Automatic Readability Evaluation of Dark Mode Websites

Joni Rautiainen
Aalto University
Helsinki, Finland
joni.rautiainen@aalto.fi

ABSTRACT

As web content largely relies on reading, effortless reading plays an important role. Dark mode websites are trending, but good readability remains an issue for many of them despite the large number of guidelines available.

A previous study has provided evidence that guideline-based automatic evaluation can be an effective way to identify some of the readability issues. Algorithms perform particularly well on evaluating many low-level legibility and text-formatting features. However, previous study lacks research on using the automatic readability evaluating on dark mode websites. As dark mode websites are becoming increasingly popular, we need tools that can automatically evaluate the readability of them.

This study proposes modifications needed to extend the system from previous study to allow effective readability evaluation of dark mode websites. Increasing demand for dark mode website designs requires specific knowledge on dark mode readability guidelines from designers. Automatic readability evaluation system could help minimize the human error, but also guide new designers to better address readability issues on trending dark mode designs.

KEYWORDS

readability, dark mode, negative contrast polarity, automatic evaluation

ACM Reference Format:
Joni Rautiainen. 2021. Automatic Readability Evaluation of Dark Mode Websites. In *Proceedings of (HCI research course)*. Helsinki, Finland, 7 pages.

1 INTRODUCTION

Readability is about how readable the text is visually but also about how readable is the writing and language, such as the structure or the complexity of the text. There are a lot of readability guidelines, which are sometimes contradicting and can be hard to apply since there are so many.

In previous research [13], authors proposed an evaluation system which tries to automate the readability guidelines with algorithms and help from design experts. As outcome, the proposed evaluation system performs better than humans when evaluating simple guidelines, such as if the text is left-aligned and if there is enough white space. Also they found out that some guidelines are difficult to automate and they require evaluation from humans, such as evaluating if titles are meaningful and if there are too complex words.

HCI research course, 2021, Aalto University

In this paper we research, how the previously proposed evaluation system needs to be modified and extended for allowing it to be used for evaluating dark mode based websites.

To approach this problem, we first define new dark mode specific guidelines. These guidelines would be used to extend the automatic evaluation system from previous study [13] to also allow the system to be used for evaluating dark mode websites. In this study we generate a model for one of the defined guidelines. We study if the selected guideline affects readability and we measure if the generated model performs as well as design experts. We first run an observation study, where we measure reading times on sample dark mode websites that implement the guideline on different level. This allows us to measure if the selected guideline affects readability. Then we create a survey, where design professionals manually rate the readability of sample dark mode websites based on the selected guideline. Then we compare the survey data to model generated data, which allows us to measure if automatic evaluation could work with equal performance as design experts.

This paper contributes a definition of 7 dark mode specific readability guidelines, and a further validation for one of the guidelines and an exploration for automating the evaluation for this specific guideline.

2 PRE-STUDY

To gain understanding into practices and preferences of potential end-users of our implementation, we conducted a semi-strucutred interview with a junior UI/UX designer who has some experience in dark mode based designs. As the final implementation is especially targeted for any starting designers, this provides us valuable information on the difficulties that are often faced while designing the readability in a dark mode based UIs. This also allows us to gain insights on the use of design tools where our final implementation could be used.

The interview was conducted with an online video call, where the subject was able to share screen and display some created designs and tools (Figma with plugins) that are used in daily work.

We found that our subject creates complete dark mode based UI designs less often, however some UI components are created often as dark mode based. The subject does not use any automatic tools for evaluating the readability, but uses some tools that evaluate some accessibility properties that partly overlap with the readability features. The designs are often validated with some test users and some feedback on readability is gathered that way.

3 RELATED WORK

To further discover the background of this topic, we need to understand the used terminology. Light and dark modes are often referred with term contrast polarity, which describes the contrast between

text and background. Positive contrast polarity refers to light text on dark background while negative contrast polarity refers to dark text on light background. In this study we mainly speak with terms light mode and dark mode.

Different readability guidelines have been researched in the past, however that research does not provide insights specifically for dark mode readability. In recent study [12] researchers propose a comprehensive set of 61 readability guidelines, which were developed with design and dyslexia experts. The set was further reduced down to 12 core guidelines that were considered as effective ones. The research does not provide guidelines specific for dark mode based websites, or distinguish different guidelines between regular light mode and dark mode. Some guidelines are light mode specific, such as "Use an off-white color for your background, like light gray or tan; use dark gray for text instead of pure black.". However, many of the guidelines possibly apply for dark mode websites as well. The set of proposed readability guidelines was later on further validated by another study [13], in where the researchers looked into how these guidelines affect the readability and also investigated their use in expert-based and automatic evaluation. They found that algorithms performed particularly well on evaluating many low-level legibility and text-formatting features. However, also that study did not look how guidelines interact with readability on dark mode websites.

Some recent research has compared the readability between light mode and dark mode. In one research [2] researchers investigated the impact of dark mode and light mode to code readability. The research compared readability on dark text on a light background to light text on a dark background. They studied task solving and eye tracking data from 30 participants. There were no significant differences in time or accuracy between the two modes. Also the eye tracking data showed no significant difference. This study did not take into account any readability guidelines.

On the contrary, there are many recent studies that suggest that light mode is often superior to dark mode when it comes to reading performance. In one study [15] researchers investigated the effect of contrast polarity on visual acuity and proofreading. There were two different tasks for the participants. In the first visual acuity task, participants were shown a symbol resembling letter "C" in different orientations, and they were asked to identify where the gap in the symbol was located. In the second proofreading task, participants were asked to read short texts. Researchers randomly assigned each participant either positive contrast polarity or negative contrast polarity. There were two groups of participants: younger adults (18-33 years) and older adults (60-85 years). The study found that light mode performed better on all situations, regardless of the participant age. However, in visual acuity task older adults did not benefit of light mode as much as the younger adults.

Another research [5] studied how text size affects proofreading performance on positive contrast polarity versus negative contrast polarity. In the proofreading task, 165 participants were randomly assigned either negative or positive polarity and each participant read 40 short texts, after which they answered to a questionnaire. Research found that participants achieved better proofreading performance with positive contrast polarity. The positive polarity advantage increased linearly when the text size decreased. However, participants themselves did not report any difference in readability between the dark mode and light mode.

One research [11] studied how the ambient lighting affects the reading performance on light mode versus dark mode. In this study 34 participants were given string of letters and they had to recognize whether it's a word or not. Text was displayed to participants with varying conditions such as in light mode or dark mode, different text size and simulated daytime or simulated night time lighting. Research found that in simulated daytime tasks were solved quicker than in simulated night time and light mode performed better overall. Also with simulated night time, light mode performed better. They also found that larger text performed better, and especially with simulated night time small text performed much better in light mode.

Overall, the previous research demonstrates that light mode most often leads to better reading performance than dark mode. However, despite the growing popularity of dark mode based websites, there seems to be very little research done on the dark mode specific readability guidelines and how each of them affect the readability.

4 APPROACH

4.1 Dark mode specific readability guidelines

To approach the research problem, we first defined the new readability guidelines that are dark mode specific and are not covered by the evaluation system proposed in previous study [13].

We researched for commonly recommended readability guidelines from research articles, literature and guides made by design experts to collect different dark mode specific readability guidelines. All collected guidelines are visible in table 1. Guidelines were mainly found from online resources that have been written by design experts and organizations. There is little previous scientific research on the collected guidelines, meaning they can be opinionated and following them in design may not actually indicate better readability for the end product.

4.1.1 Collected guidelines. Some guidelines guide the use of dark mode in general level. Some expert guides recommend that dark mode should not be used at all on text heavy websites, since dark mode is argued to decrease reading performance in many cases [14] [6]. Also previous research has showed that light mode has an advantage over dark mode when it comes to reading performance, as discussed earlier in the related work section. For this recommendation we defined a guideline (G1): Avoid dark mode on text heavy websites. Many guides recommend, that website should offer user a way switch between dark mode and light mode, to comply with user preferences and to make the website accessible to large group of users [18] [16]. We defined this as a guideline (G2): Provide user a way to switch between light mode and dark mode.

Some more dark mode specific guidelines are related to color and contrast of the text and background. Website text content is recommended to meet WCAG (Web Content Accessibility Guidelines) contrast requirements [19], which also concerns dark mode websites and can be considered as a base for all the other color and contrast related guidelines defined in this study. For this recommendation we defined a guideline (G3): Text and background contrast ratio must meet the WCAG standards. Furthermore, many guides

by design experts recommend avoiding pure black and white, but instead recommend using an off-white color, such as light grey for text and dark grey for the background [3] [1] [9]. A similar guideline but reversed was defined for light mode readability in previous research, and it was found significantly effective for typical websites [12]. We defined a guideline (G4) for this: Avoid using pure black as background color and pure white as text color. Google's Material Design guidelines recommend no to use saturated colors in dark mode designs, since they do not meet WCAG's accessibility standard and can produce optical vibrations against a dark background [9]. Similar guiding can also be found from other guides by design experts [7] [1]. We defined this as guideline (G5): Don't use saturated colors.

Some guidelines concern the font weight and sizing. Many design experts recommend using a lighter font weight for text in dark mode websites. This is because light text on dark background tend to appear more bold than dark text on light background, even though the font weight is the same. [8] [1] [17] [3] This phenomenon is caused by a well-known optical illusion, known as irradiation illusion [20]. Some experts also say that too heavy text can appear too bright for the reader [17] [1]. Some fonts are also designed to address this dark mode specific issue [4]. We defined this as a guideline (G6): Use lighter font weight to avoid optical illusion where font looks too bold. Furthermore, some experts suggest to avoid using small font sizes especially on dark mode websites, since it may lead to readability issues or decreased reading performance [18] [16]. It has also been demonstrated by previous research that small text size performs worse in dark mode than in light mode [11] [5]. We defined this as a guideline (G7): Avoid using small font sizes.

4.1.2 Study focus. For this study, we selected to focus our research only on one guideline, since this study is implemented as a part of a school course and we are limited in resources and time. This limits the conclusions that can be drawn from this study but can contribute as a base for further research. We selected to focus our research on guideline G6: Use lighter font weight to avoid optical illusion where font looks too bold. The selection was done based on the following criteria: First, guideline does not overlap with the guidelines used for previous system evaluating light modes [13]. Second, there is some scientific background indicating a possible connection to perception and readability. Third, guideline can be feasibly implemented into a model in a scope of this research and course.

4.2 Ground truth

To evaluate the affect on readability for the selected guideline (G6), we gathered ground truth data by running an observation study with text reading task and simple questionnaire, to measure the reading time on different dark mode based sample websites. In this approach non-professional participants read the content on our sample websites and evaluated the readability by answering a simple end question. This approach allows us to measure if the selected guideline has an affect on the readability of a dark mode website.

Table 1: Collected dark mode specific readability guidelines

ID	Guideline text
G1	Avoid dark mode on text heavy websites
G2	Provide user a way to switch between light mode and dark mode
G3	Text and background contrast ratio must meet the WCAG standards
G4	Avoid using pure black as background color and pure white as text color
G5	Don't use saturated colors
G6	Use lighter font weight to avoid optical illusion where font looks too bold
G7	Avoid using small font sizes

Table 2: Sample dark mode website specifications

Sample ID	Text word count	Font
S1	411	Roboto
S2	342	Open Sans
S3	319	Noto Sans JP
S4	435	Lato
S5	363	Montserrat
S6	443	Source Sans Pro
S7	307	Poppins
S8	258	Ubuntu
S9	243	Rubik
S10	318	Work Sans

4.2.1 Participants. We recruited 10 non-professional participants (6 male, 4 female) to take part in this observation study. All participants had Finnish as their native language and they had good level of technical knowledge, and they had normal vision. Participant's age varied between 23-55 years old, while the mean age was 33,8.

4.2.2 Materials. We created 10 different sample dark mode websites to be used in this study. In addition, each sample site had two variations, where one variation applied the guideline G6 and other variation did not. With variations counted in, we used a total of 20 websites in this study.

Sample sites varied only by the text content and used font. Each of the 10 sample sites contained a different text describing various dog species, in Finnish. Text length varied between 243-443 words. Each sample site used a different font, which allows us to research if the affect of guideline G6 in readability heavily depends on the font. First variation of each sample site did not apply the guideline G6 and used regular weight (400) of the font. The second variation did apply the guideline G6 and used thin weight (300) of the font. Based on the guidelines discovered in previous research, we used plain and evenly spaced sans serif fonts to achieve good readability [13]. We selected 10 popularly used sans serif fonts from Google Fonts [10], which provided the font weights required for this study. All used fonts and text lengths are visible in table 2.

On all other aspects, sample sites were similar to each other and were made as simple as possible. Purpose of this was to minimize

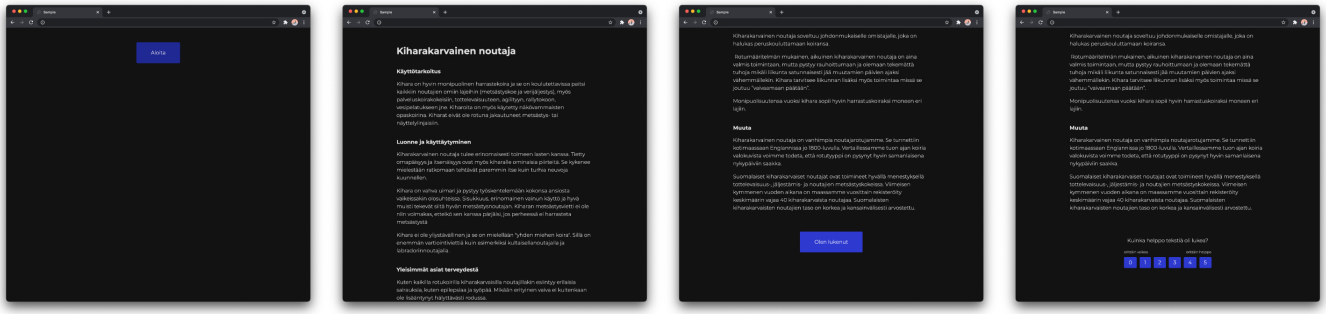

Figure 1: Images displaying a created sample website. In observation study, non-professional participants were asked to read through the website content. Users were asked to press a button to start reading and to indicate the reading was finished. After reading, participant was provided a simple questionnaire asking "How easy was this website to read?". Images here from left to right display the flow.

the affect of other readability factors interacting with the measurements. Sample sites did not include any additional elements such as images in addition to text content and questionnaire content, to allow focusing purely on the readability aspect. Sample sites followed different guidelines to make them as readable as possible. For website background color we used dark grey and for font color we used light grey, exact color values were applied by following Google's Material Design guidelines [9] and were also carefully measured to meet WCAG's standards. In addition we applied guidelines based on the previous research [13], for example we used left aligned non-justified text alignment, font size of 18pt and between-line spacing of 1.5 and the text content used clear language and headings.

For non-professional participants, every sample website was programmed to have an initial state, in which there was no text visible. Initial state showed a button labeled "Start". When participant presses the "Start" button, website content was immediately displayed and website started measuring time. At the end of the website content was another button labeled "I have finished reading". When that button was pressed, website measured the reading time and displayed a simple questionnaire which asked "How easy was this website to read?". Participant had to answer on scale 0-5, where 0 was labeled "Very difficult to read" and 5 was labelled "Very easy to read". After each questionnaire, the data (reading time and readability rate) was collected and saved. A preview of a sample website is showed in figure 1.

4.2.3 Procedure. Participants were divided to two groups and each group had 3 male and 2 female participants. First group was assigned to read samples 1-5 (S1-S5) with thin font weight (guideline G6 applied) and samples 6-10 with regular font weight (guideline G6 not applied). Second group read samples 1-5 (S1-S5) with regular font weight (guideline G6 not applied) and samples 6-10 with thin font weight (guideline G6 applied). All samples were then displayed in random order for the participant. In total each participant were asked to read all 10 sample websites. Each sample site variation were read for equal amount of times.

Each participant was guided personally so that they felt familiar with testing environment and knew how the testing process works. However, participants were not introduced with the topic of the study. They were asked to read all text content and once read to press the "I have finished reading" button, and to perform the same for each given sample site.

All participants used the same computer monitor for the test. Environment was not strictly controlled, however all tests were done during the natural day time, in a room with good lighting.

4.3 Manual evaluation

To compare the performance of the automatic evaluation done by the model with manual evaluation done by design experts, we run a simple survey with design experts. In this survey selected design professionals are asked to evaluate the set of sample dark mode websites based on the selected guideline. In survey, participants rate the readability of the sample websites, by evaluating how well the website implements the selected guideline and also by giving an overall rate for the readability. This approach allows us to compare the models performance on individual guidelines and allows us to validate if the model is in line with the professionals' view.

4.3.1 Participants. We recruited 4 design experts to take part in this survey. All participants were male, and their age varied between 23-28 years old while the mean age was 26. All participants had a varying level of expertise, two of the them were practitioners and two were students.

4.3.2 Materials. For this design expert survey, we used the same set of sample dark mode websites that was used for ground truth observation study. Sample website specifications are displayed in table 2. To meet the requirements of this survey, the time measuring system was removed from the sample websites and the questionnaire at the end of the website was modified. The questionnaire included two questions. The first question was "Does this website apply the following guideline: Use lighter font weight to avoid optical illusion where font looks too bold?" and answer options were "Yes" or "No". The second question was same as in ground truth observation study: "How easy was this website to read?", and participant had to answer on scale 0-5, where 0 was labeled "Very difficult to read" and 5 was labelled "Very easy to read".

4.3.3 Procedure. Each participant was displayed both variations of all 10 sample websites, being a total of 20 different websites. Websites were displayed on random order and participants were not displayed any information regarding the used font or it's weight. Participants were asked to examine each of the websites and they were briefly explained, that for each website they will need to evaluate, whether the guideline G6 was applied or not. We did not limit the time, and participants were able to examine the websites at their own pace. Then for each website, they were asked to mark true or false depending on whether they believed the guideline G6 was applied on the current website or not. Also, similar to non-professional participants in ground truth observation study, experts were requested to rate readability of each site on scale 0-5 (0 = Very difficult to read, 5 = Very easy to read).

All experts used the same computer monitor for the test. As in the ground truth observation study, the environment was not strictly controlled, however all tests were done during the natural day time, in a room with good lighting.

4.4 Automatic evaluation

After defining the new readability guidelines and selecting the one guideline (G6) to focus on in this study, we extend the evaluation system from previous research [13] to support the selected guideline (G6). Since we were limited in resources and time, we focused only on extending the system with one guideline. And also, since the evaluation system proposed in previous study is not provided or described in detail, we focus only in extending the system and not validating the previous work.

We created a basic deep learning model which can evaluate whether the guideline G6 is applied in the dark mode website or not. The model works with image input, which allows inputting a screen shot image of the desired dark mode website and allows the model to be integrated to other existing evaluation systems later on. The model was implemented with Python, Tensorflow 2 and Keras. The model created is a sequential model consisting of three convolution blocks with a max pool layer in each of them. On top of it there is a fully connected layer with 128 units, which is activated by a relu activation function. To improve the model further, we also added data augmentation which applied random zoom and crop for the data set images, and a 20 percent dropout.

As a data set, we used 1200 screen shot images collected from the sample dark mode websites that were created for this study. Images were divided in two classes. First class contained images from the sample websites that applied correctly the guideline G6 and used thin font weight (300). Second class contained images from the sample websites that did not apply the guideline G6 and used regular font weight (400). For each class the data set then contained 600 sample images. The model was trained for 50 epochs, with 80 percent of the data set used for training and 20 percent used for validation. Figure 2 shows the achieved training and validation accuracy and loss.

To validate the performance of our model, we measured the accuracy against two validation data sets. First data set contained 200 new screen shot images from the sample websites that were created and used for this study. The second validation set contained 210 images from completely new sample dark mode websites, that used

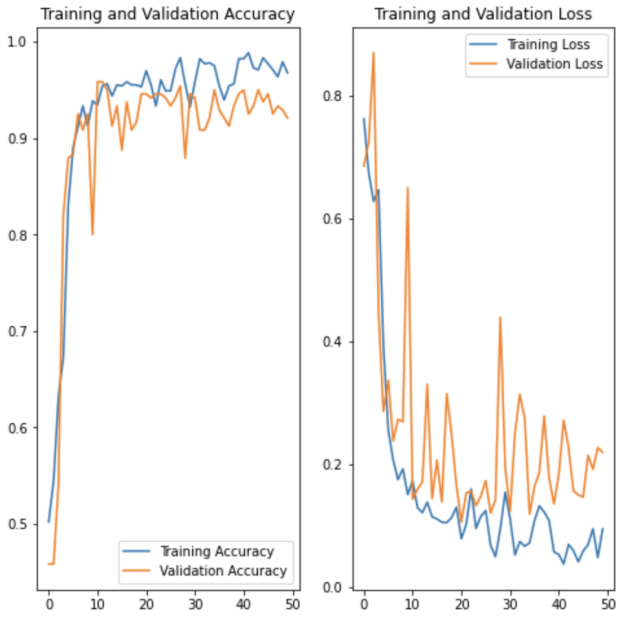

Figure 2: Model training metrics, displaying training and validation accuracy and loss.

different font and text content. Finally we compare the results from expert evaluation to the accuracy of the automatic evaluation done by the model. This allows us to measure if the automatic evaluation is as effective as the manual evaluation done by design experts, which gives us an indication whether the automatic evaluation could offer some advantage for designers.

5 RESULTS

In here we show three different pieces of data collected with an observation study with non-professional readers, a survey done with professional designers and model based evaluation data.

5.1 Ground truth

Table 3 shows the results from the observation study that contained non-professional participants reading sample websites while we observed the reading time. We also asked participants to rate the readability on scale 0-5, where 5 corresponds to very easy to read and 0 corresponds to very difficult to read.

The results show that guideline G6 did not significantly affect the readability of the dark mode website. There was no significant difference in reading time or in readability rate reported by participants.

To further analyse the data from the observation study, we can compare the reading performance and readability rate of each sample separately, as displayed in table 4. We can see that especially the mean difference in reading time varied between different samples. This could be due to different font or different text lengths. However we can not find clear correlation between the font or the text length. It also might indicate that our participant count was

Table 3: Ground truth study results. Reading time was measured with non-professional participants, who also rated on scale 0-5 how easy the website was to read (5 = very easy, 0 = very difficult).

Guideline ID	Guideline implemented	Mean reading time (s)	Mean readability rate (1-5)
G6	false	104,3	3,9
G6	true	103,2	4,1

Table 4: Sample specific data from observation study. Positive difference means an advantage for the thin font weight (= guideline G6 applied).

Sample ID	Font	Difference in time (s)	Difference in rate (0-5)
S1	Roboto	15,4	-0,8
S2	Open Sans	3,0	-0,6
S3	Noto Sans JP	20,0	0,2
S4	Lato	28,6	0,2
S5	Montserrat	13,8	-0,6
S6	Source Sans Pro	-29,0	0,8
S7	Poppins	-9,0	0,2
S8	Ubuntu	-8,2	0,6
S9	Rubik	-14,8	0,8
S10	Work Sans	-9,4	1,2

too low and thus lead to large variance in reading time with some sample sites.

5.2 Manual evaluation

Table 5 shows the results from the survey that contained expert designers evaluating whether the guideline G6 was applied in each sample site or not. We also asked participants to rate the readability on scale 0-5, where 5 corresponds to very easy to read and 0 corresponds to very difficult to read.

Results show that expert designers were able to recognise whether the guideline was applied or not with a mean accuracy of 96%. With most sample sites experts achieved 100% accuracy. However some sample sites seemed to cause more difficulty on evaluation than others, which might be related to font used. Experts also seemed to rate sites with thin font weight as slightly easier to read. This could indicate that participants were familiar with the guideline G6, or also this can be due to personal preferences since the participant group was small.

5.3 Automatic evaluation

The created model achieved a moderate accuracy even despite the small data set size. The small data set lead to some amount of variance especially in validation accuracy and loss during the training, which is visible in figure 2. However, the model performed the evaluation with over 90 percent accuracy on our both validation data sets. The first validation data set contained 200 new screen shot images captured from the sample websites that were created

Table 5: Manual evaluation data from survey, where design experts evaluated whether the guideline G6 was applied or not on the sample dark mode websites. Design experts were also asked to rate the overall readability of each site on scale 0-5 (5 = very easy, 0 = very difficult).

Sample ID	Font	Correct evaluations	Mean readability rate for thin font (0-5)	Mean readability rate for regular font (0-5)
S1	Roboto	100%	4,25	3,75
S2	Open Sans	100%	4,25	4,25
S3	Noto Sans JP	100%	4	4,25
S4	Lato	100%	4	4
S5	Montserrat	88%	4,75	4,5
S6	Source Sans Pro	100%	4	4
S7	Poppins	75%	4	4
S8	Ubuntu	100%	3,5	3,25
S9	Rubik	100%	3,75	3,25
S10	Work Sans	100%	4	4
All		96%	4,1	3,6

for this study. These images were not part of the data set used for training and validation, but used partly same text content and fonts. When validating the model on this set, it achieved a total accuracy of 93,0 percent. The second validation set contained 220 screen shot images from completely new sample dark mode websites. Those sample websites used different fonts and text content than were used for training and validation. When the model was tested against this validation data set, it achieved a total of 93,6 percent accuracy.

6 DISCUSSION

Dark mode is trending and it is becoming more and more popular while many website and app providers decide to use it or offer it as an option alongside the light mode. In this study we reviewed the existing research on dark mode, which well demonstrated that light mode is often superior to the dark mode when it comes to reading performance. However, as dark mode is not disappearing and some users still prefer it, we should seek for ways to make it as readable as possible.

In this study, we defined 7 new readability guidelines that are specific for dark mode. Guidelines were defined based on design experts' guides and previous research. There is very little previous scientific research done on how the defined guidelines affect the readability, which indicates that many of them might be opinionated recommendations from design experts. However, this study demonstrates that there is possibly a need for dark mode specific processing when building an automatic readability evaluation system for websites or for other user interfaces. Further research is needed to measure how the defined guidelines affect the readability

in dark mode and also to explore if there are more important dark mode specific guidelines.

This study focused on researching and validating one dark mode specific readability guideline (G6): Use lighter font weight to avoid optical illusion where font looks too bold. Ground truth study results indicate that the font weight does not heavily affect the readability of a dark mode website, in cases where the readability is good overall. However, there are reasons why further research is needed. Firstly, the number of participants in our study was fairly low, which can lead to some error in the collected data. Secondly, our reading time measurement system could be improved in a way that it would not require action from the participant, for example by using an eye-tracking system to get more detailed data. Thirdly, to also measure if the guideline G6 has possible connections to other elements on websites, the set of sample websites could be larger, with more variance in the content.

Manual evaluation results show that design experts were able to evaluate whether the guideline G6 was applied or not with a great accuracy. The automatic evaluation performed slightly less accurately than design experts did, however, it still achieved a moderate accuracy. Manual and automatic evaluation accuracy measures did not differ with a large margin, which gives an indication that the evaluation of guideline G6 could be automated with a deep learning model. With further fine-tuning and a larger data set, the accuracy of the automatic evaluation could possibly improved, and the model could possibly be used as a part of a readability evaluation system. This could provide help especially for starting designers, but could also serve as check-tool for more experienced designers. However, the accuracy measures discussed here can have some error due to low participant count in the design expert survey and small data set used for model training, and they should be researched further to allow more accurate comparison. Also, since the font weight did not significantly seem to affect the readability, should be further researched if this specific guideline is needed for the automatic evaluation systems. Also, to further explore the idea of automating the guideline G6, the deep learning model is not the only option. If the evaluation system would be integrated into an existing design software or a browser, the font weight values could be received straight from the software, this way allowing for more linear approach. However, the guideline G6 can be more complex than discovered in this study. The readability might not purely depend on the font weight but might also vary by font, and this is also a place for further research. This study did not research how user groups that struggle with reading, such as dyslexics or children are affected by dark mode and how the dark mode specific guidelines defined in this study concern those user groups. Future research should address these limitations of this study.

7 CONCLUSION

This study demonstrates that there are readability guidelines that are dark mode specific, and there is possibly a need for dark mode specific processing when automating readability evaluation. This study also provides us some indication, that the font weight does not heavily affect the readability of a dark mode website, in cases where the readability is good overall. However the result should not be blindly relied upon and should be researched further with more participants.

REFERENCES

[1] Adhuham. 2020. *A Complete Guide to Dark Mode on the Web.* Retrieved May 8, 2021 from https://css-tricks.com/a-complete-guide-to-dark-mode-on-the-web/

[2] Joar Rutqvist Anna Nyqvist. 2019. The Impact of Colour Themes on Code Readability. (2019).

[3] Nick Babich. 2019. *8 Tips for Dark Theme Design.* Retrieved May 8, 2021 from https://uxplanet.org/8-tips-for-dark-theme-design-8dfc2f8f7ab6

[4] Barry Collins. 2020. *The Font That Solves The Big Problem With Dark Mode.* Retrieved May 8, 2021 from https://www.forbes.com/sites/barrycollins/2020/05/27/dark-mode-font/?sh=31c924922cd2

[5] Susanne Mayr Cosima Piepenbrock and Axel Buchner. 2014. Positive Display Polarity Is Particularly Advantageous for Small Character Sizes: Implications for Display Design. *HUMAN FACTORS* 56, 5 (2014), 942–951. https://doi.org/10.1177/0018720813515509

[6] Eric Daily. 2019. *The Dos And Don'ts Of Dark Theme In UI And Web Design.* Retrieved May 8, 2021 from https://www.fyresite.com/the-dos-and-donts-of-dark-theme-in-ui-and-web-design/

[7] Microsoft Design. 2019. *Designing for Dark Mode: More Than Flipping a Switch.* Retrieved May 8, 2021 from https://medium.com/microsoft-design/designing-for-dark-mode-more-than-flipping-a-switch-c21fb38722d2

[8] Irene Fatyanova. 2020. *Dark Mode Readability Tips: How to Make Dark Mode UI Trend More Accessible?* Retrieved May 8, 2021 from https://www.templatemonster.com/blog/dark-mode-ui-trend-readability/

[9] Google. 2021. *Dark theme.* Retrieved May 8, 2021 from https://material.io/design/color/dark-theme.html#ui-application

[10] Google. 2021. *Google Fonts – Web resource.* Retrieved May 8, 2021 from https://fonts.google.com/

[11] Bryan Reimer Jonathan Dobres, Nadine Chahine. 2016. Effects of ambient illumination, contrast polarity, and letter size on text legibility under glance-like reading. *Applied Ergonomics* 60 (2016), 68–73. http://dx.doi.org/10.1016/j.apergo.2016.11.001

[12] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design Guidelines for Web Readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 285–296. https://doi.org/10.1145/3064663.3064711

[13] Aliaksei Miniukovich, Michele Scaltritti, Simone Sulpizio, and Antonella De Angeli. 2019. Guideline-Based Evaluation of Web Readability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300738

[14] Miklos Philips. 2017. *Dark UIs. The Good and the Bad. Dos and Don'ts.* Retrieved May 8, 2021 from https://www.toptal.com/designers/ui/dark-ui

[15] Cosima Piepenbrock, Susanne Mayr, Iris Mund, and Axel Buchner. 2013. Positive display polarity is advantageous for both younger and older adults. *Ergonomics* 56, 7 (2013), 1116–1124. https://doi.org/10.1080/00140139.2013.790485 PMID: 23654206.

[16] Nielsen Norman Group Raluca Budiu. 2020. *Dark Mode vs. Light Mode: Which Is Better?* Retrieved May 8, 2021 from https://www.nngroup.com/articles/dark-mode/

[17] Robin Rendle. 2020. *Dark mode and variable fonts.* Retrieved May 8, 2021 from https://css-tricks.com/dark-mode-and-variable-fonts/

[18] CPACC Sheri Byrne-Haber. 2018. *Dark UI themes are new and cool — but are they accessible?* Retrieved May 8, 2021 from https://uxdesign.cc/accessibility-and-dark-ui-themes-f01001339b65

[19] WCAG. 2018. *Web Content Accessibility Guidelines (WCAG) 2.1.* Retrieved May 8, 2021 from https://www.w3.org/TR/WCAG21/#contrast-minimum

[20] Eric W Weisstein. 2021. *Irradiation Illusion, From MathWorld–A Wolfram Web Resource.* Retrieved May 8, 2021 from https://mathworld.wolfram.com/IrradiationIllusion.html