Assessing the Reading Level of Web Texts for WCAG2.0 Compliance – Can It Be Done Automatically?

Evelyn Eika[1] and Frode E. Sandnes[1,2]

[1] Faculty of Technology, Art and Design, Oslo and Akershus University College of Applied Sciences, Oslo, Norway
[2] Westerdals Oslo School of Art, Communication and Technology, Oslo, Norway
{Evelyn.Eika, frodes}@hioa.no

Abstract. Readability of text on the web is a key prerequisite for achieving universal accessibility. The World Wide Web Consortium's Web Content Accessibility Guidelines state that general text should not require reading levels more advanced than lower secondary education. The subsequent research into readability on the web is limited. However, the literature on measuring readability and reading level is vast, but limited to simple measures of sentence length and word difficulty. This study explores the value of using other features that are harder to acquire manually, but are now readily available through computer technology. Our results indicate that the proposed features are not as accurate predictors to readability as the classic measurements. There may thus be some way to go before we have reliable automatic means of assessing texts on the web for readability.

Keywords: Readability · text on the web · Accessibility · WCAG2.0 · Universal Design

1 Introduction

The widely embraced W3C Web Content Accessibility Guidelines (WCAG2.0) give advice on how to make web content accessible to users with various disabilities [1]. Most research into accessibility on the web focuses on visual impairment and some on reduced motor abilities. Cognitive disabilities have probably received the least attention due to its complexity. However, the cognitive task of comprehending texts on the web is central to the websites' accessibility.

WCAG2.0 addresses cognitive disabilities under section 3 and readability in particular under section 3.1. Some of the recommendations that are easy to handle include 3.1.1 and 3.1.2, which state that web pages and their parts must be correctly coded with information about the language used. Sections 3.1.3, 3.1.4 and 3.1.6 address the definition of unusual words, abbreviations and the pronunciation of ambiguous words, respectively.

A more challenging recommendation is that of 3.1.5 addressing reading level, which states that

> *"When text requires reading ability more advanced than the lower secondary education level after removal of proper names and titles, supplemental content, or a version that does not require reading ability more advanced than the lower secondary education level, is available. (Level AAA)"*

The WCAG2.0 documentation further uses the UNESCO definition of lower secondary education as being equivalent to 7-9 years of schooling [2]. It further acknowledges that it is not possible to make text universally readable because text is connected to a specific language. The WCAG2.0 documentation gives a specific example for American English using the Flesch-Kincaid formula [3] where the lower secondary education level is set at 7.2.

Although actual user testing of texts can give reliable facts about its readability, and eye-tracking data can give objective data, it is not practical to employ user testing on all text at all times. Therefore, it is desirable with simple mathematical models that could roughly predict readability. A good metric will highlight which parts are satisfactory according to some standard and unsatisfactory parts that require more work. Another perspective is to give authors tools that allow them to produce satisfactory text in the first place, eliminating need of stringent post composition audits [4].

The classic readability measures are relatively easy to compute, even by hand, as they usually evolve around sentence length and word length. These measures have also been criticized for being over simplistic [5]. This study thus attempted to explore other features that are very laborious to explore manually but are easily available using computers and openly available language processing technology.

2 Background

Readability research has been conducted for more than a century. Sentence length is one of the features that are cited as affecting readability throughout most studies [5, 6, 7], where long sentences are generally considered harder to read than short sentences. However, it is not necessarily always true that short sentences are easier and thus a simple focus on shortening sentences for the sake of improving readability is not recommended.

The second factor that is also frequently mentioned is word difficulty [5, 6, 7], as texts with difficult words are harder to read than those with easy words. However, there is disagreement on what constitutes a difficult word. Many simply measure word length, using the number of syllables or number of characters in the word. If a word has more than two syllables, it is often considered a difficult word. Word frequency has also been used to quantify the difficulty level of a word, where less frequent words are considered more difficult than frequent words. Some have attempted to make lists of difficult words through manual assessment.

Sentence length and word length commonly occur in readability metrics. For example, the Flesch-Kincaid reading easy index [3] is defined as follows:

$$206.835 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right) \quad (1)$$

The Gunning Fog index uses similar variables, namely:

$$0.4\left[\left(\frac{words}{sentences}\right) - 100\left(\frac{difficult\ words}{words}\right)\right] \quad (2)$$

Coleman-Liau is another similar index [8]:

$$5.89\left(\frac{characters}{words}\right) - 29.5\left(\frac{sentences}{words}\right) - 15.8 \quad (3)$$

Yet another is the simple SMOG index [9], which is short for simple measure of goobledygood.

$$1.043\sqrt{30\frac{syllables}{sentences}} + 3.1291 \quad (4)$$

A practical and popular approach that can be used without computer is offered by the Fry-chart [10]. In Fry's method, three 100-word paragraphs are selected arbitrarily in the text; the mean number of sentences per 100 words and that of syllables per 100 words are computed. These two values are then plotted in the Fry-chart allowing the reading level to be read off directly.

Much of the readability research, including those above, target English. However, some research has also been conducted for other languages including Chinese [11] – a language very unlike English. In Scandinavia (Norway, Denmark and Sweden), the lix-index proposed by Carl Hugo Björnsson [12] is frequently used. The LIX (lesbarhetsindeksen) is defined as:

$$\frac{words}{sentences} + \frac{difficult\ words}{words}100 \quad (5)$$

Note that the number of sentences is defined as the number of sentence delimiters (punctuations, capital letters, colons, etc.). Difficult words are defined as words with more than six letters. Values in the range 25-30 indicate simple texts, while values below 25 are typical for children's books.

One weakness of many readability indices is that they can be misleading. For example, a high readability score can be obtained by rewriting a short text with difficult words into a longer text with easier words and shorter sentences. Although this gives a higher readability score, the actual readability may be worse. This point is especially valid on the web as both overall text length and word difficulty, as well as other language aspects, can affect how effectively the readers are to read. As an example, the Norwegian language council recently removed its LIX-calculator from its web site as it is deemed not useful.

2.1 Other Approaches

Factors related to style, including punctuations, prepositional phrases, verb tense and mood, may also affect readability. Other aspects such as content, formatting and organization also play a role. Among these, content is difficult to assess automatically by machine.

More aspects of language may be captured by qualitative approaches than simple readability indexes. One good example is leveling [13]. As it is believed that text resembling speech is easier to understand, leveling allows the assessment of whether a text is similar to speech.

Readable texts are useful for all readers, and particularly for groups having specific needs. Visually impaired users relying on screen readers benefit from well-structured and short document with front-loaded sentences allowing easy navigation. Documents with these characteristics are also known to be useful for dyslexic readers [14].

3 Method

The open source LanguageTool [15] was chosen for this study (version 2.0). This tool allows more sophisticated readability models to be explored. LanguageTool is an extensive grammar and spelling framework developed for several languages including English, Spanish, Polish, Danish, etc. LanguageTool is written in java; it is the default language checking mechanism used in several open source word processing packages. It can be run stand-alone and it has an API allowing arbitrary java applications to make use of its functionality. LanguageTool has therefore been integrated into various research projects [16, 17].

In this study, the evaluation is limited to English although LanguageTool also supports other languages. Three measures are proposed, namely, language problem signature, part-of-speech signatures and part of speech entropy. Each of these will be introduced in the subsequent sections.

3.1 Language Problem Signatures

The LanguageTool framework contains approximately 1,000 rules related to grammar and style in English. It is assumed that text published on the web has been proofread and therefore contains few language issues related to spelling and grammar. However, LanguageTool often triggers rules as false positives reporting issues that are not problems. The purpose of the proposed scheme is to run the framework on a text and note the type of problems that are reported. The profile of the reports may give a clue to the type of text at hand.

3.2 Part-of-speech signature

LanguageTool has a built-in part-of-speech tagger. This module can classify the individual parts of a sentence such as nouns, verbs, articles, etc. LanguageTool recognizes approximately 40 different part-of-speech classes. We further grouped these classes

into six categories, namely, nouns, verbs, modifiers (such as adjectives phrases and adverbs), linking words (e.g., however and in addition), weights (unnecessary phrases) and complex (foreign words).The mapping from the LanguageTool part-of-speech tags to our six categories is provided in Table 1.

Table 1. The part-of-speech tag to category mapping.

Category	Part-of-speech token
noun	PRP, NN, NNS, NN:U, NN:UN, NNP, NNPS
verb	VB, VBD, VBG, VBN, VBP, VBZ
modifier	CD, DT, JJ, JJR, JJS, RB, RBR, RBS, PDT, UH, PRP$
linker	CC, IN, RP, TO, WDT, WP, WP$, WRB
weight	EX, MD
complex	FW

The framework was thus used to generate a histogram for the text according to these categories. The idea is that more complex texts would contain more modifiers, linkers, unnecessary weight and descriptors than simple texts, and that the signature would be linked to the genre of writing.

3.3 Part-of-speech Entropy

Unlike the two previous measures that comprise multiple values, the last proposed measure is a single value - entropy. Entropy is a quantity measure originally used in physics to quantify the amount of information needed to know the complete state of an object. It can also be considered a measure of uncertainty or randomness. We define part-of-speech entropy by computing the entropy of the part-of-speech histogram for the text with all the 40 tags. The entropy computation [18] is defined as:

$$e = -\sum_{i} p_i \ln(p_i) \tag{6}$$

Where p_i is the ratio, or probability, of token i. Tokens that do not occur are not included in the computation. The rationale for this measure is that texts that employ more varied patterns of tokens yield higher entropy, and that the entropy somehow correlates with writing difficulty.

4 Experiments

To test the proposed measures, a set of texts were selected at various reading difficulties. These include three short children's texts at three levels, a paragraph from a disclaimer, a university strategy document and a scientific journal article. The text and their traditional reading scores are listed in Table 2.

Table 2. Text test suite with text length (number of words), mean sentence length (number of words), mean word difficulty (mean number of syllables per word) and the Flesch-Kincaid reading index.

Text	Length	Sentence length	Word difficulty	Flesch-Kincaid
Child level 1	79	8.8	1.3	87.6
Child level 2	76	12.7	1.3	82.7
Child level 3	74	12.3	1.3	81.1
Disclaimer	109	21.8	1.6	45.8
Strategy	2,235	18.2	2.0	22.5
Journal	8,231	13.4	1.7	48.1
J. abstract	178	13.7	1.7	46.6
J. introduction	437	24.3	1.7	35.8
J. discussion	887	19.3	1.7	43.9

The last column shows that the Flesch-Kincaid score correlates quite well with the texts as the three children's stories have the highest score. Moreover, the levels of the children's stories correlate with the readability index where the text at level 1 has the highest score and the most difficult text at level 3 has the lowest score of the three.

Next, the journal article is slightly more readable than the disclaimer according to the Flesch-Kincaid score, while the strategy document is the least readable. This ranking corresponds well with the authors' subjective perceived readability of the texts.

Table 2 shows that the word difficulty appears to be a relatively stable measure as it is constant for each category (children's story and journal), while the sentence length varies depending on the part of the text. Consequently, the Flesch-Kincaid measure does vary according to which part of a text that is analyzed. Clearly, the score for the introduction is significantly lower than the abstract and the discussion.

Observably, the texts vary from as little as 74 words to more than 8,000 words. The short texts are realistic test cases since texts on the web often are short. In fact, the length of the text is itself an indication of accessibility as the journal text probably is less attractive to the average reader.

Table 3. Language issues reported.

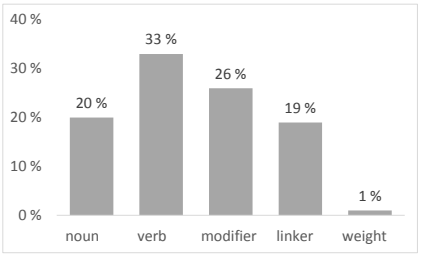
Test	issues
Child level 1	3.8 %
Child level 2	5.3 %
Child level 3	5.4 %
Disclaimer	1.8 %
Strategy	4.7 %
Journal	8.7 %
J. abstract	3.9 %
J. introduction	0.9 %
J. discussion	0.5 %

Table 3 shows the number of language issues reported for each text. The journal article triggers most issues, while each smaller part of the article triggers fewer issues,

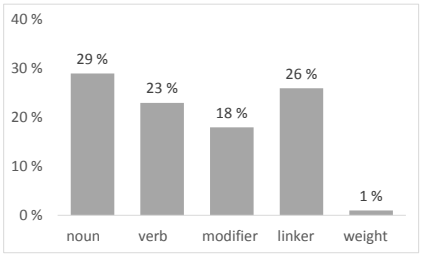
with introduction and discussion among the least. The results indicate journal abstract, introduction and discussion have few issues, but that there must be some issues in the other part of the journal article. Next, the children's stories at level 2 and 3 trigger the most errors and the number of issues correlates with the reading level. Simply looking at the percentage of issues does not appear useful, while inspecting the details of the issues reveal useful information. For instance, the children's story only contains a possible typo, while the strategy has several readability issues including redundant phrases, confusion of British and American English and overly long sentence. The following is a listing of the issues output for the journal article (the heading number indicates frequency of the problem).

329 : Possible Typo(prio=50) - Possible spelling mistake
136 : Miscellaneous(prio=50) - Whitespace repetition (bad formatting)
91 : Capitalization(prio=50) - Capitalize lowercase words ('i am')
38 : Capitalization(prio=50) - Checks that a sentence starts with an uppercase letter
35 : Miscellaneous(prio=50) - Readability: sentence over 40 words
22 : Bad style(prio=50) - Number starting a sentence
20 : Miscellaneous(prio=50) - Use of whitespace before comma and before parentheses
11 : Grammar(prio=50) - Articles: article missing before a countable noun
6 : Miscellaneous(prio=50) - Unpaired braces, brackets, quotation marks, similar symbols
5 : Miscellaneous(prio=50) - Flag passive voice
5 : Miscellaneous(prio=50) - Smart ellipsis (…)
3 : Miscellaneous(prio=50) - American words easily confused in British English
3 : Redundant Phrases(prio=50) - in order to (to)
3 : Punctuation Errors(prio=50) - Warn when the serial comma is used (incomplete)
2 : Grammar(prio=50) - Possible agreement error: numeral + singular countable noun
2 : Grammar(prio=50) - Agreement error: Non-third person verb with 'he/she'
1 : Miscellaneous(prio=50) - Repetition of two words ('at the at the')
1 : Possible Typos(prio=50) - web site (website)
1 : Miscellaneous(prio=50) - Use of 'a' vs. 'an'

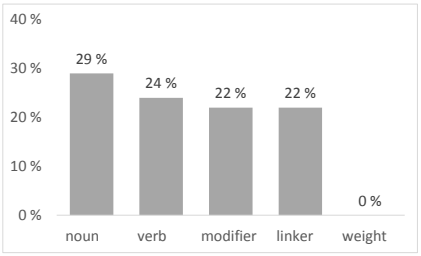
Next, Fig. 1 shows the part-of-speech signatures for the texts. The signatures show that the disclaimer and the children's story have the most verbs. However, the disclaimer has more modifiers than the children's story, and the children's story has more nouns than the disclaimer. Both the journal and the strategy have most nouns, where the journal has the most even distribution of all part-of-speech categories. The journal does not have any weight tags. The strategy document, however, has some words that add unnecessary weight and it has more linkage words than modifiers. This could indicate that the text is unreadable with many prepositional phrases, or it could indicate that text is readable with linked passages. More work is needed to clarify such ambiguities.



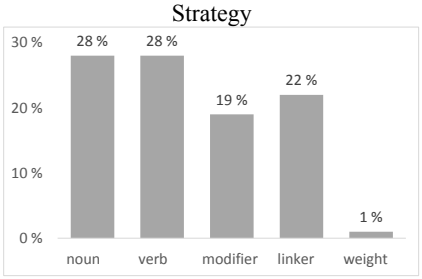
Disclaimer



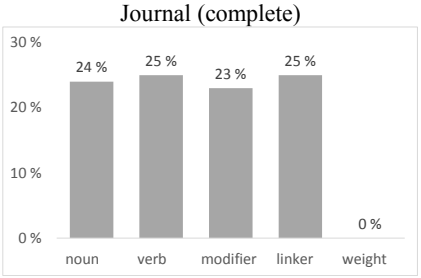
Strategy



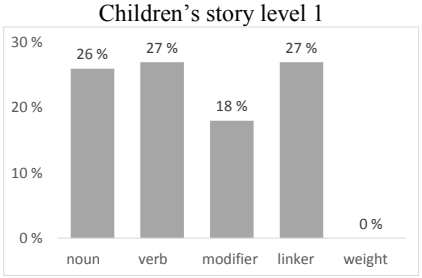
Journal (complete)



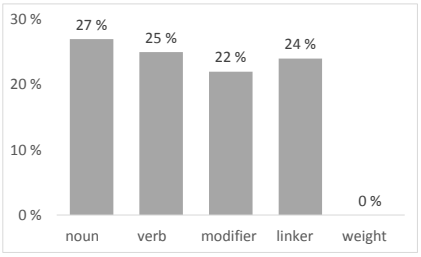
Children's story level 1



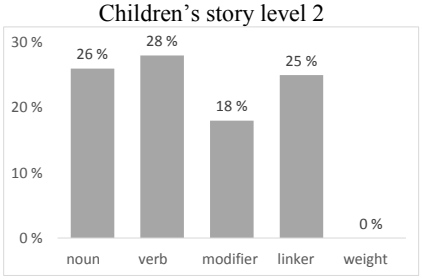
Journal abstract



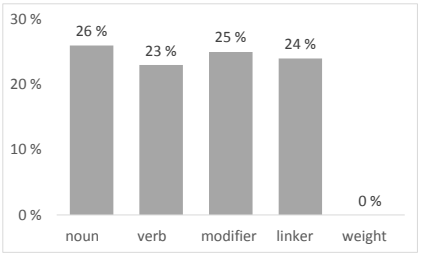
Children's story level 2



Journal introduction



Children's story level 3



Journal discussion

Fig. 1. Part-of-speech signatures for the texts according to nouns, verbs, modifiers, linker, weight and complexity.

Table 4. Part-of-speech entropy.

Test	entropy
Level 1	2.785075
Level 2	2.805429
Level 3	2.836197
Disclaimer	2.771111
Strategy	2.889141
Journal full	2.962654
Journal abstract	2.918766
Journal introduction	2.945794
Journal discussion	2.959201

Table 4 lists the part-of-speech entropy for the texts. The results suggest that the entropy correlates with the reading level and that it is robust to the text sample. Clearly, the part-of-speech entropy increases gradually with the reading level of the children's text as it is 2.79 for level 1, 2.81 for level 2 and 2.84 for the level three text.

Further, the entropy for the full journal article is 2.96, while it is 2.92 for the abstract, 2.95 for the introduction and 2.96 for the discussion. The length has some effect as longer texts result in a higher entropy. This is as expected since a longer text is likely to make use of more sentence variations than a shorter text. However, the variation is relatively small and entropy appears to converge with length.

An interesting result is that the lowest entropy is obtained for the disclaimer. Visually inspected, the disclaimer can be perceived as being complex to read due to its formal style, many long words and comparatively longer sentences than other types of text. However, the part-of-speech entropy suggests that it has the least grammatical variation. It may be that the disclaimer tends to use similar sentence patterns. Future research could investigate further the sentence type, inclusive of its makeup involving subjects, verbs and objects. More complex constructs such as relative pronouns and subordinate conjunctions could be examined within sentence types as they can affect the precision of meaning, which further complicates readability.

5 Conclusion

This study explored the use of more sophisticated models for quantifying readability using computer analysis. Three models were studied, namely, profile of reported errors, part-of-speech signatures and part-of-speech entropy. The results show that the proposed measures do not sufficiently or entirely discriminate texts according to readability and that the simple measures based on sentence length and word difficulty appear to be better predictors. Nevertheless, the measures give some alternative views on texts and could help draw authors' attention towards potential problems. As shown, it is not trivial to automate the process of determining whether texts on the web are readable or not according to given criteria. Future work should explore natural language processing framework that explores the text at a deeper level. The ideas presented herein may also be applied to language learning [19] and the teaching of writing [20].

References

1. World Wide Web Consortium. Web content accessibility guidelines (WCAG) 2.0 (2008)
2. UNESCO International Standard Classification of Education, http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm
3. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis (1975)
4. Eika, E., Sandnes, F.E.: Authoring WCAG2.0-compliant texts for the web through text readability visualization. In: Proceedings of HCI International 2016. LNCS, in press, (2016).
5. Janan, D., Wray, D.: Reassessing the accuracy and use of readability formulae. Malaysian Journal of Learning and Instruction 11, 127--145 (2014)
6. Kitchenham, B.A., Brereton, O. P., Owen, S., Butcher, J., Jefferies, C.: Length and readability of structured software engineering abstracts. Software, IET 2, 37--45 (2008)
7. Heydari, P.: The validity of some popular readability formulas. Mediterranean Journal of Social Sciences 3, 423—435 (2012)
8. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology 60, 283 (1975)
9. McLaughlin, G.H.: SMOG grading: A new readability formula. Journal of reading 12, 639--646 (1969)
10. Fry, E.B.: A readability formula that saves time. Journal of reading 11, 513–-516 (1968)
11. Pang, L.T.: Chinese readability analysis and its applications on the internet. PhD diss., The Chinese University of Hong Kong (2006)
12. Björnsson, C.-H..: Readability of newspapers in 11 languages. Reading Research Quarterly, 480--497 (1983)
13. Pearson Levelling Guide, http://www.pearsonplaces.com.au/Places/Primary_Places/Primary_English_Place/Levelling_Guide.aspx
14. Baeza-Yates, R., Rello, L.: Estimating dyslexia in the Web. In L. Ferres, M. Vigo & J. Abascal (Eds.), W4A'11 Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (article no. 8). New Work: ACM (2011)
15. LanguageTool, https://languagetool.org/
16. Viljoen, C.M., Nitschke, G.S., Van Heerden, W.S.: Evolution of a fictional dialogue. In: 2011 IEEE Congress on Evolutionary Computation (CEC), pp. 1100-1107. IEEE (2011).
17. Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.F., Davalos, S., Teredesai, A., De Cock, M.: Age and gender identification in social media. In: Proceedings of CLEF 2014 Evaluation Labs (2014)
18. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5, 3--55 (2001)
19. Jian, H.-L., Sandnes, F.E., Law, K.M.Y, Huang, Y.-P., Huang, Y.-M.: The role of electronic pocket dictionaries as an English learning tool among Chinese students. Journal of Computer Assisted Learning 25, 503--514 (2009)
20. Jian, H.-L., Sandnes, F.E., Huang, Y.-P., Cai, L., Law, K.M.Y.: On students' strategy-preferences for managing difficult course work. IEEE Transactions on Education 51, 157—165 (2008)