ACTA UNIVERSITATIS UPSALIENSIS
Studia Linguistica Upsaliensia
4

EVALUATING READABILITY ON MOBILE DEVICES

BY

GUSTAV ÖQUIST

ACTA UNIVERSITATIS UPSALIENSIS
UPPSALA 2006

Dissertation presented at Uppsala University to be publicly examined in Sal X, Universitetshuset, Uppsala, Saturday, December 16, 2006 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Öquist, G. 2006. Evaluating Readability on Mobile Devices. Acta Universitatis Upsaliensis. *Studia Linguistica Upsaliensia* 4. 80 pp. Uppsala. ISBN 91-554-6745-8.

The thesis presents findings from five readability studies performed on mobile devices. The dynamic Rapid Serial Visual Presentation (RSVP) format has been enhanced with regard to linguistic adaptation and segmentation as well as eye movement modeling. The novel formats have been evaluated against other common presentation formats including Paging, Scrolling, and Leading in latin-square balanced repeated-measurement studies with 12-16 subjects. Apart from monitoring Reading speed, Comprehension, and Task load (NASA-TLX), Eye movement tracking has been used to learn more about how the presentation formats affects reading.

The Page format generally offered best readability. Reading on a mobile phone decreased reading speed by 10% compared to reading on a Personal Digital Assistant (PDA), an interesting finding given that the display area of the mobile phone was 50% smaller. Scrolling, the most commonly used presentation format on mobile devices today, proved inferior to both Paging and RSVP. Leading, the most widely known dynamic format, caused very unnatural eye movements for reading. This seems to have increased task load, but not affected reading speed to a similar extent. The RSVP format displaying one word at time was found to reduce eye movements significantly, but contrary to common claims, this resulted in decreased reading speed and increased task load. In the last study, Predictive Text Presentation (PTP) was introduced. The format is based on RSVP and combines linguistic chunking and adaptation with eye movement modeling to achieve a reading experience that can rival traditional text presentation.

It is explained why readability on mobile devices is important, how it may be evaluated in an efficient and yet reliable manner, and PTP is pinpointed as the format with greatest potential for improvement. The methodology used in the evaluations and the shortcomings of the studies are discussed. Finally, a hyper-graeco-latin-square experimental design is proposed for future evaluations.

Keywords: Readability, Mobile devices, Text presentation, Usability, Evaluation, Adaptation, Eye movement tracking

Gustav Öquist, Department of Linguistics and Philology, Box 635, Uppsala University, SE-75126 Uppsala, Sweden

ISSN 1652-1366
ISBN 91-554-6745-8
urn:nbn:se:uu:diva-7378 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7378)

"Although to penetrate into the intimate mysteries of nature and thence to learn the true causes of phenomena is not allowed to us, nevertheless it can happen that a certain fictive hypothesis may suffice for explaining many phenomena"

- Leonhard Euler (1707-1783)

Swedish National Graduate School of Language Technology

Department of Linguistics and Philology, Uppsala University

Department of Clinical Neuroscience, Karolinska Institutet

Contents

Publications by the Author

Öquist, G. (forthcoming). Predictive Text Presentation: Using Linguistic Segmentation and Eye Movement Modeling to Improve Dynamic Text Presentation on a Mobile Phone. *Submitted for publication.*

Öquist, G. and Lundin, K. (forthcoming). Eye Movement Study of Reading on a Mobile Phone using Scrolling, Paging, Leading, and RSVP. *Submitted for publication.*

Öquist, G. (2007). Experiences and Findings from three Eye Movement Studies of Mobile Readability. In: *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, J. Lumsden (Ed.). Hershey, PA: Idea Group Publishing. *To appear.*

Öquist, G. (2006). Multimodal Interaction with Mobile Devices: Outline of a Semiotic Framework for Theory and Practice. In: *Proceedings of Wireless Networks and Systems 2006* (Setubal, Portugal), 276-283. Setubal: INSTICC Press.

Goldstein, M., Öquist, G., and Lewald, I. (2006). Evaluation of PreCodia, a Computerized Reading Aid for Readers Suffering from Dyslexia. In: *Proceedings of Human Factors in Telecommunication 2006* (Sophia-Antipolis, France), 127-134. Brighton, MA: IGI Group.

Öquist, G. (2004). Enabling Embodied Text Presentation on Mobile Devices. In: *Proceedings of Mobile and Ubiquitous Information Access 2004* (Glasgow, Scotland), 26-31.

Öquist, G., Sågvall-Hein, A., Ygge, J., and Goldstein, M. (2004a). Eye Movement Study of Reading Text on a Mobile Device using the Traditional Page and the Dynamic RSVP Format. In: *Proceedings of Mobile HCI 2004* (Glasgow, Scotland), 108-119. Berlin: Springer.

Öquist, G., Goldstein, M. and Chincholle, D. (2004b). Assessing Usability across Multiple User Interfaces. In: *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces*, A. Seffah and H. Javahery (Eds.), 327-349. New York, NY: John Wiley & sons.

Öquist, G. and Goldstein, M. (2003). Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation. *Interacting with Computers*, 15(4), 539-558.

Goldstein, M., Öquist, G. and Björk, S. (2002). Evaluating Sonified Rapid Serial Visual Presentation: An Immersive Reading Experience on a Mobile Device. In: *Proceedings of User Interfaces for All 2002* (Paris, France), 508-523. Berlin: Springer.

Öquist, G. and Goldstein, M. (2002). Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation. In: *Proceedings of Mobile HCI 2002* (Pisa, Italy), 225-240. Berlin: Springer.

Öquist, G., Goldstein, M. and Björk, S. (2002). Utilizing Gaze Detection to Stimulate the Affordances of Paper in the Rapid Serial Visual Presentation Format. In: *Proceedings of Mobile HCI 2002* (Pisa, Italy), 378-381. Berlin: Springer.

Öquist, G. (2001). *Adaptive Rapid Serial Visual Presentation*. Masters Thesis in Computational Linguistics, Department of Linguistics, Uppsala University, Sweden.

Goldstein, M., Öquist, G., Bayat-M, M., Björk, S. and Ljungberg, P. (2001). Enhancing the Reading Experience: Using Adaptive and Sonified RSVP for Reading on Small Displays. In: *Proceedings of IHM-HCI 2001* (Lille, France), 12-21. Toulouse: Cépaduès-Editions.

List of Tables

List of Figures

Figure 1. Courtesy of U.S. NIH National Eye Institute

Figure 2. Unknown creator, reprinted under the fair use doctrine

Figure 4. Reprinted from Journal of Psycholinguistic Research, 31(1), Braze, D., D. Shankweiler, W. Ni, and L. Palumbo. Readers' Eye Movements Distinguish Anomalies of Form and Content, p. 36, Copyright (2002), with kind permission of the author and Springer Science and Business Media.

Figure 6. Reprinted from Multiple User Interfaces. Edited by A. Seffah and H. Javahery, Assessing Usability across Multiple User Interfaces, Öquist, G., Goldstein, M. and Chincholle, D., p. 328, Copyright (2004), with kind permission of John Wiley & Sons.

Figure 12-14. Reprinted from Interacting with Computers, 15(4), Öquist, G. and Goldstein, M., Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation, pp. 539-558, Copyright (2003), with kind permission of Elsevier.

Figure 15, 17. Reprinted from Mobile Human-Computer Interaction - Mobile HCI 2004, Lecture Notes in Computer Science, 3160, Edited by S. Brewster and M. Dunlop, Eye Movement Study of Reading Text on a Mobile Device using the Traditional Page and the Dynamic RSVP Format, Öquist, G., Sågvall-Hein, A., Ygge, J., and Goldstein, M., pp. 108-119, Copyright (2004), with kind permission of Springer Science and Business Media.

Figure 20, 21. Reprinted with kind permission of the author (Danvall, 2004).

Figure 24-28. Reprinted with kind permission of the author (Lundin, 2006).

Abbreviations

fMRI	Functional Magnetic Resonance Imaging
CRT	Cathode Ray Tube
GLM	General Linear Model
GLS	Graeco Latin Square
GSM	Global System for Mobile Communications
GUI	Graphical User Interface
HGLS	Hyper Graeco Latin Square
MUI	Multiple User Interface
IR	Infra-Red
J2ME	Java 2 Micro Edition
LCD	Liquid Crystal Display
LIX	Läsbarhetsindex (Readability Index)
MANOVA	Multivariate Analysis of Variance
MS	Microsoft
NASA	National Aeronautics and Space Administration
PC	Personal Computer
PDA	Personal Digital Assistant
PET	Positron Emission Tomography
PTP	Predictive Text Presentation
RSVP	Rapid Serial Visual Presentation
SPSS	Statistical Package for the Social Sciences
TFT	Thin-Film Transistor
TLX	Task Load Index
WAP	Wireless Application Protocol
WPM	Words Per Minute

1 Introduction

"Language has always been the invisible partner of technology"

- Howard Rheingold

Reading and writing has played a fundamental role in the development of our culture. Compared to speech and other communication forms, written language is a fairly recent invention, yet it has had an immense impact as it offers us the possibility to store and share information over unprecedented distances in time and space. The first steps towards the literate society we live in today can be seen on clay bricks dating back to the Mesopotamian era around 7000 years ago. Writing was initially used as a method to keep account of commodities in trade, but later evolved into a general writing system for language (Schmandt-Besserat, 1996). The technologies used for mediation have evolved alongside the language, and at time also changed the language itself. The Sumerians initially used a stylus to carve symbols into the clay bricks. The bricks were later replaced by papyrus scrolls by the Egyptians who also developed an early alphabet. The Romans in turn replaced papyrus by parchments made from animal hides since the availability of papyrus at times was limited to them. Sets of parchments were found to be easier to handle if folded together into codices, the early handwritten books. Since it could take years for a scribe to make a copy of a codex, the availability of books was naturally scarce. Gutenberg's invention of the moveable type press facilitated mass production of written material on paper in the middle of the 15th century. However, it was not until the industrial revolution and the widespread availability of newspapers and cheap books in the middle of the 19th century that literacy reached the general public. The printing press may very well come to equal the invention of language in respect to its impact on society.

Today, we live in the midst of a new revolution. Information technology has radically changed both how information is shaped, how we work with it, and last but not least, how we gain access to it. Written language has ceased to be bound by the physical surface words are scribed upon and have transitioned into a virtual realm. Since almost all services and applications are

based upon text in one way or another, it is fundamental for the use of computers. With the introduction of the Internet, the greatest challenge in information access has become to find what you are looking for rather than how to gain access to it (Sahami et al, 2003). With the introduction of network connected mobile devices, such as mobile phones and Personal Digital Assistants (PDAs), any electronic text can be displayed on any screen, anywhere and anytime. The mobile Internet has for several years been predicted to be the next big thing in how we will access information, and the predictions are well founded. Today there are over 2.5 billion users of mobile phones globally; the figure is more than twice as large as the number of Internet users and growing at a rate of 40 million per year (Wireless Intelligence, 2006). In a few industrialized countries there are now more mobile phones than citizens, and in many developing countries, a mobile phone will probably be the first computational device that most people will come to own (GSM Association, 2006). However, regardless of the fact that mobile devices are readily available and most of them are network connected, the predicted success of the mobile Internet has not been realized so far.

There are many different reasons for the slow uptake of the mobile Internet. Service providers made a mistake by promising too much too soon when claiming that the Wireless Application Protocol (WAP) was the same thing as the Internet, which it never was and never will be. Producers of mobile devices made a mistake by not providing opportunities for third party developers to create applications. Companies have not been keen to invest in mobile solutions due to lacking standards and business models. Consumers have not thought of mobile phones as computers, but rather as phones for making calls. Developers, whom actually have thought of mobile phones as computers, failed to recognize that reusing desktop interaction methods might not be the ideal solution in a mobile setting. However, all of this is changing. Service providers now do offer access to the real Internet and third party developers can create applications; something that combined makes it possible for companies to build upon existing standards to extend existing, or invent new, business models. The companies can then market these services to consumers, which make them realize that their phone can be used for so much more than talk. Developers have begun to explore and utilize the novel interaction possibilities offered by mobile devices, but there is still much that remains to be done. Challenging how we do things today is the first step towards succeeding in doing it tomorrow.

When wireless phones were enhanced with computational functionality and became Internet connected they inherited the interaction methods and office metaphors of the direct manipulation paradigm, justifiably so since these tools already had proved to be extremely useful for interaction with computers (Schneiderman, 1982). Nonetheless, the usefulness of any tool is dependent on a combination of design and use, and these tools were never intended for interaction with devices based on a design derived from mobile

phones, moreover used in a nomadic environment. It is important to see that limited input and output capabilities due to smaller keyboards and screens are not optional, they are a fact since mobile devices have to be small to be mobile (Öquist et al., 2003). This thesis focuses on the conflict between how we traditionally present text, which requires a fairly large area to draw the text upon, and the limitation mobile devices put on the screen size available for this. One approach to overcome the size constraints may be to design interfaces that utilize the possibilities offered by mobile devices to dynamically work with the text and present it in a more suitable way for the user. Any such new presentation format must however still adhere to the principles for reading that has evolved over time. Moreover, to be able to see if the novel formats work, we need methods to empirically evaluate them in usability studies. The fact that readability has long been considered important as even small improvements can ease reading for large groups of people (Huey, 1908), has made the issues concerning readability on small screens progressively more important for mobile usability.

1.1 Aim of the Thesis

The aim of the thesis has been to bring forward efficient and usable text presentation formats for small screens, which adheres to the natural reading process that has evolved over time, by making the most of the opportunities for linguistic processing and novel interaction offered by mobile devices. In order to do this, eye movement tracking has been introduced as a tool in evaluation of text presentation on mobile devices.

1.2 Thesis Overview

Regardless of the device used for reading, or the format used for text presentation, the physiological and cognitive limits for reading remain the same. A natural starting point for this thesis may therefore be an overview of the reading process and a clarification of what we mean by readability. This is followed by an introduction to text presentation on small screens and the merits and pitfalls of the most common approaches. Next, previous evaluations of readability on mobile devices are presented and the methods used in these are discussed. Thereafter, the readability studies performed in the scope of this thesis are presented together with a comparative review of the results. The findings and the methods used to reach them are then discussed. Directions for future research are pointed out together with an improved experimental design, and finally, a few concluding remarks and a statement of contributions wrap up the thesis. The experimental designs that are discussed are available as a supplement on the last pages of the thesis.

2 The Reading Process

"You cannot depend on your eyes when your imagination is out of focus"

- Mark Twain

Reading is a skill that lies deeply embedded within in our mind. Researchers have proposed several different models of how the reading process works; some are highly detailed whereas others are more generalizing. Most agree that the process can be seen as a form of pattern recognition, but most also acknowledges that exactly how process works within the brain remains to be discovered (Reichle et al. 2000). Neuroimaging techniques such as Positron Emission Tomography (PET), and more recently Functional Magnetic Resonance Imaging (fMRI), have been able to show us where and when processing takes place during reading (Shaywitzet et al., 2006). This is important as it gives us a physiological understanding of the process, but it does not really answer the question about how it works. Since we do not know much about the processing, a better starting point for understanding how we read might be in the other end. By observing how the eyes move while reading we can tell how the recognition part works, if we do that we might also learn some about how the processing works as well. Next, we shall have a closer look at the eye and its physiology. This is followed by an overview of what we know about cognitive processing, and how we interpret eye movements in this respect. Finally, the concept of readability and how it lends itself to measurement will be discussed.

2.1 Physiological Limitations

The receptive part of the eye, called the retina, is essentially a panel full of photosensitive receptors located on the back of the eyeball (Ø ~42 mm) (Figure 1). The retina has two types of receptors, cones and rods. Cones register luminosity and colors whereas rods register light changes. Rods are much more sensitive to light, but they cannot detect colors and are also

slower to respond. Most of the cones are located in a tiny area at the centre of the retina called the fovea (Ø ~0,2 mm). The fovea is surrounded by the parafovea (Ø ~3 mm); in this region there are still many cones, but also an increasing amount of rods. Outside the parafovea there are few cones and a decreasing amount of rods, therefore vision becomes progressively less clear in the periphery of the retina (Procter and Procter, 1997).

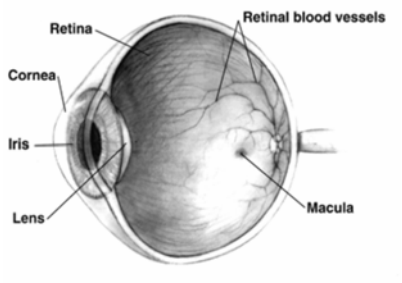

Figure 1 The eye

The fixation target must be projected on the fovea since a high concentration of cones is required for accurate recognition. Even though the retina has a 240-degree field of vision, the foveal field of vision is only 2-3 degrees wide which means that only 6-8 characters can be seen clearly in a single gaze (Robeck and Wallace, 1990). Moving centrifugally out from the fovea, the number of cones diminishes rapidly. The area immediately surrounding the fovea, the parafoveal region, further extends how much of the text that can be seen in a single fixation to around 12-14 characters (Robeck and Wallace, 1990), but beyond that the resolution is too low for recognition (Figure 2). The perceptual span is centered to the right of the fixation point, at least for readers of left-to-right languages (Just and Carpenter 1980).

Figure 2 The perceptual span

The effect of this on reading is that we have to move a very narrow focal point of vision across the text to be able to read it. Information is processed in fixations, e.g. the fixed gazes, with a duration ranging depending on what the target is. The durations have been found to vary greatly. In some studies it has ranged between 100-500 ms (Rayner, 1998), whereas it in others has

been found to vary between as much as 50-1500 ms (Just and Carpenter 1980). To move between fixations the eye performs very swift eye movements, called saccades, stretching up to 1-20 characters. The planning and execution of a saccade is based on the previous fixations and that which can be seen in the parafoveal region (Robeck and Wallace, 1990). For normal readers approximately every fifth saccade is directed backwards in the text, the reason for this is that the reader has to go back and reread a word or change position within a word. When reading a text on a page with a traditional layout like this, return sweeps are used to move between the lines and page sweeps are used to move between pages (Figure 3). You can experience this for yourself; just hold a fingertip lightly on top of an eyelid and you should be able to feel how the eye moves while you read.

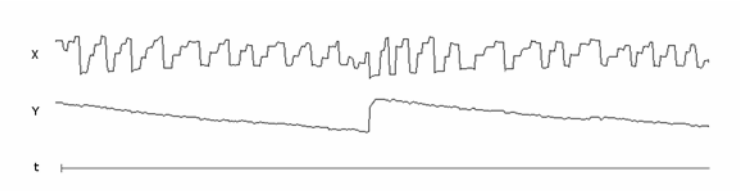

Figure 3 Time-plot of horizontal (X) and vertical (Y) eye movements over time (t) when a subject reads a text over two pages for ~30 s.

2.2 Cognitive Processing

An illuminating example of the complexities involved in the reading process is found in an empirical study performed by Braze et al. (2002). In their study, eye movement patterns from subjects reading sentences containing syntactic errors (form) and pragmatic errors (meaning) were compared to reading non-anomalous sentences. The aim with the study was to observe where and when regressions take place during reading the sentences and treat these as and indication of cognitive processing. The results showed that syntactic and pragmatic errors result in regression distributions that were distinctively different. Syntactic errors (cracking*) generated many regressions initially, with rapid return to baseline. Pragmatic errors (bite*) resulted in lengthened reading times, followed by a gradual increase in regressions that reached a maximum at the end of the sentence (Figure 4). These findings support the hypothesis that there exists a distinction between form (content) and meaning (context); moreover the study shows how eye movement tracking can be used to learn more about this difference.

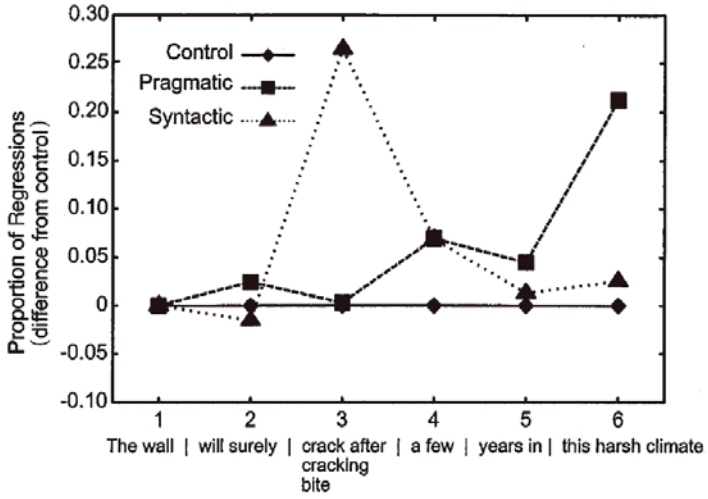

Figure 4 Proportion of regressive eye movements from each sentence region (1-6) as a difference from the control condition depending on a syntactic (cracking*) or pragmatic (bite*) error

The planning of saccades and the use of regressions for clarification seems to indicate that there is more to reading than meets the eye. The large differences observed in saccade lengths and fixation durations appear to reflect an ongoing process that changes depending on what is being read. What we know about the physiology of the eyes and their movements while reading seem to suggest that perception and recognition is highly dependent on cognitive (i.e. linguistic) processing. Eye movements can tell us surprisingly much about cognitive processing and most models of the reading process are based on empirical data of reading. The models that have been proposed can be roughly divided into either ocular motor or processing driven models (Reichle et al., 2000). The ocular motor models mostly look at the visual properties of the text (i.e. word lengths) and the physiological limits of the eye (i.e. perceptual span and saccade lengths) in order to determine the location and duration of fixations (Reichle et al., 2000). Ocular motor modeling has successfully been used to predict eye movements, but the models can never (and do not claim to) explain the whole reading process since they ignore the fact that language evidently has an impact on reading. The processing models on the other hand assign linguistic processing a very central role. The general assumption of these models is that the fixation duration is directly related to the cognitive processing whereas the fixation targets are determined by a combination of linguistic, orthographic and ocular motor factors (Reichle et al., 2000).

2.3 Modeling Reading

Just and Carpenter suggested that, "a reader can take in information at a pace that matches the internal comprehension process" (Just and Carpenter, 1980, pp. 329). From this starting point they developed Reader, the most widely known processing model. They began by observing actual gaze durations, the sum of all fixations on a word before moving to the next, made by college students reading scientific passages of text. Just and Carpenter found large variations in the duration of individual fixations as well as the duration of fixations on individual words. They also found that almost each content word was fixated and that fixation times were longer on words that were infrequent, thematically important or clarifying the interpretation of previous words. The gaze durations were also found to be longer at the end of a sentence thus indicating integrative processing. From these findings they founded their model on two assumptions. The first was the *immediacy hypothesis*, which state that each word is immediately processed when it is fixated. The second assumption is the *eye-mind hypothesis*, which state that the eyes remain fixated on a word as long as it is processed (Just and Carpenter, 1980). Both assumptions have later been criticized because they don't account for context and parafoveal preview effects (Robeck and Wallace, 1990).

The model presented by Just and Carpenter is very comprehensible but unfortunately it tries to explain the entire reading process, from fixation to long term-memory (Just and Carpenter, 1980). Although this made the model quite complex it is still disputed as it is assumed to simplify matters too much (Reichle et al. 2000). However, although the model might have tried to cover too much of the reading process it still has merits, simplifying a complex problem is not necessarily negative. If we combine Just and Carpenter's processing model with ocular motor modeling of the physiological limits of the eye and the visual properties of the text, we may get closer to a realistic definition. Fixation duration, i.e. determination of when, is governed by cognitive processing, while saccade execution, i.e. determination of where, is governed by a combination of linguistic, orthographic and ocular motor factors.

There are also other processing models available but these have had a minor impact on the work presented here. Rayner's E-Z reader is a model that is similar to the one presented by Just and Carpenter, but with a narrower scope since it does not try to account for high-level linguistic (e.g. semantic) processing. It does however account for preview and context effects (Rayner, 1988; Reichle et al., 2000). The model is unfortunately quite complex and the underlying assumptions are not as transparent as in the model presented by Just and Carpenter. Another processing model is the attention-shift model (Reilly, 1993). It utilizes two connectionist back-propagating neural networks, one for word recognition and one for planning saccades.

From a linguistic viewpoint the attention-shift approach seems a little too simple to be plausible, however the use a learning algorithm is appealing since individual differences in reading behavior are likely to be quite large.

2.4 Measuring Readability

Readability is typically referred to as the ease of "which the meaning of text can be comprehended" (Mills and Weldon, 1987, pp. 331). This is of course a very vague definition, but the assessment of readability is also affected by a multitude of factors. First, there are many differences between texts, some are very comprehensive and well written whereas others can be totally unreadable. Second, there are differences between readers; some are very experienced whereas others cannot read at all. Third, there are differences between reading situations, reading reference literature before an exam differs a lot from reading a novel while waiting for the bus. Fourth, there are differences between the presentation formats, this thesis might be comfortable to read on paper but is likely to be strenuous to read on a flickering screen with low resolution. To summarize: There are so many factors that affect readability that it is impossible to account for them all.

Since readability is hard to quantify, the solution is to use approximate measures instead. The readability estimations used in this thesis can be categorized according to their use as either ratings or measures. Ratings are used to determine readability of text based on quantitative predictions whereas measures are used to evaluate readability based on actual reader performance. Readability of text is usually rated by using readability formulas. Most readability formulas are quite simple and use a combination of word frequencies, word lengths and sentence lengths as a basis for the results. Although most formulas use purely quantitative measures they can give an indication of how hard or easy a text is likely to be to read. There are several readability formulas for English available (see Tekfi, 1987 for an overview), but for Swedish there is only one that is widely known. LIX (Läsbarhetsindex in Swedish) is a quantitative readability formula for developed by Björnsson (1968). An estimated value of the readability of a text is calculated on basis of the percentage of long words, seven or more characters, and the average sentence length. The result is a value between approximately 1 and 100 where lower values are interpreted as easier to read.

Readability has mostly been measured in terms of reading speed and comprehension (Mills and Weldon, 1987). Reading speed is often calculated as words read per minute (WPM) whereas comprehension is represented as percent of correctly answered questions. The reading speed results are mostly reliable when comparing results from different evaluations whereas the comprehension scores are unpredictable since they are highly dependant on the type of questions asked. The product of reading speed and compre-

hension scores are commonly used as a composite measure of reading efficiency (Jackson and McClelland, 1979; Rahman and Muter, 1999; Castelhano and Muter, 2001). The measure is used to avoid problems associated with assumed trade-offs between speed and comprehension (Wickens, 1992). However, since the comprehension scores are likely to be unreliable, both reading speed and comprehension must be reported separately as well if the results are to be comparable to other studies.

Since readability is an inherently subjective measure, subjective inventories have to be included in order to learn about the reading experience. The most widely used subjective measure is the attitude inventory. It is especially common to use when different text presentation formats are compared against each other. Attitude inventories are essentially a set of questions about experience and preference. Unfortunately the questions often differ between evaluations making it hard to compare the results; nonetheless they can be very illuminating for the evaluators. Another subjective measure used in evaluations, which actually is comparable, is the standardized NASA-TLX (Task Load Index) task load inventory (Hart and Staveland, 1988). The inventory is composed of six factors denoting cognitive demands that are rated by the subjects after completing a task. The inventory covers Mental, Physical, and Temporal demand, as well as perceived Effort, Frustration and Performance (Figure 5).

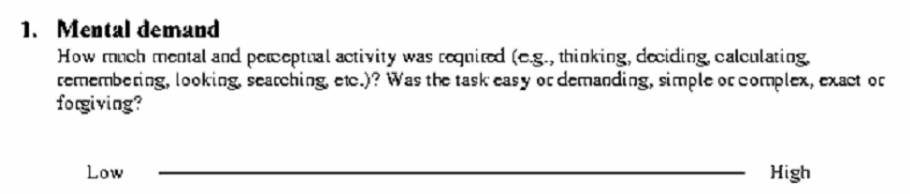

Figure 5 NASA-TLX (Task Load Index) rating scale for mental demand

What we really want to learn by measuring readability is to find the text presentation format that best support reading. A more accurate definition of readability that relates closer to how we actually read may thus be: "the ease with which the reading process can proceed" (Öquist et al., 2004a, p. 109). This is where tracking and analysis of eye movements becomes interesting as this can be used as an additional measure of how the reading process has proceeded. The fact that more difficult texts or a text read in a second language resulted in a significantly larger amount of regressions was one of the first findings made after the discovery of eye movements just now a century ago (Paulson and Goodman, 2000). Looking at eye movements and observing how they differ or conform to what we expect from reading is just as useful today as it was then since it is one of the very few objective measures of readability that we have available.

3 Reading on Mobile Devices

"Simplicity is the ultimate sophistication"

- Leonardo da Vinci

Until novel technologies such as folding screens or retinal projection prove to be viable solutions for mobile users, billions of people will read on screens with a size we see on mobile phones today. Since better readability results in increased usability, design guidelines for the improvement of readability based on empirical findings are important for mobile development. In this section, we begin by looking at mobile interaction and briefly review the challenges and opportunities that are offered for text presentation. Next, we will focus on the issues of presenting text on limited display areas by a look at how readability on desktop screens has evolved with the aim of finding parallels to the future of mobile displays. We will then have a look at the presentation formats most commonly used on mobile devices today. This is followed by a review of previous evaluations, and a few remarks on evaluation methodology.

3.1 Mobile Interaction

Mobile devices are essentially computers with much smaller form factors, e.g. smaller keyboards or screens. The result is limited input and output capabilities compared to desktop computers, but in return they are highly portable which makes them easy to bring with you. It is important to see that mobile devices differ from desktop computers, not only in regard to size but also to where and how they are used. Different devices are targeted for use in different locations, whereas different interfaces are targeted for use in different situations. But there is nothing that says that there must be a specific device for each location, nor that each interface will be used in the same way in each situation. The combination of location and situation creates different contexts of use, which are decisive of how usable different combinations of

devices and interfaces are going to be. By identifying the contexts of use, it is possible to identify which interfaces are useful in the situations and locations where an application is to be used. Weiss (2002) proposes a personal computing continuum ranging from desktop-laptop-palmtop to handheld, where portability increases as device size decreases (Figure 6).

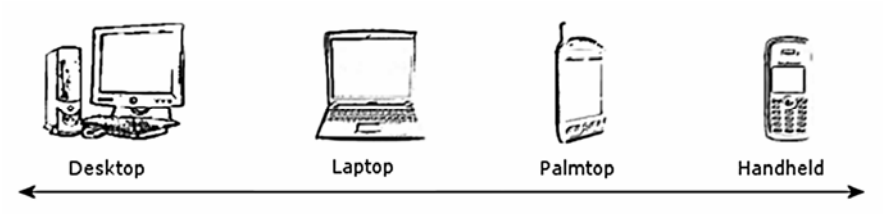

Figure 6 The personal computing continuum

In Öquist et al. (2004b), we used the personal computing continuum as a start point to outline a usability taxonomy for Multiple User Interfaces (MUIs). The taxonomy is based on four typical contexts of use: stationary (desktop), seated (laptop), standing (palmtop), and moving (handheld). Next, we characterized the different contexts of use by enumerating the four environmental factors that we believed affected usability most: portability, attentiveness, manageability, and learnability. For each of the contexts of use we could identify combinations of environmental factors that were indexical for different contexts of use. The last two of the contexts, the standing (PDA) and the moving (mobile phone), are those that pertain to the mobile devices used in this thesis.

When comparing the contexts of use, several factors distinguished the standing from the seated context. Portability increased, allowing a separation into two objects (e.g., a PDA and a stylus); moreover attentiveness has been reduced as the interfaces used in this context may form a secondary focus of attention. Manageability is more unstable, since the user holds the device with one hand and operates it with the other. The technological paradigm requires the most additional learning effort in this context (Cooper 1995); an idiomatic interface may be used (e.g. learnt once it's tried), as can a metaphorical interface (e.g. learnt by similarity), although its integrity can no longer be assured due to the degree of interface minimization. When we went from the standing to the moving context of use, the typical context of use for a mobile phone, all factors were found to change. Portability demanded that the device is a single artefact; attentiveness is minimal due to the fact that the user should now be able to move, and manageability is reduced to one hand and is therefore unbalanced. Furthermore, the paradigm of learnability is limited to the idiomatic. It may seem like that the more portable a device is, the less versatile it is. However, the more portable a device is, the more contexts can it be used in; which is a kind of versatility in itself

3.2 Text Presentation on Small Screens

Readability was a problem in the early days of computing. Reading speed was found to be 20-30% slower on screens although comprehension was roughly the same (Muter et al., 1982; Kang and Muter, 1989). These findings are not too surprising given that the first screens were primitive Cathode Ray Tube (CRT) units with low resolution and mediocre refresh rates. The designs of the early experiments have also been criticized mainly because the reading situations were quite unrealistic (Dillon 1992). The Achilles' heel of the first-generation large screens seems to have been the low resolution.

Screen technology evolved rapidly and the second-generation CRT screens offered far better resolution and also color. However, the breakthrough in readability, and usability in general, came with the introduction of the Graphical User Interface (GUI). Studies performed on computers with GUIs showed that there was in fact little or no differences between screen and paper, provided that attention was paid to such factors as screen resolution, refresh rates, anti-aliasing, text polarity, etc (Gould et al., 1987; Osborne and Holton, 1988; Muter and Maurutto, 1991; Muter, 1996). Although reading speed and comprehension does not differ much between high-quality screens and paper the users still seem to prefer reading on paper. This may be partially due to the fact that reading on a large screen requires the reader to view the text from a distance and in a fatiguing posture (Schneiderman, 1998). However, the screen must not necessarily be seen as a successor to paper but rather as a complement. There are many things that can be done with a text on a computer that is hard, or impossible, to do on paper. A common situation is also to browse for documents on the screen and then print the selected document on paper for reading. In this light the readability on the large screens of today seems quite satisfactory. However, reading on small screens is a different story.

Most mobile devices utilize flat Liquid Crystal Display (LCD) screens. The early LCD screens were monochrome and offered poor resolution, a bit like returning to a sized down version of an early CRT screen. However, LCD technology has evolved and today Thin-Film Transistor (TFT) technology offers a resolution and colour depth that is comparable or better than second-generation CRT screens. The problem with readability on small screens is not so much the resolution as it is the limitation of screen space, which restricts the amount of information that can be presented at a time. This implies a higher the rate of interaction by the user to view the text. Reading a longer text on a small screen can thus be frustrating and, to complicate matters further, users of mobile devices do not always have access to printing facilities.

Duchnicky and Kolers (1983) performed an experiment with varying window widths and heights on a desktop computer and found that a height of 20 lines only increased reading speed by a mere 9% compared to using a

height of 4 lines. Smaller window heights than 4 lines were however found to be significantly less efficient to use. The results also showed that a window width of 2/3 of a full page increased reading speed by 25% compared to using 1/3 of a full page. These widths are much larger than what the average mobile device has to offer, but the findings seem to suggest that a limited screen width decreases reading speed. Using a higher density of characters per line was also found to improve readability; using 80 characters compared to 40 increased reading speed by 30%. This is not very surprising given that a lower density implies less information in the perceptual span at a time and therefore also a lowered efficiency. Dillon et al. (1990) investigated how reading was affected by using window heights of 60 and 20 lines. The subjects who read using the smaller window height were found to perform significantly more jumps and also altered the direction of reading much more often, but the results showed that neither reading speed nor comprehension differed. This is not in line what Duchnicky and Kolers found, one reason for the difference may be that Dillon used much longer texts (~3000 words compared to ~350 words). Nonetheless, the results from both evaluations seem to suggest that a small screen space does imply a higher rate of interaction.

3.3 Text Presentation Formats

Since reading on computer screens is a recent innovation it is natural that the two most commonly used formats to present text, Scrolling and Paging, are adapted from how we traditionally have presented text on papyrus and paper. Computers do however offer us new ways to present text on screens by displaying it dynamically over time (Bruijn and Spence, 2000; Juola et al., 1995; Mills and Weldon, 1987; Rahman and Muter, 1999). The two most well know formats for doing this are Leading and Rapid Serial Visual Presentation (RSVP). Each of the four formats, Scrolling, Paging, Leading and RSVP, will now be presented in turn with focus on how the may be used on mobile devices.

3.3.1 Scrolling

Scrolling presents the text in the traditional format on a display area that may be larger than the screen. Scroll bars are usually used to indicate how much of the text that is displayed on the screen as well as current horizontal and vertical position. Using both horizontal and vertical scrolling for text presentation is neither efficient nor appreciated by users (Muter, 1996), the length of the lines are therefore often shortened to fit the screen so that only vertical scrolling is necessary. A joystick or a set of arrow keys can be used to move to a different position. In interfaces where a mouse or stylus is used, the

scrollbars may be used for navigation as well. On mobile devices, the text is usually vertically scrolled line by line. The number of lines is decided by the amount of text that can fit horizontally. The least number of interactions needed to read a text using is equal to the number of lines minus the number of lines initially displayed on the screen (Figure 7).

Figure 7 Scrolling implemented on a mobile phone

3.3.2 Paging

Paging presents the text in the traditional format, but divides it into pages that fit the screen area. A joystick or a set of arrow keys is used to move between pages. The current page number and the total number of pages are usually displayed to inform the user of the position in the text (Muter, 1996). The number of pages needed to present a text is a function of the line length, decided by how much that can fit horizontally, and the number of lines per page, decided by how much that can fit vertically. If either the line length or the number lines per page is small, as it usually is on mobile devices, even slight changes to either can dramatically change the number of pages. The least number of interactions needed to read a text using paging is equal to the number of pages minus the first page displayed. Compared to Scrolling the number of interactions is thus reduced by a factor of how many lines that fit per page on the screen (Figure 8).

Figure 8 Paging implemented on a mobile phone

3.3.3 Leading

Leading dynamically scrolls the text horizontally on one line across the screen. The text continuously moves across the screen at a certain speed that may be selected by the user. Moving the text pixel for pixel has been found to be more efficient than moving it character for character (Kang and Muter, 1989). A joystick or a set of arrow keys are used to start and stop the presentation, go forward and backward in the text, as well as increasing or decreasing the speed. A progress bar or completion meter may be used to indicate location in the text. The time required to read a text is decided by the speed of the text presentation. The size of the text display area does not affect the presentation speed, but text presented on a shorter line may be perceived as going faster. The least number of interactions needed to read a text using Leading is just one (or even none if the presentation starts automatically), but speed changes as well as interactions to go back and forward realistically adds to this number (Figure 9).

Figure 9 Leading implemented on a mobile phone

3.3.4 Rapid Serial Visual Presentation

Rapid Serial Visual Presentation dynamically presents the text in chunks of one or a few words at a time at a fixed location on the screen (Forster, 1970; Juola et al., 1982; Potter, 1984). The chunks are successively displayed at a pace that may be selected by the user (Muter 1996). Adapting the exposure time of each chunk to its length or frequency in language has been found to improve readability (Castelhano and Muter, 2001; Öquist and Goldstein, 2003). A joystick or a set of arrow keys are used to start and stop the presentation, go forward and backward in the text, as well as increasing or decreasing the speed. A progress bar or completion meter may be used to indicate location in the text and previous evaluations have shown that this is beneficial for the RSVP format (Rahman and Muter, 1999). The time required to read a text is decided by the speed of the text presentation, which usually is measured in words per minute. The size of the chunks affect the speed of the

presentation since smaller chunks requires a higher presentation pace. The time to read a text is however constant at a certain speed regardless of chunk size. In similar to Leading, the format requires a minimum of user interaction (Figure 10).

Figure 10 RSVP implemented on a mobile phone

Today most mobile phones use a joystick in combination with a set of soft keys to control the interface. The joystick is typically used to control the text presentation whereas the soft buttons are used to control the interface, e.g. switch between menus or alter settings (Figure 11). Utilizing accelerometers to control the presentation by tilting the device may prove fruitful in the future as it offers a more direct mode of interactivity (Öquist, 2004).

Figure 11 Button assignments for RSVP on a Sony Ericsson T610 mobile phone

3.4 Previous Evaluations

Most evaluations of the text presentation formats targeted for small screens have not actually been performed on mobile devices. The majority of them have moreover been directed towards exploring new possibilities to present text, usually by evaluating novel variations of the RSVP format. To make matters a bit more complicated, the implementations of the text presentation formats and the experimental designs vary considerably between experiments. It is therefore hard to compare a finding for one format reported in one evaluation to those achieved in another. Nonetheless, an overview of previous experiments may at least shed some light on what we have learnt so far about the text presentation formats in terms of readability.

Joula et al. (1982) presented shorter paragraphs of text on a CRT screen, either in the page format or in the RSVP format with text chunks of 5, 10 or 15 characters. Each text chunk was exposed for 200-300 ms, which is equal to a reading speed of approximately 300 WPM. The results showed no significant differences in comprehension between the reading conditions.

Masson (1983) evaluated how the insertion of blank windows at sentence boundaries affected the RSVP format. Masson experimented with durations of 500 and 1000 ms and found that performance increased with blank windows regardless of duration.

Cocklin et al. (1984) compared RSVP with the text divided into either idea units or ad-hoc chunks. The idea unit segmentation was performed by hand and was based on clause and phrase boundaries as well as linguistic features. Each chunk averaged 13 characters and the reading speed was approximately 300 wpm. The results showed that the use of idea units increased comprehension a little but not significantly.

Muter et al. (1988) performed experiments with self-paced RSVP and RSVP that permitted regressions. The results showed that larger regressions yielded slower reading and regressions back to the beginning of the sentence were found to be more frequent than regressions two words back. Overall the results indicated that permitting reader control was feasible but permitting regressions resulted in lower performance.

Kang and Muter (1989) compared RSVP to word-by-word, letter-by-letter and pixel-by-pixel Leading. Except for word-by-word, comprehension was as high for Leading as it was for RSVP. The comprehension scores for pixel-by-pixel leading were also found to match RSVP at reading speeds ranging from 100 to 300 WPM. The subjects in the evaluation were also found to express a significantly higher preference for pixel-by-pixel Leading.

Fine and Peli (1995) evaluated how visually impaired and elderly subjects read using RSVP and scrolled text. They found that the visually impaired read at a similar speed using both formats whereas the elderly read faster using RSVP.

Joula et al. (1995) compared Leading to RSVP on eight-character horizontal display. The results showed that sentences were read more accurately in the RSVP format than in the Leading format.

Rahman and Muter (1999) benchmarked word-for-word RSVP and sentence-by-sentence presentation, with or without a completion meter, to traditional text presentation in the page format. No significant differences were found for comprehension and reading speed but the subjects liked the inclusion of a completion meter.

Sicheritz (2001) compared reading using RSVP with three different text presentation window widths (11, 17 and 25 characters) on a PDA to reading in a paper book. The results showed that neither reading speed nor comprehension differed between the conditions. The NASA-TLX task load inventory did however reveal significantly higher task load ratings for the RSVP conditions for all factors but Physical demand. A 25-character window width was found to be more efficient, but the difference was not significant.

Castelhano and Muter (2001) evaluated the effects of using RSVP with or without punctuation pauses, variable word durations and a completion meter. They compared a few RSVP formats to traditional text presentation and sentence-by-sentence presentation. The results showed that pauses and variations made the RSVP format significantly more accepted. However, the sentence-by-sentence and traditional page format remained more popular although RSVP was just as effective.

Laarni (2002) compared reading using Scrolling, Paging, Leading and RSVP on several different screen sizes that were emulated on a desktop. The results showed that Scrolling and RSVP were the most suitable formats to use on mobile phones. RSVP was moreover found to be the fastest format when reading on a screen with the size of a mobile phone.

All the findings reported in the reviewed evaluations are valid in the sense that they are obtained scientifically. For each experiment, hypotheses are tested statistically using methods appropriate for their respective experimental design. All of the experiments are more or less repeatable; some of the software used in the experiments may be hard to come by. The main problem is however that none of the experimental designs are comparable to each other. This makes it very difficult to compare the results. The most com-

monly evaluated format in the experiments is RSVP, but not a single evaluation has implemented the format in the same way. There is little or no documentation in previous studies on exactly how the exposure times have been calculated. What is known however is that the exposure times have generally been fixed (e.g. every word or chunk has been displayed for the same time in relation to the set reading speed). All use words per minute (WPM) as a measure of reading speed, but it remains unclear if it is defined equally in terms of what the time includes. Only a few of the comprehension tests and the subjective inventories used are comparable.

The current state of affairs is understandable given that different researchers have performed most evaluations, but it is nonetheless regrettable. This does not mean that there is nothing to learn from the previous experiments, but it does mean that one has to be careful when drawing conclusions from a comparison of results. Moreover, it clearly illustrates the need for standardized evaluation measures and guidelines for evaluation of readability, whether it is performed on a mobile device or not.

4 Readability Studies

"Whether you can observe a thing or not depends on the theory which you use. It is the theory which decides what can be observed."

- Albert Einstein

In this chapter, the five readability studies that have been performed in the scope of this thesis are presented. We start with a discussion around the methodological problems that are related to study readability. The experimental design that is used in the studies is then explained. Next, each of the five studies is presented in turn.

4.1 Methodology

Evaluating readability is difficult. To start with, you need a representative number of subjects that read texts using the formats that you want to evaluate. How large the number must be is decided by the experimental design. Running readability studies is typically time consuming and designs that limit the number of subjects are usually desired. Moreover, there are a few things to keep in mind. Since the subjects cannot read the same text twice in the same way, you need one text for each text presentation format. Even if the difficulty of the texts is rated and found to be similar, they are not going to be equal. A text read in one format may for some reason fit better for that format, or it may just be that the questions on the comprehension test was easier for that text. Texts and presentation formats thus has to balanced. It may come as a surprise, but the single largest source of error in a readability evaluation can be the subjects. The difference between how persons read is generally much larger than the differences between text presentation formats or the texts themselves. To ensure that it really is the text presentation formats that are being evaluated; each subject has to read a different text using each presentation format. Finally, subjects actually do get tired when reading

even if it sounds like an easy task. The texts may not be that difficult, but the presentation formats that are being evaluated may be cumbersome to use. In order to get reliable results it is important to balance the order in which subjects read using the presentation formats.

In our experiments, we have used a within-subject repeated-measurement experimental design that meets all of the aforementioned criteria. The benefit of using such a design in readability studies is that it limits the effects of variance caused by the subjects reading performance on the results for the text presentation formats. Each subject reads a text using each of the formats, what we then look at is not how well the reader performed but rather how well the formats performed for that reader. When several measurements are taken on the same experimental unit, in this case the different presentation formats, the measurements tend to be correlated with each other. The correlations between formats can then be taken into account using a multivariate analysis of variance (MANOVA). We used the repeated-measurement General Linear Model (GLM) to test for significances. The significance (alpha) level was set to 5%. Since several hypotheses are tested simultaneously, the level of multiple comparisons has been Bonferroni adjusted (Bonferroni, 1935) (e.g. the standard of proof needed is heightened by dividing the alpha level by the number of factors).

To limit the number of subjects we employed a graeco-latin-square (GLS) design. Swiss mathematician Leonhard Euler introduced latin-squares in 1782 as a "une nouvelle espèce de quarrès magiques", a new kind of magic squares (Euler, 1782). A latin-square is a table with $n*n$ cells where every element occurs exactly once in each row and column. Common examples of latin-squares are all solutions to a Sudoku puzzle. A graeco-latin-square (also called Euler square after the inventor) is a latin-square of two sets of n elements, S and T, ordered in a $n*n$ table so that each cell contain an ordered set <s, t> and no row or column contains more than one s or one t (Box et al, 2005). If we let the formats be S = {A, B, X, Δ} and the texts T = {α, β, χ, δ}, we can create a GLS of experimental conditions (Table 1).

Table 1 Graeco-latin-square (GLS) for presentation formats and texts

Experimental conditions			
Format A / Text α	Format B / Text β	Format X / Text χ	Format Δ / Text δ
Format B / Text β	Format X / Text χ	Format Δ / Text δ	Format A / Text α
Format X / Text χ	Format Δ / Text δ	Format A / Text α	Format B / Text β
Format Δ / Text δ	Format A / Text α	Format B / Text β	Format X / Text χ

When evaluating readability we also have presentation order as a factor to take into account. Moreover, always reading the same text with the same format is not desired. To balance for this as well, the graeco-latin-square is randomized by transposition (e.g. putting the first column last and shifting

either S to T one row up or down, or vice versa) to create three additional fourth order squares. The result is a table of sixteen rows and four columns (Table 10). A subject can then be assigned to each row; the experimental conditions then become the ordered cells in that row. Using a graeco-latin-square design makes it possible to run a reliable experiment with four text presentation formats using sixteen subjects. This may sound like a small figure, and in fact it is, but the statistical model used for the experiment is intended just for such situations where it is impractical or expensive to run a large number of tests. A GLS experimental design has been used in all studies presented in this thesis. However, as we will soon see in the third study, tampering with the latin-squares can easily result in flawed results.

4.2 Experiments

Each of the five studies will now be presented in turn. Focus will be on the rationale behind the evaluations, the tools and methods involved in performing them, and of course, the key findings stemming from them.

4.2.1 Study one – Introducing Linguistic Adaptation

The aim with the first study was twofold, on the one hand we wanted to compare RSVP to other presentation formats where all conditions were performed on a mobile device, on the other hand we wanted to see if the RSVP format could be improved by using linguistic adaptation (Öquist, 2001; Öquist and Goldstein, 2002). In a previous experiment (Sicheritz, 2000), texts were read on a Personal Digital Assistant (PDA) using RSVP and a paper book. Neither reading speed nor comprehension was found to differ significantly. However, the NASA-TLX task load inventory revealed significantly higher task loads when using RSVP for most factors. One explanation to the higher task load may be the fact that the exposure times in previous RSVP implementations have been equal for all chunks given a certain speed although the reading speed actually varies (Just and Carpenter, 1980).

Adaptive RSVP (Goldstein et al., 2001; Öquist, 2001) attempts to mimic the reader's cognitive text processing pace more adequately by adjusting each text chunk exposure time in respect to the text appearing in the RSVP text presentation window. By assuming the eye-mind hypothesis (Just and Carpenter, 1980), i.e., that the eye remains fixated on a text chunk as long as it is being processed, the needed exposure time of a text chunk can be assumed proportional to the predicted gaze duration of that text chunk. Since very common, known, or short words are usually processed faster than infrequent, unknown or long words, the text chunk exposure times can be adjusted accordingly (Just and Carpenter, 1980). Further, most new information tends to be introduced late in sentences and therefore ambiguity and

references tends to be resolved there as well. A shorter sentence is also usually processed faster than a longer one since it conveys less information (Just and Carpenter, 1980). Thus, processing time differs both within and between sentences and the text chunk exposure times can therefore be adjusted accordingly as well.

On basis of these findings, two adaptive algorithms supposed to decrease task load were developed (Öquist, 2001). The first algorithm adapts the exposure time to the content of the text chunks whereas the second also looks to the context in the sentences. Both algorithms insert a blank window between each sentence if there is not enough space to begin on the next sentence in the same window, otherwise a delay is added to the sentence boundary instead. In content adaptive mode, the exposure time for each text chunk is based on the numbers of characters and words that are being exposed for the moment. Longer words are assumed to be more infrequent and take longer time to read than shorter words. A higher number of words are also assumed to take longer time to read and should thus receive more exposure time. The following formula is used to calculate the exposure time for content adaptation (Equation 1):

$$time_1 = (nwrd+nchr)/(davg*wpm/60) . \quad (\mathbf{1})$$

The formula uses the number of words (nwrd) and the number of characters (nchr) as a basis for the results. Both arguments are added and divided by the product of the average word length including delimiters (davg) and the currently set speed in words per minute (wpm) divided by 60. The result is a variable exposure time ($time_1$) depending on the content the current text chunk. The average word length including delimiters was set to 7,8 whereas the other variables were the same as those used for Fixed RSVP (Öquist and Goldstein, 2003).

In context adaptive mode the exposure time for each text chunk is based on the following: The result of content adaptation, the word frequencies of the words in the chunk and the position of the chunk in the sentence being exposed. To begin with, each word in the chunk is looked up in a lexicon with word frequencies. If the word is common it receives a weight lower than one (<1) and if it is rare or not in the lexicon it receives a weight higher than one (>1). The following formula is used to calculate how the exposure time is affected by the word frequencies (Equation 2):

$$time_2 = time_1 * ((wfrq_1+\ldots+wfrq_{nwrd})/nwrd) . \quad (\mathbf{2})$$

The formula uses the exposure time for content adaptation ($time_1$) and the word frequency weights for the words in the chunk (wfrq) as a basis for the

result. The word frequency weights are added and divided by the number of words in the text chunk (nwrd). The product is then multiplied with the content adaptive exposure time to get the weighted exposure time ($time_2$). The next step is to give the chunk less exposure time if it appears in the beginning of a sentence and more if it appears in the end. The following formula is used to calculate the text chunk exposure time depending on the position in and the length of the current sentence (Equation 3):

$$time_3 = (time_2 + time_2 * \tanh(swrd/savg))/2 \; . \qquad (3)$$

The formula uses the intermediary exposure time reached earlier ($time_2$), the number of words in the sentence exposed so far (swrd) and the average sentence length (savg). In order to get a smooth drop-off in speed along the sentence, a mean of the previously calculated exposure time and its product with the hyperbolic tangent (tanh) of the division of the number of exposed words and the average sentence length is calculated. The result is a varying text chunk exposure time ($time_3$). The average sentence length was set to 11,5 words and the word frequency weights ranged between 0,6-1,2. A lexicon with frequencies for the 10.000 most common words in Press 97, a corpus of 11,9 million words, was used to assign the weights according to a lognormal distribution (Öquist and Goldstein, 2003).

In order to evaluate the RSVP algorithms they had to be incorporated into a mobile device. Bailando was developed for the Compaq iPAQ 3630 Pocket PC, a small PDA with a touch sensitive high-resolution colour display, 240 x 320 pixels (57.6 x 76.8 mm), 0.24 mm dot pitch, 12-bit (4,096 colours) TFT LCD. It was important that the graphical interface was appealing and yet intuitive to use. The prototype had to give a professional impression since it was supposed to be compared to other professional applications for traditional text presentation. In Bailando, the text is presented at an area located slightly above the half of the screen. The text is presented at one single line that utilizes 2:3 of the screen width and the vertical alignment is similar to the text presentation area as a whole. Above the text presentation area there is a border for aesthetical reasons. Below the text presentation area there is an information area displaying the text title, the progress bar, and the current speed settings (Figure 12). The progress bar is included in order to support memory of spatial location while reading, as said earlier a completion meter has been found to increase the user preference for RSVP (Rahman and Muter, 1999).

Figure 12 The Bailando prototype on a Compaq iPAQ 3630 (left), view of the RSVP interface (right)

In the experiment we wanted to compare the adaptive RSVP formats to a non-adaptive variant (Fixed RSVP), as well as traditional text presentation formats. We moreover wanted to see how the reading differed for long and short texts. Two commercial programs were chosen for traditional text presentation, Microsoft Reader for long texts and Microsoft Internet Explorer for short texts. It would probably have been more experimentally sound to use a single program for all traditional text presentation, but it would not have been realistic. The foremost reason for including two different programs was their intended context of use; the MS Reader is custom made to present longer texts such as e-books (Hill, 2001) whereas the MS Explorer is designed to present shorter web content such as news articles (Figure 13).

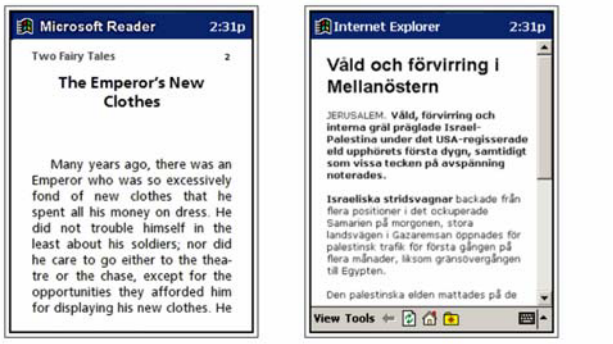

Figure 13 The MS Reader interface (left) and the MS Explorer interface (right)

The experiment took place in a dedicated usability lab outfitted with audio and video-recording facilities. While reading the subject was seated in a comfortable chair in a room separated from the experimenter by a one-way mirror (Figure 14).

Figure 14 Setup of the first study with experimenter (left) and subject (right)

A balanced within-subject repeated-measurement experimental design was employed for the experiment. Four experimental conditions were formed where each subject read one long and one short text using each presentation format. The combinations of long (A-D) and short (a-d) texts were fixed creating four text pairs Aa, Bb, Cc and Dd. Subjects thus always read the long text first and the short text afterwards. The text pairs were balanced against presentation format and order generating sixteen combinations. One subject was assigned to each of the sixteen sessions at random**.** No difference in Reading speed, Comprehension, Task load were set as a null hypothesis. The hypotheses were tested in the SPSS V10.0 software using the repeated-measurement General Linear Model (GLM). The significance level was set to 5% and the level of multiple comparisons was Bonferroni adjusted. Sixteen subjects (eight males and eight females; mean age: 25) participated in the experiment. Four long Swedish fiction texts of similar length (~4000 words) and difficulty (LIX ~31), and four shorter Swedish news text (~400 words) and difficulty (LIX ~47), were chosen to be included in the experiment. Comprehension was measured for long text by ten multiple-choice questions with three alternatives, for short texts there were five questions. The NASA-TLX inventory was used to measure Task load (Hart and Staveland, 1988). Reading speed was calculated as words read per minute based on the total time it took for the subjects to read a text including all kind of interruptions like pauses, regressions, speed changes etc.

The null hypothesis regarding no difference in reading speed between the conditions when reading short texts was rejected since the main factor for reading speed was significant ($F[3,45]=8.4$, $p<0.04$). Pair wise comparisons revealed that all RSVP conditions increased reading speed significantly ($p<0.002$) compared to using traditional text presentation with the MS Explorer (Table 2). The statistical analysis indicated no significant differences in Reading speed for long text. Comprehension was computed as percent of correctly answered multiple-choice questions. The differences between the conditions for comprehension were small for both long and short texts and the null hypothesis regarding no difference in comprehension for both was kept (Table 3).

Table 2 Results for Reading speed (WPM) in the first study

Condition	*Short texts*		*Long texts*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Explorer / Reader	157	53,2	242	80,4
Fixed RSVP	212	46,5	249	58,5
Content RSVP	213	36,8	260	51,2
Context RSVP	203	43,9	258	79,5

Table 3 Results for Comprehension (% Correct) in the first study

Condition	*Short texts*		*Long texts*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Explorer / Reader	70	21,9	73	19,6
Fixed RSVP	66	26,0	75	17,9
Content RSVP	59	19,9	76	17,5
Context RSVP	66	18,9	71	21,9

Task load was enumerated as percent of millimetres to the left of the tick mark on the NASA-TLX scale. The factors were not rated within each other so that the results would be comparable to Sicheritz (2000) findings. For short texts there were small differences in task load ratings and the null hypothesis regarding no differences was kept. For long texts there was a significant main effect ($F[3,45]=5.2, p<0.014$). Pair-wise comparisons revealed that the use of RSVP resulted in significantly higher ($p<0.014$) task loads compared to using traditional text presentation with the MS Reader for all factors but Physical demand. Content adaptive RSVP decreased task load ratings and the only factor that was rated significantly higher compared to the MS Reader was Frustration level ($p<0.002$). Context adaptive RSVP also decreased task load, but in a different way. The only significantly higher factor compared to the MS Reader was Temporal demand ($p<0.001$).

The task load ratings obtained for Fixed RSVP and the MS Reader were close to identical to those obtained for Fixed RSVP and paper-book in the Sicheritz evaluation (2000). Adaptive RSVP was supposed to decrease task load and it seems to have worked as expected for long texts. Compared to the MS Reader the only factor significantly higher for Content adaptation was Frustration level. Probably some words were not exposed for a duration that matched the time needed for cognitive processing; it is however encouraging that even the most straightforward form of adaptation actually decreased task load. In Context adaptive mode, the only significant factor compared to the MS Reader was Temporal demand. A probable cause for this is that the variations in exposure time were too large. However, the relation between what was exposed and the time for exposure was probably sound since the Frustration level decreased compared to Content adaptation. It seems that although the variations were too large they probably occurred

at the right places. Surprisingly, there were no significant differences in task load when reading short texts. When RSVP was used, the task load ratings were almost equal to using the MS Explorer although the reading speed was 33% higher. This confirms that traditional text presentation is neither a guarantee for low task load nor high reading speed and that RSVP actually can improve readability on mobile devices. In a follow up study we now wanted to learn more about how the formats affected reading.

4.2.2 Study two – Eye Movement Study on a PDA

The aim with the second study was to learn more about readability on small screens by analyzing eye movements. Using eye movements as a measure of readability connected well with the revised definition of readability, e.g. "the ease with which the reading process can proceed" (Öquist et al., 2004a, p. 109). In this experiment, we looked at readability in terms of comprehension score, reading speed, task load rating, and eye movements (Öquist et al. 2004). We wanted to compare the conditions that fared best in the previous evaluation, e.g. traditional text presentation in the Page format and dynamic text presentation in the RSVP format using adaptation (Öquist and Goldstein, 2003). The IOTA XY-1000 eye tracking system was used for eye movement detection and integrated with the Compaq iPAQ used for evaluation. The eye tracking system consists of a pair of goggles in which infrared (IR) diodes emit light onto the eyes (Ober, 1994). The IR reflections on the eyes are sensed by eight sensors, four for each eye, which may be sampled at a frequency of up to 1 kHz. The processing unit is connected to a PC running the Orbit eye trace program, which converts the eye movements into horizontal and vertical coordinates and records them. The benefit of the system is that it is reasonably comfortable to wear and can record eye movements with a high resolution as the sensors are located close to the eyes. A downside is that the recordings are sensitive to head movements (Figure 15).

Figure 15 IOTA XY-1000 system, goggles (left) and processing unit (right)

The result of a recording is a set of horizontal and vertical coordinates for the position of each eye over time. Before any recording can be made, the system has to be calibrated so that the recorded coordinates really correspond to the coordinates on the screen. To do this, a nine-point calibration

pattern was displayed on the PDA and the user was asked to look at each point in turn. Only when the coordinates recorded by the eye tracker agreed with the coordinates actually looked upon could the recording start. Next, the system had to be aligned so that a known horizontal and vertical distance on the screen was available in the recording; animating a four point cross on the mobile phone with a known distance between each point did this. Given that the distance between these points is known and the distance from the eyes to the screen is known, it is possible to calculate the position of each eye on the screen for the duration of the recording. In order to be able to calibrate and align the system with the text presentation on the PDA, a program was developed that automatically sets up an eye movement recording session, maintains synchronization with the mobile device, and enables monitoring of the recording throughout the session (Figure 16).

Figure 16 Calibration interface (left), alignment pattern (middle), XY-plot of resulting eye movements (right)

For text presentation, Microsoft Reader was used for the Page condition whereas Bailando (Öquist and Goldstein, 2003) was used for the RSVP condition. The experiment took place in a dedicated eye movement laboratory. All subjects were instructed to read at a pace that was comfortable to them and they were allowed the presentation speed at any time. While reading, the subject was seated in a comfortable chair with the head held in a fixed position by an adjustable kin support. Although this is not a very natural reading position, realism had to be sacrificed for reliable experimental data. The experimenter was seated near the subject and monitored the recordings (Figure 17). To limit the amount of data generated by the eye tracker the sampling rate was set to 100 Hz. This is considerably lower than the system can handle, but given that reading a text takes a while, it would just be to cumbersome to deal with the data if a higher setting was used.

Figure 17 Setup of the second study with subject (left) and experimenter (right)

A balanced within-subject repeated-measurement experimental design was employed for the experiment. Two conditions were formed where each subject read one text using either presentation format. The conditions were balanced against presentation order and texts, thus generating four combinations, which each were repeated four times yielding sixteen experimental sessions. One subject was assigned to each of the sixteen sessions at random**.** No difference in Reading speed, Comprehension, Task load, and Eye Movements were set as a null hypothesis. The hypotheses were tested in the SPSS V11.5 software using the repeated-measurement General Linear Model (GLM). The significance level was set to 5% and the level of multiple comparisons was Bonferroni adjusted. Sixteen subjects (eight males and eight females; mean age: 28) participated in the experiment. Two Swedish fiction texts of similar length (~2500 words) and difficulty (LIX ~30) were chosen to be included in the experiment. Comprehension was measured for each text by ten multiple-choice questions with three alternatives, the NASA-TLX inventory was used to measure Task load.

The eye movement recordings were analyzed using the JR saccade detection program (Ygge et al., 1999). The program was used to single out movements in the recordings; eye movements were defined as continual changes in the recording with durations lasting more than 10 ms independently detected in each of the four channels (e.g. horizontal and vertical movements for both left and right eye). Using this threshold, anything else than the detected movements can be assumed to be a fixation. The movements were categorized according to their function when reading based on duration, velocity, amplitude, and co-occurrence as either: Saccades and Regressions (≤4 deg. without vertical movement), Forward and Backward sweeps (>4 deg. without vertical movement or ≤4 deg. with vertical movement), Stray sweeps (>4 deg. with vertical movement), and Eye blinks (peak values caused by opening and closing the eyelids). The number of movements for each category was then normalized in respect to the length of the recordings into type of movements per minute.

The statistical analysis indicated no significant differences in Reading speed or Comprehension (Table 4). However, the null hypothesis regarding no difference in Task load between the conditions was rejected as there was a significant difference ($F[1,15] \geq 25.4$, $p \leq 0.001$). Pair-wise comparisons revealed that the use of RSVP format resulted in significantly higher ($p \leq 0.001$) Temporal demand compared to using the Page format.

Table 4 Results for Reading speed and Comprehension in the second study

Condition	*Reading speed (WPM)*		*Comprehension (% correct)*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Page format	216.9	78.7	78.1	14.7
RSVP format	191.9	45.1	74.4	20.0

The most striking differences were found in the eye movement recordings (Figure 18). RSVP was found to significantly increase the number of Regressions ($p \leq 0.001$), although it also decreased the number of Saccades significantly ($p \leq 0.006$). These findings were interesting since the advantage of the RSVP format originally was presumed to be the elimination of eye movements, which would lead to a possible reduction in cognitive load (Potter, 1984). The results show that the RSVP format does not eliminate eye movements, although it does reduce them. The reduction does however not seem to reduce cognitive load, it rather seems to increase cognitive load. The reason for this may be the increase in regressions, which can be seen as an indication of when the reading process has not proceeded with ease.

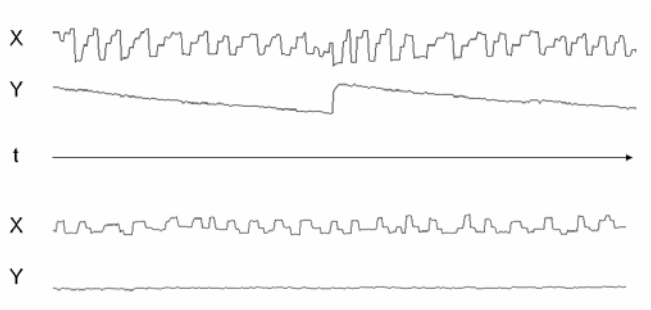

Figure 18 Time-plot of a ~30 s. excerpt of eye movements for Paging (top) and RSVP (bottom)

These empirical findings contradict the theoretical basis of RSVP, which means that we may have to reconsider the format. From these finding, it was suggested that a dynamic text presentation format like RSVP maybe shouldn't try to reduce eye movements, but rather try to stimulate an eye movement pattern similar to when reading in the Page format.

4.2.3 Study three – Verifying the Results

The primary aim with the second study was to verify that the RSVP format really could not eliminate eye movements (Danvall, 2004). The implementation we had used displayed as many words as could be fitted into a chunk of 25 characters. The decision to use this chunk size was based on the findings in Sicheritz (2000) experiment as it showed that this was more efficient than using smaller chunks. However, the most common implementation of RSVP displays one word at a time centered on the screen. For this experiment we thus developed a one word RSVP implementation that could be compared to the previous implementation. As a mean to validate the results from the previous evaluation, we also chose to include Paging in this experiment. Since the latin-square experimental design allows for four different formats, we also had an opportunity to try out a new idea. In the previous readability studies performed using Bailando on the PDA (Öquist and Goldstein, 2003; Öquist et al., 2004), some of the subjects had suggested that they would like to use a RSVP format displaying more lines than one. Therefore we implemented Buffered RSVP in which three chunks of 25 characters are displayed on the screen stacked upon each other with the most recent line at the bottom. The new formats were incorporated into the Bailando prototype; eye movements were recorded and analyzed using the same hardware and software as in the previous study (Figure 19).

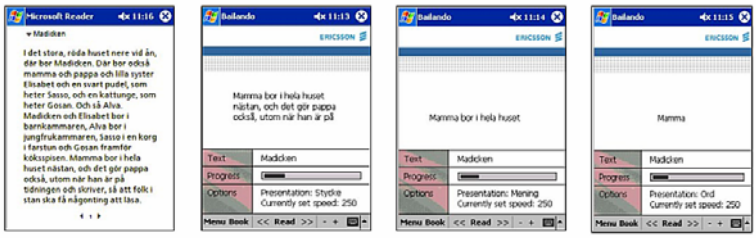

Figure 19 Text presentation formats evaluated in the third study: Paging (leftmost), Buffered RSVP (left), Chunked RSVP (right), and Word RSVP (rightmost)

A balanced within-subject repeated-measurement experimental design was employed for the experiment. Four conditions were formed where each subject read one text using either presentation format. The experiment was performed as a thesis project (Danvall, 2004) and to reduce the number of subjects required, the author of this thesis decided to let all subjects read using the Page format first. Only the RSVP formats were balanced against presentation order and text according to a latin-square design (sic!). This gave us twelve experimental sessions to which a subject was randomly assigned. No difference in Reading speed, Comprehension, Task load, and Eye Movements were set as a null hypothesis. The hypotheses were tested in the SPSS

V13.0 software using the repeated-measurement General Linear Model (GLM). The significance level was set to 5 % and the level of multiple comparisons was Bonferroni adjusted. Twelve subjects (six males and six females; mean age: 26) participated in the experiment. Four Swedish fiction texts by Astrid Lindgren of similar length (~1000 words) and difficulty (LIX ~30) were chosen to be included in the experiment. Comprehension was measured for each text by five multiple-choice questions with three alternatives, the NASA-TLX inventory was used to measure Task load (Hart and Staveland, 1988). The eye movement recordings were analyzed using the same software and metrics as in the previous study (Figure 20).

Figure 20 Setup of the second study, subject (left) and the experimenter (right)

The statistical analysis showed that the null hypothesis for Reading speed could be rejected since there were significant differences ($F[3,33]=7{,}787$, $p<0{,}001$). Pair-wise comparisons showed that the Page format was read significantly faster compared to Chunked RSVP ($p < 0.033$) and Word RSVP ($p < 0.010$). There were no significant differences in Comprehension between any of the formats (Table 5).

Table 5 Results for Reading speed and Comprehension in the third study

Condition	*Reading speed (WPM)*		*Comprehension (% correct)*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Page format	239,9	48,9	93,3	9,8
Buffered RSVP	210,2	46,9	96,7	7,8
Chunked RSVP	201,8	44,0	91,7	13,4
Word RSVP	177,0	41,1	100,0	0,0

The null hypothesis for Task load could be rejected since there were significant differences ($F[3,33]=3{,}317$, $p<0{,}032$). Pair-wise comparisons revealed significances for all factors. Mental demand was significantly lower for the Page format compared to all RSVP formats; Buffered ($p<0{,}010$), Chunked ($p<0{,}012$), Word ($p<0{,}011$). Physical demand was significantly lower for the Page format compared to the Buffered RSVP format ($p<0{,}011$). Temporal demand was significantly lower for the Page format compared to both Buff-

ered RSVP ($p<0.013$) and Chunked RSVP ($p<0.002$); Word RSVP was moreover rated significantly lower than Buffered RSVP ($p<0.002$). Performance was rated significantly higher for the Page format compared to the Buffered RSVP format ($p<0.008$). Effort was rated significantly lower for the Page format compared to Buffered RSVP ($p<0.001$), Chunked RSVP ($p<0.002$), and Word RSVP ($p<0.027$). Frustration was rated significantly lower for the Page format compared to Buffered RSVP ($p<0.001$), Chunked RSVP ($p<0.004$), and Word RSVP ($p<0.017$) (Figure 21).

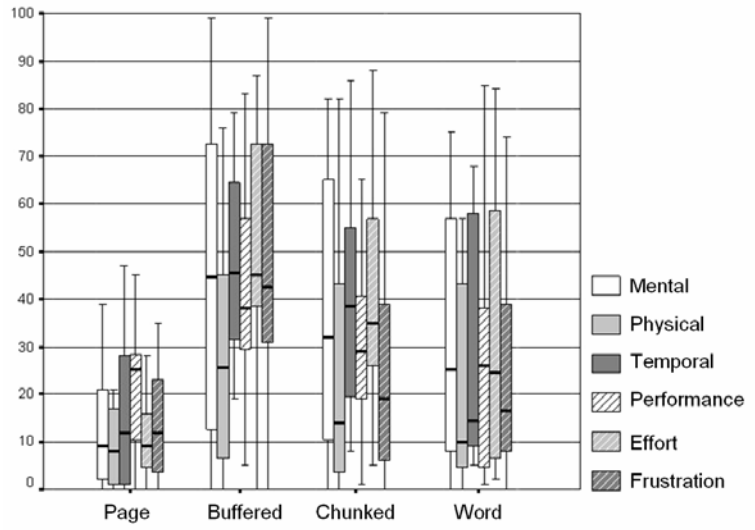

Figure 21 Box-plot of NASA-TLX task load ratings in the third study

The results for Reading speed and Task load were not inline with previous findings (Öquist and Goldstein, 2003; Öquist et al., 2004). The Page format was now significantly faster and significantly less demanding for most factors compared to the Chunked RSVP format although the implementations were exactly the same. This is probably a result of the flawed experimental design. If we had balanced all formats against presentation order according to the latin-square design, something that only had required four more subjects; the findings might have been different. The comparisons between the RSVP implementations are more reliable. The Buffered RSVP format that subjects had wished for in earlier experiments was probably not so good since the task load ratings were so high, interestingly enough it was quite fast however. The Chunked RSVP format was rated more demanding to use than Word RSVP, a quite interesting finding given that we set out to disprove the RSVP format. Now, could RSVP also eliminate eye movements?

Figure 22 Plot of ~30 s. of eye movements for the same subject superimposed over the presentation formats

The eye movement analysis showed that Word RSVP almost could eliminate eye movements (Figure 22). The findings for the Page format and the Chunked RSVP format were inline with the results from study two (Öquist et al., 2004). The Buffered RSVP format gave more vertical variation, but was otherwise similar to the Chunked RSVP format. The Word RSVP format significantly reduced the number of saccades and regressions compared to all the other formats ($p<0.001$) (Table 6).

Table 6 Saccades and regressions per minute for left and right eye in the third study

Condition	*Saccades per minute*		*Regressions per minute*	
	Right eye	Left eye	Right eye	Left eye
Page format	73,6	74,2	19,3	15,2
Buffered RSVP	40,5	40,8	36,3	33,7
Chunked RSVP	43,7	44,1	40,7	36,5
Word RSVP	0,6	0,9	0,9	0,7

The results from this and the previous studies show that the Page format in Microsoft Reader used on a PDA is hard to beat. The RSVP format has been, with a few exceptions, equally efficient in terms of reading speed and comprehension. The problem seems to be the extra task load induced by the format. Using Chunked RSVP results in a Saccade/Regression ratio close to 1:1 as opposed to 5:1 on paper, this may be a partial explanation as this hardly is a natural way to read. Using Word RSVP more or less eliminates the need for eye movements, but this does not seem to increase reading speed or reduce task load. Probably we really need to reconsider the format. Maybe a dynamic text presentation format like RSVP should try to stimulate an eye movement pattern more similar to reading on paper. Probably we really need to reconsider the format. Maybe a dynamic text presentation format like RSVP should try to stimulate an eye movement pattern more similar to reading on paper? However, there was also one more issue that had to be addressed before we could try this out. All our evaluations so far had been performed on a PDA with a screen much larger screen than those typically used on mobile devices. How would the RSVP format work out compared to the Page format on a much smaller screen?

4.2.4 Study four – Eye Movement Study on a Mobile Phone

The aim with the fourth study was to compare traditional text presentation to dynamic text presentation on a mobile phone (Öquist and Lundin, 2006). For traditional text presentation we choose to include Scrolling and Paging as these are the formats most commonly used on mobile devices. For dynamic text presentation we included Leading and RSVP, as these are the most commonly known dynamic formats. Since Word RSVP is the most common implementation we used that version of the format. To be able to evaluate the readability of using the four text presentation formats on a mobile phone and measure eye movements, a Java 2 Micro Edition (J2ME) application was developed and integrated with the eye movement tracker used earlier. The intention with the application was to keep as many aspects of the text presentation equal as possible except for the format used. The formats were moreover supposed to be generic in the sense that they should be representative of how they usually are implemented. A Sony Ericsson K750i mobile phone was chosen for the experiment since it supported J2ME and had a screen with a size typical for new mobile phones, 176 x 220 pixels (28 x 35 mm), 0.158 mm dot pitch, 18-bit (262,144 colors) TFT LCD (Figure 23).

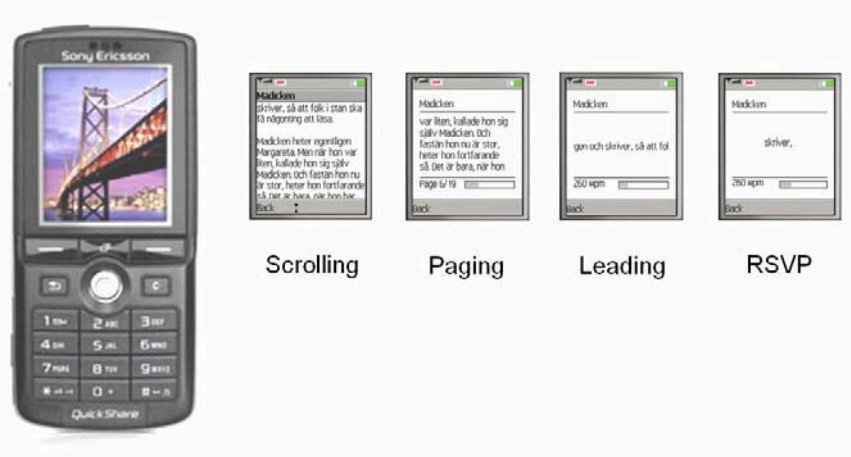

Figure 23 The Sony Ericsson K750i mobile phone used in study four (left) and the presentation formats in the same scale (right), see figure 4-7 for close-ups

For Scrolling, the native text presentation interface offered by the phone was used (20 characters x 8 lines), the joystick or the keypad could be used to scroll up or down. For the other presentation formats custom canvas interfaces were developed using the same font (Times New Roman 10 pixels) and screen settings (black on white). Each of the custom formats displayed the text title, a progress bar, and page numbers or speed settings. The Page format displayed five lines at a time and the joystick or the numeric keypad

was used to flip pages. The Scrolling format displayed the text at one line in the middle of the screen and moved the text pixel for pixel. The RSVP format presented one word at a time centered on the screen; the exposure time was calculated using the content adaptive RSVP algorithm (Öquist and Goldstein, 2003). For Leading and RSVP the joystick or the numeric keypad was used to control presentation speed (up/down), going backward and forward in the text (left/right), or starting and stopping (joystick press). The initial presentation speed for the dynamic formats was always set to 250 wpm. An updated XY-1000 system offering better resolution was used in this study. Calibration was manual using a nine-point pattern on the phone whereas the alignment process was automated (Öquist and Lundin, 2006).

A balanced within-subject repeated-measurement experimental design was employed for the experiment. Four conditions were formed where each subject read one text using either presentation format. The experiment was performed as a thesis project (Lundin, 2006), but wise from experience we did not try to reduce the number of subjects this time. All presentation formats were balanced against presentation order and text according to a latin-square design. This gave us sixteen experimental sessions to which one subject was randomly assigned. No difference in Reading speed, Comprehension, Task load, and Eye Movements were set as a null hypothesis. The hypotheses were tested in the SPSS V14.0 software using the repeated-measurement General Linear Model (GLM). The significance level was set to 5% and the level of multiple comparisons was Bonferroni adjusted. Sixteen subjects (eight males and six females; mean age: 26) participated in the experiment. The same Swedish texts used in the previous study were used in this one, but the questions in the Comprehension inventory was made more difficult. Comprehension was measured for each text by five multiple-choice questions with three alternatives, the NASA-TLX inventory was used to measure Task load (Hart and Staveland, 1988). The eye movement recordings were analyzed using the same software as in previous studies, but the analysis metrics had to be adjusted to match the updated eye tracker and the new experimental setup (Figure 24).

Figure 24 Setup of the fourth study, subject (left) and experimenter (right)

The statistical analysis showed that the null hypothesis for Reading speed was rejected since there was a significant main effect ($F [3,45] = 28.35$, $p < 0.001$). Pair-wise comparisons showed that RSVP reduced reading speed significantly compared to all other formats ($p<0.002$) and that the Page format increased reading speed significantly compared to the Scrolling format ($p < 0.002$). The null hypothesis regarding no difference in Comprehension was kept (Table 7). The null hypothesis for task load between the conditions was rejected ($F[3,45]=4.26$, $p<0.010$). Pair-wise comparisons showed that Mental demand was rated significantly higher for the Leading format compared to the Paging format ($p<0.009$). Physical demand was significantly higher for the Scrolling ($p<0.003$) and Leading formats ($p<0.005$) compared to RSVP. Temporal demand was significantly higher for the Leading and RSVP formats compared to the Scrolling ($p<0.001$, $p<0.013$) and Paging formats ($p<0.001$, $p<0.011$). Finally, Effort was found to be significantly higher for the Leading format compared to the Scrolling ($p<0.049$) and Paging ($p<0.003$) formats.

Table 7 Results for reading speed and comprehension from the fourth study

Condition	*Reading speed (WPM)*		*Comprehension (% correct)*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Scrolling	178,1	60,1	92,5	12,4
Paging	217,7	70,7	87,5	14,4
Leading	195,2	56,4	88,8	16,3
RSVP	135,4	44,3	92,5	12,4

Unfortunately, four of the eye movement recordings were too distorted to be usable for analysis. To make matters worse, the distorted recordings were from four different subjects. The cause of the distortions was probably that the manual calibration process was too imprecise; other recordings for the same subjects were fine. Running additional subjects would have been an acceptable solution, but as both equipment and interfaces had been updated after the experiment we decided to focus on upcoming experiments instead. Since the latin-square design does not allow missing cases in the statistical analysis, only descriptive statistics can be offered for the eye movements analysed in this study. The recordings that did work presented us however with some interesting data (Table 8).

Table 8 Eye movements per minute (Std. dev.) for the text presentation formats in the fourth study

Condition	*Saccades*	*Regressions*	*Eye blinks*	*Distortions*	*N*
Scrolling	74,5 (17,9)	55,9 (20,6)	19,8 (11,8)	11,83 (2,2)	12
Paging	76,0 (27,2)	48,7 (16,7)	28,6 (15,1)	11 (3,5)	12
Leading	93,4 (23,9)	88,46 (53,9)	14,5 (12,9)	9,62 (3,0)	12
RSVP	20,8 (11,8)	22,96 (10,8)	3,9 (2,8)	7,89 (2,5)	12

Scrolling resulted in more vertical variations and more regressions than Paging, but yielded for most aspects fairly typical reading movements. The Paging format resulted in very typical reading eye movements. The number of eye blinks was however much higher than the other formats. The Leading format yielded much more eye movements than any of the other formats; it seems that the subjects followed the text in smooth pursuit. The RSVP format resulted in the least number of eye movements, but far from eliminated them (Figure 24-27).

Figure 25 Scrolling eye movements on a mobile phone

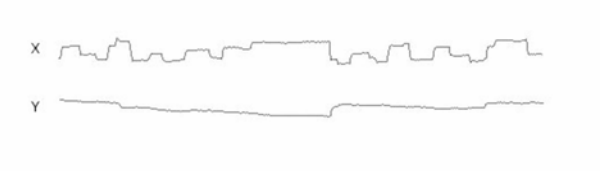

Figure 26 Paging eye movements on a mobile phone

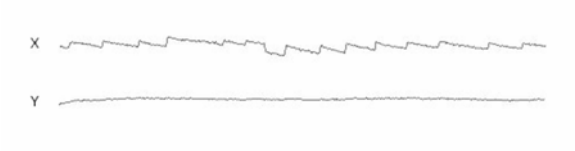

Figure 27 Leading eye movements on a mobile phone

Figure 28 RSVP eye movements on a mobile phone

For the Page format, The reading speed dropped by 25 % compared to the preceding experiment (178 vs. 239 wpm), interestingly the drop for Word RSVP format was similarly large 24 % (135 vs.177 WPM). Since the pre-

ceding experiment was not fully balanced, an educated guess would be that the decrease for the Page format would have been less if the experiment was properly balanced. Compared to study one, Reading speed for the Page format only dropped by 9 % (217 vs. 239 WPM) when used on a mobile phone as opposed on a PDA. This is interesting since this is exactly the same figure Duchnicky and Kolers (1983) found when reducing window heights from 20 to 4 lines on a CRT screen. Word RSVP on the mobile phone was 30 % slower than Chunked RSVP on the PDA (135 vs. 192 WPM). A decrease in speed is understandable for the Page format when used on a smaller screen; that the difference is smaller than 10 % is surprising.

The RSVP format was not expected to decrease at all in Reading speed since it should not be penalized by less screen space. One reason for the conflicting results is, yet again, probably stemming from the faulty design of study three. It may be the case that the bad results for the Buffered RSVP format tainted the results for the Chunked RSVP format. The Buffered RSVP format was really bad in terms of readability. Since the Chunked RSVP format was more similar to the Buffered format than the Word format it might have been read slower and received higher task load ratings.

4.2.5 Study five – Introducing Predictive Text Presentation

The aim with the fifth study was to see if an alteration of the RSVP format would improve readability on a mobile phone. Reducing eye movements to a minimum had using Word RSVP had not proved to improve the reading experience. The 25-character Chunked RSVP format used in the first two studies was probably not optimal either. The saccade/regression ratio was close to 1:1, something that we assumed increased task load ratings. However, since the 25-character chunk size actually is larger than what can be seen in the parafoveal region, this eye movement pattern might not be too strange. For this evaluation, two new formats were developed based on the limitations of the perceptual span (Figure 2). Adaptation of the exposure times depending on the content of the chunks offered an improved reading experience, what if linguistic segmentation of the chunk content and positioning of chunks based on eye movement modeling could do the same? Since the new formats do not have so much in common with how RSVP usually works, apart from being dynamic that is, we choose to call these Predictive Text Presentation (PTP) formats to make a distinction.

Two predictive formats were developed. The first is called Segmented PTP (S-PTP) and the other Moving PTP (M-PTP). The Segmented format was based on how much that could be seen in the parafoveal perceptual span (max 18: avg. ~15) characters, whereas the Moving format was based on how much that could be seen in the foveal span (max 9: avg. ~7) characters. Apart from the difference in size, the chunks were presented differently. The Segmented format always presented the chunks left justified on the screen

whereas the Moving format presented the chunks according to a simplistic eye movement model. The first chunk of a sentence was always presented leftmost on the screen; the next chunk was positioned so that 1/4 of the chunk would overlap the previous chunk, and so on until it's not possible to fit the next chunk in the display area. In this way the chunk moves across the screen from left to right in a saccadic pattern (as opposed to Leading where the text continuously moves from right to left) (Figure 29).

Figure 29 Moving PTP sequentially displaying four text chunks

For both formats, content size of each chunk was decided by linguistic segmentation. Just as in Cocklin et al. (1984), the segmentation was performed by hand and was based on clause and phrase boundaries as well as linguistic features. The main difference from ad-hoc chunking is that function words (such as the, on, in, before) are displayed in the same chunk as the content word (woman, time, progress, yesterday) more often. There are also a few other subtle differences that make the chunks more cohesive.

A balanced within-subject repeated-measurement experimental design was employed for the experiment. Four conditions were formed where each subject read one text using either presentation format. All presentation formats were balanced against presentation order and text according to a latin-square design. This gave us sixteen experimental sessions to which one subject was randomly assigned. No difference in Reading speed, Comprehension, Task load, and Eye Movements were set as a null hypothesis. The hypotheses were tested in the SPSS V15.0 software using the repeated-measurement General Linear Model (GLM). The significance level was set to 5% and the level of multiple comparisons was Bonferroni adjusted. Sixteen subjects (six males and ten females; mean age: 27) participated in the experiment. The same Swedish texts used in the previous study were used. Comprehension was measured for each text by five multiple-choice questions with three alternatives, the NASA-TLX inventory was used to measure Task load. The eye movement recordings were analyzed using the same software as in previous studies; the only difference was that the alignment sequence was longer so that an erroneous calibration could be spotted more easily (Figure 30).

Figure 30 Setup of the fifth study, subject (left) and experimenter (right)

The statistical analysis showed that the null hypothesis for Reading speed could be rejected as there was a significant main effect ($F[3,45]=13.51$, $p<0.001$). Pair-wise comparisons revealed that the Page format was significantly faster than the RSVP format ($p<0.003$). The Segmented PTP format was significantly faster than the RSVP format ($p<0.001$) and the Moving PTP format ($p<0.004$). The Moving RSVP format was moreover found to be significantly faster then the RSVP format. The null hypothesis regarding no difference in Comprehension between the formats was kept and the differences were small. The results for Reading speed were surprisingly good for both PTP formats. The results for word-for-word RSVP were more or less inline with findings from the previous studies. The Moving RSVP format was around 20% faster than RSVP, but the Segmented was in turn about 20% faster yet and was almost equal to the Page format (Table 9).

Table 9 Reading speed and Comprehension in the fifth study

Condition	*Reading speed (WPM)*		*Comprehension (% correct)*	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Paging	257,6	66,8	90	16,3
RSVP	184,1	48,4	90	12,6
S-PTP	254,4	60,0	90	21,9
M-PTP	217,8	40,0	96,3	8,1

The null hypothesis for Task load was rejected since there was a significant main effect ($F[3,45]\geq 2.89$, $p\leq 0.046$). Pair-wise comparisons showed that the Page format was rated significantly higher for Physical demand compared to Segmented PTP ($p<0.028$) and Moving PTP ($p<0.002$). Temporal demand was rated significantly lower for the Page format compared to RSVP ($p<0.001$), Segmented PTP ($p<0.024$), and Moving PTP ($p<0.002$). Finally, Effort was rated significantly lower for the Page format compared to RSVP ($p<0.026$) and Moving PTP ($p<0.042$). These findings are very interesting as the Page format and the Segmented PTP was almost equal in respect to Task load, for some factors Segmented PTP was even rated lower (Figure 31).

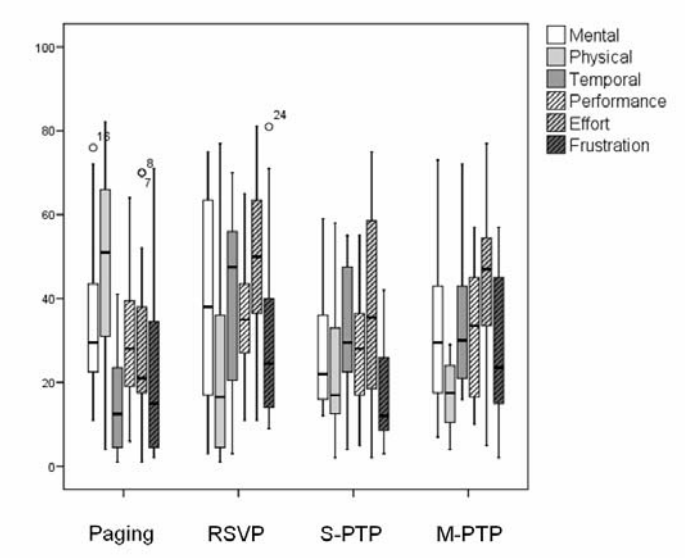

Figure 31 Box-plot of NASA-TLX ratings in study five (lower values are better)

The new calibration sequence seems to have worked better; at least all recordings contained an alignment sequence that could be used for analysis. A few of the recordings did however suffer from skewed calibrations. The consequence of this is that a horizontal movement also creates a simultaneous vertical movement. New software is currently under development that may be able to straighten the recordings. Instead of presenting results that may be erroneous, the analysis of eye movements is omitted for this study. Nonetheless, by just looking at a few snapshots of eye movements for the same subject using each format we may at least be able to make some informed speculations about what to expect from the analysis.

The eye movements resulting from reading using the Page format were similar to the recordings from study four. The lines on the mobile phone are short and usually one or two saccades suffice for reading (Figure 32).

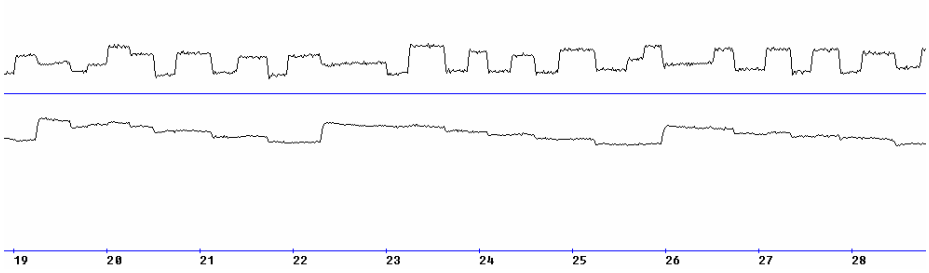

Figure 32 Paging eye movements in the fifth study

The eye movements resulting from reading using word-for-word RSVP were also similar, or actually more or less identical, to the recordings from study four. The eye movements are reduced to a minimum. At a few occasions there are small saccades or regressions, but in general the eye movements are more or less eliminated. This does however not seem to increase Reading speed or reduce Task load (Figure 33).

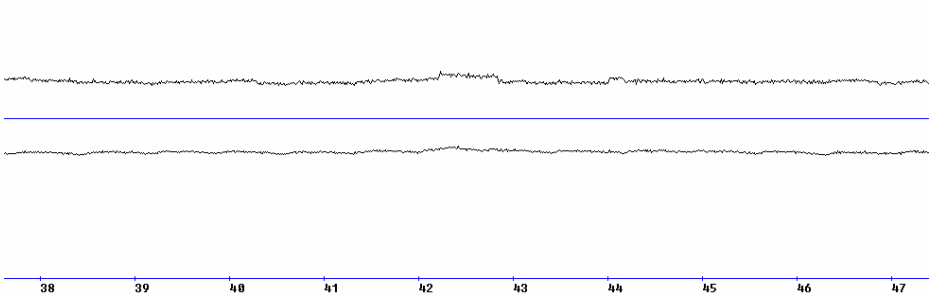

Figure 33 RSVP eye movements in the fifth study

The eye movements resulting from reading using Segmented PTP were similar to the recordings for Chunked RSVP in study two. There are frequent saccades and regressions. Given the low Task load ratings in the last study, we can probably conclude that the regressive eye movements were not the source of increased task load (Figure 34).

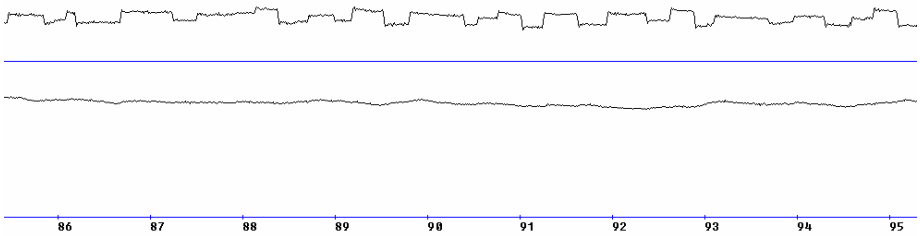

Figure 34 Segmented PTP eye movements in the fifth study

The eye movements resulting from reading using Moving PTP were most similar to the results for reading using the Page format in study two. The eyes move in a saccadic pattern. The saccades are however fairly short and at times there are regressions. Probably due to that words longer than nine characters overlap the preceding chunk to much. (Figure 35)

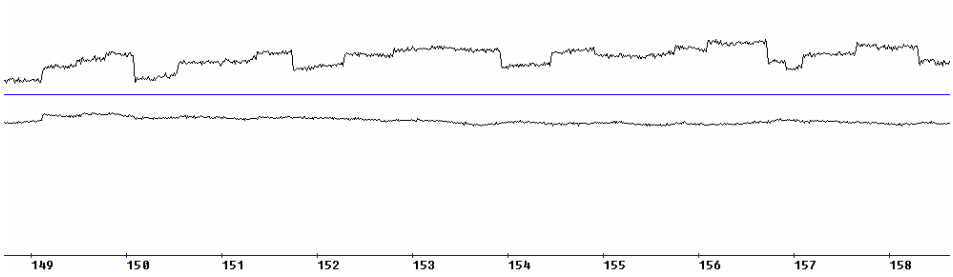

Figure 35 Moving PTP eye movements in the fifth study

5 Discussion

"A likely impossibility is always preferable to an unconvincing possibility"

- Aristotle

The discussion will focus on three issues: What we learned about readability on mobile devices based on our findings, what we learned about evaluating readability based on our experiences, and finally, what we whish to learn from evaluations of readability on mobile devices in the future.

5.1.1 Readability on Mobile Devices

Just as in most previous evaluations, we have focused on improving dynamic text presentation in the RSVP format. There is probably something with the RSVP format that attracts the attention of researchers, unfortunately however the format has not been found to live up to the expectations. It has repeatedly been found to be less likeable than other formats, even if it is just as efficient. The idea that the RSVP format by eliminating eye movements should increase reading speed and reduce cognitive load does not seem to hold, our experiments point in the opposite direction. In both studies performed on mobile phones, RSVP was far less efficient compared to the other formats. A partial explanation for the poor results in study four is that the experiment only contained one condition using RSVP. To be fair, none of the subjects participating in the experiments had any previous experience of using RSVP, or at least extremely limited compared to the other formats. There is probably a learning curve for using RSVP. The question is how much of this curve we can see in our experiments and how much training we can expect users to put in. A new text presentation format does not really let you do new things; it lets you read things in a new way. For a new text presentation format to rival existing formats it must probably offer an immediate gratification, either in terms of increased readability or something else. RSVP as it is commonly implemented today does not offer this gratification.

The Page format has in our experiments repeatedly been found to excel in terms of readability. On the PDA it was understandable since the screen was not much smaller than a pocket book and the MS Reader program was very well designed (Hill, 2001). On the mobile phone it worked surprisingly well although the reading speed dropped by 9% compared to the using it on the PDA (in the not flawed experiment that is). The figure cannot be taken for granted given the huge variation in reading speed between subjects. Just a look at the standard deviations for reading speed confirms this. However, the variations have been fairly consistent around 50 WPM. If an experimental design was used that did not take this between-subject variation into account, hardly none of the significances reported in this thesis would have been found. In the final experiment, the reading speed of the Page format was much higher than in the preceding study although exactly the same implementation of the format was used. Compared to study two (which also involved eye tracking) the reading speed in the last experiment was actually higher on a mobile phone than on a PDA (258 vs. 242 WPM). This is probably just a result of differences in reading speeds within the subjects, e.g. the subjects in the final study generally read texts faster than the subjects in study two. Compared to study four the Reading speed for the Page format increased by 9%, which suggests that the decrease in reading speed when reading on a PDA compared to a mobile phone is somewhere around, and probably less, than 10%. That is not bad given that the readability in MS Reader is more or less equal to a paper book.

In the fourth study, Paging was found to offer better readability than Scrolling. This is interesting since Scrolling is the text presentation format predominantly used on mobile phones today. The implementation of Scrolling on the phone did however only scroll line by line. A scrolling format that moved more lines at a time might improve the format. A combination of Scrolling and Paging would however probably be a good idea, the up and down buttons can be used for Scrolling whereas the left and right buttons can be used for Paging (this is also how a few third party developers, including Opera, have started to implement the format). The least number of interactions needed for Scrolling to read a text was approximately 150 as opposed to 50 for the Page format; this is however not reflected in Physical demand which could indicate that the number of interactions required does not have such a large impact on task load as could be expected. The format that resulted in the highest Task load ratings in study four was Leading. The factor for physical demand is very high, especially considered that the format is dynamic and requires very little user interaction. It may very well be the case that the high amount of unnatural eye movements actually resulted in physical strain. This is in line with the findings of Joula et al. (1995). Probably, Leading is not a suitable format for extended text presentation. However, it was surprisingly efficient in terms of speed and comprehension given that the task load was so high.

Before the last evaluation we proposed that a dynamic presentation format should not reduce eye movements but rather stimulate them. The findings from the last evaluation confirm that this was a correct assumption. Both the Segmented and the Moving PTP formats increased reading speed significantly. However, the Segmented PTP format also increased reading speed significantly compared to the Moving PTP format. Stimulating a more natural reading eye movement pattern worked, but it did not improve the reading experience to the same extent that Segmented PTP did. One of the early assumptions we had concerning the RSVP format was that the high saccade/regression ratio could be an indication of high task load. The last experiment does not support this hypothesis since both the Page format and the Segmented PTP format had a high ratio and low task load ratings. Task load has been the main problem with RSVP in all our studies. The key to lower task load ratings seems to be linguistic adaptation and chunking. Since PTP offered a readability that is comparable to the Page format in the final study, it is probably the format with most potential for improvement. The readability of the Page format is not likely to improve over time, PTP was new to all subjects and with some training they will probably read faster using that format. Moreover, both the adaptation and the chunking can be improved further. To stimulate eye movements still seem like an appealing idea and maybe a revised version the Moving PTP format could improve reading speed and task load. In either case, taking eye movements and the eye physiology into account when designing new text presentation formats seem to be rewarding.

5.1.2 Evaluation methodology

The experiments presented in this thesis have not been very realistic. Sitting in a lab and reading texts (that you did not even choose yourself) does not give the real picture of how reading works. However, the aim with the evaluations was never to find out how people read in real life situations. It was to evaluate how the text presentation formats work for reading. In order to do this in a reliable fashion the situations must be controlled, e.g. realism has to be sacrificed for reliable measurements. Obviously, the interfaces have to be evaluated in real life situations as well before we can tell if they really work or not. In the end it is the users who decide which interfaces they want to use. Nonetheless, if an interface does not work under controlled conditions, chances are fairly large that it won't work under real life conditions either. One aspect of our evaluations which more than any other limits how much the results can be generalized upon is the kin support. It had to be used to keep the subjects head still while reading since head movements affected the eye movement recordings. When reading in real life situations head movements are very common. One of the most interesting observations from the first study was how much the subjects actually moved while read-

ing. Changing position in the chair or just altering the gaze angle by a slight head movement was very common. The imposed restriction in movement is reflected in the Task load results between study one and study two. Although it impacts all formats, the question remains if it affects all formats to a similar extent.

The eye movement tracking techniques used in the studies have proved valuable. There is just no going back once you can actually see for yourself how subjects read. Not only does it offer you a rich resource of objective data to work on, it also offers you a visual representation of the differences between readers, which are large to say the least. The problem then becomes how to interpret the data. Eye movements are usually tracked for much shorter periods than we have used in our evaluations. Analyzing the recordings by hand would be more or less impossible; instead we used software that categorized the movements according to their function when reading (e.g. saccades, regressions, eye blinks, etc). The categorization criteria used in the thesis are disputable. If it had existed software for doing the analysis we would have used it to avoid defining our own measures. The recordings performed on the PDA were generally much more reliable than the recordings performed on the mobile phone. One reason for the difference may be the differences in the calibration and alignment. However, more probable is that we are pushing the limits for what the XY-1000 system can handle when using it for eye movement tracking on mobile phones. The screen of the mobile phone is only a few degrees wide from the viewing distance, whereas the screen of the PDA offers a wider angle. The XY-1000 system was not designed to be used on small screens. Increasing the sampling rate does not really help, what we need is a system with a higher resolution. Preferably such a system should not be head mounted so that we can give the subjects more freedom of movement.

As mentioned, evaluating readability is difficult. If nothing else, the experiments presented in this chapter illustrates how easy it is to go astray. To use within-subject repeated-measurement experimental designs that are balanced according to latin-square designs seems like a good methodology. If used properly, it can limit the number of subjects while still controlling the effects of text, format and order. Used wrong it can cause misleading results. It is a powerful tool but must be used with caution. In our experiments we have not realized the full potential of the model. When creating the design it is important that it is randomized. We used graeco-latin-squares made up of presentation formats and texts. These were then transposed over presentation order and text to get a randomization. A downside with transposition is that one certain type of format more often than randomly is read after a second type of format, although it happens in different positions and with different text. The same pattern occurs with the texts. The transposed randomization can however be improved. Four special sets of fourth order orthogonal graeco-latin-squares can be combined into a hyper-latin-graeco-square

(HGLS) (Box et al, 2005). Four text presentation formats can then be evaluated with sixteen subjects in a design where text, format and order is perfectly randomized (Table 11). Using such a design might have limited the eventual contamination effects of Buffered RSVP on the results of Chunked RSVP in study two. It is hard to assess the effects of the ties in the design we used, probably it has had an impact on the results. The question is how large the error is; a follow up experiment replicating the last study may shed some light on this issue.

5.1.3 Future studies

One of the most appealing properties of PTP is the extensibility of the format. In our evaluations we have used very simple or handcrafted models for adaptation, chunking, and eye movement modeling. We have worked with proof of concept prototyping rather than perfecting models. Now, if PTP works as well as the last evaluation suggests, there is much existing knowledge than can be used with the format. Adaptation is basically a language modeling problem, chunking is basically a parsing problem, and eye movement modeling is basically a question of determining the next most likely fixation point given what has been read and what can be sensed in the parafoveal region. This is obviously a gross simplification, but to use existing or future models of language and physiological modeling with the PTP format would probably improve it further. The improvements do not necessarily have to be evaluated on mobile devices; it would probably be possible to work on the formats in a simulator and then deploy them to a mobile device when they are working. The interface used on the device must however be evaluated on a real device since a simulation on a desktop does not offer the same affordances and limitations as a real device does.

Since PTP uses a language model there is nothing that says that the model must be the same for all users. Mobile devices are typically very personal and thus the resources used to adapt the text presentation could be weighted by the words that a person reads. A personal corpus that collects everything that you read would be an interesting resource, not only for text presentation but also for text entry, information retrieval, etc. Obviously there are some serious privacy concerns involved in creating such a resource, but if it would prove useful it may be worth it. Something else that may prove rewarding is to use PTP in combination with novel interaction methods such as tilt sensors. A problem with dynamic text presentation in general on mobile devices is that it requires the user's full attention, if you happen to look away the text proceeds and when you look back you are lost. A very simple implementation would be to just stop the presentation when the device is held at angles deviating from the viewing angle. It would moreover be interesting to see how the format works for elderly. Fine and Peli (1995) found that elderly

read faster using RSVP, it would be exciting to see how it would work with PTP. Using a text presentation format that does not take up much screen space then also has an additional benefit as the text size can be made much larger.

The most interesting issue to study is probably the training effects of using a dynamic format like PTP. If the text presentation can be made even more efficient than the Page format on mobile devices there is a clear benefit of using it. Evaluating how subjects read using the formats over a longer time span must be done in a different way than the evaluations presented here, yet the findings will probably be interesting. We will probably access more and more information while on the move. Not all information will be suitable to display using PTP, but the format may prove to be a valuable complement to other presentation formats. Challenging how we do things today is how we succeed in doing them tomorrow, but not without learning of the past.

6 Conclusions

"Experience is the past tense of experiment."

- Gregory Alan Elliott

As we have seen in this thesis, any new format we want to use for text presentation must conform to how we are used to read. Regardless of the device used for reading, or the format used for text presentation, the physiological and cognitive limits for reading remain the same. With a starting point in our ability to read, we have seen how readability can be defined and measured. A review of text presentation formats intended for mobile devices has been presented together with results from previous studies. A methodology for evaluating readability based on a graeco-latin-square (GLS) balanced repeated-measurement experimental design was introduced.

The GLS design was used in five studies of readability on mobile devices where novel variations of the RSVP format was evaluated against other common presentation formats including Paging, Scrolling, and Leading. Eye movement tracking was introduced used as an additional measure of readability. The results from the evaluations show that the graeco-latin-square design is useful, but must be implemented correctly. The studies moreover showed that the Page format was quite efficient, both on a PDA and a mobile phone. In fact, using Paging on a mobile phone was only about 10 % less efficient than using it on a PDA. The RSVP format did not live up to the expectations. Clearly, the elimination of eye movements does neither increase reading speed nor decrease task load. Leading was found to be efficient on a mobile phone in terms of reading speed, the unnatural eye movements required for reading does however seem to induce to much strain to be acceptable. In the last study, Predictive Text Presentation was introduced. The format is based on RSVP and combines linguistic chunking and adaptation as well as eye movement modeling to achieve a reading experience that can rival the Page format on mobile devices.

The methods used in the evaluations have been discussed and a further improvement of the GLS design, the hyper-graeco-latin-square (HGLS) de-

sign has been introduced. This thesis has shown why readability on mobile devices is important, how it may be evaluated in an efficient yet reliable manner, and finally pinpointed Predictive Text Presentation as the format with greatest potential for improving readability on mobile devices.

7 Contributions

"One of the advantages of being disorderly is that one is constantly making exciting discoveries."

- A. A. Milne

Developed Bailando, a research prototype for dynamic text presentation on Pocket PC PDAs. Bailando is available to researchers and has been used for studying multimodal interaction using sounds (Goldstein et al., 2003) and gaze detection (Åkervall and Granath, 2002; Öquist et al, 2001).

Introduced the graeco-latin-square experimental design for readability studies. The design has since been used in several evaluations where texts and interfaces have to be balanced, for example when evaluating tools for dyslectics (Nilsson and Thunholm, 2005; Goldstein et al., 2006).

Implemented two variants of Adaptive RSVP, a dynamic text presentation format that adapts the presentation speed of words and sentences to the linguistic content and context. Showed that adaptation could improve readability and was faster than Scrolling on a PDA in a readability study.

Integrated the XY-1000 eye movement tracking system with the Bailando prototype and developed BaiCom for automated calibration and alignment via Orbit. Developed EyeAlign, a suite of tools to convert and analyze eye movement recordings data from the Orbit and Tobii eye tracking systems.

Performed the first two eye movement studies of reading on a PDA using traditional text presentation in the Page format and dynamic text presentation in the Adaptive RSVP format. Disproved the notion that reducing eye movements by using RSVP improves reading speed or reduces task load.

Developed Rapido, a research prototype for evaluating traditional and dynamic text presentation on J2ME enabled mobile phones. Implemented the

Paging, Leading, and RSVP formats and integrated the Rapido prototype with the Orbit eye tracking system and the EyeAlign suite.

Performed the first eye movement studies of traditional and dynamic text presentation formats on a mobile phone. Found that Paging and RSVP was better than Scrolling and Leading, but also showed that RSVP displaying one word at a time was not optimal.

Introduced Predictive Text Presentation, a dynamic text presentation format based on RSVP that utilizes linguistic adaptation and segmentation as well as eye movement modelling. Evaluated PTP against the Page format and word-for-word RSVP and found that PTP could improve readability.

Proposed a hyper-graeco-latin-square design for future readability evaluations. The design achieves a perfect randomisation of text, format and presentation order; factors that must be controlled when evaluating readability in repeated measurement evaluations.

Suggested further improvements of the PTP format using existing language and eye movement modelling resources and introduced the notion of the personal corpus. Identified a few interesting avenues of research based on previous findings and experiences.

Acknowledgements

“Okay, Houston, we've had a problem here”

- John L. Swigert

Writing a thesis is like a trip to the moon. It all begins with stargazing and dreams, but to really get there; you must qualify for astronaut training. Then one day, you are finally launched into space. The ride is risky and bumpy. Resources onboard are scarce and equipment is prone to fail, usually when needed the most. Ever reaching the Sea of Tranquility is never certain, neither is getting back to earth again. Yet one thing is for sure; you do not get there by yourself.

My fellow astronauts, *Anna Sågvall Hein*, *Jan Ygge*, and *Mikael Goldstein*. When I felt unsure, it was the calm look on your faces that pushed me on. When I went astray, it was the curiosity in your voices that nudged me forward. When I was lost, it was the company of you that assured me that we really were discovering something new. We did this together. Without either of you this would have remained a dream.

My parents, *Louise Seimyr* and *Oscar Öquist*. Mom, you showed me that impossible is not impossible, it's just the notion of something that is more difficult than it ought to be. Dad, you showed me that imagination is the only limit, and that language is the limit of imagination. Without your love this would just not be.

My brothers, *Henrik* and *Carl-Otto Öquist*. Together we explored where the streams started. I can't think of a better company to find out where the rivers end. We have had a world of fun so far, and do you know? We haven't even started heading downstream yet.

My family, *Thomas Seimyr*, *Harriet Berthold*, *Jessica Johansson*, *Barbro Hjalmarsson, Torsten Hjalmarsson, Quynhoung Hjalmarsson, Monica Öquist*, cousins, siblings, uncles and aunts. Life is rich with you.

My friends in Trosa. *Andreas*, *Gudrun*, and *Börje Kjellén*. If I ever need comfort and assurance, I know which porch to turn to. *Peter Ahldin*, *Mattias Ågren*, *Andreas Imeryd*, *Conny Brander*, and *Carl Leijon*. You all left happy memories with me, I am sure I left some with you. I am also indebted to a host of teachers; I hope you are proud, not of me but of you.

My family in Cairo, *Fatma*, *Mohammed*, and *Khaled*. To live with you for a year teached more of life than a lifetime without you ever would. *Anu*, *Pia*, *Thomas*, *Calumn*, *Georgina*, *Benjamin*, *Martin*, *Kate*, *Jeremy*, *Diana*, and *Anna-Lisa*. Egypt was my greatest adventure; it wouldn't even have been half as exiting without you.

My friend, *Jonna Millroth*. With you I share some of my fondest memories. You made an everlasting impression on me, an impression I will always treasure. *Elisabeth Larsson* and *Linda Pagin*, you complete the perfect triumvirate for late dinners, deep discussions, and of course, a fair amount of red wine.

My friends from Stockholm, *Ida Bergman*, *Peter Vouri*, *Joel Huselius*, *Linda Hamberg, Joakim Ekström*, *Daniel Osser*, *Ingvar Åkerblad*, *Jannika Tingvall*, *Stefan Lundgren*, *Håkan Andersson*, *Linus Jogin*, *Therese Agehed*, *Olle Eriksson*, *Anu Lindquist*, *Ola Lundell*, *Siri Axelsson*, *Linda Samuelsson*, *Elin Malmberg*, *Helena Sas*, *Jörgen Berggren*, *David Lysholm*, *Tommy Lindquist*, *Mattias Königsson*, *Rebecca Huselius*, *Sanna Williams*, and *Nico Cleyndert*. To share the happiness of life with you is just as necessary as dispensing the hardships, I cherish you immensely for you being you.

My friends from the Language Engineering Programme in Uppsala, we had some great years together and I miss you all. Uppsala is just not the same place without you.

My former colleagues at Mira Network, learning by doing is always better. The experiences I got from working with you have proved invaluable, I am looking forward to the alumni dinner.

My former colleagues at Ericsson, working at the Usability & Interaction lab was an inspiration. I did not expect to keep working on the project we started there for six years though.

My colleagues at Uppsala University, after having been at the Department of Linguistics and Philology for ten years, first as a student, then as a PhD, I just want to say thank you. It has been a wonderful time.

My colleagues at Karolinska Institutet, I felt welcome at the Bernadotte laboratory the very first time I visited you, now I feel at home. It will be a joy working with you in the future.

My friends from GSLT, I probably would not have managed this thesis without the help of all of you. Supervisors and PhDs alike, I will miss going to Göteborg for the intensive weeks and meeting you all regularly.

My co-entrepreneurs, *Ingela* and *Gunnar Lewald*. I will soon start that company we are always talking about. We have some exciting projects to work on and I am really looking forward to it.

My co-experimenters, *Linda Danvall* and *Kristin Lundin*. Work was never dull with you around and both of you have made significant contributions to this thesis. My thesis students, *Stefan Nilsson*, *Arvid Thunholm*, and *Örjan Berglund*, it was a joy working with you. I also want to express gratitude to all students who have commented my work in classes at Uppsala University and Karolinska Institutet; your questions were my insights.

My editors, *Fabio Paterno*, *Ahmed Seffah*, *Homa Javahery*, and *Jo Lumsden*. It is not always easy to know how things are supposed to be done, but you have all been extraordinarily assistive.

My reviewers, thanks for all your constructive comments on submissions, and even more importantly, thanks for not accepting work that needed more thought.

My love, *Malin*. I have been to the moon and back again. There were many wonders out there, but nothing compared to the beauty of you in the distance. Now I am home. This was my dream; you are my life.

References

Åkervall, P., and Granath, R. (2002). *Smart Bailando - Eye controlled RSVP on handhelds*. MSc thesis, Department of Computer Science, Chalmers, Göteborg, Sweden.

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

Box, G. E. P., Hunter, W. G., and Hunter, S. J. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition. New York, NY: John Wiley & sons.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*. 13-60. Rome: Italy.

Braze, D., Shankweiler, D., Ni, W. and L. Palumbo. 2002. Readers' Eye Movements Distinguish Anomalies of Form and Content. *Journal of Psycholinguistic Research*, 3(1), 25-44.

Bruijn, O. Spence, R. (2000). Rapid Serial Visual Presentation: A space-time trade-off in information presentation. In *Proceedings of Advanced Visual Interfaces, AVI2000*, 189-192.

Castelhano, M.S. & Muter, P. (2001). Optimizing the reading of electronic text using rapid serial visual presentation. *Behaviour & Information Technology*, 20(4), 237-247.

Cocklin, T.G., Ward, N.J., Chen, H.C. & Juola, J.F. (1984). Factors influencing readability of rapid presented text segments. *Memory & Cognitio*n, 12(5), 431–442.

Cooper, A. (1995). About Face: The Essentials of User Interface Design. New York, NY, IDG Books.

Danvall, L. (2004). *Läsning på små skärmar – en studie med ögonrörelseregistreringsteknik.* Optician thesis, Department of Clinical Neuroscience, Karolinska Institutet.

Dillon, A., Richardson, J. & McKnight, C. (1990). The effect of display size and text splitting on reading lengthy text from screen. *Behaviour and Information Technology*, 9(3), 215-227.

Dillon, A. (1992). Reading from Paper versus Screens: A Critical Review of the Empirical Literature. *Ergonomics*, 35(10): 1297-1326.

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on video display terminals as a function of window size. *Human Factors*, 25, 683-692.

Euler, L. (1782). *Recherches sur une nouvelle esp ece de quarrès magiques*. Verh. Zeeuwsch Gennot. Weten Vliss 9, 85-239.

Fine, E.M. & Peli, E. (1995). Scrolled and rapid serial visual presentation texts are read at similar rates by the visually impaired. *Journal of Optical Society of Americ*a, 12(10), 2286-2292.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Percep. Psychophys.* 8*,* 215-221.

GSM Association (2006). *Universal Access - How mobile can bring communications to all.* GSM Association (GSMA) report, 17 October 2006.

Goldstein, M., Öquist, G. and Björk, S. (2002). Evaluating Sonified Rapid Serial Visual Presentation: An Immersive Reading Experience on a Mobile Device. In: *Proceedings of User Interfaces for All 2002* (Paris, France), 508-523. Berlin: Springer.

Goldstein, M., Öquist, G., and Lewald, I. (2006). Evaluation of PreCodia, a Computerized Reading Aid for Readers Suffering from Dyslexia. In: *Proceedings of Human Factors in Telecommunication 2006* (Sophia-Antipolis, France), 127-134. Brighton, MA: IGI Group.

Gould, J.D., Alfaro, L., Finn, R., Haupt, B. & Minuto, A. (1987). Reading From CRT Displays Can Be as Fast as Reading From Paper. *Human Factors*, 29(5), 497-517.

Hart, S.G. & Staveland, L.E. (1988). Development of Nasa-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload,* by P.A. Hancock and N. Meshkati (Eds.). Elsevier Science Publishers, B.V.: North-Holland.

Hill, B. (1999). *The Magic of Reading*. Available at: http://slate.msn.com/ebooks/ (December 2001).

Huey, E.B. (1908). *The psychology and pedagogy of reading.* Cambridge, MA: MIT Press. (Reprinted 1968).

Jackson, M.D. and McClelland, J.L. (1979). Processing determinants of reading speed. *Journal of experimental psychology*, 108, 151–181.

Juola, J. F., Ward, N.J., & McNamara, T. (1982). Visual search and reading of rapid serial presentations of letter strings, words, and text. Journal of Experimental Psychology: General, 111, 208-227.

Juola, J.F., Tiritoglu, A., & Pleunis, J. (1995). Reading text presented on a small display. Applied Ergonomics, 26, 227-229.

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), pp. 329-354.

Kang, T.J., and Muter, P. (1989). Reading Dynamically Displayed Text. *Behaviour & Information Technology*, 1989, 8(1), 33-42

Laarni, J. (2002). Searching for optimal methods of presenting dynamic text on different types of screens. *In proceedings of NordiCHI 2002*. 217-220. New York: ACM Press.

Lundin, K. (2006). *Läsning på mobiltelefon*: *En studie med ögonrörelseteknik*. Optician thesis, Department of Clinical Neuroscience, Karolinska Institutet.

Masson, MEJ. (1983). Conceptual processing of text during skimming and rapid sequential reading. *Memory and Cognition*, 11, 262-274.

Mills, C.B. & Weldon, L.J. (1987). Reading text from computer screens. *ACM Computing Surveys,* 19(4), ACM Press.

Muter, P. (1996). Interface Design and Optimization of Reading of Continuous Text. In *Cognitive aspects of electronic text processing.* H. van Oostendorp and S. de Mul (Eds.). Norwood, N.J.:Ablex.

Muter, P., Latrémouille, S. A., Treurniet, W. C., & Beam, P. (1982). Extended reading of continuous text on television screens. Human Factors, 24, 5, 501-508.

Muter, P., Kruk, R. S., Buttigieg, M. A., and Kang, T. J. (1988). Reader controlled computerized presentation of text. *Human Factors*, 30, 473-486.

Muter, P. & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behavior & Information Technology*, 10, 257–266.

Nilsson, S. and Thunholm, A. (2005). *Reading syllable-separated text – eye movements in dyslexic subjects*. Masters Thesis, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.

Ober, J. (1994). Infra-Red Reflection Technique. In *Eye Movements in Reading*. J. Ygge and G. Lennerstrand (Eds.). Elsevier Science.

Osborne, D.J. & Holton, D. (1988). Reading From Screen Vs. Paper: There Is No Difference. *International Journal of Man-Machine Studies*, 28, pp. 1-9.

Öquist, G. (2001). *Adaptive Rapid Serial Visual Presentation*. Masters Thesis in Computational Linguistics, Department of Linguistics, Uppsala University, Sweden.

Öquist, G. (2004). Enabling Embodied Text Presentation on Mobile Devices. In: *Proceedings of Mobile and Ubiquitous Information Access 2004* (Glasgow, Scotland), 26-31.

Öquist, G. (2006). Multimodal Interaction with Mobile Devices: Outline of a Semiotic Framework for Theory and Practice. In: *Proceedings of Wireless Networks and Systems 2006* (Setubal, Portugal), 276-283. Setubal: INSTICC Press.

Öquist, G. (2007). Experiences and Findings from three Eye Movement Studies of Mobile Readability. In: *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, J. Lumsden (Ed.). Hershey, PA: Idea Group Publishing. *To appear.*

Öquist, G. (forthcoming). Predictive Text Presentation: Using Linguistic Segmentation and Eye Movement Modeling to Improve Dynamic Text Presentation on a Mobile Phone. *Submitted for publication.*

Öquist, G., Goldstein, M. and Björk, S. (2002). Utilizing Gaze Detection to Stimulate the Affordances of Paper in the Rapid Serial Visual Presentation Format. In: *Proceedings of Mobile HCI 2002* (Pisa, Italy), 378-381. Berlin: Springer.

Öquist, G. and Goldstein, M. (2002). Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation. In: *Proceedings of Mobile HCI 2002* (Pisa, Italy), 225-240. Berlin: Springer.

Öquist, G. and Goldstein, M. (2003). Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation. *Interacting with Computers*, 15(4), 539-558.

Öquist, G., Sågvall-Hein, A., Ygge, J., and Goldstein, M. (2004a). Eye Movement Study of Reading Text on a Mobile Device using the Traditional Page and the Dynamic RSVP Format. In: *Proceedings of Mobile HCI 2004* (Glasgow, Scotland), 108-119. Berlin: Springer.

Öquist, G., Goldstein, M. and Chincholle, D. (2004b). Assessing Usability across Multiple User Interfaces. In: *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces*, A. Seffah and H. Javahery (Eds.), 327-349. New York, NY: John Wiley & sons.

Öquist, G. and Lundin, K. (forthcoming). Eye Movement Study of Reading on a Mobile Phone using Scrolling, Paging, Leading, and RSVP. *Submitted for publication.*

Paulson, L.J. & Goodman, K.S. (2000). Influential Studies in Eye-Movement Research. *International Reading Association*.

Proctor, R.W. and Proctor, J.D. (1997). Sensation and Perception. *In Handbook of Human Factors and Ergonomics*. G. Salvendy (Ed.). Second Edition, Wiley-Interscience, New York, 53-57.

Potter, M.C. (1984). Rapid Serial Visual Presentation (RSVP): A method for studying language processing. In *New Methods in Reading Comprehension Research.* Kieras, D.E. & Just, M.A. (Eds.). Hillsdale, N.J.:Erlbaum

Rahman, T. & Muter, P. (1999). Designing an interface to optimise reading with small display windows. *Human Factors,* 1(1), 106-117, Human Factors and Ergonomics Society.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, pp. 372-422.

Reichle, E.D., Rayner, K. & Pollatsek, A. (2000). Comparing the E-Z Reader Model to Other Models of Eye Movement Control in Reading. *Cognitive Sciences Eprint Archive* (Available at: http://cogprints.soton.ac.uk)

Reilly, R. (1993). A connectionist framework for modeling eye-movement control in reading. In G. d'Ydewalle & Van Rensbergen (Eds.), *Perception and Cognition: Advances in Eye Movement Research,* 191-212. Amsterdam: North Holland.

Robeck, M.C. & Wallace, R.R. (1990). *The Psychology of Reading: An Interdisciplinary Approach,* Second edition, Lawrence Erlbaum Associates, Hillsdale: New Jersey.

Sahami, M., Mittal, V., Baluja S., and Rowley, H. (2004). The Happy Searcher: Challenges in Web Information Retrieval. *In Proceedings of the Eighth Pacific Rim International Conference on Artificial Intelligence*, 3-12. Springer.

Schneiderman, B. (1982). The Future of Interactive Systems and the Emergence of Direct Manipulation. *Behaviour and Information Technology*, 1, pp. 237-256.

Shneiderman, B. (1998). *Human-Computer Interactio*n, 3rd edition, Addison Wesley Longman, Inc, 412-414.

Schmandt-Besserat, D. (1996). *How Writing Came About*. University of Texas Press.

Shaywitz, B.A., Lyon, R.G., & Shaywitz, S.E. (2006). The Role of Functional Magnetic Resonance Imaging in Understanding Reading and Dyslexia. *Developmental Neuropsychology*, 30(1), 613-632.

Sicheritz, K. (2000). *Applying the Rapid Serial Presentation Technique to Personal Digital Assistants,* Master's Thesis, Department of Linguistics, Uppsala University.

Tekfi, C. (1987). Readability formulas: An Overview, *Journal of Documentation,* 43(3) 261-73.

Weiss, S. (2002). Handheld Usability. New York: John Wiley

Wickens, C. D. (1992). *Engineering psychology and human performance,* 2nd edition, Chapter 8, HarperCollins Publishers Inc., New York.

Wireless Intelligence (2006). *Worldwide cellular connections pass 2.5 billion.* Wireless Intelligence survey, 7 September 2006.

Ygge, J., Bolzani, R., Tian, S. (1999). A computer based system for acquisition, recording and analysis of 3D eye movements signals. In *Transactions IX International Orthoptic Congress*, Pritchard C (Ed), Stockholm, 91-94.

Experimental Designs

These experimental designs based on latin-squares are discussed in the thesis. The graeco-latin-square is used in the evaluations and the hyper-graeco-latin-square is a proposed improvement for future studies.

Graeco-Latin-Square

A graeco-latin-square is a latin-square of two sets of *n* elements, S and T, ordered in a *n*n* table so that each cell contain an ordered set <s, t> and no row or column contains more than one s or one t. If we let the formats be S = {A, B, X, Δ} and the texts T = {α, β, χ, δ}, we can create a GLS of experimental conditions (Table 10).

Table 10 Fourth order graeco-latin-square (GLS) for readability studies

Format A / Text α	Format B / Text β	Format X / Text χ	Format Δ / Text δ
Format B / Text β	Format X / Text χ	Format Δ / Text δ	Format A / Text α
Format X / Text χ	Format Δ / Text δ	Format A / Text α	Format B / Text β
Format Δ / Text δ	Format A / Text α	Format B / Text β	Format X / Text χ

Format B / Text χ	Format X / Text δ	Format Δ / Text α	Format A / Text β
Format X / Text δ	Format Δ / Text α	Format A / Text β	Format B / Text χ
Format Δ / Text α	Format A / Text β	Format B / Text χ	Format X / Text δ
Format A / Text β	Format B / Text χ	Format X / Text δ	Format Δ / Text α

Format X / Text α	Format Δ / Text β	Format A / Text χ	Format B / Text δ
Format Δ / Text β	Format A / Text χ	Format B / Text δ	Format X / Text α
Format A / Text χ	Format B / Text δ	Format X / Text α	Format Δ / Text β
Format B / Text δ	Format X / Text α	Format Δ / Text β	Format A / Text χ

Format Δ / Text χ	Format A / Text δ	Format B / Text α	Format X / Text β
Format A / Text δ	Format B / Text α	Format X / Text β	Format Δ / Text χ
Format B / Text α	Format X / Text β	Format Δ / Text χ	Format A / Text δ
Format X / Text β	Format Δ / Text χ	Format A / Text δ	Format B / Text α

Hyper-Graeco-Latin-Square

A hyper-graeco-latin-square is a latin-square of two sets of *n* elements, S and T, ordered in a *n***n* table so that each cell contain an ordered set <s, t> and no row or column contains more than one s or one t and both s and t are orthagonal for each order. If we let the formats be S = {A, B, X, Δ} and the texts T = {α, β, χ, δ}, we can create a HGLS of experimental conditions (Table 11).

Table 11 Fourth order hyper-graeco-latin-square (HGLS) for readability studies

Format A / Text δ	Format B / Text χ	Format X / Text β	Format Δ / Text α
Format B / Text χ	Format A / Text δ	Format Δ / Text α	Format X / Text β
Format X / Text β	Format Δ / Text α	Format A / Text δ	Format B / Text χ
Format Δ / Text α	Format X / Text β	Format B / Text χ	Format A / Text δ

Format B / Text α	Format X / Text δ	Format Δ / Text χ	Format A / Text β
Format X / Text δ	Format B / Text α	Format A / Text β	Format Δ / Text χ
Format Δ / Text χ	Format A / Text β	Format B / Text α	Format X / Text δ
Format A / Text β	Format Δ / Text χ	Format X / Text δ	Format B / Text α

Format X / Text χ	Format Δ / Text δ	Format A / Text α	Format B / Text β
Format Δ / Text δ	Format X / Text χ	Format B / Text β	Format A / Text α
Format A / Text α	Format B / Text β	Format X / Text χ	Format Δ / Text δ
Format B / Text β	Format A / Text α	Format Δ / Text δ	Format X / Text χ

Format Δ / Text β	Format A / Text χ	Format B / Text δ	Format X / Text α
Format A / Text χ	Format Δ / Text β	Format X / Text α	Format B / Text δ
Format B / Text δ	Format X / Text α	Format Δ / Text β	Format A / Text χ
Format X / Text α	Format B / Text δ	Format A / Text χ	Format Δ / Text β