Readability of Wikipedia

by Teun Lucassen, Roald Dijkstra, and Jan Maarten Schraagen

Abstract

Wikipedia is becoming widely acknowledged as a reliable source of encyclopedic information. However, concerns have been expressed about its readability. Wikipedia articles might be written in a language too difficult to be understood by most of its visitors. In this study, we apply the Flesch reading ease test to all available articles from the English Wikipedia to investigate these concerns. The results show that overall readability is poor, with 75 percent of all articles scoring below the desired readability score. The 'Simple English' Wikipedia scores better, but its readability is still insufficient for its target audience. A demo of our methodology is available at www.readabilityofwikipedia.com.

Contents

- [1. Introduction](#)
 - [2. Methodology](#)
 - [3. Results](#)
 - [4. Discussion and conclusions](#)
 - [5. Demo](#)
-

1. Introduction

Since its introduction in 2001, Wikipedia has evolved to a vast collection of encyclopedic information. In an interview with Slashdot in 2004 [1], co-founder Jimmy Wales is straightforward in stating his goals for 'the free encyclopedia anyone can edit':

"Imagine a world in which every single person on the planet
is given free access to the sum of all human knowledge.
That's what we're doing."

Wales seems to be quite successful in achieving this goal. As of March 2011, the English version of Wikipedia alone contained over 3.5 million articles [2]. Furthermore, Wikipedia is available in 279 different languages.

Much research has already been carried out on the quality of information in Wikipedia (Giles, 2005; Wilkinson and Huberman, 2007) as well as the perception of quality by its users (Kubiszewski, *et al.*, 2011; Lucassen and Schraagen, 2011). However, a different concern has recently been raised, namely about its readability.

In a study on the accuracy of cancer information (Rajagopalan, *et al.*, 2010) on Wikipedia, it was found that while information on this topic was of comparable accuracy with information from a professionally maintained database, readability of Wikipedia was significantly worse. This was investigated by applying Flesch–Kincaid readability tests (Kincaid, *et al.*, 1975) to articles from both Web sites and reported by means of the Flesch–Kincaid grade level. This enables a comparison with reading levels expected at the various grades in the U.S. educational system. While poor readability was only shown for Wikipedia articles on cancer, we have no reason to expect that articles on different topics will score better.

Wikipedia does not state a particular target audience for whom the encyclopedia was built, except for Jimmy Wales' quote at the beginning of this paper. He stated that the sum of all human knowledge is made accessible for every single person. However, we hypothesize that information on Wikipedia is often not accessible to many, because the articles are too difficult to read. Walraven, *et al.* (2009) found frequent use of Wikipedia among 14-year-old high school students. These adolescents are not expected to have fully developed reading abilities and may thus have limited access to information on Wikipedia.

1.1. Flesch Reading Ease Test

Readability can be described as the ease with which a reader can understand the message conveyed by a writer. Several attempts have been made to evaluate the readability of texts automatically (DuBay, 2004). One of the earlier attempts is the Flesch reading ease test (Flesch, 1948), which incorporates two easily measurable concepts deemed important for readability in a simple algorithm. First, it is assumed that longer sentences are harder to read. This is measured by the average number of words per sentence (SL). Similarly, it is assumed that longer words are harder to read. Thus, word length (WL) is determined by the average number of syllables per word. The formula used to calculate the Flesch reading ease (RE) is:

$$RE = 206.835 - (1.015 \times SL) - (84.6 \times WL)$$

The outcome of this formula is generally between 0 and 100 (although higher and lower scores are theoretically possible) and can be interpreted as shown in [Table 1](#).

Table 1: Reading Ease Scores with interpretations.	
Reading Ease Score	Interpretation
90–100	Very easy
80–90	Easy
70–80	Fairly easy
60–70	Standard
50–60	Fairly difficult
30–50	Difficult
0–30	Very difficult

Following Table 1, we can derive that when writing for the general public (as in the standard English Wikipedia), you should aim at a standard readability, reflected by a score between 60 and 70 (Standard). When the target audience is expected to have a limited proficiency in the English language, you should aim for a substantially higher readability, perhaps with a minimal score of 80 (Easy).

Readability tests such as the Flesch Reading Ease Score, or the Flesch–Kincaid Grade Level test (which uses the same characteristics from a text) are applied for a range of purposes. The U.S. Department of Defense uses the Flesch–Kincaid Grade Level to indicate the readability of their documents (Kincaid, *et al.*, 1975). Furthermore, insurance policies in Florida are required by law to have a reading ease score of at least 45 [3].

The validity of the Flesch reading ease test to predict comprehension has always been topic of discussion. Of course, by considering the two simple measures of word length and sentence length, the whole concept of readability cannot be covered. However, strong correlations have been found between Flesch reading ease scores and comprehension tests (DuBay, 2004). Still, the reading ease scores (by any readability formula) should be interpreted with some caution.

1.2. Simple English Wikipedia

The hypothesized poor readability of Wikipedia has been acknowledged by the Wikimedia Foundation. This has resulted in the introduction of a 'Simple English' version of Wikipedia, which as of March 2011 contains about 68,000 articles [4]. The goal of this site is "to provide an encyclopedia for people with different needs, such as students, children, adults with learning difficulties and people who are trying to

learn English.” [5] This goal is pursued by using simpler grammar and a limited vocabulary. It is advised to contributors to start writing using only Basic English (Ogden, 1968), which consists of 850 words. If necessary, contributors may expand their vocabulary with words from the 1,500–word extended version of Basic English.

Several studies have examined the effectiveness of the Simple English Wikipedia. It was shown that in the first three years after its introduction in 2003, readability, in terms of Flesch reading ease scores, declined as the number of articles increased (Besten and Dalle, 2008). While in the second half of 2003, the average reading ease score was around 80 (Easy), by 2006, the average reading ease score had dropped to just over 70 (Fairly Easy). A renewed analysis of the current state of affairs concerning the readability of the Simple English Wikipedia is of interest to see if this decline has continued. If this is the case, then the Simple English Wikipedia is becoming less and less readable for users with special needs and its added value over the regular English Wikipedia is vanishing.

This study investigates the readability of both the English and Simple English Wikipedia by means of the Flesch reading ease test (Flesch, 1948). We hypothesize that the English Wikipedia is too difficult to read for many. Furthermore, we expect that the Simple English Wikipedia tackles this problem by featuring an improved readability. However, if the decline in readability found earlier (Besten and Dalle, 2008) continues, the added value of a simple version of Wikipedia drops, as people with limited proficiency in English may be unable to comprehend content.

2. Methodology

2.1. Dataset

Instead of attempting to select a representative sample of articles from both versions of Wikipedia, we performed reading ease tests on all suitable articles. By doing so, we ensure that our results are a true reflection of the readability of Wikipedia at a particular moment in time.

Articles from the English Wikipedia were obtained from the downloadable database dump in XML format of 17 August 2010 [6]. The articles from the Simple English Wikipedia were obtained from the database dump of 2 September 2010.

2.2. Processing

The XML files were loaded into a local MediaWiki using a MediaWiki data dump importer tool, MWDumper [7]. Subsequently, the content of each article was retrieved from the MediaWiki database and filtered for full sentences (see next section).

Flesch reading ease scores were calculated using php-text statistics [8], which uses the formula discussed in [section 1.1](#). The result was stored again in a local MySQL database. After all articles were processed, the reading ease scores were exported to comma-separated values (CSV) files and imported into SPSS for further analysis.

2.3. Filtering

Due to the metrics used in the Flesch Reading Ease test (sentence length and word length), only full sentences should be included in the analysis. This means that all other content had to be filtered out before applying the tests.

First, all articles marked as disambiguation pages, redirects, and lists were removed from the analysis, as well as pages solely featuring other media such as graphic, audio, or PDF files. These pages do not feature content in the form of full sentences and should therefore be excluded from the analysis.

Second, all elements in the remaining articles which were not (part of) full sentences were filtered out. Examples of these are headings, tables, and URLs. Such elements are included in Wikipedia articles using the MediaWiki markup language. This language provides simple codes to add layout to plain text. For example, a heading is created by surrounding the title of the heading with equal signs (*e.g.*, ‘== Introduction ==’).

In order to remove all irrelevant elements from the text, regular expressions were constructed in PHP to localize these markup codes. Some elements were removed from the text entirely, such as headings, tables and images. Other elements were modified in order to remove all contextual information about these elements, which are not part of the text itself. Examples of these modified elements are links, references and citations. The remainder of those elements, which were a part of full sentences, were not removed.

Not all Wikipedia articles were suitable for the application of reading ease tests. Some articles, for instance, did not contain any full sentences after applying the filter. A large number of these articles were lists, for example containing links to related topics. Moreover, in some cases other markup than the MediaWiki language was used by the authors of a given article, using, for instance, HTML. This yielded problems in our filter, which was predominantly designed for MediaWiki markup. The filter was designed so that articles which yielded errors were excluded from further analysis.

For the English Wikipedia and Simple English Wikipedia, respectively 88 and 85 percent of all articles featuring full sentences were successfully filtered and tested.

3. Results

3.1. English Wikipedia

A total of 2,955,210 articles were available for analysis after filtering for full sentences. However, when analyzing the articles on the English Wikipedia, it immediately becomes apparent that many of the articles contain only a few sentences. As can be seen in [Figure 1](#), about 40 percent of the articles contain only five sentences or less.

Figure 1: Distribution of the number of sentences in articles in the English Wikipedia.

[Figure 2](#) shows that the readability of articles tends to stabilize around articles with more than five sentences. Moreover, the variation in readability is much larger for shorter articles ($SD = 24.32$ for articles with five sentences or less, $SD = 13.84$ for articles with more than five sentences). A small peak in average readability can be observed around 55 sentences, caused by numerous entries on American cities added in batch (Ortega, *et al.*, 2009). They all feature a similar readability and about the same number of sentences.

Figure 2: Reading ease scores for articles with a varying number of sentences in the English Wikipedia.

We argue that such short articles are not representative of Wikipedia articles in general. Therefore, we exclude articles with five sentences or less from further analysis.

[Figure 3](#) shows the distribution of reading ease scores for the remaining 1,710,752 articles with more than five sentences in the English Wikipedia. The average reading ease score was 51.18 ($SD = 13.84$). We found that 73.5 percent of all articles scored below the suggested goal of 60 (Standard), whereas 45.0 percent could even be qualified as difficult or worse (<50).

Figure 3: Histogram of reading ease scores in the English Wikipedia.

3.2. Simple English Wikipedia

Of the 57,422 articles available for analysis of the Simple English Wikipedia, an even larger proportion contains only five sentences or less (see [Figure 4](#)). In this version, the number of short articles amounts to over 60 percent.

Figure 4: Distribution of the number of sentences in articles in the Simple English Wikipedia.

Similarly to the regular English Wikipedia, the readability of articles tends to stabilize around articles with more than five sentences on the Simple English Wikipedia (see [Figure 5](#)). Short articles show a larger variation in readability ($SD = 19.52$) than articles with more than five sentences ($SD = 13.00$).

Figure 5: Reading ease scores for articles with a varying number of sentences in the Simple English Wikipedia.

[Figure 6](#) shows the distribution of reading ease scores of articles with more than five sentences in the Simple English Wikipedia. A total of 21,366 articles were analyzed leading to an average reading ease score of 61.69 ($SD = 13.00$). We found that 94.7 percent of all articles scored below the suggested goal of 80 (Easy), whereas 42.3 percent did not even reach a reading ease score of 60 (Standard), which is the desired goal for the general public, instead of a public with special language needs.

Figure 6: Histogram of reading ease scores in the Simple English Wikipedia.

3.3. Comparison

It is unfair to directly compare the average reading ease scores of both versions of Wikipedia as they differ widely in the number of available articles. Therefore we only selected articles with exactly the same page title on both versions ($N = 9,603$) for a direct comparison. Moreover, as discussed before, only articles with more than five sentences were considered.

Due to the reduced number of considered articles, the average reading ease of the English Wikipedia was 49.27 ($SD = 11.33$), compared to an average reading ease score of 61.46 ($SD = 12.95$) of the Simple English Wikipedia. For the comparable article set, the Simple English Wikipedia had significantly higher reading ease scores than the English Wikipedia ($t(9602) = 99.63, p < .001$).

4. Discussion and conclusions

In this paper, we investigated the current state of affairs concerning the readability of the English Wikipedia. We have shown that a high number of articles scores well below the desired readability standard. The proposed solution in the form of the Simple English Wikipedia features better readability; however, this edition should aim for a much better readability as it specifically aims at individuals with limited reading abilities. Moreover, we have shown that the decline in readability in the Simple English Wikipedia, since its introduction in 2003, has continued.

The results of this study show that the readability of the English Wikipedia is overall well below a desired standard. Although the average score of 51.18 does not seem far from the desired goal, nearly 75 percent of all articles scored below 60 in the Flesch reading ease test. Moreover, half of the articles can

be classified as difficult or worse. This finding confirms our hypothesis that numerous articles on Wikipedia are too difficult to read for many people.

This readability problem was already demonstrated for articles on cancer (Rajagopalan, *et al.*, 2010). For articles on this topic, a Flesch–Kincaid grade level of 14 was found, which reflects a reading ease score of about 30. Fortunately, we showed that the average readability of the English Wikipedia is considerably higher than this score, although numerous articles (about seven percent, which amounts to about 12,500 articles) scored 30 or below. This indicates that readability may vary heavily between different topics on Wikipedia. A topical analysis (similar to Halavais and Lackaff, 2008) can now reveal which articles most urgently need to improve readability. Entries on technical topics may be less readable than articles on other subjects, such as popular culture.

The urgency of improving different articles also depends on usage statistics. Articles with many page views need to be improved more quickly than those on arcane subjects seeing few readers. This can also be related to users with a limited proficiency in English, as they might be more interested in certain topics over others. The use of openly available data on page views for each Wikipedia article could help in prioritization topics for improvement.

A solution to this problem was proposed by the Wikimedia Foundation in the form of the Simple English Wikipedia. This version of Wikipedia performed much better on readability tests than the English Wikipedia. However, the results show that readability dropped 10 points in the last four years (Besten and Dalle, 2008), which means that it now falls in the Standard category. This is a worrying result, as it seems that the Simple English Wikipedia has lost its focus. Instead, this version now seems suitable for the average reader, instead of aiming at those with limited language abilities. One way to check the success of the Simple English Wikipedia is to test how many of words in it actually appear on the 850– and 1,500–word Basic English lists (Ogden, 1968).

A possible explanation for the poor readability of both versions of Wikipedia rests with the contributors to Wikipedia. It can be expected that these are predominantly well-educated individuals, as it requires a substantial level of domain knowledge to be able to add content to Wikipedia which is not already available. It may be possible that these contributors write entries for their peers, instead of considering the broad and generic audience of Wikipedia. Even if they pay special consideration to their audience, as in the case of the Simple English Wikipedia, they seem to overestimate their audience's reading skills.

It would be interesting to see how contributors of Wikipedia could be encouraged and supported in improving the readability of their entries. Perhaps it could be achieved by implementing a tool in the edit environment of Wikipedia, which alerts contributors to long sentences or difficult words. Suggestions to break up sentences or to use easier wording could be given in the course of editing. The nature of Flesch readability tests makes it possible to have an online readability measure available for contributors to Wikipedia.

It should be noted that the Simple English Wikipedia is still rather underdeveloped. This is not only reflected by the relatively low number of entries but also by their average length. During our analysis, we noticed that about half of the articles consisted of only three sentences or less. This observation gives extra motivation to shift the focus of contributors towards enhanced readability, especially at this stage. We expect that it is more effective to write articles in an easy way initially, rather than to improve readability in later versions.

Furthermore, in this study we applied only one (automated) measure of readability. Although the usefulness of the Flesch readability test is widely acknowledged, a renewed validation of this and other measures would be of interest, in order to ensure that the scores correlate with comprehension and user perceptions of readability.

5. Demo

An online tool has been created to demonstrate the methodology of applying Flesch reading ease tests to Wikipedia articles. Using this tool, the readability of live Wikipedia articles can be tested by entering their titles. Articles on both the English and Simple English Wikipedia can be tested. In addition to an absolute reading ease score for each article, the reading ease relative to the samples used in this study is given by means of a percentile score. Moreover, visitors may enter their own text to test readability. The online demo is accessible at <http://www.readabilityofwikipedia.com>.

About the authors

Teun Lucassen is Ph.D. candidate in the Department of Cognitive Psychology & Ergonomics at the University of Twente in Enschede, The Netherlands. He received his Master's degree in the Department of Human Media Interaction at the same university. Currently, Teun is researching trust in collaborative repositories for his Ph.D. thesis, with a special interest in Wikipedia.

Roald Dijkstra is owner of Babbletics, a company which focuses on the development of intuitive Web applications. He acts as a freelance interaction designer and is currently also a Master's student in the Department of Human Media Interaction at the University of Twente.

Jan Maarten Schraagen is Senior Research Scientist at the Netherlands Organization for Applied Scientific Research TNO and Professor of Applied Cognitive Psychology at the University of Twente. His research interests include task analysis, team decision-making, trust in collaborative repositories, and adaptive human-computer collaboration. He was co-editor of *Cognitive task analysis* (Mahwah, N.J.: L. Erlbaum Associates, 2000) and *Naturalistic decision making and macrocognition* (Aldershot, England: Ashgate, 2008). Dr. Schraagen holds a Ph.D. in cognitive psychology from the University of Amsterdam, The Netherlands.

Notes

1. <http://interviews.slashdot.org/article.pl?sid=04/07/28/1351230>.

2. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

3. <http://law.onecle.com/florida/insurance/627.4145.html>.

4. <http://simple.wikipedia.org/wiki/Special:Statistics>.

5. http://en.wikipedia.org/wiki/Simple_English_Wikipedia.

6. <http://download.wikimedia.org>.

7. <http://www.mediawiki.org/wiki/Mwdumper>.

8. <http://code.google.com/p/php-text-statistics/>.

References

Matthijs L. den Besten and Jean-Michel Dalle, 2008. "Keep it simple: A companion for Simple Wikipedia?" *Industry and Innovation*, volume 15, number 2, pp. 169–178. <http://dx.doi.org/10.1080/13662710801970126>

William H. DuBay, 2004. "The principles of readability," at <http://www.nald.ca/library/research/readab/readab.pdf>, accessed 20 August 2012.

Rudolph Flesch, 1948. "A new readability yardstick," *Journal of Applied Psychology*, volume 32, number 3, pp. 221–233. <http://dx.doi.org/10.1037/h0057532>

Jim Giles, 2005. "Internet encyclopaedias go head to head," *Nature*, volume 438, number 7070 (15 December), pp. 900–901, and at <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>, accessed 20 August 2012.

Alexander Halavais and Derek Lackaff, 2008. "An analysis of topical coverage of Wikipedia," *Journal of Computer-Mediated Communication*, volume 13, number 2, pp. 429–440. <http://dx.doi.org/10.1111/j.1083-6101.2008.00403.x>

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, Brad S. Chissom, 1975. "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel," Millington, Tenn.: Chief of Naval Technical Training, Naval Air Station Memphis; Springfield, Va.: distributed by National Technical Information Service (NTIS).

Ida Kubiszewski, Thomas Noordewier, and Robert Costanza, 2011. "Perceived credibility of Internet encyclopedias," *Computers & Education*, volume 56, number 3, pp. 659–667. <http://dx.doi.org/10.1016/j.compedu.2010.10.008>

Teun Lucassen and Jan Maarten Schraagen, 2011. "Factual accuracy and trust in information: The role of expertise," *Journal of the American Society for Information Science and Technology*, volume 62, number 7, pp. 1,232–1,242.

Charles K. Ogden, 1968. *Basic English: International second language*. New York: Harcourt, Brace & World.

Felipe Ortega, Jesus M. Gonzalez–Barahona, and Gregorio Robles, 2009. "Quantitative analysis of the top ten Wikipedias," In: Joaquim Filipe, Boris Shishkov, Markus Helfert, and Leszek Maciaszek (editors). *Software and data technologies: Second International Conference, ICSOFT/ENASE 2007, Barcelona, Spain, July 22–25, 2007, revised selected papers. Communications in Computer and Information Science*, volume 22. Berlin: Springer–Verlag, pp. 257–268.

M.S. Rajagopalan, V. Khanna, M. Stott, Y. Leiter, T.N. Showalter, A. Dicker, and Y.R. Lawrence, 2010. "Accuracy of cancer information on the Internet: A comparison of a Wiki with a professionally maintained database," *Bodine Journal*, volume 3, number 1, article 8, at <http://jdc.jefferson.edu/bodinejournal/>, accessed 20 August 2012.

Amber Walraven, Saskia Brand–Gruwel, and Henny P.A. Boshuizen, 2009. "How students evaluate information and sources when searching the World Wide Web for information," *Computers & Education*, volume 52, number 1, pp. 234–246. <http://dx.doi.org/10.1016/j.compedu.2008.08.003>

Dennis M. Wilkinson and Bernardo A. Huberman, 2007. "Cooperation and quality in Wikipedia," *WikiSym '07: Proceedings of the 2007 International Symposium on Wikis*, pp. 157–164, and at <http://www.hpl.hp.com/research/scl/papers/wikipedia/wikipedia07.pdf>, accessed 20 August 2012.