Wiki Workshop (11th edition) – June 20, 2024

Do People Perceive Differences in the Readability of Wikipedia Articles?

Indira Sen
University of Konstanz

Mareike Wieland
GESIS

Katrin Weller
GESIS

Martin Gerlach
Wikimedia Foundation

Abstract

We study the alignment between human and automated assessments of Wikipedia articles' readability using surveys and interviews. We find that people's assessment of readability is subjective, depending on factors like the rater's education.

Keywords: Text Simplification, Knowledge Gaps, Readability, Surveys, Simple Wikipedia

Introduction

Automatic Readability Assessment, one of the prerequisites for text simplification, aims to computationally measure the difficulty level of texts. Assessing the readability of Wikipedia articles is of prime importance due to Wikipedia's value in educational contexts. Previous research has used automated readability metrics on English Wikipedia articles, and found that their readability level is too high, often catering to highly educated people while remaining inaccessible to wider populations (Lucassen et al., 2012). To address these knowledge gaps, the Simple Wikipedia project[1] attempts to create simpler, more readable versions of English Wikipedia articles. However, it is unclear if automated readability metrics are well-suited for assessing the readability of Wikipedia articles. Particularly, we know little of whether people's perceptions of readability align with these automated measures, especially for Wikipedia and it's Simple edition.

Therefore, to better understand how readers gauge the readability of Wikipedia articles, we conduct surveys to obtain quantitative and qualitative measures of perceived readability. By studying the alignment between people's perception of readability of Wikipedia articles and automated readability metrics, we shed light on the utility, applicability, and limits of these automated measures.

In this extended abstract, we discuss preliminary results of people's assessment of readability based on pilot surveys and cognitive pretesting (Koskey, 2016). Specifically, we sample pairs of articles from English and Simple Wikipedia, and ask participants which article snippet they find easier to read and understand, without telling them which snippet comes from which Wikipedia edition. We also obtain the participants' relevant background and demographic information, such as their education levels, level of fluency in English, etc. Finally, we complement the findings from this survey using cognitive pretesting.

We find that the agreement between survey participants is lower than fair agreement, indicating that people fail to come to a consensus about which snippet is simpler to read and understand. We also find varying and opposing views on what facets of text drive readability assessments.

Data and Methods

Survey Layout. We used LimeSurvey[2] to construct our survey. To control for confounds due to the topic of the articles, each survey question consists of pairs of snippets that came from the same article, but in different readability levels, i.e., pairs of snippets where one article was from Simple Wikipedia (easy, labeled 'simple') and the other from English Wikipedia (difficult, labeled 'en'). We also added an optional free-text question at the end of the survey for the participants to describe what type of strategies they rely on when judging the ease of reading. Since we opt for a descriptive annotation paradigm (Röttger et al., 2022), i.e., our goal is to understand how people perceive readability without priming, we do not provide an explicit definition of what 'simpler' means. Figures 2 and 3 show the layout of our survey.

Data Selection and Preprocessing. We used the Media API to collect 103,971 articles and their texts, where each text was taken from the same article in two different versions (Simple Wikipedia and English Wikipedia). We retained those pairs which have at least three sentences for both versions. Since we can only show a limited amount of text in the survey, we truncated all the articles to 750 characters to obtain article snippets. We then randomly selected 40 pairs of snippets for the pilot. We calculated the automated readability score of all snippets with the Flesch Reading Ease (FRE) formula.[3]

Survey Participants. We recruited 15 participants from Prolific,[4] who were either fluent in English or had English as their first language. Each participant was shown 10 randomly selected pairs out of the total 40

[1] https://simple.wikipedia.org/wiki/Main_Page

[2] https://www.limesurvey.org/

[3] https://pypi.org/project/textstat/

[4] https://www.prolific.com/

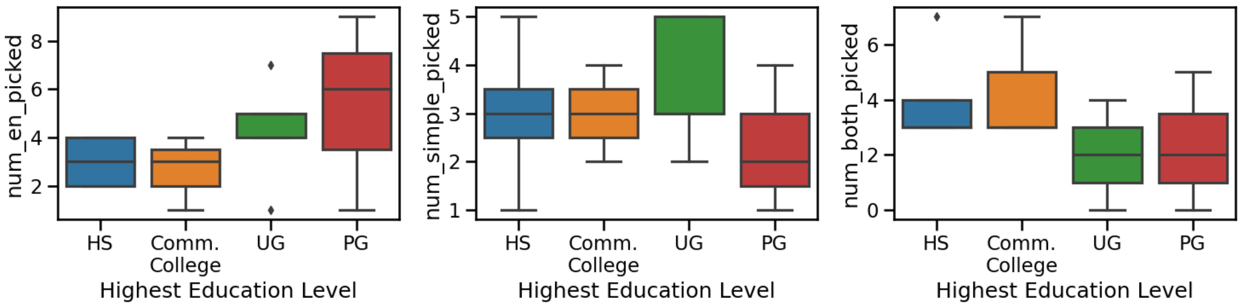


Figure 1: **Association between Education Levels of Participants and their Perception of Readability.** People with higher education levels, especially those with graduate degrees, tend to pick the 'en' version, i.e,, the more difficult English Wikipedia version as more readable. ('HS': High School, 'Comm. College': Technical or Community College, 'UG': Undergraduate Degree, 'PG': Graduate Degree)

snippet pairs and each pair had at least three ratings.

Cognitive Pretesting. The pretesting was done with three participants who took the same survey taken by the Prolific participants, but this time while engaging in a 'think-aloud' session with an interviewer, where the interviewee explained why they rated a snippet to be simpler.

Results

Analyzing the results from survey, we find that out of a total of 150 ratings of the 40 snippets, the Simple Wikipedia version was picked to be more readable 58 times, i.e., only 38.6% of the time. Interrater-agreement, measured using Krippendorf's α is 0.15; this indicates low or weak agreement. We find no correlation of agreement with difference in FRE scores of the pair of articles. The raletively low rate of picking the Simple Wikipedia version and the low agreement indicates that there is some subjectivity in people's perception of readability. However, one of the factors driving this subjectivity could be the raters' education levels; people with higher education levels tended to pick the English Wikipedia (and not the Simple Wikipedia) snippet as more readable (Figure 1).

Qualitatively analyzing the free-text answers about strategies for assessing readability, we find that raters have variable preferences — some prefer shorter sentences, while others think that too-short sentences break the text's flow. During the pretesting interviews, we find that raters struggle to identify the more readable snippet when the two snippets in a pair diverge in terms of content, indicating a specific issue with Wikipedia English-Simple pairs and document-level readability assessment vs. the more common sentence-level assessments.

In conclusion, our preliminary results from the pilot reveal that readability is subjective and may depend on several characteristics of the raters (e.g., education level), of the content (e.g., short vs. long sentences, complex words), and possibly their combination.

Discussion

Given the importance of Wikipedia for educational and informative purposes, it is crucial to measure the readability level of Wikipedia articles in a reliable and sound manner. Automatic Readability metrics are widely used, however they might not align with people's perceptions (Alva-Manchego et al., 2021). Furthermore, there are very few studies measuring people's readability assessment of Wikipedia articles. This work attempts to bridge this gap by using a mixed-methods approach, relying on surveys and pretesting interviews. Our ongoing work and preliminary results surface the subjectivity of assessing readability and point to the need for studying reader demographics as a mediating variable. We will follow-up on this work to conduct a survey with a larger pool of participants and more article pairs. Our work has important implications in the context of personalizing Wikipedia articles for specific and diverse audience groups, based on their motivations, backgrounds, and sociodemographics.

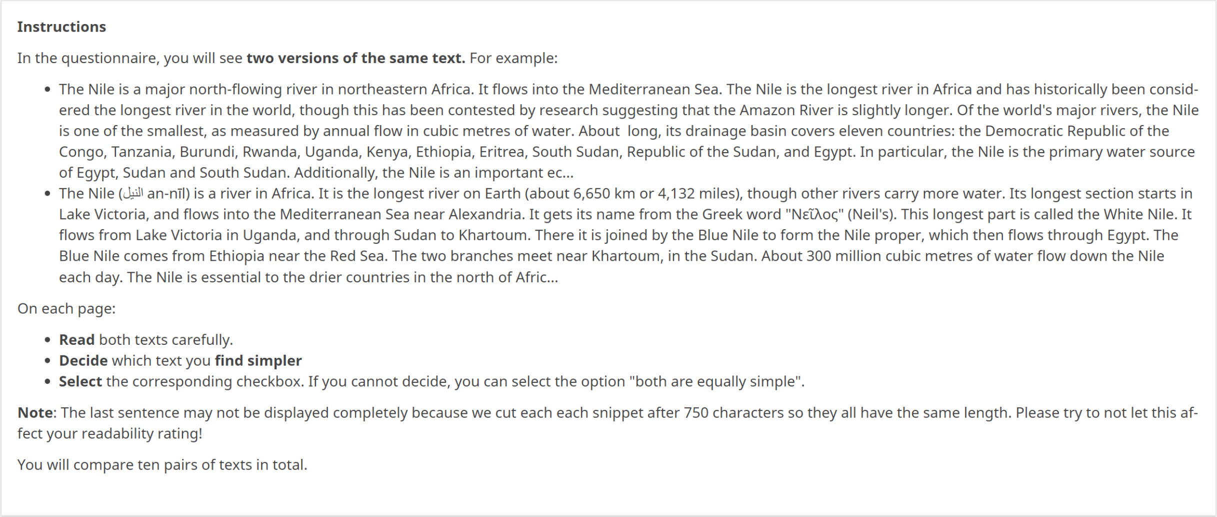
References

[Alva-Manchego et al.2021] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*.

[Koskey2016] Kristin LK Koskey. 2016. Using the cognitive pretesting method to gain insight into participants' experiences: An illustration and methodological reflection. *International Journal of Qualitative Methods*.

[Lucassen et al.2012] Teun Lucassen, Roald Dijkstra, and Jan Maarten Schraagen. 2012. Readability of wikipedia. *First Monday*.

[Röttger et al.2022] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *NAACL*.



Instructions

In the questionnaire, you will see **two versions of the same text.** For example:

- The Nile is a major north-flowing river in northeastern Africa. It flows into the Mediterranean Sea. The Nile is the longest river in Africa and has historically been considered the longest river in the world, though this has been contested by research suggesting that the Amazon River is slightly longer. Of the world's major rivers, the Nile is one of the smallest, as measured by annual flow in cubic metres of water. About long, its drainage basin covers eleven countries: the Democratic Republic of the Congo, Tanzania, Burundi, Rwanda, Uganda, Kenya, Ethiopia, Eritrea, South Sudan, Republic of the Sudan, and Egypt. In particular, the Nile is the primary water source of Egypt, Sudan and South Sudan. Additionally, the Nile is an important ec...
- The Nile (النيل an-nīl) is a river in Africa. It is the longest river on Earth (about 6,650 km or 4,132 miles), though other rivers carry more water. Its longest section starts in Lake Victoria, and flows into the Mediterranean Sea near Alexandria. It gets its name from the Greek word "Νεῖλος" (Neil's). This longest part is called the White Nile. It flows from Lake Victoria in Uganda, and through Sudan to Khartoum. There it is joined by the Blue Nile to form the Nile proper, which then flows through Egypt. The Blue Nile comes from Ethiopia near the Red Sea. The two branches meet near Khartoum, in the Sudan. About 300 million cubic metres of water flow down the Nile each day. The Nile is essential to the drier countries in the north of Afric...

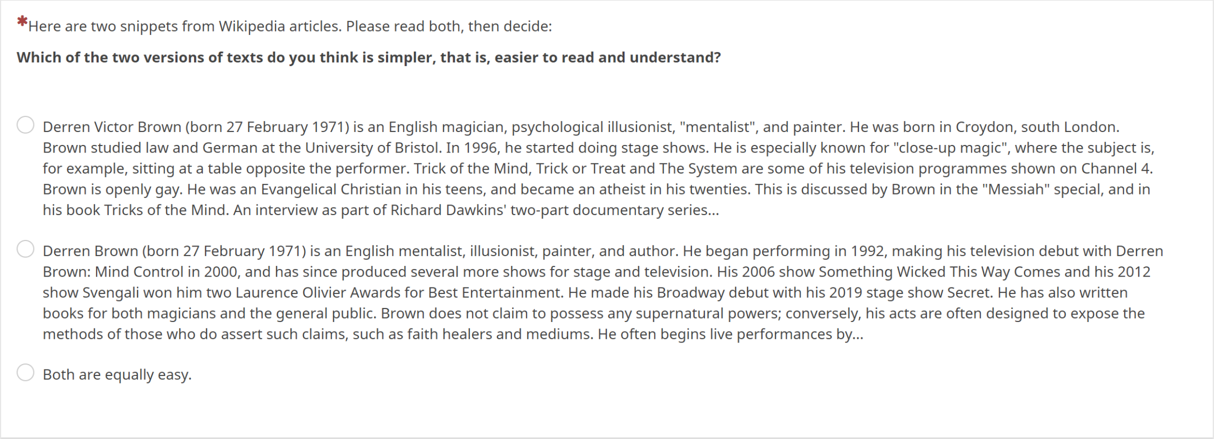
On each page:

- **Read** both texts carefully.
- **Decide** which text you **find simpler**
- **Select** the corresponding checkbox. If you cannot decide, you can select the option "both are equally simple".

Note: The last sentence may not be displayed completely because we cut each each snippet after 750 characters so they all have the same length. Please try to not let this affect your readability rating!

You will compare ten pairs of texts in total.

Figure 2: **Instructions given to participants in the Readability Assessment Survey.**



*Here are two snippets from Wikipedia articles. Please read both, then decide:

Which of the two versions of texts do you think is simpler, that is, easier to read and understand?

○ Derren Victor Brown (born 27 February 1971) is an English magician, psychological illusionist, "mentalist", and painter. He was born in Croydon, south London. Brown studied law and German at the University of Bristol. In 1996, he started doing stage shows. He is especially known for "close-up magic", where the subject is, for example, sitting at a table opposite the performer. Trick of the Mind, Trick or Treat and The System are some of his television programmes shown on Channel 4. Brown is openly gay. He was an Evangelical Christian in his teens, and became an atheist in his twenties. This is discussed by Brown in the "Messiah" special, and in his book Tricks of the Mind. An interview as part of Richard Dawkins' two-part documentary series...

○ Derren Brown (born 27 February 1971) is an English mentalist, illusionist, painter, and author. He began performing in 1992, making his television debut with Derren Brown: Mind Control in 2000, and has since produced several more shows for stage and television. His 2006 show Something Wicked This Way Comes and his 2012 show Svengali won him two Laurence Olivier Awards for Best Entertainment. He made his Broadway debut with his 2019 stage show Secret. He has also written books for both magicians and the general public. Brown does not claim to possess any supernatural powers; conversely, his acts are often designed to expose the methods of those who do assert such claims, such as faith healers and mediums. He often begins live performances by...

○ Both are equally easy.

Figure 3: **Example of Questions in the Readability Assessment Survey.** Each question consists of one of the 10 snippet pairs and the question soliciting the readability rating for it, including the “both are equally easy” option for cases perceived to be ambiguous.