

Domain Specific Highlighting

Faculty of Humanities

Programme: Communication and Information Sciences

Specialization: Human Aspects of Information Technology

Thesis counsellors:

Prof. Dr. A.P.J. van den Bosch

Drs. A. M. Bogers

Naomi Warnert

18 July 2008

Domain Specific Highlighting

Naomi Warnert

HAIT Master Thesis series nr. 08-02

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN COMMUNICATION AND INFORMATION SCIENCES,
MASTER TRACK HUMAN ASPECTS OF INFORMATION TECHNOLOGY,
AT THE FACULTY OF HUMANITIES
OF TILBURG UNIVERSITY

Thesis committee:

Prof. dr. A.P.J. van den Bosch
Drs. A.M. Bogers

Tilburg University
Faculty of Humanities
Department of Communication and Information Sciences
Tilburg, The Netherlands
July 2008

©July 2008 Océ-Technologies B.V.

The writer of this report was given the opportunity by Océ-Technologies B.V. to do an investigation that was used as a foundation for this report.

Océ-Technologies accepts no responsibilities for the correctness of the stated figures, considerations and conclusions in this report, for which the writer will be held responsible.

All products mentioned in this report are claimed as trademarks or righted trademarks of the respective companies.

This report is the graduation thesis of Naomi Warnert.

“One ought, every day at least, to hear a little song, read a good poem, see a fine picture and, if possible, speak a few reasonable words. “
Goethe

Preface

Before you lies my last thesis before the start of my working career. I have spent a total of five years at the University of Tilburg. I did my bachelor in three years and followed two Master programmes, each with its own master thesis. This is the final one, of the mastertrack Human Aspects of Information Technology.

I had the opportunity to do this experiment at one of the major companies in the proximity of my hometown, which has an outstanding reputation. I have not only been able to improve my research skills, but also learned a lot of the business, myself, and my future. My two counselors within the company, Jan Jacobs and Rob van den Tillaart, not only helped me to find my way within the firm, but they also helped me to find my way outside the company. For that, I extend my special gratitude to them.

Also, I would like to thank my professor and supervisor, Prof. Dr. Van den Bosch for his guidance and advice. He pointed me in the right direction when I lost the overview and didn't know where to begin.

I would also like to thank my family and closest friends for their patience during this thesis project; I worked a lot, and was sometimes a bit touchy. Nevertheless, they never stopped supporting me and for that I am grateful. I praise myself for having such a good "supportgroup".

Abstract

This research aims at investigating the domain of automatically emphasizing text. This is a new domain which has its roots in the field of automatic summarization. The main goal of this research is to find the preferences of users in automatic emphasis. The users are in this case knowledge workers who read professionally, such as analysts, researchers, or broadly “expert readers”.

Most adults read in the same way that they learned as children. People who read as part of their professional activities may want to approach reading from a different angle. Experts want to read faster and more productively. This can potentially be achieved with the help of an automatic emphasiser. This application emphasizes parts of a text in such a way that by reading those parts, one can grasp the meaning (state, importance, content) of a document faster.

There are many ways to emphasize parts of text; in this investigation only a limited number of possibilities were tested: highlighting, underlining and variation in font size. These possibilities were chosen on the basis of their use in daily life, because most people are used to them. Titles, headings or important words are frequently bigger or underlined in text. The results of this study are that most people liked the highlighting annotations, especially when the highlighting was done in yellow. The underlined and variations in font size were least preferred.

This research also aimed at investigating what parts of structured documents people wanted emphasized. For this goal, 24 participants annotated 6 different documents. With the help of ROUGE, precision, recall and F-score were measured between the different annotators, in order to obtain an inter-annotator agreement score. These scores indicate that people highlighted the same parts of a text to a reasonable extent. The structure of these overlapping parts were anchored in rules. With the help of these rules the same documents as the participants annotated, were annotated and again, precision, recall and F-score were calculated. We observed that these rules were at least as good as the inter-annotator agreement scores.

Preface	
Abstract	
1 Introduction	8
2 Background	10
2.1 The reading process: Basics	10
2.2 Paper vs digital	13
2.3 Automatic summarization	16
3 User study 1: Preferences	20
3.1 Subjects	21
3.2 Material	21
3.3 Procedure	21
3.4 Results	22
3.5 Analysis	23
4 User study 2: Discovering rules	26
4.1 Subjects	26
4.2 Material	26
4.3 Instrument	27
4.4 Procedure	27
4.5 Results	28
4.6 Analysis	30
5 Conclusion	34
References	38
Samenvatting	37

1 Introduction

Reading is a daily activity for most people in the Western World. Most people do it largely subconsciously and assume that other people do it in the same way. They learn it at an early age and continue doing it for the rest of their lives. Yet, a vast majority does not feel the need to refine and sharpen their reading skills, whereas actually this is possible. To some extent, reading could be compared with athletics. People learn to walk at an early age, and it also is an activity which most people do subconsciously. However, to run faster, longer, or with obstacles, like an athlete has to do, training is required. People need to practice and sharpen their technique to their final goal. Techniques are usually based on the distance an athlete runs. Sprinting for 500 meters requires explosive power, but running a marathon requires a duration strategy. Something equivalent has to be done while reading. To become a “professional” reader one must refine and sharpen their reading techniques to obtain the goal in a preferred time. Reading a novel requires quite a different technique than reading a manual of a dvd-recorder, or reading a professional report.

Nowadays most people have to read such large quantities of text that the term “information overload” comes to mind. There are a number of ways to reduce this information overload. This study tries to provide one way of reducing the quantities of information with the help of a computer application.

This investigation aims to investigate user preferences for highlighting documents, for example, for intelligence analysts. Intelligence analysts look for information, but do not have a clear idea what kind of information they seek until they find it. Tailored computer applications may be developed to help people screen large amounts of information before they read it. Screening can be done with the help of summaries, highlightings, or other features. These activities used to be performed by humans exclusively, but nowadays these activities can be done with the help of computer applications. There are certain advantages of computer applications. One of these advantages is that an application will tend to be objective. People have the tendency to be biased about certain topics or authors, while an application is only biased when it is programmed to be.

The main questions presented in this study are, what are user preferences about emphasizing text? Which parts of documents do users prefer to be emphasized, and in what form the emphasis should be presented? Different modes of presentations are tested on the basis of of paper documents. In the future these preferences should also be tested with electronic documents displayed on computer monitors¹.

In this research the two words *emphasis* and *highlighting* frequently occur. With *emphasis* the author means that a text element is singled out as being important, in a

¹ This research is done at Océ Technologies BV in Venlo at the Research department.

certain way. Highlighting refers to the area in a text which is marked in a different background or foreground color than the original text.

In the first section, theories underlying reading and summarization are outlined. The section also explains what the basic differences are between reading from paper versus reading from screen. In the second section, the first experiment is outlined. This experiment investigates the user preferences about visualisations of emphasis. In the fourth section the second experiment is explained. This experiment has the goal to investigate the user preferences about what parts of text need to be emphasized. In the fifth section the conclusions are drawn, and in the final section discussion points are raised, and possibilities for future research are discussed.

2 Background

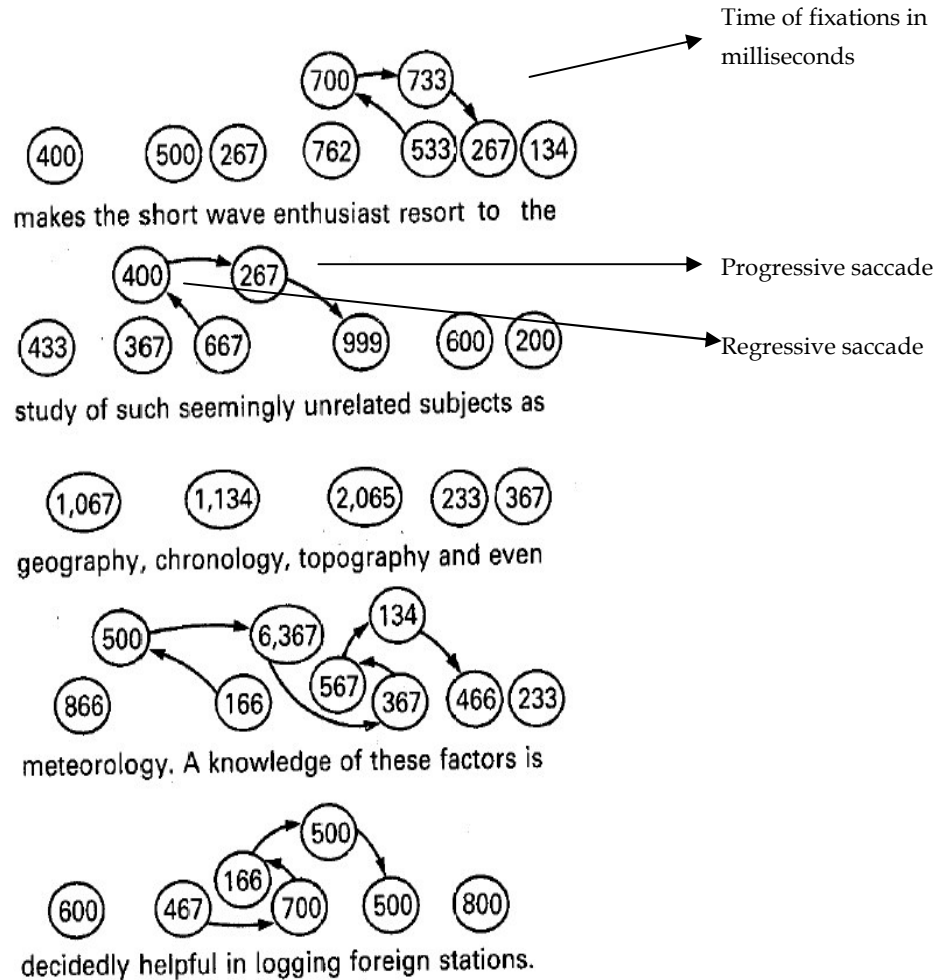
In this chapter, the theory behind this research is outlined. Choices that were made in the study described in this thesis are based on previous research and findings. In the first section basic theories underlying reading are summarized. In the second section the difference between text from paper and text from a screen are discussed, and in the final section an overview of the field of automatic summarization is given.

2.1 The reading process: Basics

Reading is the comprehension of written language, according to Just and Carpenter (1987). Reading is a skill that most humans in first and second-world countries acquire in their early childhood in elementary school. At first, children learn to recognize individual characters, and later they learn how to read groups of these individual characters that together form words. Eventually the skill is expanded to reading sentences. This technique is one of the basic techniques in most Western elementary schools. Reading is thus a sequential process which is done word after word, line by line, from top to bottom. Yet again, the meaning of the words and their interrelated information does not necessarily come in a sequential way. Words may have meaning-bearing relations with words a few positions or even entire paragraphs away. Relations are not always explicitly expressed. To understand how people read, aggregating the exact meaning of words and thereby sentences is not sufficient (Just, Carpenter, 1987). Context is important to understand texts and to increase the level of comprehension of a reader.

Reading is not only a mental process which provides meaning to a text; it is also a physical process. One of the first steps of reading are the movements of the eye. Reading starts with saccades, eye movements. These saccades follow each other rapidly. Each saccade has an average duration of about 0.25 seconds. Before or after these saccades, people fixate their gaze briefly; the foci of these fixations are called fixation points. Fixations occur when the eye stands still, between saccades. People can have fixation points on an object like a painting, eyes of a person and many other objects. In reading, these fixation points are typically textual elements such as characters or words. Fixations have an average duration of 0.6 seconds and have an average span of seven tokens. These seven tokens can be characters, numbers or other symbols. The fixation durations are the periods in which people actually read the words. However, word processing, deriving meaning from words, is done after fixation. Generally, a text is of an easy level for the reader if each fixation sequentially follows the other.

Figure 1: Eye movements: saccades and fixation points



Source: Just & Carpenter, 1987

When a text is of a difficult level for the reader if there are a lot of difficult words in it, long sentences and with a difficult sentence structure. When the reader reads a difficult word, the reader tends to make regressive saccades. It is inevitable that these regressive saccades slow down the reading process (Just, Carpenter, 1987). The process of these eye movements is illustrated in Fig. 1. The numbers state the time of a fixation point in milliseconds, the arrows indicate the saccades. As can be seen in Fig. 1, there are progressive saccades and regressive saccades, and fixation time varies between 0.134 second and 2.065 seconds.

As was stated earlier, word processing occurs after fixation. This implies that the physical recognition of printed words and their comprehension does not occur at the same time. Comprehension is the process of reaching a level of understanding of a passage of a text. The non-sequential physical reading processes can progress in such speed that readers are not aware of their sequence (Just, Carpenter, 1987). Yet, it can be tested with, for instance, the 'Stroop task'. Participants in this test have to read different (known) colour terms. Each word is printed in a different colour than the referent of the word. For example, the word 'yellow' is printed in red and the word 'blue' is printed in green. If participants are asked to read the word aloud and ignore the ink colours, they have almost no difficulty, but if they must read the color of the ink and ignore the word that is printed, they have a much slower reading pace and make significantly more mistakes (Jensen & Rowher, 1966). This implies that recognizing a word is a process that is hard to suppress.

Recognizing a word takes some time. These durations can be relatively brief, but can also be long. One of the factors that plays a significant role on these recognition times is the word frequency effect. This effect states that high frequency words are processed faster, and therefore recognized faster, than low frequency words. The frequency effects are independent of word length. Small low-frequent words are recognized at a significantly slower pace than small frequent words (Rayner & Duffy, 1986). There are two main facilitations which explain the frequency effect. The first facilitation is the abstractionist view. This view suggests that effects result from "primed" abstract word meanings, often called logogens. When a logogen of a word is activated once, the second time it must be activated is done easier. The second view is the episodic view. This states that frequency effects are due to memory for particular episodes and processes. In this view the context of a word is more important. When the context is the same, the frequency effect occurs (Raney, 1995). Evidence based on several studies support both views. Carr et al. (1989) concluded in his study that context was relatively unimportant but Carlson, Alejano and Carr (1991) concluded in their study that context was in fact important.

Another factor that influences recognition time is the context in which the word is presented. Words that are preceded by a related context are recognized more quickly than words preceded by an unrelated context (Becker, 1979; McDonald, 1980).

Yet another factor is that the first word of a text is read relatively slower than the last word, even if they are the same. This is because of the fact that people have certain expectations about word occurrence when reading a text. For instance, the word 'man' is expected to follow 'the' whereas 'walking' has a much lower expectancy rate. If a word meets these expectations, the reading speed increases. This familiarity also comes in mind when reading a text about a known topic. If, for instance, an economist reads a text about Cephalhematoma, the first time this word occurs takes a relatively long fixation time. The second time this word occurs it takes a little less time and so on (Just, Carpenter, 1987). But when the same text is read by a biologist, he reads the text

at a much faster pace from the start, because he is familiar with the topic and has a lot more background information about it.

To speed up reading, it is important that people become aware of their reading technique such as the span of their fixation points and their place. Not every word has to be fixated on; well trained speed readers have the tendency to fixate only on words containing four or more tokens (Just & Carpenter, 1987). Usually they even make a difference in content and function words. Content words contain nouns, verbs and adjectives. These classes are open, and it is relatively easier to expand these classes. Content words have relatively low user frequencies. Function words are prepositions conjunctions and determiners. These word classes are relatively more fixed. Function words also have the tendency to contain fewer characters than content words, and are indeed less fixated on (Just & Carpenter, 1987). Usually function words have high frequencies, meaning that most people use them most of the time. Function words are also essential for learning a language (Shi, Werker & Cutler, 2006). Those words help to determine the appropriate role and meaning of the content words they surround. Because of their frequent appearance, function words are most of the time more familiar than content words, so fixation can occur at a faster pace or they can be left out of the fixation points.

2.2 Paper versus digital text

This investigation explores the possibilities to speed up reading pace from paper with the help of highlighting methods. There are obvious indications that research about this topic in the future should be performed with digital media as well. People have the tendency to read differently when comparing paper reading versus reading from a screen. To explore these differences, for the present study and future studies, this section is included.

There are a number of differences in ways of representation between paper and digital representations. Digital text is, more than text on paper, useful for more dynamical uses. Symbols such as logos, animations, and pictographs require different graphical and visual rhetoric. These symbols also challenge the human visual literacy (Maes, 2005).

There are a number of ways in which differences in paper and digital text can be categorized. In an influential study, Dillon (2004) explains differences in representation. He names three factors that explore the different ways of representation.

The first factor is that there are physical differences between a digital screen and a piece of paper. They can vary on a number of aspects like compatibility, size, ways in which humans can manipulate the representation, orientation and the aspect ratio (the proportion of width and depth). Typical paper sizes are higher than they are wider, while screens are typical wider than they are higher (Dillon, 1992). These observations

can affect eye movements and therefore may affect other factors and reasons for differences.

The second factor is the fact that there are different perceptions about the quality of digital text versus text on paper. Especially in the beginning of the computer era, most humans did not have good experiences with reading from a screen. The quality of the image was much worse than text on paper. Also, when considering the distance, height and the angle in which a screen can be positioned relative to the reader, there are more limitations than reading from paper. For example, the visual angle is significantly larger for screens (Gould & Grischkowsky, 1986). Most types of screens have a near-vertical tilt, and do not allow laying flat on a table, whereas people can hold a book in much more different ways and can adjust the physical distance much easier (Dillon, 2004). There are a few exceptions to this factor. So-called e-books have the advantage that people can hold them in their hand just like a real book.

As mentioned in the part about reading, reading starts with eye-movements. Eye movements also differ when reading from a screen versus reading from paper. A study by Gould et al. (1987) shows that reading from a screen records 15% more forward fixations per line than reading from paper. But this resulted in only 1 extra fixation per line. It seems therefore that there aren't great differences in terms of different presentation media (Dillon, 1992)

A third factor is that there are cognitive reasons for reduced performance of reading from a digital screen. Several studies show different reasons. A reason is that there is less visual memory for the location of information because of the rapid transitions between screens. Another reason is that screen reading has less conventions. Books or other paper media have strict conventions about text positioning, sequential presentation of information and the effort that people must make to find the right information (Dillon, 2004).

Also one of the most common findings is the fact that reading from screen is significantly slower than reading from paper (Kak, 1981, Muter et al, 1982, Smedshammer et al, 1989). The exact outcomes of the different studies vary because of different measurements. But overall it is said that performances slows down between 20% and 30% (Dillon, 1992). However, neither of these studies can explain why reading speed is slowed down (Dillon, 1992).

Another issue is that accuracy measures differ in reading from screen versus paper. Studies done by Creed et al (1987) and Wilkinson & Robinshaw (1987) show that on the task of proofreading, reading from screen scored poorer on the accuracy measurement. It must be remarked that accuracy scores cannot be measured in such a way reading speed can. The question: When does somebody understand a text? is not trivial to answer. Cushman (1986) and Kak (1981) found no significant results in comprehension. Also Belmore (1985) noted, after further analyses, no significant differences in comprehension. These studies show that there is no proven evidence that

screens decrease or increase comprehension. It should be mentioned that comprehension, as accuracy, is difficult to measure.

Yet another factor is that fatigue scores differ in media. Several studies indicate that reading from a screen as for example, Cushman (1986). He found that reading from a Visual Display Unit (VDU) leads to greater ocular discomfort and was more fatiguing than paper. However, several other studies show that VDUs in themselves do not lead to greater ocular discomfort and fatiguing effects (Starr et al, 1982, Sauter et al (1983). As concluded can be stated that it really depends on the quality of the screen. It may even be stated that only performance levels are more difficult to sustain over time. It may be possible that in time screen standards increase and that this problems disappears (Dillon, 1992).

As said, there are many ways of presenting text, analog or digital. People can use different font types, different sizes, colors and more. Most of the time ordinary black characters are used and to indicate a different section, important word(s) etc. different representations are used such as bold or italics. In paper, underlining or highlighting can refer to important information whereas underling or highlighting in digital text can refer to a hyperlink. Hyperlinks are not always important for comprehension but can point to background information, different opinions etc. etc. (Maes, 2005).

Highlighting is a method that people often use to indicate whether a word, phrase, sentence or paragraph is important or relevant for the user's interests. In a research conducted by Chi, Hong, Heiser, Card & Gumbrecht (2006) they created a program which highlighted text to let users skim through a text. As an outcome was that participants indeed found the highlighted parts a great help for skimming the text (Chi et al. 2006).

Another factor that is relevant for digital text is that user have a more or less regular reading trajectory. Most user read websites in a F-kind of shape. The first reading movement is horizontal which creates the upper bar of the 'F'. The second reading movement is vertical and the third reading movement is horizontal again. These three movements create the upper part of the 'F'. The last reading movement is a vertical skimming/reading movement which completes the 'F-shape' (Nielsen, 2006). It is possible that the 'F-shape' is converted into an 'E-shape' or something like an inverted 'L' (Fig. 3). These results indicate that first lines are important for users because the horizontal movements start usually at a new block/paragraph/bullet etc. etc. These results also imply that most information is read in the first few text blocks. This can indicate that users want the first line(s) highlighted in a document because of their reading patterns on screen.

Figure 2: Reading patterns on screen



(source: http://www.useit.com/alertbox/reading_pattern.html)

2.3 Automatic Summarization

One natural language processing task that can be performed to some extent by or with the help of a computer is automatic summarization. The goal of automatic summarization is:

“To take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application needs.”
(Mani, 2001)

This goal is similar to the goal of this study, except that highlighting certain content implies that the context of the highlighted words is maintained.

In the field of automatic summarization there are two main approaches: One can make an extract or an abstract. An abstract is a newly formulated document that contains the information of the source document, yet formulated in new sentences. Abstracts are not related to this research, therefore this topic will not be treated further. Extracts, however, are documents which are generated on the basis of sentences or other fragments that stem from the source document (Mani, 2001). Automatic highlighting refers best to extracting. For automatic highlighting no paraphrases have to be made because the highlighted text is presented in the original source document.

There are a few basic notions for automatic summarization that can be important for automatic highlighting. The first notion is that extracts can be indicative or informative. Indicative summaries provide reference pointers for selecting documents for more in-depth reading (Mani, 2001). These summaries are aimed at helping a reader decide whether or not to read this document.

Informative summaries are aimed at providing the reader all salient information in the source at some level of detail. This notion is somewhat vague because the question whether information is salient or not is very difficult to answer. One answer would lie in the reader. Some people would like to know every detail and others do not. If a text is relatively old, most people would not want to know every detail, whether if a text has a very new topic, most people would like to know every detail (Mani, 2001). A third distinction can be made but is not often practiced, namely, critical evaluative summaries. These summaries not only give a short summary of the source but express also the view of the summarizer (Lancaster, 1991).

Automatic highlighting refers best to indicative summaries because highlighting can be done to give the reader an idea about the topic of the source document. When a reader reads the highlighted parts he knows more or less what the topic of the document is and if this document is worth further reading.

Another notion is that for automatically formulating extracts computers can be relatively easy replace humans because the application does not have to create new text. They do not have to have deep knowledge of difficult natural language processes like sentence structures, ontologies and knowledge of the world (Mani, 2001). Creating a well formulated document can be a hard task for humans let alone for computers.

A third notion that can be applied to automatic summarization is the compression rate. The compression rate indicates how much of the source text is presented in the summary or is highlighted. When the compression rate is set at 25%, it should also be stated what elements of text the compression refers to. Is the remaining 25% composed of individual words, sentences, or complete paragraphs? One can make this distinction, and it is likely that different outcomes are produced (Mani, 2001). Whereas compression rates for extractors are most of the times not lower than 15%, they will be typically lower for automatic emphasis (Mani, 2001). When emphasizing certain words or sentences the words or sentences are not placed in a new document, or in other words, context. This can help the reader to derive meaning from words, for instance, the name of the author. When one sees a name in an extract it is not obvious that that is the author. Whereas emphasized in the original document one sees in an instance that this is the author because of the form and place in the document.

The fourth notion is the audience. Is the application developed for a generic audience or for a specific target group? When developing for a generic group the application should produce summaries that do not detail the subject, but if the application is developed for a specific group the outcome should be tailored for that group (Mani, 2001). This can also be said about automatic emphasis. When one develops an application tailored for a specific user group one can be much more specific about the emphasized parts of text. When, for example, the user of the application are all biologists, names of specific genes can be emphasized so the users know what genes are important in this document. Novel users would not have any idea

about gene names, so in that case a whole section of the text, e.g. which includes an explanation, should be emphasized (Mani, 2001).

For automatic summarization one can distinguish two approaches. The first approach is a “Shallow” approach. This means that the application does not venture beyond a syntactic level of representation of the document, although different elements may be represented at different levels. Words in a source document may be analyzed to a certain semantic level, but sentences will be analyzed to a syntactic level. The second approach is a “Deeper” approach. The applications which uses this approach are potentially suited for producing genuine abstracts. Applications of this type need to have some knowledge of generating natural language. As was stated above, they need to have knowledge of ontologies, semantics, syntax, world knowledge etcetera (Mani, 2001).

In the field of automatic summarization it is possible to speak about the “Edmunsonian paradigm”. Edmundson (1969) defined a framework which nowadays is still used. He used a corpus of about 200 scientific papers on chemistry. The features he considered where cue words, title words, key words and sentence location. Cue words where extracted from a training corpus, whereas the other words where derived from the source document. He excluded words that occurred in a stop word list. He created programs for each of the features, and evaluated these features. He found that location was the best individual feature to use (Mani, 2001).

The work of Edmundson (1969) must be kept in mind when defining rules for automatic emphasis. If the location of certain sentences provides a good indication for extracting, perhaps this is also the case with automatic emphasis.

3 User study 1: Preferences

In this section the first user study is described. The goal of this study is to get a general idea of the preferences of humans, as there is not much research performed in this area. The first question that underlies our study is what the constraints are that future studies should keep in mind. In our first user study, six different methods of emphasizing text were tested: pink, yellow, blue and green background highlighting; underlining; and variation in font size. See also Fig. 3 below. The alternative color highlighting methods were based on the available color markers in stores and thus the markers which the participants used during their daily activities. The underlining method was selected on the grounds of experiences of the participants and the author. The idea to vary in font sizes was borrowed from experiences with relatively new web 2.0 textual visualization possibilities, such as tag clouds. Important information was printed in a larger font size than less important information.

Figure 3: representations used in study 1

contributing to some of the most exciting findings of the past decade. Her work—and that of several other neuroscientists—has made clear that **new neurons are produced in certain areas of the adult brains of**

mammals, including primates. Moreover, these cells can be killed off by stress and unchallenging environments but thrive in enriched settings where animals are learning, and they may play a role in memory.

Until recently, **dogma held that mature brains were static:** no cells were born, except in the olfactory bulb. One of the cornerstones of this understanding came from studies by Pasko Rakic of Yale University, who examined macaque

Republican presidential **hopeful John McCain is a well-known critic of frivolous government spending** otherwise known as pork: those pricey projects that legislators routinely—and surreptitiously—slip into appropriations packages to benefit their own districts and bring them coveted votes. But scientists charge that an important study of grizzly bear DNA has gotten caught in the crosshairs as the veteran Arizona lawmaker attempts to showcase his creds as a crusader against wasteful government spending.

It is unclear why McCain, who has taken a firm stand on some other environmental issues—he believes more needs to be done to curtail global warming—considers the research to be a waste of time and money, and his press office did not respond to repeated e-mails and phone calls for comment. Yet, he is apparently so "outraged" that he takes a dig at it in a campaign TV spot in which an announcer declares:

Currently the front-runner for the GOP nod, McCain also hits the research in speeches on the stump, cracking jokes about bear paternity tests and criminal investigations. "I don't know if it was a paternity issue or criminal, but it was a waste of money," McCain railed last month during a campaign stop in Clawson, Mich. Scientists, however, are not amused: **They insist that the study is not only worth every penny but that the \$3-million price tag cited in the ad is, in a word, wrong.**

In fact, Congress over the past five years has forked over a total of \$4.8 million to study the genetic

"This is not pork barrel at all," says Richard Mace, a research biologist with Montana Fish, Wildlife & Parks (FWP). "We have a federal law called the Endangered Species Act and [under this law] the federal government is supposed to **help identify and conserve** threatened species."

The grizzly has been listed as a threatened species since 1975 and scientists say that it is essential to get a handle on the population to preserve it. But, **according to Kendall,** until the feds decided to invest in this grizzly bear DNA study, researchers lacked the funds to conduct research at the scale necessary to get a reliable measure.

The team collected 34,000 samples of bear hair over a 14-week period in 2004, which it sent over the border to the Wildlife Genetics International laboratory in Nelson, British Columbia. By extracting and analyzing DNA in the strands, researchers were able to pinpoint the species (grizzly or black bear), gender, and individual identity of host bears. **It took two years to analyze the large swath of samples** and another to compile the data and conduct statistical analyses to estimate the size, distribution and genetic structure of

Still, for many Americans who have never seen and probably never will see a grizzly bear, the question remains: Why should one bear population merit millions in taxpayer money?

The reason, grizzly expert Servheen says: **the bears are a threatened species.** He estimates that only about 1,500 still reside in the 48 contiguous states, compared with some 50,000 before the arrival of Europeans in the 15th century (a 97 percent population decline). The once far-reaching grizzly habitat, which **stretched**

3.1 Subjects

A total of six subjects were questioned for this study. Subjects were randomly recruited from the Océ Research & Development department. Each subject was male and had finished a master of science or was a university graduate. One of the participants had color-defective vision. Participants were between 27 and 55 years old and were all native speakers of Dutch. The articles were written in English, so they were not written in their native language of the participants. All subjects were familiar with English, however, as in the company which they work for English is used as a second language. The literature and other documents they used are usually written in English, so reading articles written in English was not unusual for them.

3.2 Material

The six documents were all taken from the news section of the online magazine Scientific American². Scientific American was originally a printed magazine since 1845. Nowadays the printed magazine still exists, while the website is more up to date about current events. The lengths of the six documents are presented in Table 1. All articles had current events as their topics.

Table 1: Document size

Document	Number of words
1	1265
2	1387
3	759
4	1160
5	1067
6	766
Average	1067

3.3 Procedure

Every subject saw three variations of one of the highlighted, underlined or various font-sized documents. Their first task was to read the annotated documents at a regular pace. Reading time for each article was also recorded. Reading times were

² www.sciam.com

measured to get a global idea of the average reading pace of the employees. After each article subjects were asked to write a short summary to sum up what the most important features of the particular article were. This was done to ensure that participants read the whole article. After that question subjects were asked to assign a score to the representation of each article between 1 and 10. When the participants had read all three articles, they had to place them in order of preference. The first one they picked was preferred best and the second one was preferred second and the last one was preferred least. It was also possible for the participants to comment on the representations and articles afterwards.

3.4 Results

Per subject the average reading time was measured in words per minute. Then for all subjects average reading time was measured. The average reading time of the participants was 270 words per minute. The dispersal of these scores was 64.5 words ($SD = 83.6$).

In Tab. 2 below, the results of the preferences are displayed. The text with yellow highlighting was preferred best. Second preference was blue highlighting. The document with the underlined text and the document with the different font sizes were least preferred. Yellow was preferred best by all three participants who had to rate the yellow annotated document. The document with the pink highlighting was preferred second by all participants who saw this representation. Underlining was least preferred by all three participants who saw the underlined annotated document.

Table 2: User preferences

article	pink	yellow	green	blue	font	underlining
1	2	1				3
2			1	2		3
3	2	1			3	
4	2			1	3	
5		1	2			3
6			3	1	1	
Points	6	3	6	4	7	9
Ranking	4	1	4	2	5	6

Reading times were measured as a control variable. Results are listed in Tab. 3. Although the highlighted document in yellow was preferred best, reading times were not the fastest. Underlined text was least preferred, but reading times in words per

minute were not the slowest. Reading times, as tested in this relatively minor study with a small number of documents, appears to have no relation with annotation.

Table 3: reading times, w/pm, per representation form

	green	Fontsize	blue	yellow	Underlined	Pink
	281		218			153
				290		256 311
		185		187		164
		287	190			207
	323			240		305
	449	396	506			
Average	351	289	305	239		238 227

3.5 Analysis

The key result is that users prefer highlighting in yellow³. As second best, the color blue is preferred for highlighting. Green and pink follow in the joint fourth place. The annotated documents which represented the variation in font size and the underlined text were least preferred. The fact that the yellow highlighted document is most preferred has several possible explanations. The first is that yellow has a unique role in human colour perception. From a psychological point of view, one can distinguish four different colors: yellow, red, green and blue. All other colours are variants or mixtures of those four colors, and are psychologically linked to one of these colors (Sheppard, 1968).

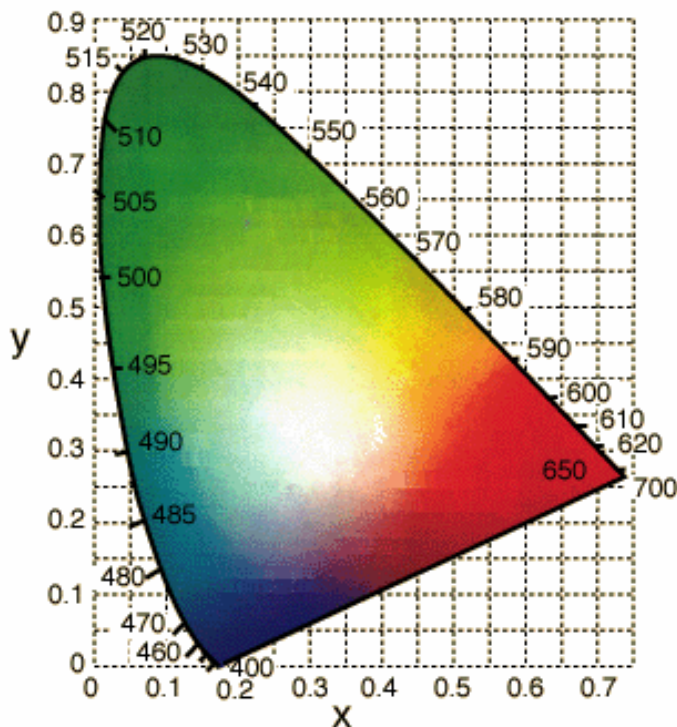
The first possible explanation for the fact that yellow was preferred best is that to perceive the colour yellow, one uses significantly more effort than red, green or blue (Boyton, 1956). Usually, the fact that one must produce more effort for completing a task, usually has a negative effect. Why this fact, in this case produces a positive effect is still a question for further study.

The second possible reason is that yellow has a unique relationship with the color white, which is called the white anomaly. When one sees the four colors, people never perceive the color yellow as white, whereas red, green and blue can in fact be perceived as white (Boyton, 1956). When one sees the color red at wavelength $m\mu = 520$,

³ Yellow is also sold best by Stabilo. A Stabilo spokesperson told, in personal communication, that 40% of the soled markers was yellow, 10% was orange, 10% was pink and 10% was green. The other 30% were all the other colours.

red is perceived as white and even green and blue are perceived as white at that wavelength. Yet, yellow is perceived at that same wavelength as blue, green and blue-green. Researcher cannot explain this effect, but only report that this effect exists. This may provide an explanation why the other colors were less preferred than yellow. This effect can be illustrated with a chromaticity diagram. In Fig. 3 a chromaticity diagram is displayed. A chromaticity diagram is a two-dimensional plot of possible colors. It divides the concept of color in two dimensions: brightness and chromaticity. Chromaticity is an objective specification of the quality of the color. For example, the colour white is a bright color, while the color grey is considered to be a less bright version of the colour white. So, the colors white and grey do not differ in chromaticity (Sheppard, 1968).

Figure 4: Chromaticity Diagram



Source: <http://hyperphysics.phy-astr.gsu.edu/hbase/vision/cie.html>

The variations in font size was preferred less as people reported to be relatively unfamiliar with this representation style. Some participants found the small font size irritating. To vary in font size is technically not a problem, but one must take into account that the bigger the font sizes get, the more paper is needed. As for underlining, people thought the highlighting was not clear enough to stand out visually. The fact

that the underlined text did not contrasted enough with the non-underlined text was one of the main reasons that participants disliked this representation.

4 User study 2: Discovering rules

In this chapter the second user study is described. The goal of this study is to get insight into what words, sentences, or phrases are generally important to readers, and to distill rules from this information. Information analysts, patent researchers and generally most knowledge workers today, are people who generally read, in comparison with other people, a lot of textual material. The quantities of information they must read grow rapidly. To process all this information they need help. A computer is an excellent tool for processing large quantities of information, in principle, and it could be used to help humans with their daily activities, for example reading. However, to be a good help, a computer must have high-quality guidelines, rules, or other constraints, which must be programmed or otherwise established before being put in production mode. To discover rules for automatic highlighting in a particular knowledge worker domain, this user study was conducted.

4.1 Subjects

A total of twenty four subjects were recruited and asked to highlight some of the selected texts with color markers. Subjects were chosen on behalf of their occupation at the Océ group: they all work in either the department of Corporate Patents, Information Management, or Research. They all had seen patents in the past for professional purposes. Each subject was a native speaker of Dutch, and they reported to have good command of English.

4.2 Material

One general article and five patents were selected for the user study. The article was randomly chosen from the Scientific American website. The topic of the article was “virtual markets”. The five patents were randomly chosen from the web⁴. All patents and the article were written in English. Each patent had a different topic. The only restriction was that patents were not part of the participant’s direct expertise, so none of the patents had anything to do with Océ or Océ’s domain directly.

⁴ www.freepatentsonline.com

4.3 Instrument

For an objective comparison of the different highlightings made in documents by different subjects, the evaluation metric ROUGE was used. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE was originally created to compare different summaries created by humans. There are different measurements like, ROUGE-N, ROUGE-s and ROUGE-W but for this study the ROUGE-L package was used. ROUGE-L counts the longest common subsequence. ROUGE-L was originally created to identify cognate candidates during construction of N-best translations from parallel text. This metric was the best option for comparing the different annotations of the participants. This evaluation metric can also be used as an approximate string matching algorithm, which is what has to be done in this experiments (Lin, 2004). Every participant highlighted one or more words in the documents. These annotations are the strings in this experiment. There are a number of strings; each string contains one or more words. A string could also be a part of a sentence or a entire sentence. Each of these strings has to be compared with the strings from the other participants which highlighted the same document so an overall agreement score can be conducted. This score can be taken as an estimate of the inter-annotator agreement. For running ROUGE-L, parameters were copied from the way ROUGE is used in the Document Understanding Conference competition. Not every parameter is relevant for this study; however, a key selected parameters is “-t”, which tells ROUGE to count with tokens rather than sentences. This was done because participants had the option to highlight only words, not just sentences. The parameter “-f A” was selected as this parameter ensures that the “average” option was specified. Other options assume one best solution, but this experiment has no gold standard solution.

4.4 Procedure

It is important to have some idea about the preferences for an application which facilitates speed reading. Information analysts, patent researchers and most knowledge workers today are people who generally read, in comparison with other people, a lot of material. To have an idea what kind of information those people generally wanted to have emphasized, this pilot was conducted. Eight people had to read two information carriers: a United States patent and an article from the Scientific American. They had to highlight that particular phrases, words or sentences that helped them to get to know the general outlining of the text in their opinion. For the United States Patent they had to highlight the phrases, words or sentences that would be important to know professionally. This could include names, dates, numbers, concluding remarks, notes

etcetera. Subsequently, sixteen people were requested to read two different patents. They had to do the same as the first eight people. A total of 48 documents were highlighted: the Scientific American article was highlighted eight times, and the five patents were each highlighted eight times as well.

4.5 Results

For each document the ROUGE-L scores were computed. Also the average score per document was computed, to be interpreted as the inter-annotator agreement. Below the results are displayed, Table 3-Table 8. The parameters for running ROUGE were copied from DUC 2002 (Lin, 2005). To form an inter-annotator agreement score, averages were calculated of the recall, precision and F-scores. Each annotated document was a reference when the rest of the seven annotated documents were candidates. Each patent was highlighted by eight participants so this had to be repeated for only each set of the same patent. This was done with every single document and the averages were calculated. In ROUGE it is important that a reference summary is inserted and a candidate summary is inserted. The recall reflects the proportion of words in the reference summary that is also present in the candidate summary. The precision is the amount of words that are present in the candidate summary and are present in the reference summary. The F-score is the weighted harmonic mean of recall and precision. In Eq. 4a the three formulae are displayed as used in ROUGE (Lin, 2004). LCS means : longest common subsequence. This is the sequence of words that two documents have in common.

Equation 4a: formulae for recall, precision and F-score

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F^{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

LCS = longest common subsequence

R= recall

P= precision

F= F-measurement

X = subsequence in one document

Y = subsequence in other document

m = length of subsequence X

n = length of subsequence Y

In this research $\beta=1$. The last formula can than be simplified into the formula below in Eq. 4b.

Equation 4b: Formule for F-measurament

$$F_{lcs} = \frac{2(R_{lcs}P_{lcs})}{R_{lcs} + P_{lcs}}$$

Each document was used as candidate summary were al the other summaries were used as reference summaries. The average scores were calculated as well as the Standard Deviation. In Tab. 4 the results of these calculations are displayed. Also the average number of words was counted which the participants highlighted. The scores can be seen in Tab. 4 below.

Table 4: Number of words highlighted

Document	Words	Number of words highlighted	Percentage of words highlighted
Article	3942	444,6	11,28 %
Patent 1	5492	505,5	9,20 %
Patent 2	6154	495,1	7,47 %
Patent 3	6278	514,4	8,19 %
Patent 4	5188	389,3	7,50 %
Patent 5	8068	513,1	6,36 %
Average	5854	477	8,33 %

Precision, recall and F-scores are displayed in Tab. 5. Standard deviations were not calculated for the F-score because those were based on the average recall and average precision.

Table 5: ROUGE-scores

Document	Average recall	Average precision	Average F-score
Article 1	0.495 (<i>SD</i> = 0.166)	0.396 (<i>SD</i> =0.238)	0.44
Patent 1	0.486 (<i>SD</i> = 0.145)	0.649 (<i>SD</i> = 0.211)	0.556
Patent 2	0.574 (<i>SD</i> = 0.176)	0.574 (<i>SD</i> = 0.191)	0.574
Patent 3	0.512 (<i>SD</i> = 0.207)	0.586 (<i>SD</i> = 0.230)	0.547
Patent 4	0.491 (<i>SD</i> = 0.153)	0.485 (<i>SD</i> = 0.180)	0.488
Patent 5	0.529 (<i>SD</i> = 0.177)	0.530 (<i>SD</i> = 0.229)	0.529

4.6 Analysis

The average recall and precision scores indicate that people do have some sort of agreement of what to highlight. Average recall and precision scores are a bit more than 0.5. This means that all highlighted documents had an overlap of about 50%. Participants indicated before the beginning of the task that they expected they would do it very differently from other participants. The average number of words which was highlighted is about eight percent. This was not as high as expected. When comparing highlighting with summarization, one can see in summarization studies that typically about twenty percent is used in a summary, or the summary has a length of twenty percent of the original document.

This experiment was conducted with the goal to develop rules to produce the appropriate highlighted phrases. A total of ten rules were developed. Each rule was selected on grounds of the fact that several subjects in the study highlighted a particular span of text that could be circumscribed exactly. In Fig. 5 a patent is displayed to illustrate the application of the first 5 rules as described below.

1. The first rule is that the patent number must be highlighted. This number is always placed in the upper right corner on the first page of the patent and is an identification of the document.
2. The second rule is that the "Date of Patent" must always be highlighted. This is because then people can place the document in its context. An old document is not worth reading most of the time. It is also possible that a date can place a document in a certain context which can be relevant or irrelevant for the reader.
3. The third rule is that the title of the patent should always be highlighted. In twenty-four out of forty patents which participants highlighted the title is, or a part of, is highlighted. This implicates that most people find it necessary to highlight the title. So a computer application must do the same.
4. The fourth rule is that the "assignee" name must be highlighted, if a name is mentioned. In eighteen out of forty documents the "assignee" is highlighted so this

indicates that a number of people find this important for highlighting. An assignee can be the author of the designated patent but it can also be a corporation, research institution or a group of people.

5. The fifth rule is that the abstract must be highlighted. In thirty-four cases the abstract, or a part of the abstract, is highlighted. This implies that most people read the abstract or find it an important part of the highlighting.
6. The sixth rule is that the first paragraph of the background must be highlighted. In most cases, 24 out of forty, one or more words of this paragraph are highlighted by the participants.
7. The seventh rule is that the first paragraph of the summary must be highlighted. In 22 out of forty annotated documents this paragraph is highlighted. This gives an indication that many people find this first paragraph important.
8. The eighth rule is conditional. It states that the first paragraph of the embodiment must be highlighted. Yet, not every patent has an embodiment.
9. The ninth rule is also an optional rule. This rule states that if there is a section which is called "Technical Field", this section must be highlighted.
10. The final rule states that sentences with corporation names and patent numbers must be highlighted. Not in every patent these sentences appear but when they do, they must be highlighted.

Figure 5: an example of a patent with application of the first five rules

US007342502B2

(12) **United States Patent**
Harkins et al.

(10) Patent No.: **US 7,342,502 B2**

(45) Date of Patent: **Mar. 11, 2008**

(54) **WIRELESS SHORT RANGE COMMUNICATION SYSTEM**

(75) inventors: **Donald Harkins, Eugene, OR (US); Chad Minter, Coburg, OR (US); Benton Ullm, Coburg, OR (US)**

(73) Assignee: **Consort, LLC, Eugene, OR (US)**

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 321 days.

(21) Appl. No.: **11/155,128**

(22) Filed: **Jun. 16, 2005**

(65) **Prior Publication Data**
US 2006/0286933 A1 Dec. 21, 2006

(51) **Int. Cl.**
G08B 23/00 (2006.01)

(52) **U.S. Cl.** **340/573.1; 455/41.2; 455/90.1; 340/692**

(58) **Field of Classification Search** **340/573.1, 340/539.1, 539.11, 539.26, 693.5, 692; 128/200.24; 455/90.1-90.3, 569.1, 41.2**
See application file for complete search history.

(56) **References Cited**
U.S. PATENT DOCUMENTS
3,069,511 A 12/1962 Gibson et al.
3,908,168 A 9/1975 McMahon
4,072,831 A 2/1978 Joscelyn
4,491,699 A * 1/1985 Walker 455/90.1

OTHER PUBLICATIONS
Daniel Ceperley et al., LifeLine: Improved Communication and Informatics . . . , Computer Society Int'l. Design Competition, May 2002.
Talk-Around Specification Card, Questions & Answers, Product Highlights, Scott Technologies Company, Circa 2001.
* cited by examiner
Primary Examiner—Thomas Mullen
(74) Attorney, Agent, or Firm—Hancock Hughey LLP

(57) **ABSTRACT**
A wireless communications system defining an wireless personal area network enables reliable communications between members of workers in a group wearing compatible systems. Each member of a group is fitted with a communications system having a microphone, a transceiver and a speaker.

32 Claims, 4 Drawing Sheets

31

All rules are applied to the same patents the participants highlighted. After that the number of highlighted words were counted to check whether the rule-based highlightings did not deviate markedly from the 8,33% that the participants highlighted on average. In Tab. 6 the results are displayed. As can be seen, no more than 6,37% was highlighted on average with the rules

Table 6: Documents highlighted with rules

Document	Words	Number of words highlighted	Percentage of words highlighted
Patent 1	5492	606	11,03 %
Patent 2	6154	248	4,03 %
Patent 3	6278	337	5,37 %
Patent 4	5188	333	6,42 %
Patent 5	8068	405	5,02 %
Average	6236	385,8	6,37 %

After that, the recall, precision and F-score were calculated. The annotated document, which was annotated according to the rules, was taken to be the reference and the corresponding highlighted documents by the participants were the candidates. Scores were calculated for each of the five patents. This means that the dispersal of the scores was higher than between the inter-annotator agreement scores. In Tab. 7 and 8 below, the results are displayed. For the average F-scores no standard deviations were calculated because these scores were based on the recall and precision.

Table 7: Recall, Precision and F-score of documents annotated with rules

Document	Average recall	Average precision	Average F-score
Patent 1	0.372 (SD=0.214)	0.742 (SD=0.227)	0.496
Patent 2	0.827 (SD=0.157)	0.560 (SD=0.234)	0.668
Patent 3	0.483 (SD=0.229)	0.490 (SD= 0.263)	0.486
Patent 4	0.497 (SD=0.175)	0.536 (SD=0.187)	0.516
Patent 5	0.559 (SD=0.225)	0.613 (SD= 0.218)	0.585

* numbers in green are higher than the participants' inter-annotator agreement score

In the last step the differences of the inter-annotator agreement scores and the reference scores using the rules-based highlightings as reference were calculated. They are displayed in Tab. 7 As can be seen, the scores that were lower for the document annotated by the rules, show a less negative trend than the scores that were better.

Table 8: Differences between participants' and rule-based annotations on: Recall, Precision and F-score

Document	recall	precision	F-score
Patent 1	- 0.1118	+0.093	-0.06
Patent 2	+ 0.253	-0.014	+0.094
Patent 3	-0.029	-0.096	-0.061
Patent 4	+0.006	+0.054	+0.028
Patent 5	+0.03	+0.083	+0.056

* average +: 0.07744

* average -: 0.06197

5 Conclusion

In this section, conclusions are drawn, and at the end of this section the discussion points of this study are described.

Review

The first two sections explored the main issues in the field of automatic emphasis. Little research has been done in this area before this study. First, theories on reading were reviewed. How people read can have a significant impact on how people perceive highlighting. The fact that people read the way they do may require an automatic emphasis module to emphasize elements of a certain preferred or minimal size, such as words, sentences or entire paragraphs. People need to comprehend the basic ideas of a text when they use an emphasis module. Comprehension and reading does not necessarily have to occur at the same time, but to comprehend certain elements and make certain connections, one needs to have read more than just single words without context.

Subsequently, the differences from reading from paper versus reading from screen were explained. Reading from paper is fundamentally different because of different aspects. One of the aspects is the physical aspect. Screen sizes vary different than paper sizes. People also have different perceptions between paper and screens. In some way most people find text on a screen of a lesser quality than on paper (Gould & Grischkowsky, 1986).

Another factor is, that there are several cognitive reasons for differences between screen reading and paper reading. One reason is that there is less visual memory for screen reading because of the relative rapid transitions between screens (Dillon, 2004).

Yet, other differences have to do with reading patterns. Reading patterns on Visual Digital Units (VDU's) have most of the times F- or E-shapes. These shapes have a lot to do with location. First sentences, beginnings of text blocks, dates and names, have everything to do with these reading patterns. These thoughts were kept in mind with the deduction of the rules from the second experiment.

The theory behind automatic summarization was also examined because this field has several notions that more or less overlap with the field of automatic emphasis. One notion is that automatic emphasis refers best to indicative summaries. Another notion is that the concept of extracts refers best to what is achieved by automatic emphasis, because of the fact that text is maintained as it was written in the source text. A subsequent notion is the compression rate. A typical compression rate in the field of summarization is about 20 %. The studies in this research indicated that around 8 % of emphasized parts is typical for the type of documents used as focus of the study

(patents). This has something to do with the fact that one maintained the context. When one extracts sentences in an extract summary, one loses context. This hinders the comprehension of the sentences. When one sees highlighted parts in the source document, location and context give also information about the content. This has a relation with the Edmunsonian paradigm, the fact that location can best be used as a feature to extract sentences for extracts (Edmundson, 1969).

This research was conducted to address the main question whether it is possible to develop a program that automatically selects parts of a text for emphasis. To have the right parts one needs user preferences. Also, user preferences are needed to get an idea of what most humans find best for representation. To get answers on these questions two user studies were conducted. Each of these studies had their own goals and methods.

The first user study had the goal to investigate which of six representations of documents was preferred best. The six representations were respectively: pink highlighting, yellow highlighting, blue highlighting, green highlighting, underlining and variation in font sizes. The six representations were chosen on grounds that they represent most variations used in everyday life. Six participants saw each three representations and then rated each representation. The yellow highlighted document was preferred best and the underlined document was least preferred. A possible explanation for the fact that yellow is preferred best is the yellow anomaly. Yellow has a peculiar place in the human perceptual color scheme. Different studies indicate that humans do not treat and perceive yellow in the same way as other colors. Reading times were also measured but gave no indication or other relation with the preferences of the participants.

The second study was conducted to get an overall idea of which parts of a structured text people preferred highlighted. One of the main outcomes was that people preferred about eight percent highlighted. This was less in comparison with the field of summarization. In this study, 24 people highlighted each 2 documents. The documents were five patents and one article, so everyone annotated 1 article and one patent or two patents. In this format eight annotations of the same document were established.

All sets of eight annotations were scored by ROUGE. ROUGE is an evaluation metric for scoring summaries and translations. In this case each set of eight was scored on recall, precision and F-score. To get inter-annotator agreement scores. Each annotation was used as a model where the other seven were used as peers. This was done for every summary and averages were calculated on each set of eight annotations.

From all the annotations of the structured documents, rules were extracted. A total of ten rules were extracted based on frequent occurrence in the human highlightings. Most rules refer to location in the document. There were also a few rules that were optional because not every document had exactly the same structure. These rules were then applied to the five patents and recall, precision and F-score were again

calculated. It appeared that the rule-annotated document scored in most cases higher on recall, precision and F-score. Standard deviations were also calculated and they were higher than between the inter-annotator agreement scores. This indicated that the scores had a bigger dispersal. Nonetheless the scores indicate that the rules as extracted in this study can be used as rules for a computer application.

The fact that most rules are based on location in the document indicates that the Edmundsonian paradigm (Edmundson, 1969) is valid in this particular context of domain-specific texts. This paradigm indicates that of most features that can be used as indicators for selecting sentences for extracts of structured documents, location is the most important indicator.

Also the 8,3 % rate of human highlighting was reasonably close to the 6.3% average percentage highlighted by the rules. This agreement is important because when a computer application exceeds the 8% mark with a large margin, this may indicate that user may not like to use this application because of the surplus of text that is emphasized. However, the application's user preferences should always allow the user to insert the percentage that he or she wants emphasized.

Points for future research

User preferences are very important in this study and program developers must never loose these preferences out of sight. One of the main findings is that user prefer yellow highlighting but it is also important that users must have choices. Not everyone is the same and so therefore they may prefer other representations. As a default the setting of the program can be yellow highlighting but other options may be given as well. While variations in font size is at this moment least preferred, together with underlined annotated documents, this may change in the future, whereas different web 2.0 applications work with "tag clouds". A "Tag cloud" is a visual depiction of user-generated tags, or simply the word content of a site. Most of the times they are listed alphabetically, and the importance of a tag is shown in font size or color⁵. User may get familiar with these representations and may get to like them. When this happens, this representation should be offered as a choice in the application.

In the future, other representations may be tested. In this study, different colours were used but in reality in most offices one prints at a black and white printer. If anyone must print something in colour, one goes to another printer. This has to be kept in mind when developing future studies. For instance, different grayscales may be used. The phrases of text that must be highlighted can be black and other, less important information can be of a lighter shade. You can work with four or five different shades of grey in this way. Another representation that is not tested in this

⁵ www.wikipedia.org

investigation is the use of blurred text. Important text can be present in a clear view whereas less important information can be presented in a blurred view.

In this study relatively few rules were extracted. Nonetheless results with these rules are obtained. Here lies a gap for future studies. These studies may go deeper into user preferences concerning emphasized parts of structured documents.

Also there is another point of consideration. This research was conducted in a specific domain, with people who were familiar with the structured documents. No findings may be extracted to other domains, nevertheless can implicate restrictions for further research. The fact that in this investigation location is also one of the decisive conclusions of Edmundson (1969) must be kept in mind when setting up future studies.

Another point of consideration is the fact that the evaluation metric which is used in this research was originally developed to evaluate summaries and translations. This may influence the results because in these documents, whole sentences are used. In the area of automatic emphasis, single words or parts of sentences may be emphasized. This has everything to do with the fact that the context is maintained and also gives information about the context. The evaluation metric, as used in this study, does not keep this fact in mind and thus may influence the outcome of the tests.

The last discussion point is the fact that results if the second study could be higher. In this study, seven out of the ten results of recall and precision, were higher and this may well be eight or nine when conducting future in depth investigations. Future studies may go deeper into figuring out what make a word, phrase or sentence important for highlighting.

This study was performed to get an overall idea of the field of automatic emphasis. There are a few limitations that must be kept in mind when conducting future study in this field. Nonetheless a few basic guidelines came to surface in this investigation like the principle for location.

References

- Aldis, H., G.(1916). The printed book. Cambridge, University press.
- Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 252-259.
- Belmore, S. (1985). Reading computer presented text. *Bulleting of the psychonomic Society*, 23(1), 12-14.
- Berns, R.S. (2000). *Principles of Color technology*. Munsell Color Science Laboratory. Rochester Institute of Technology.
- Boynton, R. M. (1956). *Rapid Chromatic Adaptation and the Sensitivity Functions of Human Color Vision*, *J. Opt. Soc. Am.*, Vol. 43, 552.
- Buzan, T. (2001). *Snellezen*. Boch & Keuning, Baarn.
- De digitale economie* (2006) Centraal bureau voor de statistiek, OBT bv, Den Haag.
- Creed, A., Dennis, L. & Newstead, S. (1987). Proof-reading on VDUs. *Behaviour and Information Technology*, 6(1), 63-73.
- Chi, E., Hong, L., Heiser, J., Card, S., Gumbrecht, M. (2006). *ScentIndex and ScentHighlights: productive reading techniques for conceptually reorganizing subject indexes and highlighting passages*. *Information Visualization* (2007) 6, 32-47.
- Cushman, W. H. (1986). Reading from microfiche, VDT and the printed page: subjective fatigue and performance. *Human Factors*, 28(1), 63-73.
- Dillon, A. (1992). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, 35(10), 1297-1326.
- Dillon, A. (2004) *Designing Usable Electronic Text*. (2nd Edition ed.). London etc: CRC Press.
- Edmundson, H.P. (1969). New methods in automatic abstracting. *Journal of the Association for Computing Machinery* 16 (2): 264-285. Reprinted in *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury (eds.), 21-42. Cambridge, Massachusetts: MIT Press.
- Frederiksen, J., & Kroll, J. (1976). *Spelling and sound: Approaches to the lexicon*. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 361-379.
- Forster, K., & Chambers, S. (1973). *Lexical access and naming time*. *Journal of Verbal Learning and Verbal Behaviour*, 12, 627-635.
- Garama, P. & de Man, P. (2008), *Finding Images Fast With Folksonomies*. Master thesis University of Tilburg.
- Gould, J. D. & Grischkowsky, N. (1986). Does visual angle of a line of characters affect reading speed? *Human Factors*, 28(2), 165-173.

- Jensen, A. R., & Rowher, W. D. (1966). The Stroop color-word test: A review. *Acta Psychologica*, 25, 36-93.
- Just, M.A., & Carpenter, P.A. (1987). *The psychology of reading and language comprehension*, Carnegie-Mellon University. Massachusetts.
- Kak, A. V. (1981). *Relationships between readability of printed and CRT-displayed text*. Proceedings of Human Factors Society – 25th Annual Meeting, 137-140.
- Lancaster, F. W. (1991). *Indexing and Abstracting in Theory and Practice*. Champaign, Illinois: University of Illinois Graduate School of Library and Information Science.
- Lin, C. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain.
- Lin, C. (2005) *A Brief Introduction of the ROUGE Summary Evaluation Package*. University of Southern California/Information Sciences Institute.
- McDonald, D. D. (1980), *Natural Language Production as a Process of Decision Making under Constraint*, PhD thesis, MIT, Cambridge, Mass.
- Maes, A. (2005). [Een multimodale kijk op informatie](#). In H. Van Driel (Ed.), *Digitaal Communiceren*. (219-258). Amsterdam: Boom.
- Mani, I. (2001). *Automatic summarization*, The MITRE Corporation, John Benjanmins Publishing Company.
- Muter, P., Latremouille, S. A., Treurniet, W. C. & Beam, P. (1982) Extended reading of continuous text on television screens. *Human Factors*, 24(5), 501-508.
- Sheppard, J. J. (1968). *Human Color Perception: A Critical Study if the Experimental Foundation*, New York, Amrican Elsevier Publishing Company.
- Smedshamar, H., Frenckner, K., Nordquist, C. & Romberger, S. (1989) *Why is the difference in reading speed when reading from VDUs and from paper bigger for fast readers than for slow readers?* Paper presented at WWDU 1989, Second International Scientific Conference, Montreal.
- Shi, R., Werker, J., & Cutler, A. (2006). *Recognition and Representation of Function Words in English-Learning Infants*, *Infancy*, 10, 187-198.
- Starr, S. J. (1984). Effects of video displays terminals in a business office. *Human Factors*, 26, 347-356.
- Wilkinson, R. T. & Robinshaw, H.M. (1987) Proof-reading: VDU and paper text compaired for speed, accuracy and fatigue. *Behaviour and Information Technology*, 6(2), 125-133.
- Zielke, W. (1984). *Sneller lezen – goed onthouden*, Intermediair bibliotheek. VNU Business Publications BV.

Samenvatting

In deze scriptie wordt een onderzoek beschreven wat als doel heeft om het gebied van automatisch accentueren te onderzoeken. In dit onderzoek is vooral de nadruk gelegd op de voorkeuren van mensen, en dan met name mensen die voor hun werk veel moeten lezen zoals analisten en onderzoekers, zogenaamde “experts”.

Mensen lezen vaak zoals ze dat vroeger hebben aangeleerd, maar voor mensen die lezen als hun werk hebben komt er meer bij kijken. Mensen willen sneller en productiever zijn en dit kan men mogelijk bereiken door teksten te accentueren zodat een gebruiker niet de hele tekst hoeft te lezen maar alleen de gemarkeerde stukken. Mocht uit deze gemarkeerde stukken blijken dat de tekst wel degelijk van belang is, zal er dan grondig gelezen dienen te worden.

Er kan op vele manieren worden geaccentueerd maar in dit onderzoek zijn maar enkele manieren onderzocht: highlighten, onderstrepen en varieëren in fontgrootte. Deze methodes zijn gekozen uit gewenning en uit het feit dat deze methodes in veel teksten al worden gebruikt. Titels, tussenkopjes, links zijn vaak groter of juist kleiner of onderstreept. Uit dit onderzoek is gebleken dat mensen highlighten het fijnste vinden in teksten op papier. De kleur geel was het meest geliefd. Onderstrepen en varieëren in fontgrootte waren de minst geliefde methodes.

Uit dit onderzoek is ook gebleken dat mensen bij gestructureerde document vaak hetzelfde markeren. Door middel van ROUGE zijn recall, precision en F-score berekend op de documenten die mensen gemarkeerd hebben en daaruit bleek dat er wel degelijk overeenstemming is tussen annotatoren. Deze overeenstemming duidt aan dat mensen met een bepaalde structuur markeren. Deze structuur is getracht vast te leggen in regels. Met behulp van ROUGE is nogmaals gekeken naar alle annotaties van de documenten én van de documenten die zijn gemarkeerd met de regels. Nu bleek dat de documenten die waren geannoteerd met behulp van de regels even hoog, of soms zelfs hoger scoorden als de menselijke annotaties.

