Comp4Text Checker: An Automatic and Visual Evaluation Tool to Check the Readability of Spanish Web Pages

Ana Iglesias¹, Ignacio Cobián¹, Adrián Campillo¹, Jorge Morato¹, and Sonia Sánchez-Cuadrado²

¹ Computer Science and Engineering Department, Universidad Carlos III de Madrid, Madrid, Spain
aiglesia@inf.uc3m.es
² Library and Information Science Department, Universidad Complutense de Madrid, Madrid, Spain

Abstract. One important requirement for a web page to be accessible for all, according to the current international recommendations from the W3C Accessibility Initiative is that the text should be readable and understandable to the broadest audience possible. Nowadays, unfortunately, the information included in the web pages are not easy to read and understand to everybody. This paper introduces the Comp4Text online readability evaluation tool, which is able to calculate the readability level of a web page based on classical linguistic measures (sentence to sentence) and detect unusual words and abbreviations. Moreover, it provides recommendations to solve the readability problems and show everything in a very visual way. Thanks to this tool, the web page designers and writers could improve their sites, being easier to be read and understand for all. Currently, Comp4Text is based on the Spanish language, but it can be easily extended to other languages if the readability metrics and easy-to-read rules are well-known.

Keywords: Web pages readability · Automatic tool · Accessibility

1 Introduction

Currently, and increasingly, we are immersed in a digital society that pretends to make our lives easier. However, unfortunately, not everybody is able to access to the digital information and services provided yet because of accessibility barriers. This proposal is focused on one of the bases of accessibility: the readability of the web pages’ text, according to the current international recommendations from the W3C Accessibility Initiative (WAI)¹.

¹ WAI accessibility principles: https://www.w3.org/WAI/fundamentals/accessibility-principles/ (accessed on June ‘20).
The accessibility barriers on the Internet become especially important and worrying when they are related to the individual’s rights, like public information from the Government, or services as education, health, employment or social activities, among others. That is why digital inclusion is a world-wide strategy nowadays, to ensure that everybody, regardless of their functional abilities/disabilities (physical or cognitive), can contribute-to and benefit-from the digital economy and society.

This paper presents Comp4Text, an online readability evaluation tool which can help web pages’ designers and writers to improve their documents and make them readable for all.

2 Literature Review

2.1 Readability Tools

There exist different types of tools which try to test or adapt the documents to make the text more readable according to the reader’s needs.

Some of the tools are focused on making the text more perceivable and readable by customizing the visual presentation of the contents. Typically, this type of tools are called legibility tools. This kind of text adaptation is very useful for people with visual impairments who usually experience eyestrain, fatigue or headaches among others when reading. Examples of these tools are the ATBar2 (by the University of Southampton), and the Accessibility Enabler3 (by Hike Orders). Both tools are toolbars to help users to customize the way they view and interact with web pages by changing the background, font style, font size, line spacing, etc. of the web pages.

Some other efforts have been focused on providing a better understanding by customizing the text to the user’s needs by modelling and graphically representing more than one version of text through Variant Graphs, like the TadER project [1] or StemmaWeb project [2]. Web applications were developed following this proposal, as the CollateX4 web tool or StemmaWeb5 tool, showing directed acyclic graphs, horizontally aligned, with the different versions of the text fragments.

Other research projects are focused on computing and showing the text readability based on typical linguistic measures to check whether the content is easy-to-read. More information about linguistic readability measures are provided in the next section. The linguistic measures are highly language dependent, and most of them are developed for English language, as Readability Text Tool6 (by webFX), the Document Readability Text Tool7 (by Online-utility.org), the Readability Grader8 (by Jellymetrics) or the

2 ATBar: https://www.atbar.org/ (accessed on June ’20).
3 Accessibility Enabler: https://hikeorders.com/accessibility/home/ (accessed on June ’20).
4 CollateX: https://collatex.net/ (accessed on June ’20).
5 StemmaWeb: https://stemmaweb.net/ (accessed on June ’20).
6 Readability Text Tool: https://www.webfx.com/tools/read-able/ (accessed on June ’20).
7 Document Readability Text Tool: http://www.online-utility.org/english/readability_test_and_improve.jsp (accessed on June ’20).
8 Readability Grader: https://jellymetrics.com/readability-grader/ (accessed on June ’20).
Reading Effectiveness Tool\textsuperscript{9} (by Clear Language and Design - CLAD). A tool for Italian language can also be found in the literature, the Vâmola\textsuperscript{10} (by Regione Emilia-Romagna).

However, specific readability tools have not yet been developed for the Spanish language. This paper presents a new visual proposal for a readability testing tool for Spanish language: Comp4Text. The next subsection introduces a state of the art related to typical and new linguistic measures in readability. Many of these measures have been considered and some of them implemented in the CompText tool, as we detail in the next sections, where the tool is introduced.

2.2 Readability Measures

Metrics to estimate readability of texts in English emerged in the 1950s in the educational field applied to level the text at the corresponding school age. These metrics were based on characteristics such as the number of characters, number of syllables, number of words, sentences, and paragraphs. Different metrics were created based on the size of different parts of the text, or by selecting two or three combinations of indicators (Flesh-Kinkaid, Gunning-Fog, Coleman-Liau Index, SMOG Index or ARI). In fact, more than 70 indexes have been developed \cite{3, 4}, including some for Romance languages such as Spanish, French or Italian. Main metrics proposed for Spanish language were Inflesz, Legibilidad µu (Lµ), Gutiérrez Polini’s readability metric, Flesh-Fernández Huerta, and Szgriszt-Pazos among others \cite{5}. In general, these formulas penalize writing with polysyllabic words and long and complex sentences. Texts score improve when simpler and shorter words and sentences are used.

Researchers Dale and Chall \cite{6} proposed a derived approach, based on the frequency of use of a word. Their hypothesis assumes that the frequency of use of a word shows that there is a community of users that understands its meaning, and therefore reflects the familiarity that readers have regarding the term. They proposed a list of simple words that included a value on the familiarity of the word with which they could presumably estimate the educational level that would correspond to the text.

The result of applying the comprehensibility metrics provides a value that establishes the text difficulty on a scale of five or seven levels depending on each formula. Flesh index proposed five values: 1) very easy, 2) easy, 3) standard, 4) difficult, and 5) very difficult. Flesh Fernández-Huerta index created seven levels: 1) very easy, 2) easy, 3) fairly easy, 4) standard, 5) fairly difficult, 6) difficult, 7) very difficult. These metrics have successfully served their purpose for decades.

2.3 Discussion

After an in-depth analysis of the current readability checker tools in this section, we have realized that none of them is specific for the Spanish language. Therefore, due to the high language dependency of the readability linguistic measures, Comp4Text

\textsuperscript{9} Reading Effectiveness Tool: http://clad.tccld.org/measuring-readability/ (accessed on June ’20).

\textsuperscript{10} Vâmola: http://www.validatore.it (accessed on June ’20).
checker covers this issue, based on the linguistic indexes described in the previous subsection.

Moreover, the new tool introduced on this paper will try to maintain the accessibility and improve the usability of the readability checker tools for English and Italian language mainly in four different points: showing graphically and directly in the web page the main readability errors or warnings found, reducing the location problems that other tools have (indicating, for instance, a line code); making recommendations for improving the readability of specific sentences and unusual words; allowing to show and download summary reports in standard formats (pdf and json); and facilitate user interaction by allowing users to navigate the website dynamically (following the links) and checking the readability of other web pages online.

Inclusive methodologies, as the Accessibility Conformance Evaluation Methodology [7] and Universal Design Principles [8] have been taken into account from the very beginning of the Comp4Text’ design in order to obtain an accessible and usable product.

As a first step, a study of the current readability tools for Spanish language and other languages was conducted, finding that most readability tools were developed for English language. In this step, an automatic readability checker tool was selected to be the focus of the study, taking into account and trying to solve the differences with the previous ones.

Moreover, an in-depth study of the best linguistic measures to take into account for the Spanish language and the adaptation of classical measures to the new type of text, as well as study of recommendations to improve the text based on linguistic research works was performed.

The first prototype of the Comp4Text tool, introduced in this paper, was designed and implemented in a laboratory environment (at the University). Then, two accessibility experts made a heuristic accessibility evaluation of the application according to the w3c methodology [7], defining the evaluation scope, exploring the target website, selecting and auditing a representative sample and reporting the findings. Automatic accessibility checker tools were also used during the developing process, as WAVE11 and aChecker12 web accessibility evaluation tools. Then, all these findings were fixed to ensure an appropriate accessibility level. As future work, an in-depth user evaluation is going to be performed for the Comp4Text tool, before its public web upload, to ensure its accessibility for all.

3 Comp4Text Application

The Comp4Text application is an online readability evaluation tool which provides a visual representation of readability issues within the document. The aspect is similar to WAVE accessibility tool (see Fig. 1). At present, it calculates the readability for Spanish language of web pages (providing the URL). In the near future, it will be able

---

11 WAVE web accessibility evaluation tool: https://wave.webaim.org/ (accessed on June ’20).
12 aChecker web accessibility evaluation tool: https://achecker.ca/ (accessed on June ’20).
to calculate the readability of documents (including the most common documents formats: .txt, .doc, .pdf, etc.) or plain text.

Classical readability measures for the Spanish language have been used to calculate the sentences readability: Flesch-Kincaid index (Fernández-Huerta Index) [9] in 1958 and mu index [5] in 2006. Both indexes are based on the number of years of education that a person needs to be able to understand the text easily on the first reading and they are very similar according to attributes used in their construction. However, the two of them are used to check the accurateness of traditional measures within the time (there is a difference of more than 50 years between them).

On the other hand, to estimate the readability of the words in the document (detecting unusual words and abbreviations), word frequency and familiarity measures have been used. Each word in the text is associated with a frequency value from Current Spanish Reference Corpus (CREA for its Spanish acronym) [10]. According to that frequency value, the words in the text are classified into three different levels (very frequent, frequent and not frequent words) using the Zipf’s law and Goffman’s transition point approach [11]. Moreover, the application is able to provide a list of synonyms from the Spanish version of Wordnet lexical database13 to suggest other options to improve readability.

Figure 1 shows the interface of the readability tool. On the right frame, the web page evaluated is shown visually. On the left frame, a summary of the web page readability is shown, using the Flesch Kincaid Index -that measures the understability-, and the Mu Index-that measures how difficult is the web page to read. Moreover, scrolling down in the left frame of the web page, the tool details the number of complex

---

13 Wordnet lexical database por Spanish language: https://wordnet.princeton.edu/ (accessed on June '20).
sentences (marked with a square icon in red within an x letter), the number of easy sentences (marked with a round icon in green within a tick symbol), the number of sentences which medium readability (marked with a round icon in yellow within an exclamation symbol) and the number of sentences that weren’t evaluated (marked with a light blue icon within an exclamation). Figure 2 shows this information.

By clicking any tag, on the left frame, the sentences on that category are marked on the web page right-side. For instance, in Fig. 1 we can see how the first sentence has been framed with a red and dashed line (complex sentence) and the corresponding squared and red icon. Moreover, the other sentences and non-common words are also marked with their corresponding readability icon (red-green-yellow-blue icons), where a number is used to identify each one of the icons on the text, needed to identify the readability problems in the final report.

A more detailed information of the readability evaluation is popped-up by clicking on any error or warning icon on the right frame. Figure 2 shows a screenshot of the application when un icon has been clicked, popping-up detailed information about the problem and how to solve it. For instance, if a complicated word or sentence is detected, suggestions for what to do to improve its readability are shown. Moreover, the user can feedback this information, rating it, including a new recommendation and how sure is of this recommendation.

One of the most important differences to other readability evaluation tools is that, similar to WAVE application, the application introduced in this paper annotates errors and warnings in-line and visually on the document, taking into account accessibility, but focused on readability evaluations. It underlies instantly and visually the elements (sentences or words) on the document that are causing the possible reading problems after sliding the mouse over it or its paper tag.

![Fig. 2. Screenshot 2 of Comp4Text. Details of readability evaluation for a sentence. (Color figure online)](image-url)
The application has been designed according to the W3C recommendations, accomplishing the WCAG 2.0\textsuperscript{14} accessibility guidelines and, at the same time, using colourful error and warning icons in order to be easily identified in the document, even if the document is a colourful web page.

The application also allows building a final readability evaluation report of the web page, allowing to download a .json report. This report includes a summary of the readability problems found, within a list of sentences that are suggested to be rewritten and a list of unusual words to be changed to improve readability and how to fix the readability issues are provided in the report.

4 Conclusions and Planned Activities

This paper introduces Comp4Text checker, a new web readability evaluation tool for Spanish web pages, based on classical linguistic measures and making recommendations to improve the readability of the web pages in a very visual way.

This new tool take into account accessibility and usability issues: it shows graphically and located in the original webpage, the main readability errors or warnings found, reducing the location problems that other tools present; it makes recommendations for improving the readability of specific sentences and unusual words; it shows and allows to download summary reports in standard formats (.pdf and .json); and it makes easier the interaction of the users, allowing them to navigate dynamically on the website (following the links).

In the near future, API services and extensions for the most Chrome and Firefox browsers are going to be developed. Currently, it calculates the readability for Spanish language of web pages (providing the URL), but it could be easily adapted to other languages, considering their specific linguistic measures and tools to provide recommendations. Moreover, Comp4Text will soon be able to calculate the readability of documents (including the most common documents formats: .txt, .doc, .pdf, etc.) or plain text.

We are also working now on an in-depth accessibility evaluation of the Comp4Text checker, being necessary to involve final users in the accessibility evaluation of the tool before publishing it. Moreover, an in-depth user evaluation of the tool effectively will be performed, checking if the tool contributes to render texts easier to understand.

Acknowledgment. This work has been partially funded by the CSO2017-86747-R Spanish project.

References

1. TAdER – Text Adaptability is Essential for Reading. http://www.tader.info/scrolling.html. Accessed April 2020

\textsuperscript{14} WCAG 2.0 guidelines (W3C): https://www.w3.org/TR/WCAG20/ (last access on June, 2020)
2. Jänicke, S., et al.: TRAViz: a visualization for variant graphs. Digit. Scholarsh. Humanit. 30 (suppl_1), i83–i99 (2015)
3. Social Science Consulting, “TextQuest”, http://textquest.de/shop/product_info.php?products_id=3. Accessed April 2020
4. Morato, J., Sánchez-Cuadrado, S., Gimmelli, P.: Estimación de la comprensibilidad en paneles de museos. EPI 27(3), 570 (2018)
5. Baquedano, M.M.: Legibilidad y variabilidad de los textos. Boletín De Investigación Educacional [Artículo De Revista] 21, 13–25 (2006)
6. Dale, E., Chall, J.S.: A formula for predicting readability. Educ. Res. Bull. 27, 11–28 (1948)
7. Web Accessibility Conformance Evaluation Methodology (WCAG-EM) v1.0. https://www.w3.org/TR/WCAG-EM/. Accessed April 2020
8. Story, M.F.: Maximizing usability: the principles of universal design. Assist. Technol. 10(1), 4–12 (1998)
9. Fernández-Huerta, J.: Medidas sencillas de lecturabilidad. Consigna 214, 29–32 (1959)
10. Real Academia Española: Banco de datos (CORDE). Corpus diacrónico del español. http://www.rae.es. Accessed April 2020
11. Urbizagastegui Alvarado, R., Restrepo Arango, C.: La ley de Zipf y el punto de transición de Goffman en la indización automática. Investigación Bibliotecológica 25(54), 71–92 (2011)