–154, 2005.
- Siddharthan, Advaith; Nenkova, Ani, and McKeown, Kathleen. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842., 2011.
- Smith, Christian and Jönsson, Arne. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010)*, Riga, Latvia, 2011.

Copyright

Svenska

Detta dokument hålls tillgängligt på Internet - eller dess framtida ersättare - under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns det lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om *Linköping University Electronic Press* se förlagets hemsida <http://www.ep.liu.se/>

English

The publishers will keep this document online on the Internet - or its possible replacement - for a period of 25 years from the date of publication barring exceptional circumstances.

The online availability of the document implies a permanent permission for anyone to read, to download, to print out single copies for your own use and to use it unchanged for any non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional on the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the *Linköping University Electronic Press* and its procedures for publication and for assurance of document integrity, please refer to its WWW home page: <http://www.ep.liu.se/>