DOI: 10.35011/icchp-aaate22-p1

ICCHP-AAATE 2022 Open Access Compendium "Assistive Technology, Accessibility and (e)Inclusion" Part I

Impact of the Layout on Web Comprehension

Martín-Calpena, Beatriz Yoldi

Linz, 2022

Persistent Link: https://doi.org/10.35011/icchp-aaate22-p1-25

urn:nbn:at:at-ubl:3-12984

Impact of the Layout on Web Comprehension

Beatriz Yoldi Martín-Calpena, Ana María Iglesias Maqueda, and Jorge Luis Morato Lara

Computer Science and Engineering Department, Universidad Carlos III Madrid, Madrid, Spain

Abstract. In this day and age everybody makes use of the Internet and all the content it provides, but there are groups of people that may face some problems to consume this information because of their disabilities, age or even education. Some entities have tried to help solve this problem by providing guidelines and laws to improve the readability and accessibility of web pages and its content. In this paper we have studied these recommendations in search of some web elements that may help improve content comprehensibility. Moreover, we have collected a corpus made up of a great variety of different web pages and extracted those web elements. Thanks to machine learning algorithms we have obtained a classification model, using all the elements extracted from the corpus, that helps classify a web page according to its comprehension difficulty.

Keywords: Web Comprehension, Readability, Accessibility, Machine Learning.

1 Introduction

Nowadays, many daily activities require the Internet, from simple administrative procedures to the need to satisfy information queries. However, unfortunately, today's web sites are not as readable as they should be, and many people have problems understanding their content. Among others, people with cognitive disabilities, the elderly, poorly educated people, etc. It is crucial to make the web accessible to everyone, so that the content of the web can be understood by anyone. The difficulty in understanding does not only depend on the language used, but also on other elements such as the layout or that the content is properly structured within the page.

The objective of this paper is to examine the impact of web page design and structure, and their relationship with the webpage content comprehensibility (readability). To this end, the presence of elements related to the webpage structure (such as links, bold or italic letters, number of words per paragraph, etc.) will be analyzed and how they influence the web readability

This paper is organized as follows. In the introduction the motivation for this work is presented, as well as a brief description of the objectives of this study. In the next section, we discuss the state of the art, where we will describe existing readability guidelines, as well as other analogous research that analyzes the impact of design on text comprehension. In the methodology we will explain the approach followed in the

work to analyze the impact of layout on web comprehension. Finally, we will find the conclusions obtained and comment on possible future work.

2 Literature Review

In this section, the main laws and guidelines dealing with text comprehension and research work on readability are going to be described. Next, some relevant works on the topic are discussed.

2.1 Laws and Guidelines

Currently there are different European and international entities that have proposed guidelines and recommendations to facilitate the development of easy-to-read documents. These guides are aimed at different population groups, from people with cognitive or physical disabilities to general recommendations addressed to any user, regardless of their physical, cognitive, or cultural situation.

Examples of these guidelines are:

1. Recommendations intended for the public in general, regardless of their personal circumstances: Public Law 111 of the United States Government [2] proposes a series of "Plain Language Guidelines" [3]. The law aims to improve access to government and administrative information for all types of documents, including digital publications and web pages. The standard insists on how to make texts more accessible, clear and concise, but does not mention how the layout should be for the web.
2. Aimed at people with physical disabilities: There are other guidelines that establish a series of recommendations aimed at facilitating accessibility, such as those proposed by the W3C [4] which includes the WCAG guidelines [5]. WCAG (Web Content Accessibility Guidelines) aim to promote a common standard for the accessibility of information on web pages and web applications, especially for the visually impaired population. In other words, they emphasize aspects more associated with legibility (e.g. text size or contrast) than readability. After reviewing the WCAG guidelines we can confirm that, although there is a wide variety of guidelines defined to help design more accessible pages for people with any disability, no rules are defined that deal with the structuring of a page at the HTML level.
3. Regarding population with cognitive disabilities: the Easy-to-read guideline of Inclusion Europe [1]. It should be noted that although the focus is on cognitive disability, the recommendations are applicable to other population groups. For example, to segments of the population that are not native speakers of the language or have a low cultural level. Similar to the W3C WCAG standards, work is being done on a compilation aimed at the cognitively impaired population, COGA [6]. In this case, although there are some standards related to HTML structuring, this is not its focus.

2.2 Research Works on Readability

In addition to the guides on accessibility recommendations and easy reading, there are also other works that have studied the comprehension of web pages. One of the works carried out is a tool for analyzing web pages called Comp4Text Checker [7]. The main function of this tool is to calculate the readability level and to show the problems of readability and comprehension of the information on a web page, but it is only focused on the Spanish language. The tool does not analyze the complete design of the web page, for example it does not take into account the number of links. Although there are already applications that perform the same function, none of them is specific to Spanish and this is something to consider since readability metrics are very language dependent.

Another work, carried out by Peter Williams [8], studies the preferences that people with disabilities have in web pages. From a study involving a group of 25 people, they concluded that some aspects such as large font size, horizontal structuring of the home menu and the use of images are favorable elements for pages. Although this work has taken into account the opinions of a group of people with cognitive disabilities, it is true that the study remains a superficial and visual analysis of the design of the pages, and many of the problems encountered by the participants were related to navigation and the large amount of information presented on the pages.

Unlike previous work on readability and comprehension of Web sites, in our work we focused on the structuring and layout of the information rather than on the information itself. Furthermore, instead of focusing on a superficial view of the web, as in the work of Williams [8], we have gone into the body of the pages, i.e., their HTML and CSS files. With this we want to find elements that facilitate the design of pages to make them more understandable for people with cognitive disabilities.

3 Research Methodology

In order to carry out this research, the following steps were followed:

1. First, characteristics related to the structure of the web document that could have an impact on the comprehension of the document were identified. These features also incorporate characteristics of the style files. Special care has been taken to analyze the aspects included in the accessibility and readability guides, although additional elements have been incorporated.
2. Then we collected a total of 640 web pages from public administrations in order to create a corpus with a great variety of templates. Special care has been taken to diversify the provenance of these pages, as the analysis of the underlying template is critical.
3. Using Python libraries, the features identified in the first step have been extracted.
4. Next, a set of 68 web pages has been selected from the corpus for the learning phase of the machine learning process. These pages have been classified by readability experts into two classes: easy and difficult to understand. This classification has been made considering the guidelines for easy reading and web accessibility.

5. Using Machine Learning algorithms and the data from the learning collection of the previous point, a classification model was obtained, which was subsequently validated with documents from templates not used in the learning process.

4 Evaluation

4.1 Design

Goals. The aim of this work is to identify design features of web pages that can help determine the level of comprehension of the information contained in the page. For this purpose, the structure of the different templates of the pages that form the corpus has been analyzed in depth to find possible design elements in their HMTL and CSS files. In addition, these elements have been searched to see if they have any relation with the comprehension of the information. For example, the structuring of text using headers is known to clearly influence the comprehension of information, as can be seen in the COGA [6] and Plain Language [3] recommendations. But there are other page elements that could influence text comprehension and whose identification is not clearly outlined in the readability guidelines. This motivates this research work since it seeks to find the relationship of style and content elements (HTML and CSS) that may influence comprehension and that are not currently included in these standards and guidelines.

The decision to include style elements, CSS, has been taken considering that these also influence the layout of the pages. Some of the elements analyzed are included in the recommendations on accessibility and easy reading in the guides mentioned above, but other elements have been chosen because we believe that they may be relevant for comprehension, despite not being clearly outlined in the regulations on readability.

In short, we have tried to take as many relevant web page design elements as possible to see which of them have the most impact when evaluating the difficulty of a web site.

Corpus. In order to carry out this work, it was necessary to obtain a robust corpus with a sufficient variety of templates to avoid the results being conditioned by a few designs. A total of 640 web pages have been collected to build the analyzed corpus. Many of the web addresses (URLs) that make up the corpus are from the same domain and therefore share the same design template. However, this does not invalidate its usefulness since within the same web domain the construction of some of its pages may vary and it is also necessary to take this into account. For example, there are cases where in a page of a domain we find longer paragraphs than in other pages, or that include more links, or that use different hierarchies of headings, etc. In Table 1 we can see the number of root domains that formed the corpus divided in terms of comprehension difficulty.

Table 1. Web domains distribution.

Difficulty of comprehension	Web domains
Easy	46
Difficult	58
Total	104

In addition, as in this work we do not focus on the content of the information displayed on the pages, page domains in different languages have also been included to enrich the corpus, since language does not affect the layout. It should also be noted that all the web addresses are from public administrations.

Of a total of 640 pages that make up the corpus, 310 have been classified as easy to understand and the rest as difficult. This first classification has been done manually considering the recommendations proposed by the previously mentioned guides on readability and accessibility. The reason why it has been decided to classify the corpus following these guidelines is because a page that does not follow any of the proposed indications to facilitate the reading of a page will make it more difficult for a person with cognitive problems to understand it without problems. Other considerations that we have taken into account for the classification are the appearance on the pages of the easy-to-read logos or W3C, since they indicate which ones comply with the readability and accessibility guidelines, and the purpose of the web site, that is, to see if its purpose is to simplify administrative procedures of other official pages.

Future Analysis. The template elements analyzed were based on the extraction of 30 HTML and CSS features. These features were considered relevant to our study, either based on heuristics based on page compression or already proposed in the literature (guides and research papers). Table 2 shows some of these features.

Table 2. Features extracted from guides.

Inclusion Europe [1]	Plain Language [3]	WCAG [5]	COGA [6]
Mean size of font in paragraphs	Percentage of words in paragraphs	Alt attribute in images. Alternative text in tooltips	Mean size of line spacing in paragraphs
Percentage of words in paragraphs	Presence of text with bullet points	Mean size of line spacing in paragraphs	Heading labels with a short sentence
Percentage of words in lists' elements	Presence of tables		Presence of text with bullet points
Percentage of underlined words			Number divs
Average font size in the paragraphs			
Heading labels with a short sentence			
Percentage of words in bold			
Presence of images			

Classification Model. The machine learning process consists of two parts: a first learning phase, where the classification algorithm is trained with some input data, and another classification phase, where the algorithm performs a classification of some new data from the previous learning [9].

For this study, the use of a classical supervised classification algorithm using decision trees has been chosen, since the aim is to analyze the priority in decision making in the tree, when one element influences readability more than another. The algorithm used in the evaluation was the J48 algorithm provided by the Weka tool. The J48 algorithm is an implementation of the ID3 or C4.5 algorithm [10] and is commonly used in data mining applications.

For the learning part we took a first set of 68 pages from the corpus, of which 37 were classified as easy and the rest as difficult manually, as explained in the Corpus section. With this first set, different training tests were performed varying the input attributes (or items). A first training test was done with all 30 input attributes obtained in the analysis of the web structure and following recommendations. Then a second test was performed with the same set of pages, but two attributes were excluded for this test, the number of videos and tables that appeared on each page, because they didn't provide any relevant information for the classification. For the last test, in addition to those two attributes, we also excluded the number of paragraphs and the total words from each page because we considered that these absolute values wouldn't be as significant as other attributes such as the percentage of the words that were in paragraphs or the average number of words per paragraph.

For the classification phase, different datasets were tested, and the classification model was tested with these additional documents. The new datasets included pages from domains that had not been included in the learning set as well as others that were from the same domain but did not share the same template.

5 Results

As a means to obtain the classification model, several different datasets were tested for the training phase, but from all the tests made with each one we concluded that the dataset formed by 68 pages was the one that gave the best results. The difference between this dataset and the others tested was the variety of templates included.

The results of the three training tests mentioned in the previous section are shown in Table 3, and as we can see the first two had the same result, which prove that the two excluded attributes were indeed irrelevant. However, in the third test there is a slightly improvement in the results thanks to having excluded the other two attributes mentioned before from the input set. The accuracy measures the frequency on which the classifier does a correct prediction, and it is more valid the more balanced the training set is. In the case of this tests, we have 37 easy pages and 31 difficult ones so we can state that the accuracy results are very reliable. Nevertheless, we also must consider on the precision and recall indexes that, as we can see in the table, are slightly better on the third test as well.

Table 3. Training tests' results.

Test	Input attributes	Accuracy (%)	Precision	Recall
#1	31	80,8824	0.800 (easy) 0.821 (difficult)	0.800 (easy) 0.821 (difficult)
#2	29	80.8824	0.800 (easy) 0.821 (difficult)	0.800 (easy) 0.821 (difficult)
#3	27	82.3529	0.857 (easy) 0.788 (difficult)	0.839 (easy) 0.839 (difficult)

Therefore, the classification model we picked was the third one and as we can see it in Figure 1 and its classification tree in Figure 2. As we can see the more important attributes for this model are the ones in Table 4.

Table 4. Description of attributes from final classification model.

Name of attribute	Description
p_mean_size	Average font size in paragraphs
cells/table	Average number of cells per table
per_h1	Percentage of total words found within the h1 tag
alt_img	Percentage of images that have alt text
per_li	Percentage of words found within the li tag, compared to the total of words.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

p_mean_size <= 13.451613: difficult (18.55/1.0)
p_mean_size > 13.451613
|   cells/table <= 6.5
|   |   p_mean_size <= 16.896552
|   |   |   per_h1 <= 0.15528
|   |   |   |   alt_img <= 40: easy (6.0/1.0)
|   |   |   |   alt_img > 40
|   |   |   |   |   per_li <= 0.479435: easy (2.0)
|   |   |   |   |   per_li > 0.479435: difficult (11.91)
|   |   |   per_h1 > 0.15528: easy (8.0)
|   |   p_mean_size > 16.896552: easy (16.54/0.54)
|   cells/table > 6.5: easy (5.0)
```

Fig. 1. Classification model.

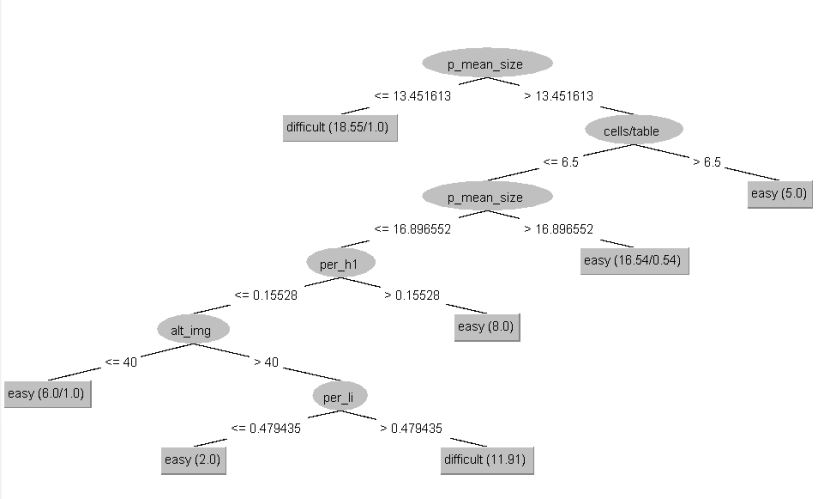

Fig. 2. Classification model's decision-making tree.

6 Conclusions and Further Research

This study has shown an analysis of a set of public administration websites to see which HTML and CSS elements have an impact on their comprehension. The starting point was a study of some guidelines of recommendations on easy reading and accessibility to try to extract some common characteristics to classify web pages according to their difficulty of comprehension.

With a series of tests made from machine learning on data obtained from a set of web pages we have obtained results that seem to be promising. The readability aspects have been widely studied, but always under non-web formats, the main contribution of this study is to consider the web dimension. This study shows that there are HTML elements that really can help to design web pages who are easier to comprehend for people with cognitive disabilities.

In this work we have avoided introducing the elements of classical readability, such as the richness of the vocabulary. However, it is worth considering that the present variables and models interact in some way with these variables.

One aspect of interest at the beginning of the study was to know to what extent the recommendations adequately reflected what was observed for the web, and the conclusion is that these recommendations do reflect the main guidelines for online publication.

References

1. Inclusion Europe. Easy-to-read - Inclusion Europe. http://www.inclusion-europe.eu/easy-to-read/
2. US Gov. Public Law 111 - 274 - Plain Writing Act of 2010 - https://www.govinfo.gov/app/details/PLAW-111publ274/summary

3. Federal Plain Language Guidelines - https://www.plainlanguage.gov/law/
4. W3C Accessibility Initiative - https://www.w3.org/WAI/
5. W3C. Web Content Accessibility Guidelines (WCAG) 2.1 - https://www.w3.org/TR/WCAG21/#robust
6. W3C. Cognitive and Learning Disabilities Accessibility (COGA) https://www.w3.org/TR/coga-usable/
7. Herramienta de evaluación de legibilidad en línea Comp4Text - https://gigabd2.uc3m.es/comp4text/index.php
8. Peter Williams (2017). Eliciting web site preferences of people with learning disabilities. University College London. https://nasenjournals.onlinelibrary.wiley.com/doi/epdf/10.1111/1471-3802.12099
9. Mitchell, T. (1997). Machine Learning, McGraw Hill.
10. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
11. Muñoz, B, M. y Muñoz, U., J.M (2019). Legibilidad Mu. Viña del Mar, Chile. Recuperado de http://www.legibilidadmu.cl
12. Fernández Huerta (1959). Medidas sencillas de lecturabilidad. Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS) 1959; (214): 29-32