Title: Interactive Exploration of the Readability of Science Authors

Authors:

Russell Jarvis, PhD¹, Patrick McGurrin, PhD², Rebecca Featherston, PhD³, Marc Skov Madsen, PhD⁴, Shivam Bansal, MS⁵, and Bradley G. Lusk, PhD⁶

¹School of Life Sciences, Arizona State University, Tempe, AZ, USA.

²National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA.

³Research Fellow, Department of Social Work Monash University, Clayton, AU

⁴Data, Models and Analytics Ninja. Lead Data Scientist Developer at Ørsted, Børkop, Denmark.

⁵Senior Data Scientist, H2O.ai, New York, NY, USA.

⁶Science The Earth; Mesa, AZ, USA.

Keywords: readability, science communication, science writing, Natural Language Processing

Abstract

Here we present a new text analysis tool that consists of a text analysis service and an author search service. These services were created by using or extending many existing Free and Open Source tools, including streamlit, requests, WordCloud, TextStat, and The Natural Language Tool Kit. The tool has the capability to retrieve journal hosting links and journal article content from APIs and journal hosting websites. Together, these services allow the user to review the complexity of a scientist's published work relative to other online-based text repositories. Rather than providing feedback as to the complexity of a single text as previous tools have done, the tool presented here shows the relative complexity across many texts from the same author, while also comparing the readability of the author's body of work to a variety of other scientific and lay text types. The goal of this work is to apply a more data-driven approach that provides established academic authors with statistical insights into their body of published peer reviewed work. By monitoring these readability metrics, scientists may be able to cater their writing to reach broader audiences, contributing to an improved global communication and understanding of complex topics.

Introduction

To ensure that writing is accessible to the general population, authors must consider the length of written text, as well as sentence structure, vocabulary, and other language features (Kutner et al., 2006). While popular magazines, newspapers, and other outlets purposefully cater language for a wide audience, there is a tendency for academic writing to use more complex, jargon-heavy language when publishing their work in scientific journals (Plavén-Sigra et al., 2017). This is a trend that is becoming more evident over time (Ball et al., 2017).

In the age of growing science communication, this tendency for scientists to use more complex language can carry over when writing in more mainstream media, such as blogs and social media (Yong, 2010). This can make public-facing material difficult to comprehend, undermining efforts to communicate scientific topics to the general public (Shulman et al., 2020). Information outflow from scientists that is not accessible can contribute to a general misunderstanding of scientific concepts and a disconnect from scientists (Schulman et al., 2020). To date over half (57%) of the world's population do not think they know much – if anything – about science. We speculate that this may be due to challenges when attempting to read and understand complex scientific topics.

To address the increasing difficulty in reading science, we created a data-driven tool to provide authors with insights into the readability of their published open access work with respect to a known collection of science documents called the ART corpus (Soldatova et al., 2007). An additional collection derived from the work “the Readability of Science is Declining over time,” has been shown to have complexity that is comparable to that of other scientific journals (Plavén-Sigra et al., 2017).

The tool we describe consists of a text analysis service and an author search service. These services were created by using or extending many existing Free and Open Source (FOS) tools, including streamlit, requests, WordCloud, TextStat, and The Natural Language Tool Kit (NLTK). The tool has the capability to retrieve journal hosting links and journal article content (both html and PDF) from APIs and journal hosting websites. Several Python libraries helped with querying and gaining access to open science scholarly research documents. The library python-requests were used to obtain content from three different APIs, including [dissemin](#), [semantic-scholar](#), and [unpaywall](#).

This large hierarchy of software dependencies might act as an impediment to reproducibility. To address this issue we did two things. First, we deployed a [live version](#) of the application. Second, we created a Docker file and associated Docker container that acts as a self-documenting software environment clone. All code used for this tool can be found on [GitHub](#), which can be run by downloading the github repository and

then either building a docker container or running it using the terminal (see below for details).

Reference Texts used for Analysis

By entering an author's name into the search box, a system of API calls and web resource requests collects and analyzes text written by that author from the dissemin API. Our tool contrasts the queried authors readability with expected science readability scores because it has pre-processed and quantified the existing corpus of publicly licenced scientific texts contained in ART Corpus - (creative commons license), as well as a number of available text repositories and common resources with varying complexity (Table 1).

Text Source	Mean Complexity	Description
Upgoer 5	7.0	library using only the 10,000 most commonly occurring English words
Readability of science declining	9.0	example of a scientific article discussing writing to a broad audience in an academic context
Science of writing	14.0	example of a scientific article discussing writing to a broad audience in an academic context
Wikipedia	14.9	free, popular, crowdsourced encyclopedia generated from self-nominating volunteers
Post-Modern Essay Generator	16.5	generates output consisting of sentences that obey the rules of written English, but without restraints on the semantic conceptual references
Art Corpus	18.68	library of scientific papers published in The Royal Society of Chemistry

Table 1. Description and Reading Level of a number of available text repositories and common resources with varying complexity

Running the Application

The application can be accessed and used without code using the url:

<https://share.streamlit.io/mcgurrgurr/scienceaccess/app.py>

Downloading the application from Github

```
git clone https://github.com/mcgurrgurr/ScienceAccess.git  
cd ScienceAccess
```

Building and running the container using Docker Ensure that you have Docker downloaded on your machine. Scraping is memory and CPU intensive. Default docker resources are not suitable for running the application and thus are manually defined in the code below.

```
docker build -t wcomplexity .  
docker run --shm-size=3gb --cpus=2.0 --memory=1g --memory-swap=1g --rm  
wcomplexity
```

This last command will print two hyperlinks to the terminal. Click on the hyperlink that does not contain the substring local host. This will take you to the application in a browser window. Running the application from the terminal

Manually download and install the Mozilla geckodriver binary.

```
which python3
```

Place output from above command in

```
sudo [output] setup.py install  
streamlit run app.py
```

By either build method, the result is a web interface that prompts the user to enter an author name (see Figure 1). After this, a text scraping process occurs.

Search Reading Complexity of an Author

Enter Author Name:

Entering a middle initial may improve search accuracy.

Fig 1. Interface for entering an author's name on the application.

Results

After querying an author, the readability of the author's work is evaluated by a variety of pre-established metrics. We used “text_standard,” a readability consensus score that returns the average estimated school grade level required to understand the text. Results from the dissem query include the number of documents on which the readability score was based. Each mined text is presented to the user in a table with an active link to the source text. This allows the user to verify that the results are correct and also provides individual readability scores for each text source.

Fig 2. Boxplot output showing the readability score of the searched author against the text repositories of the ART Corpus and Readability of Science Declining Over Time. .

The mined text is compared against the ART Corpus and the collection derived from “the Readability of Science is Declining over time” as a side-by-side boxplot of readability scores (see Fig 2.), as well as data points of each individual readability score. This allows the user to easily review the readability of their work with regard to these other sources of scientific text. Lastly, two word clouds are generated that show the most commonly used words and the largest words found in the mined text. These

results are presented first for the abstracts of the mined text only, and subsequently for the full texts.

Conclusions

Here, we outline a new tool to analyze the complexity of a scientist's work relative to other online-based text repositories. Rather than providing feedback as to the complexity of a single text as previous tools have done, the tool presented here shows the relative complexity across many texts from the same author, while also comparing the readability of the author's body of work to a variety of other text types.

The goal of this work, in contrast to previous work, is to apply a more data-driven approach to provide established academic authors with statistical insights into their body of published, peer reviewed work. This will expand upon current readability metrics by computing a more detailed and comparative look at the complexity of written text, which will allow scientists and other experts to use a data-driven approach to monitor the complexity of their writing relative to other text types. By monitoring these readability metrics, scientists may be able to cater their writing to reach broader audiences, contributing to improved global communication and understanding of complex topics.

The focus of this work on scientific texts is driven by a desire to increase public accessibility to scientific journal articles. However, with some modifications we hope that this tool can apply more broadly to other fields, including philosophy, politics, law, and art. We encourage others to use the linked source code to hypothesize, explore, and implement changes to improve the tool and cater it for other fields.

Scientific Application

This work belongs to an emerging scientific discipline that seeks to quantify and understand scientific writing. Readability metrics have improved in robustness since the publication of the document "the Readability of Science is Declining over Time" (Plavén-Sigra et al., 2017). The data set obtained from this resource used a custom implementation of the Flesch reading grade metric, which sometimes yielded negative results, and may have under-reported very high readability results. Given the more recent availability of more robust readability metrics, it is important to test and calibrate the newer readability metrics against current scientific documents. Text-stats standard obtains a readability metric by averaging over 9 different readability metrics. The presence of frequency word clouds, and large word word-clouds, and hard passages make it possible to check the text-stat metrics as applied to scientific documents. Word clouds act to validate higher readability metrics.

In general, other science text scraping tools might seek to achieve our results by scraping Google Scholar, an approach which almost usually leads to an un-gratifying experience for the app user, who is forced to solve captchas, in exchange for services. Furthermore, the google scholar robots.txt file prohibits scraping.

We present new and reusable methods for mining scientific literature by utilizing many free and newly available APIs dedicated to the responsible and free dissemination of open access science documents. The tool we described in this document is extensible such that code from various modules could be re-applied to ask more specific questions about readability of science documents.

Because a user can download results from the science accessibility app. The tool makes it possible to compare two different scientific authors in the same field, and to compare their readability metrics. A comparison between two or more authors of the same field could be used to clarify if some fields really are more difficult to understand, or if some authors in the same field do indeed write at a level that is more challenging to read.

Although we agree with the established conclusions from The Readability of Science is Declining Over Time (Plavén-Sigra et al., 2017), a complicating factor that needs to be addressed is that the readability metric of choice is also changing over time. It is therefore very important to understand readability assessments of science historically and generate reproducible code workflows. Although the code for the analysis the readability of science is decreasing over time is freely available on GitHub, the code can no longer be run due to dependencies on python packages that are no longer supported. We believe that our methodology will allow our platform to be reproducible, allowing future iterations to consider novel readability metrics in the future.

Limitations and Future Work

In the context of this work, there are two important caveats. Firstly, science writing consists of two types of complexity: irreducible complexity, and reducible complexity. Irreducible complexity in science writing stems from the fact that science is the act of communicating abstract and elaborate techniques that interrogate nature. Written instructions for how to repeat a science experiment should not be less complex than the experiments that the instructions encode. Reducible complexity in science writing is the not strictly necessary complexity in science writing, that comes from the introduction of difficult language (long jargon heavy sentences with many modifiers and written in the passive voice). This extra fraction of complex writing is added to texts by the science writers due to the nature of sharing experiment details. In this document we are only concerned with methods that help writers to identify reducible complexity in their writing.

It is very rare for a single scientist to be a sole author of a publication since teams of scientists often coauthor a publication. The data mining and NLP approach here involves querying an author's written work and analysing the co-authored documents that are discovered in this process. Therefore, the writing complexity discovered by this tool reflects the written complexity of the group of scientists in the whole co-authorship network. This tool assumes that the written contribution of the author can be approximated by averaging over multiple written documents, although we acknowledge that this approximation may decrease the accuracy of the reported readability score for an individual author.

By sharing the project source code we hope that other scientists or engineers will help to expand and improve this tool. Improvements include filtering of mined texts and increasing the content that returns from the text analysis, such as a more detailed breakdown of the returned scientific text. For example, it should be possible to provide highlighted snippets of the most difficult to read passages for a given author. One other consideration is the idea of calculating the readability score for groups of co-authors. Since scientific documents are often co-authored, the output metrics from the scrapper tool reflect the contributions from a network of co-authors rather than a single individual. Capturing the writing contributions of the author searched for and their closest co-authors may help to understand readability with a broader scope.

References

- [1] Kutner, Mark, Elizabeth Greenberg, and Justin Baer. "A First Look at the Literacy of America's Adults in the 21st Century. NCES 2006-470." The National Center for Education Statistics. (2006).

- [2] Plavén-Sigra, Pontus, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. "The readability of scientific texts is decreasing over time." *Elife*. (2017).

- [3] Ball, Philip. "It's not just you: science papers are getting harder to read." *Nature News* (2017).

- [4] Shulman, H. C., Dixon, G. N., Bullock, O. M., & Colón Amill, D. (2020). The Effects of Jargon on Processing Fluency, Self-Perceptions, and Scientific Engagement. *Journal of Language and Social Psychology* (2020).

[5] Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. "Derivation of new readability formulas (Automated Readability Index, FogCount and Flesch Reading Ease Formula) for Navy enlisted personnel".The Institute for Simulation and Training, (1975): 8–75.

[6] Soldatova, Larisa, and Maria Liakata. "An ontology methodology and cisp-the proposed core information about scientific papers." JISC Project Report (2007).

[7] Kuhn, Tobias. "The controlled natural language of randall munroe's thing explainer." International Workshop on Controlled Natural Language. Springer, Cham, (2016).

[8] Bulhak, Andrew C. "On the simulation of postmodernism and mental debility using recursive transition networks." Monash University Department of Computer Science (1996).

[9] Gopen, George D., and Judith A. Swan. "The science of scientific writing." American Scientist 78, no. 6 (1990): 550-558.

[10] Antonin Delpeuch; a3nm; Ryan Lahfa; translatewiki.net; Xia Li-yao; Steph. No.; Lucas Verney; Evarin; p4bl0; jibe-b; Stefan Beck; nemobis; Virgile Andreani; Martin Monperrus; Kenji Maillard; Jill-Jênn Vie; Golls, "Dissem Software" Zendo, 2019