MAUU5900

MASTER THESIS

in

Universal Design of ICT

May 2019

Impact of Sentence-length on the Readability of Web for Blind Users

Bam Bahadur Kadayat (s310223)

Department of Computer Science

Faculty of Technology, Art and Design

OSLOMET

Preface

This Master thesis is the final step for acquiring the Master in Universal Design of ICT at Oslo Metropolitan University (OsloMet) located in Oslo, Norway. It has been administrated and supervised by the university under the faculty of Technology, Art and Design.

I would like to thank my sincere gratitude to my supervisor, **Evelyn Eika** for her endless encouragement and helpful suggestions. Additionally, I am really thankful to **Norun Christine Sanderson** for her valuable comments and suggestions in the process of completing this research. I am really appreciated to my colleagues for their participation and providing their precious time for the experiment.

Finally, I take this opportunity to express my gratitude towards my parents and to my wife and my daughter for their unconditional love, support and encouragement in the completion of this thesis.

Bam Bahadur Kadayat

Oslo, Norway / 15th May, 2019

Abstract

This research investigates how the length of sentence impacts on the readability of the web for blind users and explores the appropriateness of the range of sentence to make the web content readable and understandable. Readability on the web is a challenging task for blind users because blind users rely on a screen reader to read the text with a speech synthesizer and they often experience difficulties in re-tracking the beginning of the sentence as the screen reader does not read it back. This research assessed if the length of a sentence affects the readability of the web in terms of workload for the blind users, and if so, what would be the appropriate sentence length to make the web content readable and understandable. An experiment was performed with twenty-one participants and employing a within-subject design for assigned conditions to the participants. A one-way repeated measure ANOVA was used to find out the significance of variance among the five prototypes used. The researcher also applied a comprehension test to indicate whether the users understand the content of the prototype to investigate the appropriate length of the sentence. The finding reveals that there is a significant impact of sentence length on the readability of the web for blind users. The majority of participants performed better in prototype B which had 16-20 word-length in a sentence. Since participants needed less mental and temporal demand, they had less frustration that resulted in better performance in prototype B. The limitations of the research are that the participants were not the actual blind users; instead, they were blindfolded during the experiment. This study suggests that 16-20-word length in a sentence can be appropriate for the blind user to perform readability on the web without much workload.

Keywords

Readability, workload, sentence length, accessibility, universal design, web accessibility

Table of Contents

List of Tables

List of Figures

Abbreviations

ACCB: Adapted Content Code Blurring

ANOVA: Analysis of Variance

ARI: Automated Readability Index

BBC: British Broadcasting Corporation

DSC: Document Slope Curve

GHZ: Gigahertz

ICT: Information and communication technology

NASA: National Aeronautics and Space Administrations

SPSS: Statistical Package for Social Science

TLX: Task Load Index

SMOG: Simple Measure of Gobbledygook

WCAG: Web content accessibility guidelines

WAI: Web Accessibility Initiative

NVDA: Non-visual Desktop Access

Chapter 1 - Introduction

Readability is the measure of ease or difficulty with which the text can be read and understood by an intended reader who is reading for a specific purpose (Pikulski, 2002). It depends on several factors such as content, structure, readers' knowledge, vocabulary, layout, and design (Owu-Ewie, 2014). It makes the task easier which reduces readers' frustration and the workload. Although it is a challenging task for blind readers to read the content because they access the web through screen reader software. And they have difficulties to re-track the reading content as the software does not read it back. They need more workload to memorize the words of the sentences, and it creates comprehension problems to understand the content of the webpage. Hence, it can be said that the impact of sentence length is the responsible factor for readability on the web for blind users.

This research investigates the impact of sentence length on the readability of the web for the blind user which affects readability on the web. This research also aims at finding out what the appropriate sentence length can be for blind users on the web. So, this chapter presents the overviews of the research, and it tries to clarify for the reader about the research context, disability, and universal design on the web, web accessibility, statement of the problem, research aims and research questions, and the structure of the thesis.

1.1 Research context

1.1.1 History of readability

According to Eika and Sandnes (2016, p. 2), "readability research has been conducted for more than a century". Most of the studies mentioned that the sentence length affects the readability where long sentences are normally measured harder to read than shorter sentences (Eika & Sandnes, 2016). But it is not frequently believed that shorter sentences are easier because it is not recommended for improving readability.

Moreover, word difficulty is another factor that is commonly mentioned, a sentence with tough word is harder to read compared with easy texts. The sentence and the word length are the two attributes that occur in readability measurement (Eika &

Sandnes, 2016). For example, one of the readability formulae, Flesch-Kincaid reading ease index is defined as follows:

$$206.835 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right)$$

Figure: 1.1 Flesch-Kincaid index formula cited from (Eika & Sandnes, 2016)

The formula is designed to show how difficult a reading passage is to comprehend. The score ranges between 0 and 100, where a high score indicated easy to read and low score revealed harder to read (Gottron & Martin, 2012). But, these classic readability "measures have also been criticized for being over simplistic" and mostly used in the printed text and they are less popular these days (Eika & Sandnes, 2016, p. 2). Also, readers read text on web differently compared to the printed text because the content in the web is presented differently compared to a written book (Gottron & Martin, 2012, p. 2).

1.1.2 Blind users and the web

Since the invention of the internet, people easily get reading materials through an online digital platform. They read more and more text that appears as part of websites (Gottron & Martin, 2012). Through some significant research, human got fundamental inventions and ideas which enabled visually challenged people to become as competent as their peers, or more (Rony, 2017b). So, visually impaired readers rely mostly on the auditory feedback of screen readers technology to read webpage information (Guerreiro & Gonçalves, 2015).

There are "45 million blind people around the world", and they are interconnect with websites through screen reader technology (Babu, Singh, & Ganesh, 2010, p. 4). Reading is a significant part on the web and reading through web applications is a challenging task for different ability people. This is especially due to a lack of accessibility and readability of text on the web. An accessible and readable web content allows a blind user to access and understand its information. Also, the unawareness of developers and designers about non-visual web content often hinders the accessibility of Web sites (Babu et al., 2010). WCAG 2.0 offers a broad set of recommendations for to make the web content accessible and readable (World Wide Web Consortium, 2008). "Why blind users face problem during Web readability?" is the major question on the digital world. Understanding the content of the sentence for the blind user is crucial for readability on the web. Although WCAG 2.0 exists to assist developers and designers to minimize this problem, it does not guarantee effective accessibility for the blind users (Babu Rakesh, 2009, p. 4).

1.2 Disability and Universal design on the web

According to the World Health Organization (1980), "disability is any continuing condition that restricts everyday activities" that include impairments, activity limitations, and participation restrictions. The resources that we found in the World Wide Web need to be accessible to offer the opportunities for different able individuals to deliver the content (Rowan, Gregor, Sloan, & Booth, 2000). Still there is a huge improvement needed in the web-based information to make the content readable and understandable to participate them equally. The information on the web should be readable and understandable for readability on the web.

In addition, when the webpage and other technologies were poorly designed, they create additional difficulties for people to use them. The Americans with Disabilities Act describe if any individuals has physical or mental impairment which limits them from daily activities is the disability (Thompson, 2015). In an ICT context, there are many diversities of web users and, they have their own abilities to access the web. Following are diversity abilities:

a) Auditory: According to W3C Web Accessibility Initiative (2017), auditory is known as "central auditory processing disorder". The people who have the auditory disorder have the hearing loss in one or both ears. The hearing loss might be a major and uncorrectable hearing loss in both ears.
b) Cognitive and neurological: People with cognitive and neurological disabilities have mental health and behavioral disorders (W3C Web Accessibility Initiative, 2017). It may effect on hearing, moving and speaking and comprehending the information but does not affect the person intelligence.

c) Physical: This includes muscular control such as paralysis, joint disorders, pain in movement and missing limbs (W3C Web Accessibility Initiative, 2017). For example, barriers when Websites and web browsers don not provide full keyboard support.
d) Speech: This includes trouble in creating speech that is understood by others or by "voice recognition software" (W3C Web Accessibility Initiative, 2017). For example, Web-based services depend on interaction using voice only. So, this the difficulties for individuals with voice disabilities.
e) Visual: This includes the disabilities when people have loss in one eye or in both eyes. Some people have difficulties to distinguish the colors and have problem with bright colors (W3C Web Accessibility Initiative, 2017). They relied on changing the appearance of web content. For example, text-size and images of the web content.

According to National Disability Authority (2014b), universal design is the design of any products and environments that is "accessible and usable by all", without the need for transformation. If any product has a good design, it is accessible, usable, convenient, and pleasure to use (Story, Mueller, & Mace, 1998). However, the universal design principles were originally created for architecture and appliances, not for the websites (Lela Kodai, 2015). But slowly its approaches were implemented in the websites design to make the web content accessible and readable for all readers. Following are the "seven principles of universal design" with their example in terms of web design.

a) Equitable use: The webpage should be designed accessible and readable for people with who have different abilities (Story et al., 1998). This principle provides equal access features and information to all user with diverse abilities (National Disability Authority, 2014a). For example, high contrast helps users with weak vision, alt text helps screen reader to describe information, a user who cannot move their hand can access via keyboard.
b) Flexibility in use: This principle provides the choose option for users while they access features. User has enough flexibility and they will have a more pleasant experience (National Disability Authority, 2014a). For example, resize text, navigate via keyboard, language option and adjust color settings.

c) Simple and instinctive: This principle makes the web design simple to comprehend in spite of the users knowledge and language skills (National Disability Authority, 2014a; Story et al., 1998). Also, a good design is the design where the user can achieve their goal easily. For example, error message is shown on red text color and e-commerce site menu items that categorize products for man and women.
d) Perceptible information: The design makes the webpage content perceptible to the user regarding their sensory abilities (National Disability Authority, 2014a; Story et al., 1998). Nowadays, websites are the most common way of conveying information for the user. So, data must be perceived by the diverse user on the web. For example, complex information needs to break down into easily digestible pieces that will make web content more accessible, and by using graphics, user can easily perceive information.
e) Tolerance for error: It reduces the threats and the unfavorable "consequences of accidental or unintended actions" (National Disability Authority, 2014a; Story et al., 1998). In web, the user could accidentally click on a wrong move that could do irreversible damage to their information. So, to prevent this type of accident, web developer and designer need to add confirmation dialog box before triggering the action. If any action was wrong and there should be undone option to make correction and process again.
f) Low physical effort: The web design could be used easily by using little physical effort (National Disability Authority, 2014a; Story et al., 1998). For instance, contents on the web should have minimum distance; it allows users to minimize hand movement while navigating content through an input device. The page content should not be lengthier where user needs to scroll the page.
g) Size and space: Suitable size and space are provided according to the user's physical size, and posture (National Disability Authority, 2014a; Story et al., 1998). For instance, action targets for one-handed mobile device use which helps to make targets large enough to click or tap easily and put primary action within easy reach.

1.3 Web accessibility

According to Web Accessibility Initiative (1999), web accessibility means the webpages and its contents need to be designed and developed so that the user with impairment can use them. Also, the disorder people can understand and access its content to interact with the web and the web content should be accessible to all users inclusive of race, nationality, religion, and disability (Paciello, 2000).

The web makes our daily life a lot easier but, different ability people not able to the benefit of the web like blind people. They require assistive technology to interact with the web. Fortunately, there is a rapid improvement on the web and, the visually impaired people are able to access the web as abled users (Bergman & Nygren, 2009). But some websites still have barriers with accessibility which creates difficulties for the people with disabilities. It is an important aspect that all kinds of websites need to be accessible such as education, government, health, business, technology other sectors as well. Web developers, designers, content writers need to focus on accessibility to make the website usable for all. The Web Accessibility Initiative has the roles to develop guidelines that explains solutions of accessibility for Web developers and designers (Web Accessibility Initiative, 1999). It depends on several factors like content type, the size and the development tools and developing environment. It is necessary that the Web should be accessible and the different ability people get equal opportunities to use them (Web Accessibility Initiative, 1999). The Web content should follow WCAG, which is developed through the World Wide Web Consortium process. It is a standard guidelines for Web content accessibility where individuals, organizations, and governments need to meet the needs (World Wide Web Consortium, 2008). The WCAG describes about the mechanisms for making web content usable and accessible for people with all kinds of disabilities. The WCAG is for page authors, site designers, developers and designers to make the webpage in a standard way to maintain accessibility.

Web Accessibility initiative provides some tips on writing for Web accessibility. It introduces that the content needs to be clear and concise (Web Accessibility Initiative, 1999). The WAI also suggest that the content writer and designer need to use simple language and formatting to make the content clear and concise, and paragraphs need to be short and clear. Complex words and phrases which are not necessary should be avoided. Writers need to expand acronyms and provide a glossary for readers if they do not know (Ibid.).

1.4 Statement of the Problem

Since the creation of the Internet, it provides a new turn of opportunities to access the information (González, Macías, Rodríguez, & Sánchez, 2003). Throughout this invention, the information and data are available in the world at the same moment. The different abilities people can also interact with the web through assistive technology. But, still all the information on the web are not readable and accessible for all. People with disabilities such as blind users often face difficulties while retrieving information from the web through screen reader technology. For the readers with visual impairments, reading and comprehending content needs more effort and time as compared to regular readers. For instance, if a sentence is too long, blind users cannot remember the beginning of the sentence after they finish reading/listening (to) it. A sighted user can track and re-track the reading content back and forth while reading long sentences on the web, but this becomes challenging for the blind user who solely depends on screen reader technology. Blind users must listen to all the words in the sentence to comprehend its meaning. In this case, the same sentence length poses additional memory workload for the blind users compared to the sighted since they are unable to track sentences' content like those who can see. The appropriate length of the sentence is, therefore, essential on the web to make the content readable for the blind user. Hence, this study examines the impact of sentence length on the Web and the appropriate length of the sentence on the readability of Web for blind users.

1.5 Research aims and research questions

The final goal of this research is to investigate whether the length of sentence impacts on the readability of web for blind users and explores the appropriate length of sentence to make the web content readable and understandable. To accomplish this goal, first, the researcher designed five prototypes (webpages) with a similar design but with different content. The screen reader was used to read the prototype content. Further, the experiment was carried out by using five prototypes which were based on the comprehension test and perceived workload of participants. The NASA-TLX tool is used to measure the subjective workload.

The expected outcome of this research suggests that readable sentence length makes it easier for the blind reader to understand the web content and to reduce memory workload and frustration.

Research questions:

- Does the length of sentence affect the readability of the web in terms of workload for blind users?
- What would be the appropriate sentence length for blind users to make the web content readable and understandable?

1.6 Structure of the Thesis

This chapter provides an explanation of the organization of the thesis. Chapter 1 introduces the topic, background, disability, and universal design, web accessibility, statement of the problem, the aim of the research and the research questions. Chapter 2 is a literature review that presents relevant research. Chapter 3 presents the methods for data collection and analysis used in this research. chapter 4 explains data findings and results in details. Chapter 5 discusses these findings and what limitations can be applied to the findings. Chapter 6 presents the conclusion and future work with concrete measures to resolve the identified barriers in this research. Reference lists and appendices can be found after the conclusion section.

Chapter 2 - Literature Review

2.1 Readability on the web

Gottron and Martin (2009) researched on Readability and the Web by employing the current content extraction algorithms to determine the readability of web documents. Similarly, they also examined about how a web-based corpus statistic could be applied to determine the readability in a linguistic independent way. They have explained two readability indices that are Flesch Reading Ease and SMOG in the first section of their paper. The advantage of using these two readability indices was to measure the complexity of written text for example, word and sentence length and it shows the easiness to comprehend a text. Furthermore, they have used the content extraction process that determines HTML documents parts and it represents the main textual content. Under Content Extraction process, they have used document slope curves (DSC) and adapted content code blurring (ACCB). DSC identifies most of the text and it excludes the tags and it also locates scattered and interfered amounts of the document. Similarly, CCB has the benefit of visual attributes of the further contents. It also ignores anchor-tags in the process that contributes to outcome better results. In addition, they have applied content extraction algorithms to find better evaluations for the webpage readability. For this purpose, the researcher had examined 1114 documents from five distinctive web sites, and they structured only the main content. They used the web site like BBC News, The Economist, Herald Tribune, MSNBC News and Yahoo News. Then, they compared readability formula among with their key article and the head article was isolated through the “ACCB or DSC” algorithm. Their observation showed that SMOG and the Flesch Reading Ease index are more precise compared with content extraction. Their results showed that “corpus statistics” can be employed on the web to achieve language independent methods for readability.

Another research conducted by de Heus and Hiemstra (2013) on the Readability of the Web on around one billion web pages. They used the Automated Readability Index (ARI) to find out the mean grade level that needed to understand a website clearly.

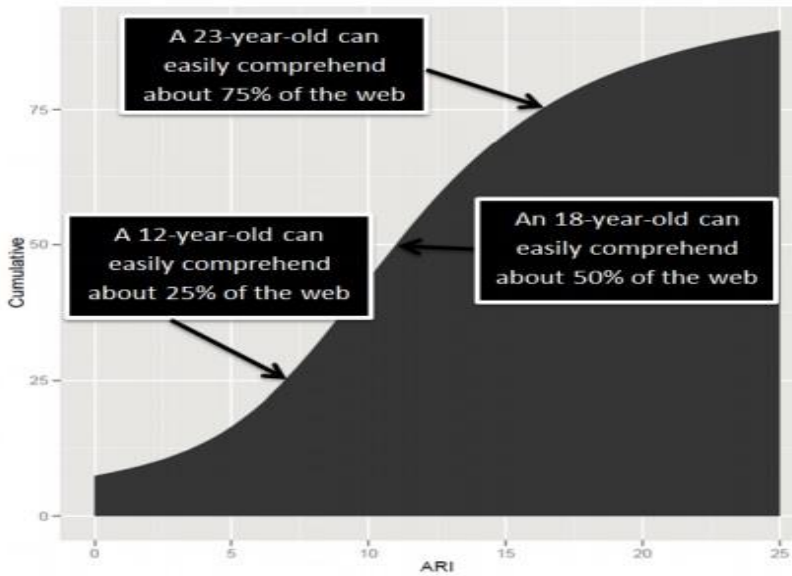

Figure: 2.1 Cumulative results (de Heus & Hiemstra, 2013)

They had used “MapReduce” program, which designed for real-time calculation of readability of digital typewriters, but it does not use the number of syllables. It was used on more than a billion webpages. The datasets called “Common Crawl” includes 61 million domains names, 92 million PDF documents, and seven million Word documents. Their 60 percent of information originated from commercial, organization and network websites. They did not filter the websites which has the content not written in English. The program uses the website text without using HTML tags and, it also processes the Automated Readability Index of the text. Their cumulative result shows that 12-year-old age, 23-year-old and 18-year-old can comfortably comprehend 25 percent, 75 percent and 50 percent of the content on web respectively.

Temnikova, Vieweg, and Castillo (2015) mentioned in their research paper how to write the post about disaster/crisis on social media in a clear and easy way so that the readers could understand easily. In their research, they had collected 100 of tweets and analyzed the data which were published through 15 several crises in native English tongue countries between the year 2012 and 2013. They also describe the factors that influence understanding negatively and review how to improve comprehension. Firstly, they have examined readability of short posts (tweets) published on Twitter. Tweets messages are about 140 characters. They have collected a data from CrisisLexT26. It was a free gathering of tweets from different 26 crisis events, which was held in 2012 and 2013. They have asked the participants to provide one tweet out of three options.

- Easy to understand (Clear)
- Requires slights IMPROVEMENT to make understandable (Partially clear)
- Difficult to understand (Not clear)

They have shown an initial investigation on the readability of posted tweets by institutes, NGOs, and media in a crisis that occurred in English speaking countries. They also suggest some recommendations for writing simple and clear tweets to understand text at time of calamity events.

- Message length: The tweets messages length should have 1 or 2 main points and it should have a brief, concise sentence.
- Vocabulary: Tweeter post should have use modest and straightforward words and Use acronyms and shortenings with prolonged form.
- Twitter particular elements: It is good to use “hashtags” while ending the tweets and avoid mentions @ users. Use of one hashtag is preferred to be better in tweeter posts.

Another research conducted by Rello and Bigham (2017), broadly discussed about increasing the readability for the users with dyslexia with the usages of colors. Their paper discussed the user study among 341 participants with 89 dyslexia; which find out the influence of using background colors on-screen comprehensibility. By reading the speed of participants along with mouse tracking, the readability was measured. In their experimental design, they have used independent and dependent variables. They have used different colors such as “blue, blue-grey, green, grey, orange, peach, purple, red, turquoise and yellow”. These colors were chosen as recommended and surveyed in past research works about dyslexia. Dependent variables use three measures such as reading time, mouse distance and mouse-tracking. Their experiment result shows that there was a significant impact of background colors in readability of text for both kind of user groups. Background colors like as orange and peach was useful for readability.

However, two colors such as blue-grey, and green reduced the text readability. The result shows that, while reading the text on screen, the people with dyslexia had considerably greater usage of the input device in the basis of the gap traveled by the mouse.

A study carried out by Banerjee, Majumdar, Pal, and Majumdar (2011), in which forty young participants were participated in the study. They have used Serif fonts and Sans serif fonts. The serif fonts are Georgia, Courier and Times New Roman while Sans serif fonts are Tahoma, Arial, and Verdana. They assigned 10, 12, and 14 fonts sizes. Eighteen passages were assigned to participants and all the passages were taken from Microsoft digital library. All the passages have same length and they were written about the same level of words difficulty. The font color was black, and its background was white. While conducting the experiment, participants were assigned a task to read in aloud from one of the passages as quickly as correctly as possible. Participants were assigned some comprehension questions to confirm their reading and understanding. They employed NASA-TLX tools to measure the subjective workload. This tool has six dimensions, but they employed only four factors such as “mental demand, performance, effort and frustration”. The other two remaining aspects like physical and temporal demand were not integrated to their evaluation scale. They have measured reading time, ranking and mental workload.

Their finding shows that there was better readability for Serif font compared to Sans serif because of minimum reading time for “Courier new 14 point”. Their research suggests that 14 font size is for reader to read in a computer screen. Also, they recommended Courier New in terms of reading time, performance, and mental workload rating of participants.

Hussain, Sohaib, Ahmed, and Qasim Khan (2011) have discussed on eight readability features like line and white spacing, color contrast, font size /style, text width, animation, graphics and headings. They compared these eight factors on how different age group users act with web applications by varying on these eight factors. They have proposed an idea based on the literature review on how to make the web more readable for all age groups people like teenagers, children, and old users. Similarly, around 83% of U.S. youths are using the internet, so information on the web must be design with the need for this large user group. They have discussed the readability design in one aspect of web design. It is crucial to develop a suitable design method for a designer for better web usability. They evaluate the web readability for Old age user by the following ways: Color contrast: Color contrast is very significant for web design. The user mostly reacts to text color and the elder's user react with the soft colors. Research shows that color patterns like grays, blues, and browns without flagrant contrast can fatigue the eye of older adults.

- White spaces: It is essential to isolate text and visual. Old age groups the user has several problems like eyesight, memory so text should be clear and simple.
- Line spacing: It makes the text more visible and ease to read. Improper line spacing makes the text overlap and challenging to understand for old users.
- Font style: Decorative font style is harder to read for the old age people.
- Font size: 12-point type is the default text size for the web, but the old user has poor eyesight, so the researcher suggests that font size of 14-18 is suitable for the reader of old age.
- Text width: It is better to narrow the field of view to boost the readability for old age also with lower literacy people.
- Headings: Proper heading and sub-headings help increase readability. They prefer a range between 14 and 30 points for headlines.
- Animations and Graphics: The animated text and its graphics distract elderly people which influence readability.

Similarly, they have discussed on web readability for children on eight factors:

- Color contrast: Children love bright color and they interested in reading mixture colors
- White spaces: They recommended white spaces among block of text.
- Spacing in line: It is crucial to make the text content simple.
- Font style: Children love to read stylish font and sans font.
- Font size: 14pt font size will be suitable for children.

- Text width: Text should be divided into small columns.
- Headings: It divides the text to make ease of understanding.
- Graphics and animations: Children like graphic and animation content on the web and picture assists them to comprehend the content easily.

Their results show that diverse age groups people have their own preferences on the web readability. They have concluded that readability factors effect on color contrast, font style, white space, font size, text, heading, graphics and animation for the understandability of content, text congestion and vocabulary. All these factors have significant influences on different age groups. From their literature reviews, they analyzed that, both color contrast and graphics have same impact for all age groups people.

Karmakar and Zhu (2010) have researched on visualizing text readability where they have developed a method that allows readers and writers to identify the distribution of complex words and sentences across a document. Using visualization method users can quickly compare not only the sentence lengths but also the syntactic structures of a sentence. They measured the word complexity by the characters per word, “number of syllables per word”, and vocabulary. It was easy to count the numbers of characters per word, but it was difficult to count the number of syllables. They have used Dale and Chall method for word complexity.

Similarly, they measured the sentence complexity by dividing the sentence structure into three levels such as clauses, words, and phrases. They found that if a sentence has two or more terms and the sentence often connected by conjunctions, then the sentence will be more complicated. They have used the “Stanford Natural Language Parser” tool to analyze the sentences and back to its grammatical structure automatically. That tool classifies words into subject or object. Finally, they demonstrated in their case studies that text readability could be adequately visualized. The writing styles of various authors are visible clearly. In their future work, they will expand the visualization of syntactical sentence complexity to the phrase level.

The researchers Chung, Min, Kim, and Park (2013) researched on the "readability of text-based web documents for deaf people". They proposed a newscasting display technique which converts difficult sentences into easy sentences, and it gives the relation between them with the help of visual illustration. Their live "news display system" that changes the real news with a simple syntactic structure and a graphical representation. They applied a clause relocation method which helps the deaf people to comprehend the hard sentences. They have developed a system that takes "a news article as an input text" which simplify the difficult sentences into simple sentences and it also make the sentence shorter too.

They have developed a system that consists of two sub-modules: "a graphical representation module and structural simplification module". The advantage of these modules was to represents the relationships between simple fragmented sentences, and the disadvantage is low literacy deaf people being not friendly with use. These modules are cheap and smaller than a sign language animation. Hence, their paper discovered the method of further improving "the readability of textual information" for deaf people in a web.

Jarrett, Petrie, and Summers (2010) have mentioned the design to read for people who do not read easily in their paper. They had focused some issues that cause people to have a problem for reading are vision, motor, cognitive problem (memory loss from aging), low-literacy and try to read in a second language. Also, they had included some unique issues for the people even those of us who are a skilled reader like lack of time, fatigue, stress, technology (reading on a mobile phone), learning in a crowded environment. Other barriers or obstructions to read are due to poor design, layout, and use of language in a web environment. They included some factor that makes the reading difficult like skilled reader use recognizing word shapes, letters in familiar chunks like "ing" or "tion". They read words in clusters which are relevant for meaning. They scan and skim the words that result they were easily disrupted from the physical level. Likewise, less skilled readers misrecognize words on the web page, and they struggle through words by letter. It creates confusion rather than understanding. So, they recommended in her studied that a web designer should have followed the accessibility set of guidelines to make the web content readable for all.

Another readability study has been done on the web page on bilingual and site readability by two researchers Lau and King (2006), they have used two languages English and Chinese for a bilingual evaluation in Web readability. They have conducted an experiment with actual data to estimate their outline. They had carried out two experiment for their university website to measure readability. They had run them all tests on Pentium 4 3.4 GHz as the operating system. In their first observation most of the web pages has scores between from 0 to five. They found that the pages which has title with hyperlinks to entire pages got low readability score compared to passage pages. Because the lengthy sentence which has no separators like full stop reduce readability. Their experimental findings showed that page readability is not just signifying difficulty, but it might be a suitable indicator of low text content such as “index and multimedia pages”.

Janan and Wray (2014) researched on the reassessing accuracy and the use of readability formulae. The purpose of their study was to review readability formulae and made the comparison of a variety of texts by six readability formulae. In their methodology, they have selected a total of 64 texts in English, and they have also chosen children between six and eleven years and all of them were native English speaker. They have used six simple readability formulae through the words count website that provides automated readability indices. A statistical test was carried out by using SMOG, Gunning Fog Index, Flesh-Kincaid, and Dale-Chall, and the consistency check was carried out among the six formulae in terms of their prediction's levels of text difficulty.

Moreover, they have demonstrated different readability indices for the same text by using different readability formulae. In their findings, some of them were consistent in the ranking of texts, but they were not consistent in the grading of every text. The significant advantage of their findings was readability formulae essential to support teacher's judgments about text difficulty. The reasons for using these six readability formulae were their popularity, and they are open standard over time, and they can be applied to any material without any cost. The aim of this methodology was it could be used as a benchmark index to guide the selection for children. The research was discussed and was carried out with native English-speaking children, and they have raised the question about is its relevance to the Malaysian context? They mentioned that they do not have enough resources in Malay language comparing to the English language. They have concluded that teachers need to weigh their professional judgments, in terms of their knowledge and give emphasis to the suitable text for the children to read. Readability formula indices can be implemented in a busy classroom, but it cannot only be the source of information about text difficulty.

Liu, Selker, and Lieberman (2003) discussed on visualizing the structure of a text. They introduced a graphical visual effect on a text document. They presented four-story documents with four users for example news article and novel. They had requested to their user to performed two times for every, and the result shows the speed of the task by an average of 27% for the known story and 36% for unknown documents. The results show that still there were not so many improvements for known stories, where users were more depends on their knowledge order in events navigation. Colors were used to represents the emotions and are series into a color bar, and the progression shows through the text document. In an effective structure of materials, they used in two stages. In the first stage, they have used every sentence in a report into one of the six basic emotion groupings like happy, sad, disgusted, fearful, neutral and surprised. Similarly, in the second stage, they have applied a heuristic smoothing algorithm for the sentence level for larger units, and it helps to purify the affective structure of the document.

2.2 Readability for blind users

A researcher Guerreiro and Gonçalves (2015) experimented with 30 participants who were visually impaired, they compared between “faster speech rate against the concurrent speech”. They combined these methods by gradually rising the speech rate from 1 to 3 voices. In methodology, they did a pilot study with five participants, where they gently increased the speech rate with voices to find out their highest values. They have used different spatially separated sounds to enhance speech segregation. In One-Voice, they used a female voice with one channel. In second voices, their speech were separated by 180 degrees, both female and male communicated to the users' right and left ear. Similarly, the third voices were unconnected by 90 degrees, two males' voices were spoke to the users' right and left ear. For data collection, they have focused on the relevance assessment. They have gathered two-hundred news from a Portuguese news portal. All the news were classified into three dissimilar topics such as "sports, politics, economy, television, celebrities, and arts". The duration of the experiment was 50 minutes, and it was carried out in a single gathering for the blind and visually impaired people. Their research result shows that default speech rate was marginally slower that Two and Three-voices that enables dramatically quicker scanning for relevant information. Their findings suggest that Two-voices with rate 1.75Xdefault-rate enables the appropriate scanning without loss in participants performance.

A research conducted by Aaron W. Bangor (1998), where they have described how the visually impaired users influence by the reading of computer text and how to improve readability for readers who had low vision. Author had used a "2X5X2X3" mixed-factor, experimental design to determine if there any changes in text size, contrast could improve reading rate and minimize the error rate for the people with low vision. They have experimented with unimpaired and impaired vision participants where eight unimpaired and eight impaired participants. They are intended to collect data from an equal number of females and males. The experimented was carried out at laboratory and participants used only a one-button Apple mouse to regulate the experimental program. The researcher used the three factors in the experiment to resolve readability difference among the unimpaired and impaired vision participants. Five different font sizes (12, 14, 18, 24, 30) were presented for the participants, and Sans-serif font of Helvetica was used.

The result shows that modification in the text does not affect sighted readers while surprisingly improving the reading abilities of the impaired participants. Font sizes, 12 and 14 were discovered to be too small for the visually impaired users.

2.3 Optimal sentence length

Ling and Van Schaik (2006) studied on the impact of “font type and line length” in web pages. From two experiments, their reports described that the impact of font type and line length based on the range of participants performance. They had used within-subjects factor which has four levels of line length: 55, 70, 85 and 100 character per line and font of the display text was the between-subject factor. They used Times New Roman 12-point and 10-point Arial, and they had recruited 72 participants for the experiments. They had used mock webpages and Microsoft Internet Explorer as the material and apparatus. Their 2-way ANOVA results show that there was a substantial effect of line length where shorter lengths were better based on preference and longer were faster for searching information although only in visual search. In their experiment 1, 70 character per line was better in terms users’ performance.

Jaan Mikk (2008) studied to find the range in sentence length for 17-18-year-old students in his paper. He has taken 30 Cloze tests where the students need to fill in the blanks. The Cloze test results were compared with the percentage of sentences over the boundary line. A total of 37 students participated in his research, and 30 texts were taken from favorite scientific books as a material. The author used within-subjects design as a procedure. The test was carried out with the texts with blanks, the participants needed to fill-up the blanks with deleted words. The sentence length was measured by Microsoft word. Their result shows that about 50-130 characters were appropriate range for 17-18 years old students. Their findings also demonstrated that the too short and too long sentences are not suitable for memory workload.

Martin Cutts (2013), ignored to recommend an upper limit of the sentence length, although if the sentence length exceed 40 words regularly that results undoubtedly weary and discourage readers. The better goal for an average sentence length of 15-20 words. The researcher argued that the key word is ‘average,’ so it is not necessary that all the sentences need to be in this range. So, other plenty of range could be possible.

Also, there is no lower limit on sentence length: a sentence can be just one word or two words.

Similarly, Björnsson (1983) have measured the readability of the regular daily papers in the 11 languages that included in the curricula for Swedish schools. He took 200 sentences from a newspaper randomly. He used different sentences length from different languages. For example, Swedish sentence length 17, Norwegian 20, Danish 22, English 25, French 23, German 22, Italian 30, Spanish 35, Portuguese 36, Finnish 14, and Russian 18.

2.4 Impact of sentence length on readability of the Web

Oelke, Spretke, Stoffel, and Keim (2012) presented a tool which is designed to assist the author to make their writing more comfortable to read. This tool indicates the complex paragraphs and sentences which are harder to comprehend. In their paper, they have introduced the tool "VisRa" which is developed to help the writer for correcting a text. When the text was loaded, this "VisRa" tool provides the user complete comment on sentences, and passages, and it indicates what is challenging to read and comprehend. The feedback not only shows the problem in a sentence, but it also explains the reason why the sentence is difficult to read. The feedback shows not only the issues on sentences, but it describes the reason behind the complex sentences to read.

"VisRa" has the following features:

- Word length: It measured the mean of characters.
- Vocabulary difficulties: It measured the terms in percentage that are not integrated in common list
- Nominal Forms: It measures the noun ratio.
- Sentence Length: It measures how many words in a sentence.

The advantage of their research they introduce a tool that visual analysis readability and support the author in refining a document, and it improves readability.

Another researcher DuBay (2004b), started to educate the literature from a “historical and statistical” point of view. The author compared between older writer and modern writer. The study observed that shorter sentences are progressing better over time and the statistical data began to count average sentence length. The studies also show that how length of the sentences reduces over time. The studies explain how sentence length averages shortened over time. Author also discussed that 50 words per sentence in before Elizabethan eras. The sentence length was shortened in Elizabethan times at 29 words per sentence. Again, the sentence length was reduced to 23 words per sentence in Sherman’s time. Hence, in modern time, the sentence length goes down by 20 words per sentence and the shorter sentences increase readability.

Another study examined by Kemal Zeki Zorbaz (2007) about variation of sentence length and level of readability of stories’ in text book of Turkish primary schools. The readability level of text was solved by employing “Atesman’s” formula. The formula was based on sentence length and word and revised by using Flesch’s reading Formula.
In their study, 46 textbooks with 66 texts were partially selected by random sampling from the Turkish lesson textbooks, and they were based on word-sentence lengths and readability. The 46 textbooks with 66 text were chosen randomly from “Turkish” lesson textbooks in the study. The stories were interesting to attracts the children concentration. For data collection, one hundred words of the stories were chosen to provide reliability. Further, the readability formula was used to determine the readability level of story. The data were analyzed by using mean of sentence length and word for every grade level. Similarly, a t-test was conducted to find out the differences among the grades.

The result of their study shows that there is no significant difference among the word lengths as per grades. Regarding to the sentence length, the findings illustrates that there was a very small differences among the stories of first and fifth grade. Finally, t-test result demonstrated only first, and second grade had significant differences.

2.5 Summary

Overall, from the literature review related to the readability on the web, it can be concluded that despite the available of variety of conducted researches, related to different readability factors, the impact of sentence length on the readability of the web for blind users was not sufficiently investigated.

Unfortunately, the number of studies discussed in literature review does not have any original research explored in the context of blind users on the web. Those studies raised an issue about font types, visual text, background color, sentence length and word complexity, and their finding suggested for sighted readers to improve readability but not for actual blind readers.

Hence, this research will be addressed above literature review issues to improve better readability for the screen reader users (blind) on the web by employing new methodology and data collection procedure.

Chapter 3 – Methodology

This chapter presents the experimental research methodologies employed for data collection and data analysis. The quantitative data was collected from the participants by conducting a controlled experiment. Quantitative data included the workload perceived by participants while reading the content of each prototype and performing comprehension test from each prototype content. The NASA-TLX tool was used for assessing the subjective workload. The one-way repeated measures ANOVA was employed to investigate whether the impact of sentence length on the readability of the web for blind users. It was assessed based on the perceived subjective workload that participants experienced while reading the content of five prototypes through a screen reader technology.

3.1 Participants

In this research, participants were gathered from Oslo Metropolitan University. All participants read English fluently, but they were not a native speaker, and most of them were from the Master of Universal design of Information and Communication Technology program, and few of them were from another educational background. Thirteen male and eight females participated in the study (for a total N=21) with a mean age of 28. Participants were chosen randomly based on their knowledge and familiarity on the web and accessibility. At least each participant had a master or bachelor’s degree, and all had an experience with computer and internet. None of them was recognized as a blind participant. All the participants were blindfolded and firstly briefed and provided information sheet. Majority of participants had no experience of using screen readers technology although they were concisely informed to use them prior the test. Participants were assigned to the experimental conditions randomly. All the participants had been reported before the beginning of the experiment.

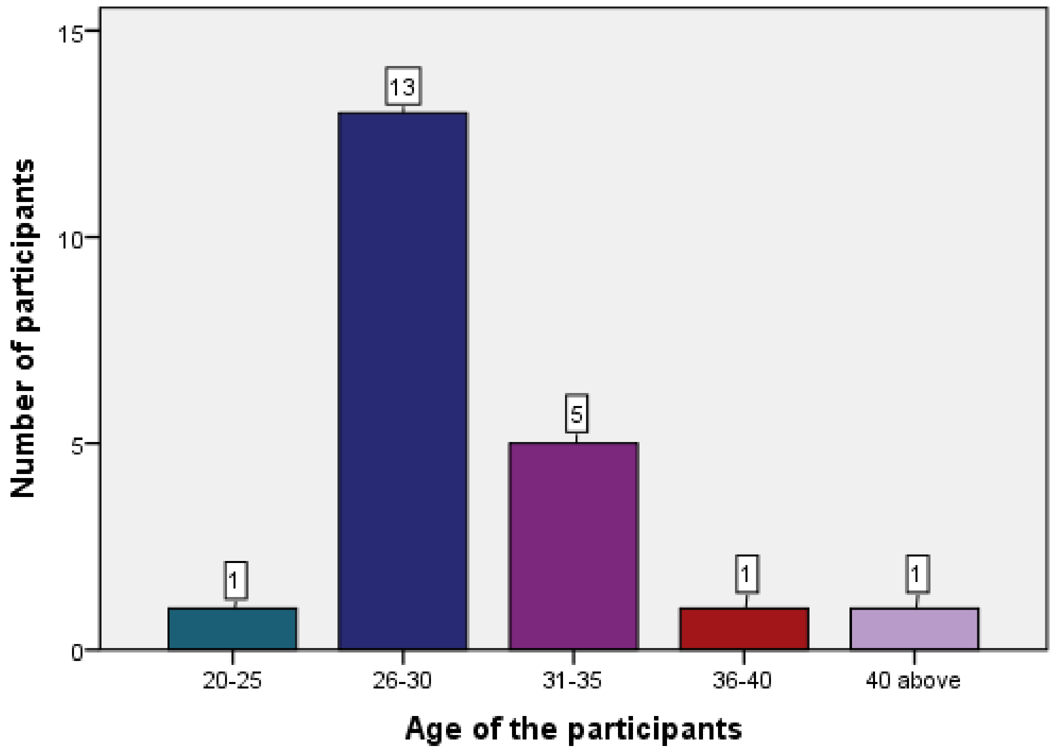

Figure: 3.1 Distribution of sample by participants age

Figure 3.1 shows the number of participants according to their age. There were a total of 21 participants in which they came from the age groups 20-25, 36-40 and above 40 consisted of only one participant. Though age limitation was not the requirements for participation in the study, all of them were above 20 years age, and all the participants were pursuing their university education. Most participants (N=13) were in the age group of 26- 30; 5 participants were between age 31-35 and while only one participant was over age 40.

Similarly, figure 3.2 shows the number of participants in terms of their gender, level of education and English proficiency. Among total participants (N=21), 13 were male, and eight were female. The highest number of participants (N=20) were master's degree students, and only one was a bachelor's student. Most of them were from ICT background students and a few of them from other faculties. Moreover, language proficiency was based on subjective self-reporting, 20 participants had an advanced level of English proficiency, and only 1 had an intermediate level, but none of them was the native English speaker. All participants participated in the experiment and performed the task within an estimated time boundary.

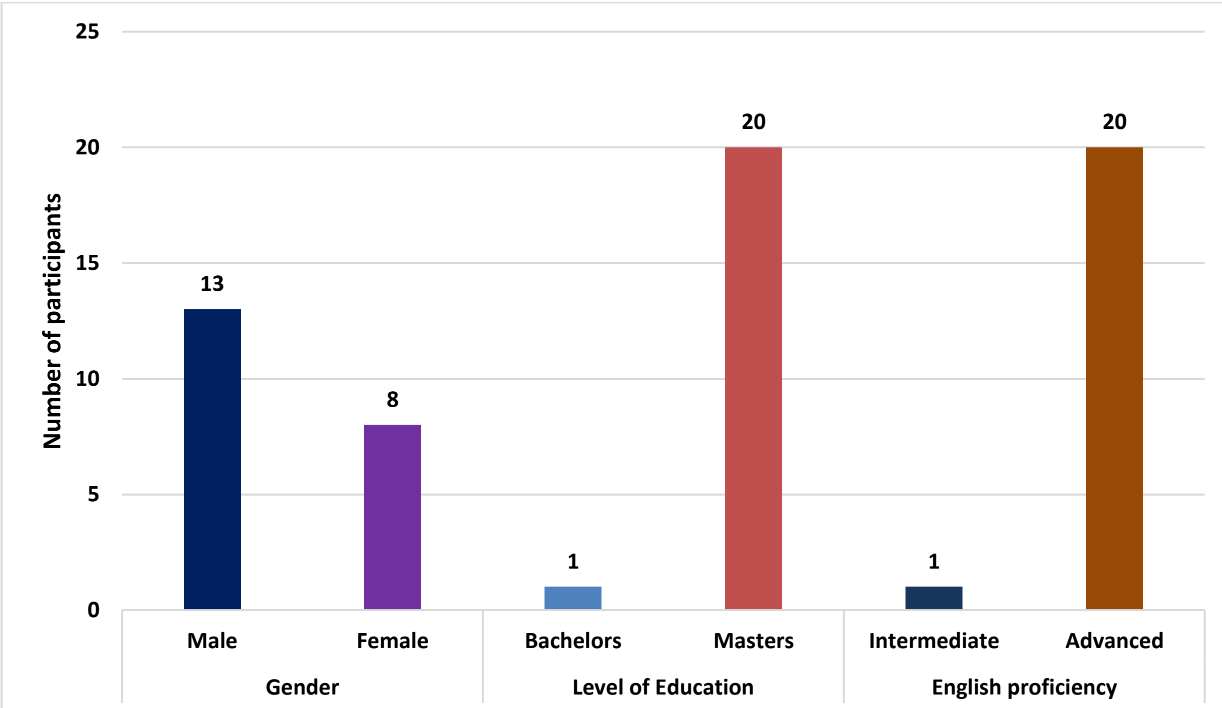

Figure: 3.2 Number of participants based on gender, level of education and English proficiency

3.2 Materials

In this research for the experiment process, we had designed five webpages as a test prototype, comprehension test as a task for the participants, NASA-TLX tools to measure the workload of the participants and NVDA (non-visual desktop access) screen reader as assistive technology.

3.2.1 Prototype

According to Moggridge and Atkinson (2007), a prototype is a representation of a design, made before the final solution exists. It is crucial to plan for what types of data need to be obtained in the testing. Many considerations need to be made to decide on the nature and scope of the prototypes to be developed (Eladhari & Ollila, 2012). In order to test the impact of sentence length on the readability of the web for blind users, five webpages were developed as a test prototype. Each webpage consists of the same layout, but the different content. The following are the prototype design and its phases of the process.

I. Understand users (task, goals, context): It is important to involve the guidelines when designing a specific prototype (Petrie & Bevan, 2009). It should be designed in terms of user needs.
II. Concept design: After understanding users, the idea of an initial design can be described. Also, a design space should be explored to consider alternative design and to meet users' needs (Petrie & Bevan, 2009). Hence, the author of this research selects news website as a concept for the prototype design, in which the majority of web readers are familiar with. Therefore, it would be easy to find likely users. According to the World Health Organization (2017), around 8.1 million people in the world are visual impairment, and most of them use the internet for reading online news. Thus, the researcher preferred to select the online news website for the prototype design in this experiment.
III. Prototype: Hence, after the design concept was completed, the prototype was designed. The designed was completed by using Photoshop version cs5. First .psd extension file was designed with different layers and by adding content and images. Then the .psd file was converted into HTML template. A complete webpage was designed by using CSS, and JavaScript. Figure 3.3 showed the PSD template that was designed in Adobe Photoshop.

3.2.1.1 Equipment for prototype

This section includes both hardware and software technology that was used in designing five prototypes.

3.2.1.1.1 Hardware Technology

A Laptop, Windows 10 Pro 64-bit Operating system, processor Intel(R) Core (TM) i5-4210U CPU @1.70GHZ (4 CPUs), 4096 MB RAM, 14 Inches of screen size with the touchpad mouse.

3.2.1.1.2 Software Technology

- Adobe Photoshop cs6: It is an image editing, image creation and graphic design software (Caruso & Postel, 2002). In this study, Photoshop was used to design the web page template for the prototype, where the design of the image was saved in Photoshop format (PSD). It preserves all Photoshop features such as layers, effects, and masks.

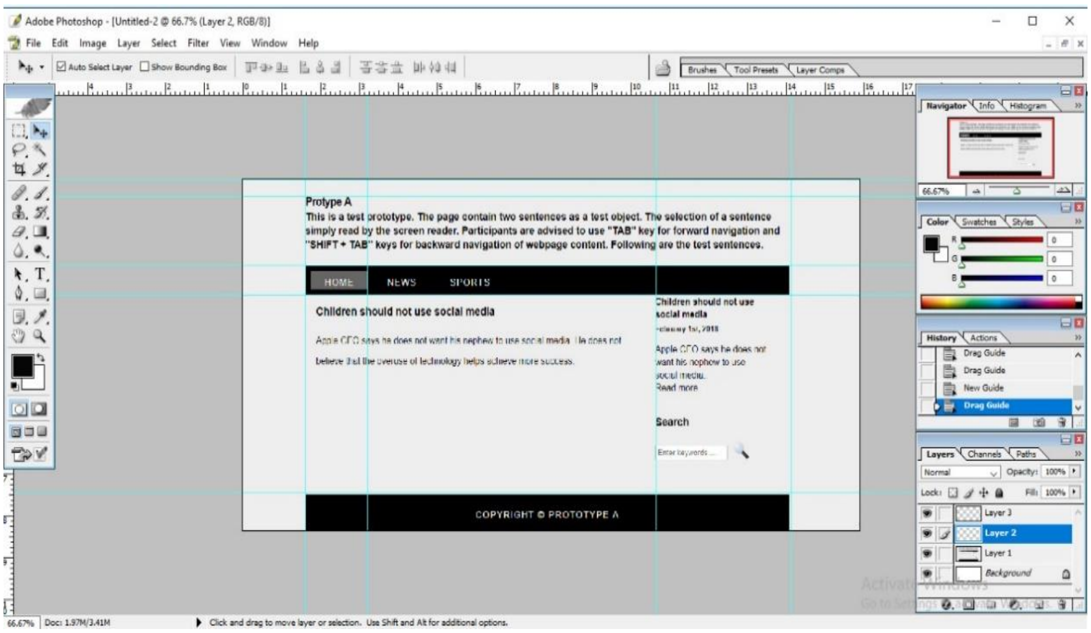

Figure: 3.3 A Photoshop design (PSD) of prototype

- Sublime Text 3.0: Sublime Text was used as a code editor to write HTML and style sheet. It is cross-platform with a Python application programming interface (Kinder Ken, 2013). It is an open source and free-software for download. It has several features which make easy to quick navigation to files, make the same interactive changes to multiple selected areas simultaneously. It can be installed in Windows, macOS, and Linux and it is compatible with many languages.

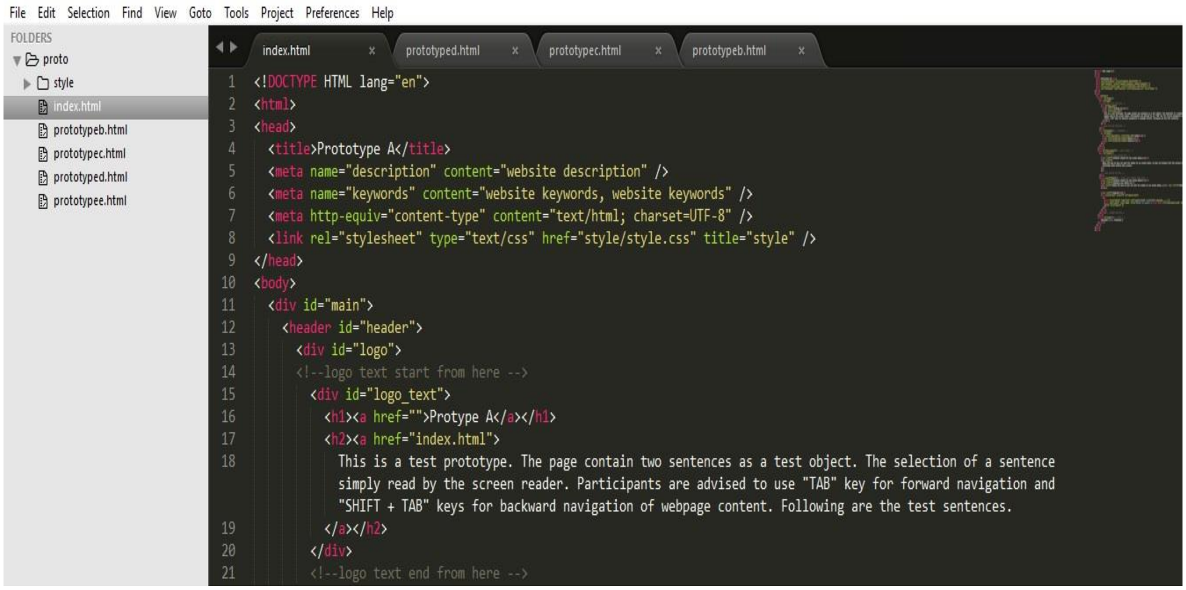

Figure: 3.4 Prototype design code in sublime text editor

- CSS: In this study, CSS is used to define styles for five prototypes, including the design, layout, and font. It set rules that determine how a webpage should look.

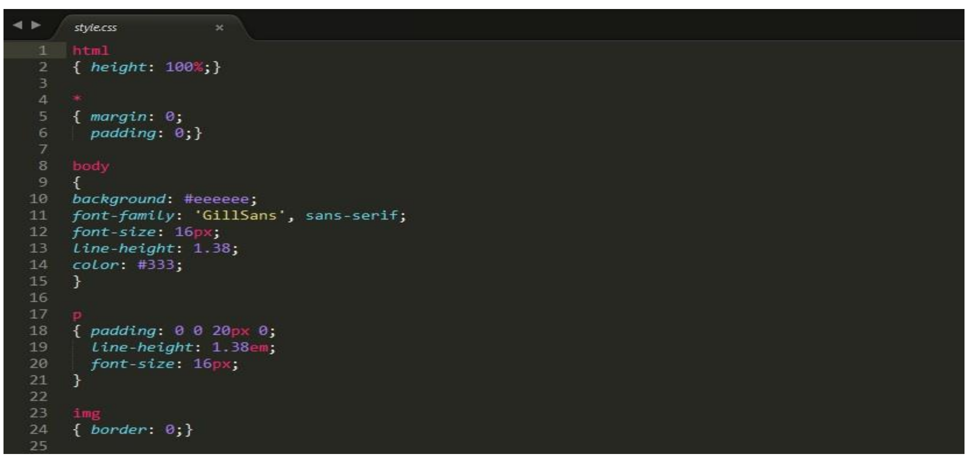

Figure: 3.5 CSS code in prototype

The external stylesheet is used while designing the prototype. A CSS file was written in another file with .CSS extension and it was linked with the index.html (home page).

- HTML: It was used to publish content with heading, text, tables, list, photos, etc. It supports video clips, sound clips, and other applications directly in documents (Raggett, Le Hors, & Jacobs).

```
    <h2>This is a test prototype.</h2>
```

Figure: 3.6 Heading structure of Prototype

```
<nav id="menubar" role="navigation"><!-- navigation -->
    <ul id="menu">
        <li class="selected"><a href="index.html">Home</a></li>
        <li><a href="#">News</a></li>
        <li><a href="#">Sports</a></li>
    </ul>
</nav>
```

Figure: 3.7 Navigation List

```
<p>
  Apple CEO says he does not want his nephew to use social media. He does not
  believe that the overuse of technology helps achieve more success.
</p>
```

Figure: 3.8 Paragraph structure

Moreover, when designing a prototype, the researcher followed the Web content accessibility guidelines to make the content accessible and readable by the screen reader technology. All the prototype was accessible just by using the correct HTML elements for the right purpose. The prototype design has a content structure of headings, lists, and paragraphs.

3.2.1.1.3 Prototype contents

Based on a literature review, the researcher decided to choose different sentence length (Prototype A, 10-15 word; Prototype B, 16-20 word; Prototype C, 21-25 word; Prototype D, 26-30 word; and Prototype E, 30 words or above) in five prototypes (Webpages). The content of all prototypes were taken from different online news portals such as BBC, Yahoo, Norway today, New York Times and The Local. The content that were taken from these online news portals were without any modification; however, all were from different categories like technology, education, and entertainment. They were chosen for a test procedure for blindfolded users to examine how the impact of sentence length on the readability of the web for blind users. Each prototype has a similar layout but variance in sentence length. There were two sentences of the same length in each prototype content.

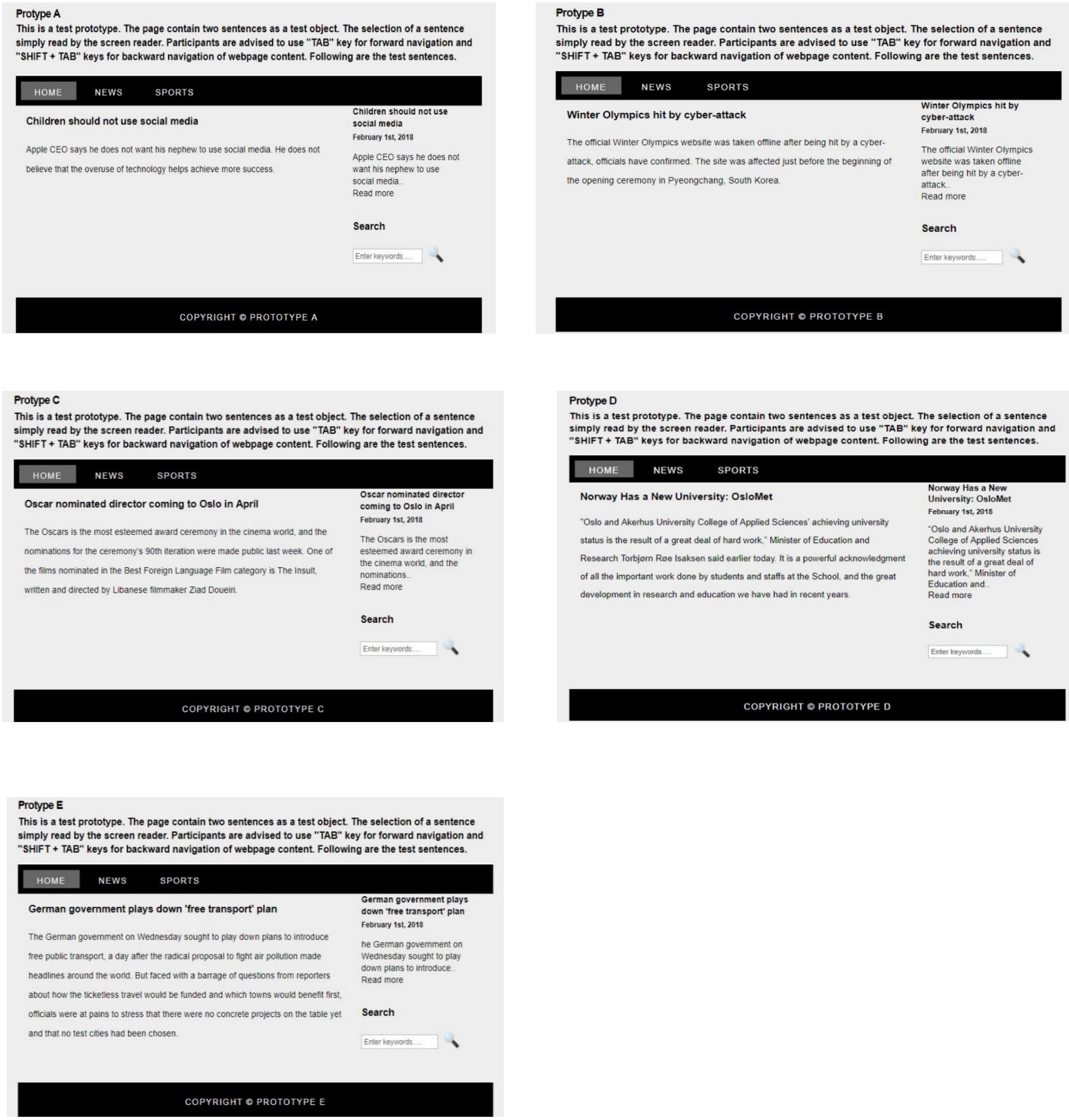

Figure: 3.9 Home page of five prototype, which consists of different sentence length, a)10-15 sentence length, b) 16-20 sentence length, c) 21-25 sentence length, d) 26-30 sentence length, and e) 30 above sentence length

3.2.2 Comprehension test

In this research, the Comprehension test assessed a participant's ability to read and understand the content of the prototype. The aim of this test was whether the participants understand the prototype content or not. The test was carried out with 21 blind-folded participants and they had to read the content through the screen reader (NVDA) software. Each participant was assigned by two multiples choose questions from every prototype, and they had to choose to answer accurately. There are a total of five prototypes and each prototype had two sentences of the same length as a content. The content of the prototype sourced from different online news portals randomly. If the participants could not concentrate on reading and do not able to answer the question, they had the option to guess or leave it as their wish.

3.2.3 NASA-Task Load Index

The researcher employed the NASA-TLX tool to measure the perceived workload of participants in the experiment process. It is the standard tool used by many studies in Human Factors and Ergonomics (Sandra G. Hart, 2006). It is a highly reliable and validate a tool to measure perceived the workload of the participants (Longo Luca, 2018). Moreover, it is broadly used for subjective multidimensional workload assessment (Hart & Staveland, 1988). It allows the researcher to determine the perceived workload of a participant while performing a task. Mental workload vary among every individuals hence it is challenging to distinguish the amount of workload in specific person in terms of their characteristics' (De Alwis Edirisinghe, 2017).

Though, it is likely to determine the level of mental workload. For instance, it is possible to measure a person high mental workload or low mental workload based on their experience. The NASA-TLX rates the performance of the participants through six dimensions to discover an average workload rating. It is the subjective multi-dimensional assessment scale that evaluates subjects' workload on six subscales: "Mental Demands (Md), Physical Demands (PD), Temporal Demands (TD), Own Performance (OP), Effort (EF), and Frustration (FR)" (Hart & Staveland, 1988). A rating scales definition is described in Appendix C.

The NASA-TLX employed 2-part of assessment procedure that consists of weights and ratings. Each participant had to evaluate to his/ her contribution of each factor in terms of workload during the task. It measures subject workload by asking participants to describe the workload they experience while performing the task. It does not measure the details about the nature and objective of task, it only focuses participants' feelings about their workload.

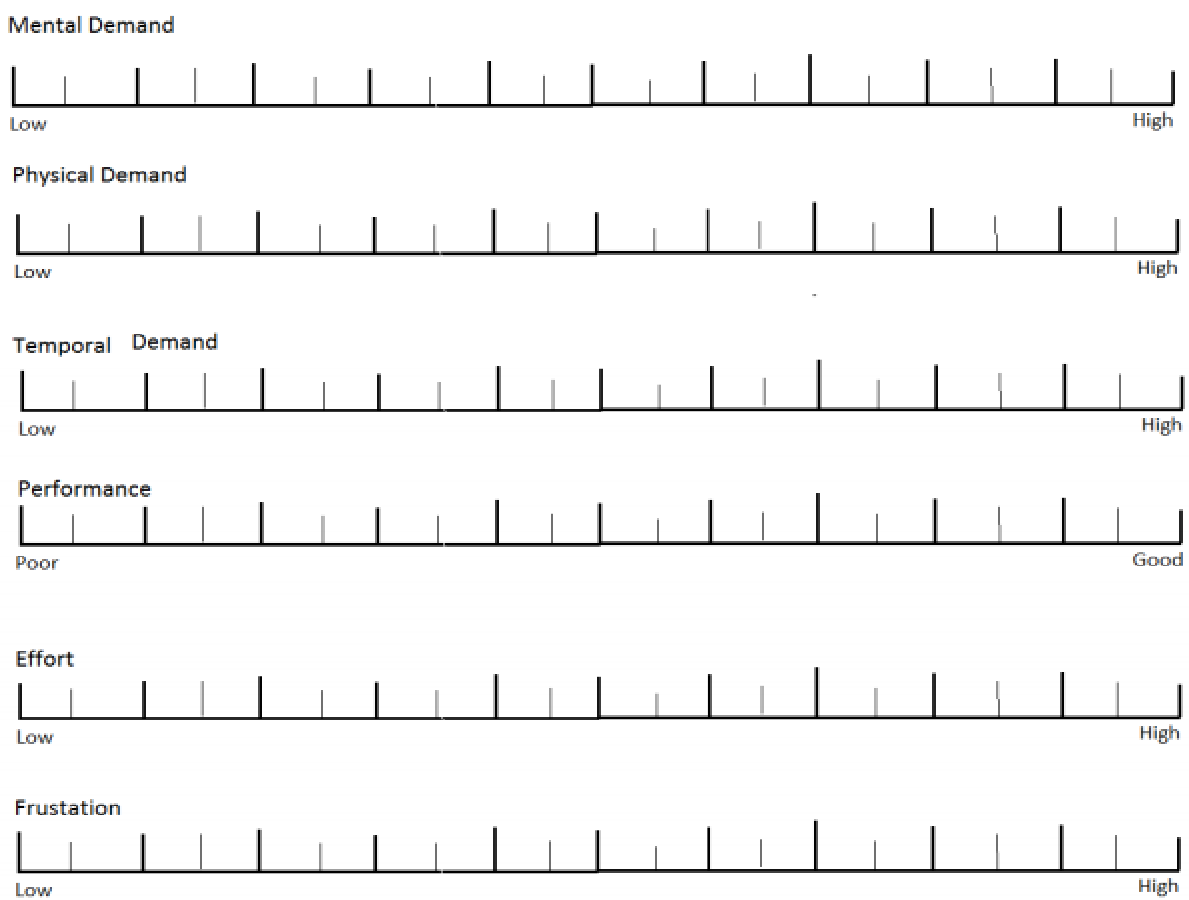

Figure: 3.10 NASA-TLX rating scale

Twenty step bipolar scales (semantic differentials) were applied to get ratings for the dimension (Hart & Staveland, 1988). Bipolar is a specific type of rating scale characterized by a range between two opposite endpoints. A score range from 0 to 100 (allocated to the closest point 5) was taken on each scale (Ibid.). After the participants' ratings, 15 possible pair-wise comparisons were carried out in terms of six scales (Hart

& Staveland, 1988) (See Appendix B). Subjects circled the scales from the pair where they had contributed more workload during the task. The tallies range is from 0 to 5 where 0 denotes not relevant, and five denotes more crucial than any other factor. The number of times that the subject selects each element is counted. The average workload score of individual subjects was calculated by multiplying each rating by the weight. (See Appendix F and G for a sample Tally Sheet and Worksheet).

3.2.4 Non-Visual Desktop Access (NVDA)

NVDA is open source software which is free of cost to download, and its code is available to everyone (NV Access, 2019). It allows developers and translators to contribute its further improvement. It can be installed in the Windows operating system and other software applications to access and interact for blind and vision impaired users. NVDA supports popular web browsers such as Mozilla Firefox and Google Chrome, email client's office programs like Microsoft Word and Excel (NV Access, 2019). It supports more than 50 languages, and it is developed in speech synthesizer. It announces automatic when the mouse hovers the text. Moreover, it supports several braille displays which have a braille keyboard. It is portable and can be operated from USB flash drive without installation.

3.3 Experimental Research

3.3.1 Research Hypotheses

The following statistical hypotheses has been generated based upon the assumption about impact of sentence length and appropriate sentence length in the context of readability of the web.

Hypothesis 1:

H_0: There will be no significant impact on sentence length on the readability of the web for blind users.

H_1: There will be a significant impact on sentence length on the readability of the web for blind users.

Hypothesis 2:

H_0: Minimum sentence length on the web will not be appropriate for the blind user in terms of subjective workload.

$H_{1:}$ Minimum sentence length on the web will be appropriate for the blind user in terms of subjective workload.

3.3.2 Dependent and Independent variables in Hypothesis

Dependent variables: The comprehension score and NASA TLX score are the dependent variables that were measured to find the understandability of a sentence and its impact of sentence length on the readability on the web. A dependent variable is a variable that was tested and measured in an experiment. The word “dependent” is used to indicate a variable is dependent on a participant’s behavior (Lazar, Feng, & Hochheiser, 2017).

Independent variables: The variable that is manipulated by the researcher, and its conditions are the major factor being examined (Gergle & Tan, 2014). The independent variable is manipulated in a series of an experiment by the researcher or experimenter in order to see the changes on dependent variables; therefore, each independent variable should have at least two levels of treatment variables or conditions (Lazar et al., 2017). The hypothesis in this research consisted of one independent variable i.e. prototype (website), and the independent variables have five levels of treatments or experimental conditions that would make changes in comprehensibility and workload of the participants – which included: prototype consisting of sentence length 10-15 (prototype A), 16-20 (prototype B), 21-25 (prototype C), 26-30 (prototype D), and 30 above (prototype E). The tasks such as a multiple-choice question and NASA-TLX workload rating performed by each participant over five different treatments – is not considered as an independent variable or dependent variable in this experiment, due to the interest of each hypothesis which was only on the impact of the type of websites (prototypes) on dependent variables. The effects on the dependent variable are measured in terms of changes on types of websites but not the task type or complexity. For example: - the five prototype websites will be tested to determine the impact of sentence length and the appropriate length of the sentence on the website, for this, the participant's performances and their workload will be measured in order to investigate changes while working over different prototype websites.

3.3.3 Within-subjects design

A within-subjects design was used in this research because each participant was measured in each prototype (condition) and their scores were compared. The main advantage of using this design was it does not require a large number of participants. Moreover, each participant test under one condition was repeated in the other environments too, which effectively allows them to serve as their control (MacKenzie, 2002). Another benefit of using this design is all prototype tested by the exact same participants (Raluca Budiu, 2018). It reduces the random noise, for example, every participant has their own behavior, backgrounds, and context. Subjects might be tired after traveling and another might be bored and might be influenced by sad news (Raluca Budiu, 2018).

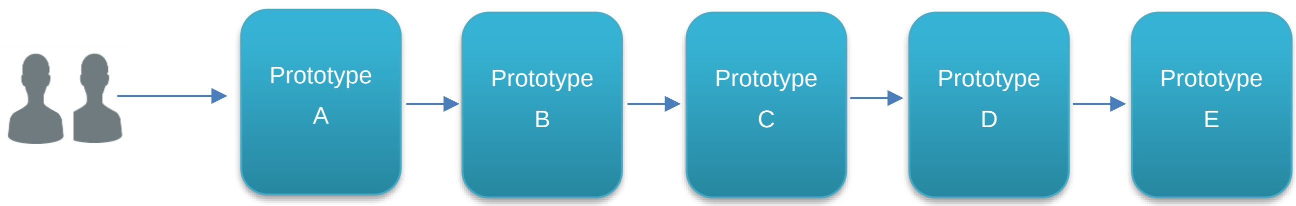

Figure: 3.11 Within-subjects design (participants test all prototypes)

It is an appropriate method for testing participant for data collection procedure because if an experiment seeks to investigate over multiple sessions and reduction in error variance associated with individual differences (MacKenzie, 2002).

In this research, first, comprehension test was carried out, and then NASA TLX tools were used to test the perceived workload of the participants. For the independent variable, the researcher decided to employ five different prototypes: Prototype A, Prototype B, Prototype C, Prototype D, and Prototype E. Each participant had to complete comprehension test before taking NASA-TLX workload. The allocation of equal numbers of participants to each of the five conditions (prototypes) was randomized.

3.3.4 Randomization

Randomization is vital for within-subjects design because it contracts possible order effects and minimizes transfer and learning across conditions (Raluca Budiu, 2018). This method assured that each participants has an equal opportunity of assigning any of the condition under the study and it eliminates the selection bias (Suresh, 2011).
Each subject had the same chance of being assigned to either involvement or control. It removes examiner bias because no one including the examiners themselves can describe a possible condition to which a person who was part of the study is going to be assigned (Suresh, 2011). Within-subjects experimental design has been implemented where each participant has tested under each condition over five experimental conditions. The randomization method overcome the issues that might affect the behavior of the subject's such as fatigue or outside factors. Hence, the researcher used randomization technique, which reduced the examiner bias and order of conditions.

Table 3.1 Randomization

Partisipants ID	**Prototype**				
1	A	B	D	C	E
2	C	E	B	A	D
3	D	C	A	E	B
4	A	B	C	E	D
5	A	D	E	C	B
6	D	B	C	A	E
7	E	C	D	B	A
8	B	A	D	C	E
9	C	E	B	D	A
10	E	A	C	D	B
-----	-----	-----	-----	-----	-----
21	A	B	D	E	C

In this study, the researcher recruited 21 participants and assigned two multiple-choice questions to one of the five conditions (prototypes). During each task, the participant completed a comprehension test where they had to answer two multiple choice questions. The examiner started anywhere in the random condition. It continued and repeated the process until all 21 participants were assigned to specific conditions. The researcher used online randomization software resources to generate it — the primary reason to use randomization techniques that experimental research can achieve its goal. For example, as shown in Table 3.1, the first participant with ID1 and the second participant with ID 2 had performed the same experiment condition but with random experimental conditions. Similarly, the same tasks were also assigned to different order to other participants as well.

3.5 Experimental Procedure

Participants were informed before one week about the test condition when they received the information sheet. The experiment was conducted with an individual participant at their own recommended place where they feel comfortable to execute an experiment. The researcher as an investigator and the participant were located in a room where nobody is allowed to enter during the experiment period. Laptop and the desktop computer were used and the investigator with the experiment report form. The estimated experiment time was 40 minutes for each participant. The experiment procedure was based on within subject-design and randomization techniques to reduce bias. The experiment process was segmented into three sections:

I. First section: In this section, participants were asked to read the information sheet and when the participant agrees to the purpose of the study, procedures, and to be aware of risks while taking the test. Then the participants were provided with the demographic questionnaires.

II. Second Section: In this section, the researcher clearly explained about a screen reader technology, and experimental procedure. Firstly, the participants were familiarized with the NVDA screen reader application and then they were taught to use it. A total of twenty-one participants took part in the experiment and were blindfolded during the experimental procedure. Participants were asked to put an eye-mask to cover their eyes while taking the experiment. As the participants were blindfolded, the researcher was responsible to open the prototype and begin the process. The participants read the content of the prototype with the help of the screen reader. After that, they were allowed to open eye-mask and see the questionnaire of each prototype after reading the content through a screen reader. They were encouraged to speak out what they thought about the prototype and its content. The number of attempts each participant make while reading the content during the test was recorded in an observation sheet. The entire experiment was based on within-subject design. There are five different prototypes and each prototype has the same design, but with different sentence/content length. The source of content was from various online news portals. Five participants at a time were used for the experiment and were introduced about the purpose of the study.

III. Third section: After completion of reading the prototype content, the participants were asked to take the comprehension test. In the comprehension test, the participant should answer two multiple-choice questions from each prototype. Then, immediately they need to fulfill the paper and pencil versions of NASA-TLX ratings. Each fulfillment of the questionnaire, the participants were asked to take a rest for 1-2 minutes and then again started the experiment along with another prototype.

3.6 Ethics

According to Resnik David BJ (2011), ethics are the norms or standard that differentiate between right and wrong. It helps to conclude the difference between acceptable and unacceptable behaviors. This study is quantitative nature and it involves the hypothesis, data collection, analysis, and interpretation. It is crucial in research that to protect the misrepresenting of information and therefore, encourages the recreation of knowledge and fact (Resnik David BJ, 2011). Researcher need to clarify the objective of the study and the usage of the fetched information. In this research, all the participants were invited to take part in experiment process, therefore, all received a written information sheet about the description and the goal of the study (see Appendix A). The aim of this information sheet was to make clear about the purpose of participation. A written information sheet was sent to all participants before conducting the experiment. Further, participants confirmed that they have accept the information and agreed the purpose of the study.

To address ethical considerations in an efficiently manner voluntary participation was conducted in this research. Also, the participants had rights to revoke from the study at any time as their wish. The information sheet provides the adequate information and assertions about taking part to permit individuals to comprehend the consequences of participation, and they have the right to decide whether or not to do. The contact information such as name, signature, and workplace were not included in the information sheet. To gather the information of participants, the researcher recorded the age, education, and language proficiency. During the experiment, no video or audio recording was used, but only notes were taken. No personally identifying information was recorded on the notes. So, this research is not subject to notification to Norwegian Centre for Research Data.

3.7 Pilot Study

According to E. R. Van Teijlingen and Hundley (2001), a pilot study is a research study conducted before the real investigation to test research procedure, data collection, and apparatus before beginning the real test. It is a crucial phase in research and is conducted to identify potential problem and limitation in the study before implementation, but on a smaller scale. In this research, the researcher conducted a pilot study among two participants to make sure that it might provide caution about where the primary research task could fail, where the research process could not be followed, or in case, the planned methods or tools are inappropriate or too complex (E. Van Teijlingen & Hundley, 2010). Other intentions for running pilot studies were to find out possible practical issues in developing the research technique, assessing the feasibility of a study and collecting preliminary data. Based on within-group experimental design, each participant was assigned two multiple-choice questions from each condition (prototype) to perform a comprehension task and immediately paper, and a pencil and pencil-based NASA-Task Load Index was assigned to each test over five conditions. It was analyzed from the experiment that there was potential bias due to experimental procedures, such as a large number of tasks; the whole session took longer time than expected.

Furthermore, participants felt fatigued during the experiment period. As the participants were blindfolded, they lost concentration while reading the first text through screen reader technology. The experiment was started from a prototype A (short sentence length) to prototype E (longer sentence length), that made bias on the result so to overcome this problem randomization test was assigned in actual data collection. All instruments such as paper, pencil, laptop, screen reader software, the webpage of the prototype were opened, and an observation sheet for recording data was ready before beginning the actual data collection. Similarly, some changes were made before the real data collection for making the data collection procedure smooth.

3.8 Data Analysis

The collected data were analyzed statistically using descriptive Statistics. The Statistical Package for Social Science (SPSS) for Windows version 24.0 (Green & Salkind, 2016), was used to perform the descriptive analysis while evaluating the results obtained in the experiment. Descriptive statistical techniques such as mean, standard deviation, correlation coefficient, percentage, and other analytical tools were used in this study. The data was imported from Microsoft Excel to the SPSS package and was analyzed. Based on the results, rejecting or accepting the null hypothesis was done by checking the significance of the mean difference of the prototypes between different workload dimensions.

The one-way repeated-measures of ANOVA was used to compare the subjective workload scores over five different prototypes in an experiment. The descriptive statistics, Multivariate tests, tests of within-subjects' effects, estimated marginal means and pair-wise comparisons were used to test the variance in terms of participants' average workload. The results were produced with confidence intervals of 95%, and significance was assessed to be present if the probability was less than 0.05.

Chapter 4 – Results

This section presents the results of an empirical study conducted on the impact of sentence length on the readability of the web for blind user. The objective of this investigation is to discover the appropriate length of sentence for blind users to make the content readable and understandable. The first section includes the comprehension test score based on different prototypes content. Similarly, the second section shows the outcomes of one-way repeated measures of ANOVA (within- subjects). Finally, the third section demonstrates the result of the average workload in every five different prototypes. The result obtained from the data analysis aims at answering the above (section 1.5) research questions and hypothesis (section 3.3.1).

4.2 Comprehension test score based on different prototypes

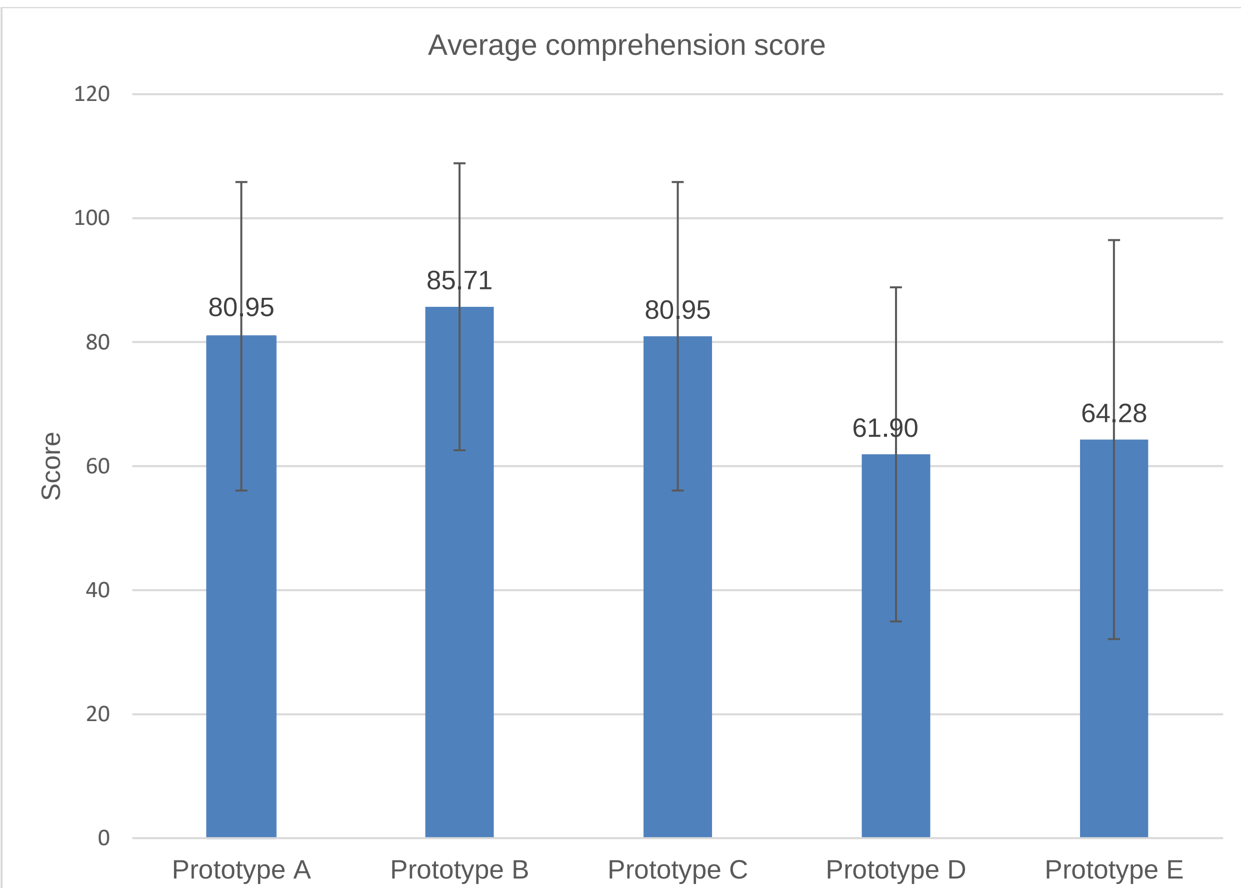

Figure: 4.1 Average comprehension test score with error bars on five prototypes

The reading comprehension test had been conducted before taking the NASA-TLX workload. The main goal of this test was to explore the reading ability of participants and their reading strategy on the web. Each content had two questions that measured reading comprehension level. The participant's analysis score was based only on the reading comprehension questions. The whole test contained ten multiple-choice reading questions, but each prototype had two questions of each content. Each content had two sentences of different length, and the content was taken from various online news portals. The material was related to education, social media, entertainment, and government transport plan. But it was not the mandatory requirement in an experiment, but a different category of content boosts participants prior knowledge and understandable level. Participant’s response to the test received 50 points for a correct answer and zero points for an incorrect response and the maximum score was 100.

The figure 4.1 showed the average comprehension test score of each prototype. The experiment consisted a total of five prototypes (From A to E) with different sentence length, and each prototype has two sentences as content. There were two comprehensions multiple choice questions from each prototype. The comprehension score was estimated by adding up the correct responses from the questionnaire. All participants (N=21) responded to the comprehension test. The total score was 100, and all participants scored above 50% in each test. There was a significant difference in comprehension test score among prototypes B, D, and E but prototype A and C has the same mean score. The result shows that prototype B has the highest mean score (mean= 85.7) among five prototypes. It has a word length between 16-20 whereas the prototype D has the lowest mean score (mean= 61.9) and it has a word length between 26-30.

4.3 One-way Repeated Measures ANOVA

Table 4.1: Descriptive statistics

	Mean	Std. Deviation	N
Prototype A	41.22	15.30	21
Prototype B	40.15	18.45	21
Prototype C	54.60	14.79	21
Prototype D	64.76	16.51	21
Prototype E	65.79	17.77	21

Table 4.1 provides basic descriptive statistics for the five-level (from prototype A to E) of the independent variable. We can see in table 4 that there is a mean difference on an average workload of five prototypes. Participants used less workload in Prototype B (M= 40.15; SD= 18.45) compared to other prototypes and highest workload in prototype E (M=65.79; SD=17.77). Moreover, the prototype A (M=41.42; SD=15.30) has a slightly higher workload than prototype B. Similarly, the prototype C (mean=54.60; SD=54.60) and prototype D (M=64.76 SD=16.51) has the mean difference of 10.14. From this descriptive table we can see that, on the average mean of the NASA TLX workload, prototype B has the lowest workload mean among other prototypes. But to find out whether these observed mean differences are significant, it needs to look at the tests of within subject's effects. This study involved five prototypes (conditions) that required to be compared. Because of variations in the data, the researcher could not directly compare the five prototypes workload means and declared their means are different. Instead, the researcher used a "statistical significance test" to calculate the differences that can be described by independent variables and the variations that cannot be explained by them.

Table 4.2 Multivariate tests result

Multivariate Tests[a]							
Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Sentence length impact	Pillai's Trace	.69	9.63[b]	4.00	17.00	.00	.69
	Wilks' Lambda	.30	9.63[b]	4.00	17.00	.00	.69
	Hotelling's Trace	2.26	9.63[b]	4.00	17.00	.00	.69
	Roy's Largest Root	2.26	9.63[b]	4.00	17.00	.00	.69

The table 4.2 shows the Multivariate tests result, the within subject's-design for sentence length impact portrays that the four tests including "Pillai's Trace, Wilks' Lambda, Hotelling's Trace and Roy's largest Root" had significant results at 5% level of significance as *p* values correspond to 0.00 in each case. The one-way repeated measures MANOVA was statistically significant. There was a difference in the workload score of five prototypes. MANOVA was employed to protects against Type I errors that could occur.

Table 4.3 Mauchly's Test of Sphericity

Mauchly's Test of Sphericity							
					Epsilon		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Sentence length impact	.217	28.17	9	.00	.56	.64	.25

Table 4.3 illustrates the output of Mauchly's Test of Sphericity, and the key columns are the that including the significance value, Greenhouse-Geisser, and Huynh-Feldt. The significance value is 0.00, which is less than 0.05, so it rejects the null hypothesis and it indicates that there was the significant differences between levels. But the assumption of Sphericity was not met. To be assumed, the Greenhouse-Geisser row was employed (if e $<$ 0.75) which corrects the degree of freedom of the repeated measures ANOVA and only the corrected results were reported which shows in “Test of within-subjects Effects” (Greenhouse & Geisser, 1959; Huynh & Feldt, 1976) . The Mauchly's test results signified that the assumption of Sphericity was violated, $\chi^2(9) = 28.17$, $p=0.001 \leq 0.05$.

Table 4.4 Interpret the One-way repeated measured ANOVA results in the "Greenhouse-Geisser" rows of the Within-Subjects Effects tests.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Sentence length impact	Sphericity Assumed	12764.79	4	3191.19	19.77	.00	.49
	Greenhouse-Geisser	12764.79	2.26	5639.31	19.77	.00	.49
	Huynh-Feldt	12764.79	2.56	4972.13	19.77	.00	.49
	Lower-bound	12764.79	1.00	12764.79	19.77	.00	.49
Error (sentence length impact)	Sphericity Assumed	12913.53	80	161.41			
	Greenhouse-Geisser	12913.53	45.27	285.25			
	Huynh-Feldt	12913.53	51.34	251.50			
	Lower-bound	12913.53	20.00	645.67			

The one-way repeated measure ANOVA was used to determine whether there was a significant impact of sentence length between five prototypes (conditions) in terms of participants workload while reading the content on the webpage. The table 4.4 showed the result of within-subjects' effects. It shows that there was a significant impact of sentence length in terms of average workload of NASA-TLX, $F(2.26, 45.27) = 19.77$, $p = 0.00 \leq 0.05$). In this case, the null hypothesis (H_0) is rejected. Since Mauchly's test of sphericity was violated, and the Greenhouse-Geisser correction was employed.

Table 4.5 Pairwise comparisons

(I) Impact of sentence length	(J) Impact of sentence length	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	1.06	2.51	.67	-4.16	6.30
	3	-13.38*	3.68	.00	-21.07	-5.69
	4	-23.54*	4.16	.00	-32.22	-14.86
	5	-24.57*	4.50	.00	-33.96	-15.18
2	1	-1.06	2.51	.67	-6.30	4.16
	3	-14.44*	4.69	.00	-24.24	-4.65
	4	-24.60*	4.49	.00	-33.99	-15.22
	5	-25.63*	5.08	.00	-36.24	-15.02
3	1	13.38*	3.68	.00	5.69	21.07
	2	14.44*	4.69	.00	4.65	24.24
	4	-10.15*	3.27	.00	-17.00	-3.31
	5	-11.19*	2.43	.00	-16.27	-6.11
4	1	23.54*	4.16	.00	14.86	32.22
	2	24.60*	4.49	.00	15.22	33.99
	3	10.15*	3.27	.00	3.31	17.00
	5	-1.03	3.38	.76	-8.08	6.01
5	1	24.57*	4.50	.00	15.18	33.96
	2	25.63*	5.08	.00	15.02	36.24
	3	11.19*	2.43	.00	6.11	16.27
	4	1.03	3.38	.76	-6.01	8.08

Table 4.5 shows mean differences across five prototypes in terms of workload. The pairwise comparison table has been generated as part of the post hoc analysis as repeated one-way measures of ANOVA results indicated a significant difference among the five prototypes. The post hoc test was conducted to figure out which two particular prototypes differed significantly in terms of workload as outlined by NASA-TLX. It is evident from the table 4.5 that there exists an absolute mean difference of 13.38 unit between 1 (prototype A) and 3 (prototype C), 23.54 unit between 1 (Prototype A) and 4 (Prototype D), and 24.57 between 1(Prototype A) and 5 (Prototype E). These differences are statistically significant as $p \leq 0.05$. However, the mean difference in between 1(Prototype A) and 2 (Prototype B) is only 1.06 unit which is not significant at 5 percent level of significance.

Moreover, an absolute mean difference between 2 (prototype B) and 3 (prototype C), between 2 (Prototype B) and 4 (Prototype D), and between 2 (Prototype B) and 5 (Prototype E) are 14.44, 24.60, and 25.63 which are significant at 5 percent level of significance. Also, an absolute mean difference between 3 (prototype C) and 4 (prototype D), and between 3 (Prototype C) and 5 (Prototype E), correspond to 10.15, and 11.19. They are significant at 5 percent level of significance. Finally, an absolute mean difference between 4 (prototype D) and 5 (prototype E) is 1.03 and not statistically significant at $p = 0.76 \leq 0.05$.

Hence, these results represented statistically significant differences of mean workload across the five prototypes except in between prototype A and Prototype B, and Prototype D and Prototype E.

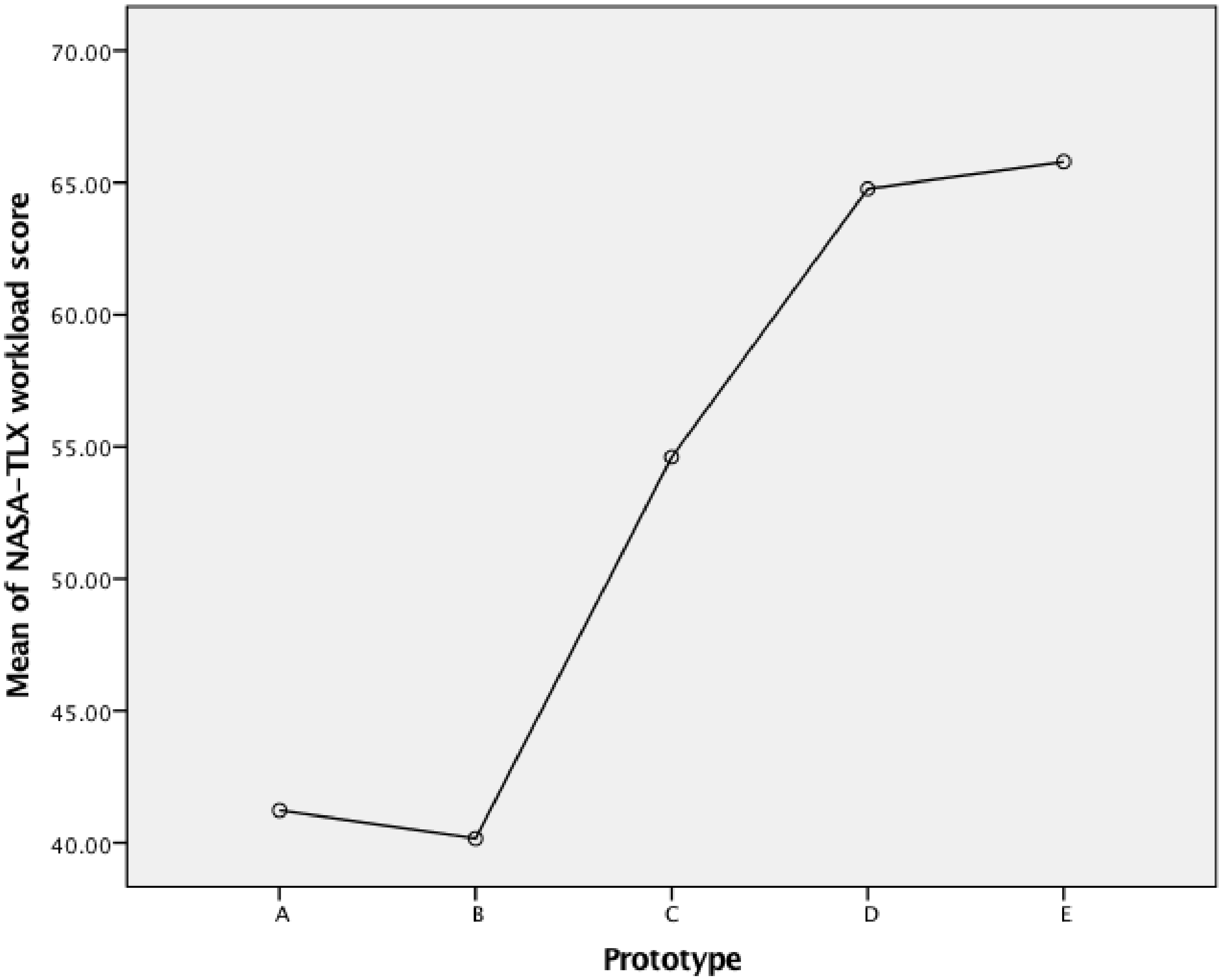

Figure: 4.2 Mean plot of NASA-TLX workload score

Finally, the overall findings illustrate that participants experienced significantly less workload in prototype B while reading sentence on web compared to Prototype A, C, D, and E. This evidence supports the hypothesis that the significant impact on sentence length on readability of the web to the blind users.

4.4 Average NASA-TLX workload ratings in different prototypes

Looking at individual dimensions for the NASA-TLX, there is a significant impact of sentence length on the readability of web among the participants in different prototype (from A to E) as evidenced from a test of hypothesis. Thus, accordingly, this section presents the glimpse of average workload ratings in each prototype. The TLX scores ranged from 0 to 100, in which 0 signifies no workload and 100 signifies maximum workload. The Task Load Index score was computed by adding the scores for the subscales together. The average subjective workload ratings for each subscale for prototype A, B, C, D, and E is shown through the graphical compositions in the figures from 4.3 to 4.7.

Overall prototype B with word length 16-20 recorded the desired preference for reading by the participants in which mental demand (MD) is best recorded (39.16) on an average compared to these in prototype A, B, C, D and E. Prototype E naturally had highest mental demand (MD) rating as it contained the word length from 30 and above. Similarly, in terms of physical demand (PD), prototype B witnessed the least average rating (22.41) compared to the other prototypes. Although, physical demand (PD) ratings shows smallest workload in all prototypes. This might potentially be due to less body movement of this respondent while reading the text. Moreover, on the basis of temporal demand, prototype B has the lowest temporal demand (38.33) among other prototypes. Participants used less time stress while reading and completing the task. The Performance (OP) recorded the highest mean (55.41) workload score in prototype C. This is due to the content is about entertainment and participants were interested while reading and completing task.

Furthermore, participants used less efforts (37.5) in Prototype B in contrast with other prototypes. Whereas, participants used highest efforts in prototype E since it has longest sentence length i.e. above 30 words length. So, they needed more efforts to remember the text while reading through screen reader.

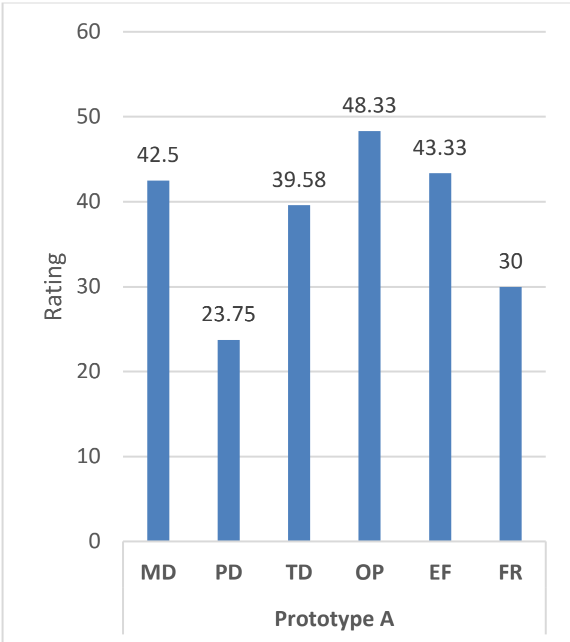

Figure: 4.3 Average rating of Prototype

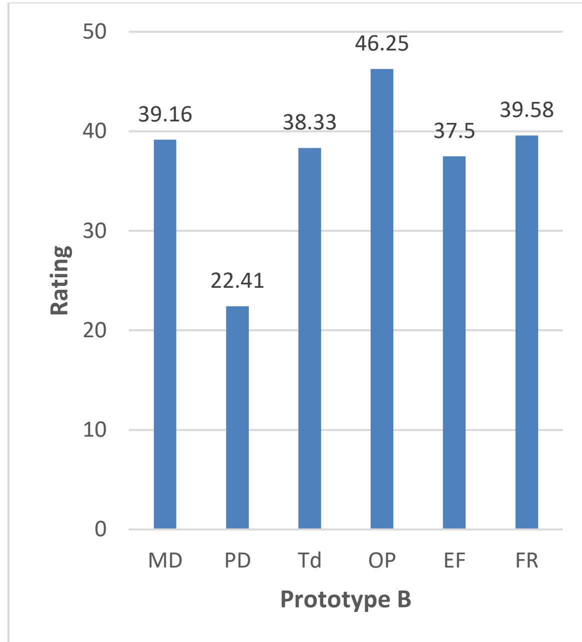

Figure: 4.4 Average rating of prototype B

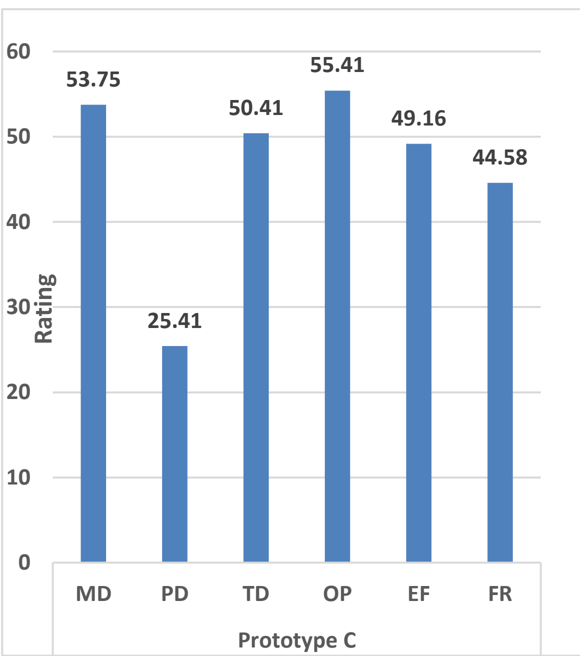

Figure: 4.5 Average rating of prototype C

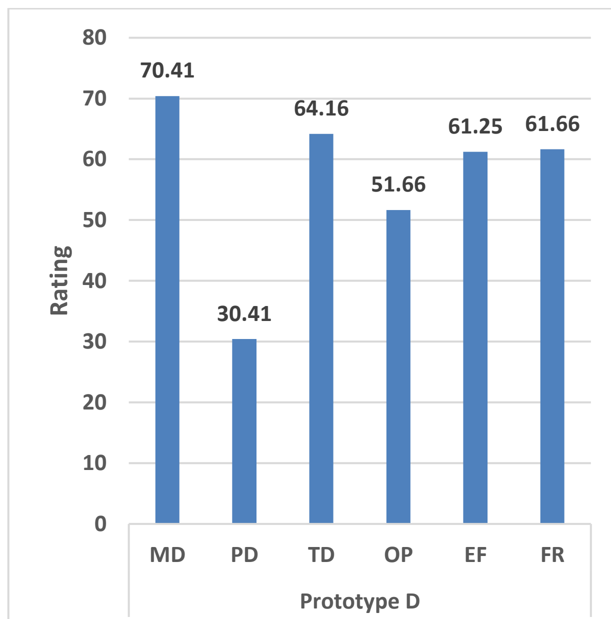

Figure: 4.6 Average rating of prototype D

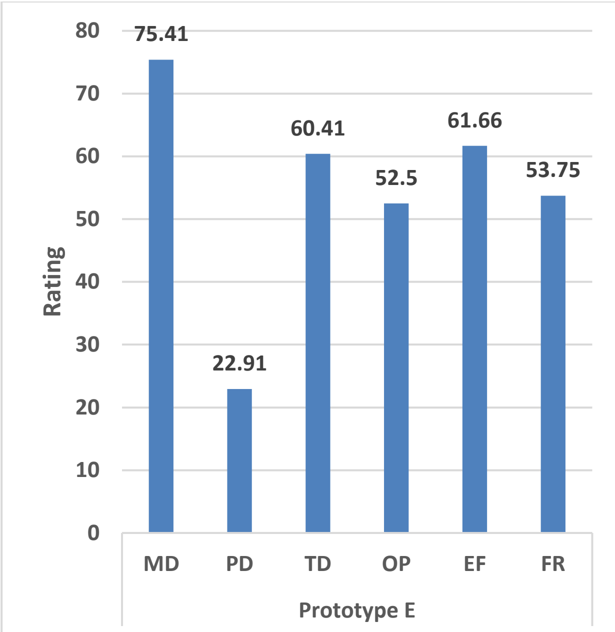

Figure: 4.7 Average rating of prototype E

Participants experienced less frustration (30) while reading content in prototype A. The content was short (10-15 words), so they experienced less frustration. But participants were more frustrated in prototype E (61.66) compared to other prototypes. This was due to the contain was token from Norwegian news portal and participants has to remember the university name as well as person name.

These results indicate overall less workload on prototype B among other prototypes. The average rating on each dimension of prototype B (figure 4.4) had less workload compared to others.

Chapter 5 - Discussion

This chapter elaborates the significant findings of the current study and compares the key differences between the previous researches and current research. It also presents the researcher's opinions on the possibility of proving some theoretical assumption. The chapter begins with a summary of results, and comparative discussion between previous studies and the current research. Further, the limitations of the study are described to show areas of possible improvement for future researches.

5.1 Summary of findings

The main aim of this research was to investigate how the length of the sentence impacts on the readability of web for the blind users and the second goal was to find out the suitable sentence length. The researcher recruited twenty-one blind-folded participants by applying NVDA screen reader, comprehension test, NASA-TLX tool, and five prototypes, i.e. the prototype A, 10-15-words; prototype B, 16-20-words; prototype C, 21-25-words; prototype D, 26-30-words; and prototype E, 30 above word. The within-subject design was used as an experimental design because each participant was measured in each prototype (condition) and their scores were compared, and randomization was used to provide an unbiased evaluation and random assignment of treatments to the participants.

The comprehension test was carried out through a screen reader before measuring the workload of participants. The comprehension test checked whether a participant could read and comprehend the prototype contents through multiple questionnaires.
All the questions were multiple choice, and the contents of the prototype were sourced from different online news portals. The test aimed at finding out the task before taking the workload of the participants. In the beginning of the test, participants failed to concentrate to the screen reader because they only had the knowledge of the screen reader but never had experienced before. As the test was based on the randomization so they did not have option to choose the prototypes. The comprehension test findings showed that majority of the participants understood the prototype contents and their

understandability score was above 60 percent. Most of the participants were from Master of Universal design of ICT background and they had the knowledge of the screen reader. Their higher education level might have affected the result of the comprehension test. However, they had difficulties to memorize all the words of the sentence as a screen reader reads the content once. They experienced even more challenging for the longer sentences to comprehend the content. The highest number of participants understood the prototype B (16-20-word length) content, the content was related to “The official Winter Olympics website.” This could possibly be attributed to greater attention of the participants towards the cyber-attack of Olympics website. It’s because youths are highly influenced to games and sports.

Further, the One-way repeated measures ANOVA was carried out to determine whether there was a significant difference in workload scores under five conditions (prototypes). The ANOVA results showed a significant impact of sentence length over five prototypes. There was a considerable mean difference found across five prototypes in terms of subjective workload. The results based on the subjective workload reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) in first hypothesis. It indicates that blind users are highly influenced by sentence length on the web. So, the researcher suggests that the content writer and the author should write appropriate sentence length for the blind user to understand the content of the web to improve readability. The one-way repeated measures only signified the mean differences of five prototypes workload scores but could not specify which prototype has a significant workload or which is the appropriate sentence length in terms workload of participants. Hence, the researcher further applied the pair-wise comparisons as a part of post hoc test. The post hoc test was conducted to figure out which two particular prototypes differed significantly in terms of workload as outlined by NASA-TLX. The prototype A and B had the minimum mean differences across other prototypes as shown in figure 15. The increasing mean difference across prototype A through prototype B, C, D and E signifies an increasing degree of complexity in readability of the text.

Overall prototype B with word length 16-20 recorded the desired preferences for reading by the participants in which the highest workload came under mental demand. All the participants witnessed the least workload score in physical demand. This might potentially be less body movement of the participants. The temporal demand was also the lowest recorded in prototype B in compared to other prototypes. Also, a scree plot offered an indication that prototype B with word length 16 to 20 witnessed the lowest workload thereby featuring an utmost readability. But, the sentence length of Prototype B is not the minimum length of sentence among all prototypes. Hence, this evidence does not support the second hypothesis that the “minimum sentence length will be appropriate for the blind users on the readability of the web”. The researcher could not support the null hypothesis in this case. Therefore, the researcher would suggest that the length of sentences on the web for the blind users makes big influences on readability and the appropriate length of sentence depends on the readers’ comprehension level and memory workload. This was evidenced by the two-tailed t-scores produced during analysis of mean differences.

5.2 Comparative discussion between previous studies and the current research

The literature review is based on the readability of font types, visual diagram, web documents, line spacing, sentence length, word complexity, and vocabulary. A lot of researchers are focusing on readability to make the web content readable and understandable. Despite a sufficient amount of studies in the readability, the researchers still claim the importance of further investigations in this field for visually impaired readers. The author Mikk (2008) suggested that the 50-130 characters, i.e. 10-25-words in a sentence were the appropriate range of sentence for 17-18-year old students and the sentence length over 25-words overload the working memory. Similarly, the researcher DuBay (2004a) discussed that in our modern time, the sentence length goes down by 20 words per sentence and the shorter sentences increase readability. Also, Ling and Van Schaik (2006) findings further explained 1-70 characters per line was better in terms of users’ performance. Though, these findings (DuBay, 2004a; Ling & Van Schaik, 2006; Mikk, 2008) are not correspond to current research. These studies measured the readability of the text based on the traditional readability formulas. Most of those studies had used textbook content and their range of sentence were different, and participants were not visually impaired. Although the current research employed different methodology and distinctive data collection procedure. Moreover, readers read the text on the web is different compared to the text in the printed book because the content on the web is presented differently and the screen reader read the web content. Also, the readability formula could not measure whether the readers understand the webpage or not (Nielsen, 2011). However, all findings suggested that too long and too short sentences are not suitable for the memory workload of participants to improve readability.

A researcher Cutts (2013), mentioned the limit of sentence length below 40 words and the better goal of an average sentence length of 15-20 words. While the findings in this current research showed that the prototype B had the appropriate sentence length (16-20 words) for blind people to improve readability which comes close to the sentence length recommended by Cutts (2013) to improve readability.

Though, the amount of research available related to the readability are general, with a focus on the impact of a sentence length on the readability of the web for blind readers was rare. Majority of studies focused for sighted readers while this research emphasized for blind readers on the web because physical disability cannot be believed to be considered as a defining parameter to hinder one’s right to access the information on the web. Blind users read through the screen reader on the web, and they face trouble to track long sentences. This research suggests that readable sentence length makes it easier for the blind reader to understand the content and to reduce frustration, memory workload and effort. This research used a different range of sentences that were employed in the prototype and distinctive design, blind-folded participants, NASA-TLX tool, comprehension test and NVDA screen reader were applied. Since, the materials and the data collection procedure were different in comparisons to previous studies. The current research finding determined that there was a significant impact of sentence length and the appropriate sentence length 16-20-word on the web for blind users.

Therefore, this research concluded that the findings compared to previous studies were different in the context of an impact of sentence length and appropriate sentence length on the web for blind users to improve readability.

5.3 Limitations of the study

Despite the findings and results, there are still some limitations to this study. This section addresses the different limitations this study.

5.3.1 Participants

Firstly, recruiting fully blind participants was a challenging task in this study due to various reasons. The researcher, therefore, experimented the study with sighted participants by blindfolding them with the eye-mask. The findings in this study may have resulted differently if actual blind users were experimented as blind people depend mainly on their auditory and touch senses to substitute their lack of visual hints while interacting with the environment (Rony, 2017a).

5.3.2 Complexity of NASA TLX measurement scales

Another limitation of this research was the complexity of measurement scales. NASA-TLX workload measurement scale is considered to be one of the problematic scales to use. As the measurement scales used in this study relied on subjective perceptions of participants, there might be differences in understanding of measurement scales among the participants. In this research, some participants experienced difficulties in NASA TLX bipolar scales (20 steps) ratings from 0 to 100 scores. Immediately, participants needed to circle the factor which had affected more to the workload of the task. In this process, participants found it uneasy and awkward to be blindfolded. For instance: a skilled and experienced participant found it easy to perform task whereas others might have found it challenging to achieve the same task in the same situation. Some participants find physical demand is not necessary for this study because they had not used any physical factors during the experiment. Thus, the researcher assumes that this might influence the outcomes of the research.

5.3.3 Time spent in the experiment process

In this research, each participant had 40 minutes to complete the experiment. But some participants took longer than the estimated time. It would have been more accurate results if the participants had given more time to familiarize them with NASA TLX tools and its procedure. They would then get a better understanding of how to utilize the tools and use them more efficiently.

Chapter 6 - Conclusion and Future work

In this research, we investigated the impact of sentence length on the readability of the web for the blind users. We employed a screen reader to read the content over five prototypes. Each prototype had two different sentences. The comprehension test was used to determine whether the users comprehend the content of the prototype and NASA-TLX was used to measure the participant's mental workload. Answering the first research question “Does the length of sentence affect the readability of the web in terms of workload for blind users?”, the one-way repeated measure ANOVA results indicated that there was a significant difference in the workload of participants over five prototypes (websites). The findings also supported that there was a substantial effect of sentence length on the web for the blind users.

Similarly, answering to the second research question about the appropriate length of sentences for blind users to make the web content readable and understandable, it was not mandatory for sentence length to be minimum to read and comprehend the sentences for the blind users. The scree plot offered an indication that prototype B with word length 16 to 20 witnessed the lowest workload thereby featuring an utmost readability. It is because the majority of participants understood the prototype B content, and they experienced lowest workload. Based on our findings, it is suggested that 16-20-word length in the sentence can be appropriate for the blind users to perform readability on the web without much workload. Hence, the sentence length with 16-20 words can have a better readability to comprehend the content of the web for the blind readers.

There are still some spaces for the future scholars who could possibly address the issues and peripheral contents which this research could not incorporate. Some specific changes could be made in terms of sample size, nature of the sample, inclusion of higher number of prototypes with a varied word-length including others in order for assessing the readability and comprehensibility of the web contents. Further improved studies might be possible with an in-depth analysis of NASA-TLX tools because this study applied 20 steps bipolar rating from 0 to 100 scores and it was not convenient for the participants under consideration. Modified studies can be expected with different instrumentations such as by applying a 5-point or 7-point Likert scale for the NASA-TLX ratings. In addition, one could opt to use open-ended questionnaire technique to collect data for assessing the content readability instead of using NASA-TLX. This study was based on experiment with blindfolded participants, so the future research can be repeated with the actual blind individuals to enhance the content validity, reliability and generalizability of the study.

Chapter 7 -References

Aaron W. Bangor. (1998). Improving access to computer displays: Readability for visually impaired users.

Babu, R., Singh, R., & Ganesh, J. (2010). Understanding blind users' Web accessibility and usability problems. *AIS Transactions on Human-Computer Interaction, 2*(3), 73-94.

Babu Rakesh. (2009). Developing an Understanding of the Nature of Accessibility and Usability Problems Blind Students Face in Web-Enhanced Instruction Environment. 12.

Banerjee, J., Majumdar, D., Pal, M. S., & Majumdar, D. (2011). Readability, subjective preference and mental workload studies on young indian adults for selection of optimum font type and size during onscreen reading. *Al Ameen journal of medical sciences, 4*(2), 131-143.

Bergman, V., & Nygren, R. (2009). Visually impaired and websites: how to improve websites to support the aiding devices of the visually impaired. In.

Björnsson, C.-H. J. R. R. Q. (1983). Readability of newspapers in 11 languages. 480-497.

Caruso, R. D., & Postel, G. C. (2002). Image editing with Adobe Photoshop 6.0. *Radiographics, 22*(4), 993-1002.

Chung, J.-W., Min, H.-J., Kim, J., & Park, J. C. (2013). *Enhancing readability of web documents by text augmentation for deaf people.* Paper presented at the Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics.

Cutts, M. (2013). *Oxford guide to plain English*: OUP Oxford.

De Alwis Edirisinghe, V. (2017). *Estimating Mental Workload of University Students using Eye Parameters.* NTNU,

de Heus, M., & Hiemstra, D. (2013). *Readability of the Web: A study on 1 billion web pages.* Paper presented at the DIR.

DuBay, W. H. (2004a). The Principles of Readability.

DuBay, W. H. (2004b). The Principles of Readability. *Online Submission*.

Eika, E., & Sandnes, F. E. (2016). Assessing the Reading Level of Web Texts for WCAG2. 0 Compliance—Can It Be Done Automatically? In *Advances in Design for Inclusion* (pp. 361-371): Springer.

Eladhari, M. P., & Ollila, E. M. (2012). Design for research results: experimental prototyping and play testing. *Simulation & Gaming, 43*(3), 391-412.

Gergle, D., & Tan, D. S. (2014). Experimental research in HCI. In *Ways of Knowing in HCI* (pp. 191-227): Springer.

González, J., Macías, M., Rodríguez, R., & Sánchez, F. (2003). *Accessibility metrics of web pages for blind end-users.* Paper presented at the International Conference on Web Engineering.

Gottron, T., & Martin, L. (2009). *Estimating web site readability using content extraction.* Paper presented at the Proceedings of the 18th international conference on World wide web.

Gottron, T., & Martin, L. (2012). Readability and the Web. *Future Internet, 4*(1), 238-252.

Green, S. B., & Salkind, N. J. (2016). *Using SPSS for Windows and Macintosh, Books a la Carte*: Pearson.

Greenhouse, S. W., & Geisser. (1959). On methods in the analysis of profile data. *24*(2), 95-112.

Guerreiro, J., & Gonçalves, D. (2015). *Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech.* Paper presented at the Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183): Elsevier.

Hussain, W., Sohaib, O., Ahmed, A., & Qasim Khan. (2011). Web readability factors affecting users of all ages.

Huynh, H., & Feldt. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *1*(1), 69-82.
Janan, D., & Wray, D. (2014). Reassessing the accuracy and use of readability formulae. *Malaysian journal of learning and instruction, 11*, 127-145.
Jarrett, C., Petrie, H., & Summers, K. (2010). *Design to read: designing for people who do not read easily.* Paper presented at the CHI'10 Extended Abstracts on Human Factors in Computing Systems.
Karmakar, S., & Zhu, Y. (2010). *Visualizing text readability.* Paper presented at the Advanced Information Management and Service (IMS), 2010 6th International Conference on.
Kinder Ken. (2013). Sublime text: one editor to rule them all? *, 2013*(232), 2.
Lau, T. P., & King, I. (2006). *Bilingual web page and site readability assessment.* Paper presented at the Proceedings of the 15th international conference on World Wide Web.
Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*: Morgan Kaufmann.
Lela Kodai. (2015). Embrace 7 Principles of Universal Design of better Website Design. Retrieved from https://www.bitovi.com/blog/embrace-7-principles-of-universal-design-for-better-website-design
Ling, J., & Van Schaik, P. (2006). The influence of font type and line length on visual search and information retrieval in web pages. *64*(5), 395-404.
Liu, H., Selker, T., & Lieberman, H. (2003). *Visualizing the affective structure of a text document.* Paper presented at the CHI'03 extended abstracts on Human factors in computing systems.
Longo Luca. (2018). On the Reliability, Validity and Sensitivity of Three Mental Workload Assessment Techniques for the Evaluation of Instructional Designs: A Case Study in a Third-level Course.
MacKenzie, I. S. (2002). Within-subjects vs. Between-subjects Designs: Which to Use? *Human-Computer Interaction: An Empirical Research Perspective, 7*, 2005.
Mikk, J. (2008). Sentence length for revealing the cognitive load reversal effect in text comprehension. *34*(2), 119-127.
Moggridge, B., & Atkinson, B. (2007). *Designing interactions* (Vol. 17): MIT press Cambridge, MA.
National Disability Authority. (2014a). The 7 Principles. Retrieved from http://universaldesign.ie/What-is-Universal-Design/The-7-Principles/
National Disability Authority. (2014b). What is Universal Design. Retrieved from http://universaldesign.ie/What-is-Universal-Design/
Nielsen, J., ,. (2011). Cloze Test For Reading Comprehension. Retrieved from https://www.nngroup.com/articles/cloze-test-reading-comprehension/
NV Access. (2019). NVDA Features. Retrieved from https://www.nvaccess.org/about-nvda/
Oelke, D., Spretke, D., Stoffel, A., & Keim, D. A. (2012). Visual readability analysis: How to make your writings easier to read. *IEEE Transactions on Visualization and Computer Graphics, 18*(5), 662-674.
Owu-Ewie, C. (2014). Readability of comprehension passages in Junior High School (JHS) English textbooks in Ghana. *Ghana Journal of Linguistics, 3*(2), 35-68.
Paciello, M. (2000). *Web accessibility for people with disabilities*: CRC Press.
Petrie, H., & Bevan. (2009). The Evaluation of Accessibility, Usability, and User Experience. *1*, 1-16.
Pikulski. (2002). Readability. *20*, 2014.
Raggett, D., Le Hors, A., & Jacobs, I. HTML 4.01 Specification. W3C Recommendation 24 December 1999, 1999. In.
Raluca Budiu. (2018). Between-Subjects vs. Within-Subjects Study Design. Retrieved from https://www.nngroup.com/articles/between-within-subjects/
Rello, L., & Bigham, J. P. (2017). Good Background Colors for Readers: A Study of People with and without Dyslexia.

Resnik David BJ, N. I. o. E. H. S. (2011). What is ethics in research & why is it important. 1-10.
Rony, M. R. (2017a). *Information Communication Technology to support and include Blind students in a school for all An Interview study of teachers and students' experiences with inclusion and ICT support to blind students.*
Rony, M. R. (2017b). Information Communication Technology to support and include Blind students in a school for all An Interview study of teachers and students' experiences with inclusion and ICT support to blind students.
Rowan, M., Gregor, P., Sloan, D., & Booth, P. (2000). *Evaluating web resources for disability access.* Paper presented at the Proceedings of the fourth international ACM conference on Assistive technologies.
Sandra G. Hart. (2006). *NASA-task load index (NASA-TLX); 20 years later.* Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.
Story, M. F., Mueller, J. L., & Mace, R. L. (1998). The universal design file: Designing for people of all ages and abilities.
Suresh, K. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences, 4*(1), 8.
Temnikova, I., Vieweg, S., & Castillo, C. (2015). *The case for readability of crisis communications in social media.* Paper presented at the Proceedings of the 24th International Conference on World Wide Web.
Thompson, A. E. J. (2015). The Americans With Disabilities Act. *313*(22), 2296-2296.
Van Teijlingen, E., & Hundley, V. (2010). The importance of pilot studies. *35*, 49-59.
Van Teijlingen, E. R., & Hundley, V. (2001). The importance of pilot studies.
W3C Web Accessibility Initiative. (2017). Diverse Abilities and Barriers. Retrieved from https://www.w3.org/WAI/people-use-web/abilities-barriers/
Web Accessibility Initiative. (1999). Web content accessibility guidelines 1.0. *World Wide Web Consortium. Available online at: http://www. w3. org/TR/WCAG10/(accessed 28 September 2005)*.
World Health Organization. (1980). International classification of impairments, disabilities, and handicaps: a manual of classification relating to the consequences of disease, published in accordance with resolution WHA29. 35 of the Twenty-ninth World Health Assembly, May 1976.
World Health Organization. (2017). Vision impairment and blindness. Retrieved from http://www.who.int/mediacentre/factsheets/fs282/en/
World Wide Web Consortium. (2008). Web content accessibility guidelines (WCAG) 2.0.
Zorbaz, K. (2007). An evaluation on the word-sentence lengths and readability levels of tales in Turkish Textbooks. *3*(1), 87-101.

Appendix A

OSLO METROPOLITIAN UNIVERSITY

Department of Computer Science Faculty of Technology, Art and Design

Participants Information Sheet

Title of Study: **Impact of sentence length on the readability of the web to the blind users**

Investigator:

Name: Bam Bahadur Kadayat Department: Computer Science

E-mail: s310223@oslomet.no

Introduction and Purpose

You are invited to participate in an experiment over five prototypes (websites), all websites have the same design and layout but different sentence/content lengths.

Description of the Study Procedures

In this study, we will ask you some multiple-choice question for comprehension test. Afterword's, you will ask to mark your opinion in NASA-Task Load Index, about your mental demand, temporal demand, physical demand, performance, effort and frustration. The whole session will take about 40 minutes. Please indicate with checkbox below that you understand your rights and agree to participate in the experiment.

All information will be kept confidential and your name will not be associated with any research findings.

☐ I have read the process about the research and I confirmed that I agree to participate in the research experiment.

Supervisor:

Name: Evelyn Eika

E-mail: hualiji@oslomet.no

Appendix A1

Participant ID: ___

Questionnaires based on personal Information

Thank you for participating in this research. It should take about 5-10 minutes to complete. In this section, I would like to ask few questions in the context of your personal information. Your privacy is protected, your name will not appear on this questionnaire so, please answer them as truthfully as possible.

Information about You:

1) Your gender: ☐ Male ☐ Female
2) Age: ________
3) Education Qualification:

   ☐ Doctoral ☐ Masters ☐ Bachelor

4) English proficiency:

   ☐ Advanced ☐ Intermediate ☐ Beginning

Appendix B

Participant ID: ___

Comprehension test.

Prototype A

<u>Question 1</u>

Who said that his nephew does not use social media?

a) Windows CEO

b) Samsung CEO

c) Apple CEO

d) Facebook CEO

<u>Question 2</u>

What does the CEO not believe about the overuse of technology?

a) It increases creativity

b) It helps concentrate more

c) It reduces memory power

d) It helps achieve more success

Participant ID: ___

Participant ID: ___

Prototype B

Question 1

Which website was taken offline after being hit by a cyber-attack?

a) Summer Olympic

b) FIFA World Cup

c) Common-Wealth

d) Winter Olympic

Question 2

When was the website of winter Olympic affected in South Korea?

a) Day Before of ceremony

b) Opening of the ceremony

c) Mid time of the ceremony

d) Closing of the ceremony

Participant ID: ___

Prototype C

Question 1

When was the nominations for the ceremony's 90th iteration made public?

a) last month
b) last week
c) last Friday
d) last Wednesday

Question 2

What is the name of the nominated film in the best foreign language category?

a) The Road
b) The Prophet
c) The kite
d) The Insult

Participant ID: ___

Prototype D

Question 1

What status did the Oslo and Akerhus University college achieve?

a) Research expert

b) University

c) Experimental spot

d) Training expert

Question 2

Whose effort is praised by the minister for the success?

a) students and staffs

b) media

c) teachers

d) technicians

Participant ID: ___

Prototype E

Question 1

What does the German government plan to introduce on Wednesday?

a) free education

b) free health services

c) free public transport

d) free housing facilities

Question 2

What has not been decided for the plan of free public transport yet?

a) Budget

b) Sample city

c) Survey

d) Recruitment of manpower

Appendix C

Rating Scale Definitions

Title	**Endpoints**	**Descriptions**
MENTAL DEMAND	Low/High	How much mental and perceptual activity was required for thinking and remembering the task?
PHYSICAL DEMAND	Low/High	How much physical (pushing, pulling, turning, controlling) activity was required for completing the task?
TEMPORAL DEMAND	Low/High	How much time pressure did you feel for completing the task?
PERFORMANCE	Good/poor	How satisfied were you with your performance in completing the task?
EFFORT	Low/High	How difficult was it for you to perform the task?
FRUSTRATION LEVEL	Low/High	How irritated, stressed or annoyed did you feel during the task?

Appendix D

Participant ID: ___ Task ID: _______________

NASA-Task Load Index (TLX)

There are six rating scales which are meant for evaluating your experience during the experiment. Please, evaluate the task by marking “X” on each of the six scales at the point which matches your experience. The scale ranges from “low” on the left to “high” on the right. Please note that the Performance scale goes from "Good" on the left to "Poor" on the right.

Mental Demand

Low High

Physical Demand

Low High

Temporal Demand

Low High

Performance

Good Poor

Effort

Low High

Frustration

Low High

Appendix E

Participant ID: ______   Task ID: _______________

PAIR-WISE COMPARISONS OF FACTORS:

Circle the Scale factor of each pair that represents the more important to your experience of workload in the task(s) that you just performed in experiment. Please consider your choices carefully and make them consistent with how you used the rating scales.

Effort **or** **Performance**	**Temporal Demand** **or** **Frustration**
Temporal Demand **or** **Effort**	**Physical Demand** **or** **Frustration**
Performance **or** **Frustration**	**Physical Demand** **or** **Temporal Demand**
Physical Demand **or** **Performance**	**Temporal Demand** **or** **Mental Demand**

Frustration **or** **Effort**	**Performance** **or** **Mental Demand**
Performance **or** **Temporal Demand**	**Mental Demand** **or** **Effort**
Mental Demand **or** **Physical Demand**	**Effort** **or** **Physical Demand**
Frustration **or** **Mental Demand**	

Appendix F

Participant ID: _______ Task ID: ______________

Sources-of-workload Tally Sheet		
Scale Title	**Tally**	**Weight**
Mental Demand		
Physical Demand		
Temporal Demand		
Performance		
Effort		
Frustration		

Total count = _________________

(NOTE: - The total count is included as a check. If the total count is not equal to 15, then something has been miscounted. Also, no weight can have a value greater than 5)

Appendix G

Participant ID: _______ Task ID: _______________

WEIGHTED RATING WORKSHEET			
Scale Title	**Weight**	**Raw Rating**	**Adjusted Rating** **(Weight X Raw Rating)**
Mental Demand			
Physical Demand			
Temporal Demand			
Performance			
Effort			
Frustration			

Sum of "Adjusted Rating" Column = _________________

WEIGHTED RATING = [i.e. (Sum of adjusted ratings) / 15]

NASA Task Load Scoring Instructions

1. In the tally column, record a mark for each time a participant chose a scale on the evaluation cards (e.g., each time the participant circled "Mental Demand" on a comparison card, the experimenter puts a mark on the "Mental Demand" row of the tally column).
2. Sum the number of tally marks for each scale in the tally column and record the number of marks in the weight column. Weights cannot equal more than 5.
3. Sum all weights and record this number in the "Total Count" box. The total count must equal 15. If it does not equal 15, a miscalculation has occurred.
4. In the Raw Ratings column, record the responses from the Rating Sheet for each scale. The Rating Sheet provides a vertical line anchored at 0 and 100 and divided into intervals of 5 for each scale. To determine the number associated with a response, count the number of intervals from the left if the left most bar is NOT counted, and multiply by 5 (e.g., if the participant marked an "X" on the fourth interval bar from the left, as below, the score would be 4 X 5 = 20).
5. If a participant marks between two interval bars, the value of the right bar is used (i.e., round up). The maximum Raw Rating for any one scale is 100.
6. Multiply the Raw Rating by the Weight for that scale. Record this number in the Adjusted Rating column.
7. Sum the Adjusted Ratings and record the total in the Sum "Adjusted Rating" box.
8. Divide the number in the Sum "Adjusted Rating" box by 15 to obtain the overall weighted workload score. Record the resulting quotient in the WEIGHTED RATING box.