Quantifying Scientific Jargon

S. D. Willoughby, K. Johnson, and L. Sterman

Willoughby, Quantifying scientific jargon, 'Public Understanding of Science' pp. 096366252093743. Copyright © 2020. DOI: 10.1177/0963662520937436.

DOI: 10.1177/0963662520937436.

Public Understanding of Science
XX(X):1–19
©The Author(s) 2020
DOI: 10.1177/0963662520937436

SAGE

Quantifying Scientific Jargon

S. D. Willoughby[1] K. Johnson[1] and L. Sterman[2]

Abstract

When scientists disseminate their work to the general public, excessive use of jargon should be avoided because if too much technical language is used, the message is not effectively conveyed. However, determining which words are jargon and how much jargon is too much is a difficult task, partly because it can be challenging to know which terms the general public knows, and partly that it can be challenging to ensure scientific accuracy while avoiding esoteric terminology. To help address this issue, we have written an R script that an author can use to quantify the amount of scientific jargon in any written piece and make appropriate edits based on the target audience.

Introduction

If scientists want to communicate their findings or their research to an audience outside their sub-discipline, there are a number of techniques for improving reader cognition. They should ensure that the message does not contain excessive jargon (Nation 2006),

[1] Department of Physics, Montana State University
[2] Renne Library, Montana State University

Corresponding author:
Shannon Willoughby, Department of Physics, 264 Barnard Hall Montana State University, Bozeman, MT 59717

Email: shannon.willoughby@montana.edu

includes a story arc to catch and keep the attention of the audience (Shanahan et al. 2019), and considers the cognitive load of the audience (Bullock et al. 2019).

It can be difficult to determine exactly which terms are jargon, develop a compelling narrative, and keep cognitive load in mind while communicating novel ideas. To address the first of these issues, we have written a freely available script in an open source language R. The script takes any written text and calculates how much scientific jargon it contains. This number, 'jargonness' can be compared to benchmarks to evaluate the appropriate level of jargon for the intended audience. It also outputs a list of words that are likely scientific jargon so that the writer can consider different word choices. Finally, the script can also be used to create word clouds, which can be used to visually evaluate word frequency and appropriateness.

This paper is organized as follows: Literature Review; Use of the Algorithm; Limitations; Results and Discussion; and Future Works. In the Appendix, the Methods section discusses jargonness calculations, and the verification of our calculations.

Literature Review

Scientists generally agree that "readability" is a key component for public comprehension of literature (DuBay 2004), though how to achieve that quality is the topic of continued debate. DuBay, in 2004, states that there are over 200 readability measures already and notes several factors that influence readability, including "format, features of organization and content".

While the factors of style have been debated for as long as communication has existed, we focus on one aspect of comprehension: the use of jargon in scientific communications. The level of general adult language literacy in the United States has been measured by the National Center for Education Statistics, to determine three distinct areas of literacy–the prose, document, and quantitative literacy of the average American. They found that about two thirds of Americans function at a basic or intermediate literacy level, where about 13 percent are considered proficient and around 20 percent are considered below basic (Woodworth 2003). Based on this assessment, research that is aimed at the general public should significantly differ from communications within a discipline. For clarity in how the communication should differ, it is useful to examine the tools of reading education. Additionally, research in readability measures is grounded in work with K-12 students.

Similar to the work of scientists seeking to clearly communicate research findings, educators seek to pair readers with text that they can understand to facilitate learning comprehension and engagement. While this may be apparent for second grade students, it is also true for readers of all ages. In one common measure, sentence length, percentage of familiar words, and word lengths within a text (Lexile Framework (Stenner 1999) are assessed to pair readers with texts that they can understand with at least 75% comprehension. While many readability formulas focus on set criteria, these measures are approximations and should be used as estimations rather than absolutes (Benjamin 2012). However, the structure of a text is one component that can greatly influence comprehension.

Research shows that laypeople or experts from other disciplines struggle with the format, structure and quantity of information in scientific communication (Cook and Jacobs 2014; Bromme et al. 1999; Scharrer et al. 2017), yet the communication of science to the public is critical for public support of science and the adoption of current, evidence-based technologies and practices. When a scientist is preparing a speech or publication, it is prudent to consider their audience so as to align the structure and vocabulary of the text with the targeted group.

A clear example of the necessity of communicating research findings to a general audience can be found in the healthcare decision making process (Park 2010). In a face to face communication, clinicians are able to measure factors including whether the medical staff "talked in terms I could understand", "checked in to be sure I understood everything", and "encouraged me to ask questions" (Makoula et al. 2007). These interactions highlight the importance of scientific understanding when making medical decisions, as understanding medical options is crucial to consent, and represents an advantage of oral versus written communication (Stableford and Mettger 2007): medical staff are able to have a conversation and adjust terminology to reduce or define jargon as they communicate.

In contrast, for written or static communications, authors must work to predict the comprehension of readers. In 2006, Nation (Nation 2006) suggested that if more than 2% of the words used in a text are jargon, then the message becomes difficult for a lay audience to understand. Bullock *et al.* conclude that excessive amounts of jargon lead to excessive cognitive load on the part of the reader. As a result, readers are more likely to think that advances in science may be more risky than someone who read about the same technology in a piece that uses less jargon (Bullock et al. 2019). However, determining what is and is not jargon can be challenging (Rakedzon et al. 2017), and determining the effect of jargon on comprehension can also be difficult (Woodward-Kron 2008).

One issue with Nation's 2% suggestion is the fact that not everyone agrees on which words count as jargon, and which words are thought to be generally understood. Additionally, some attempts to decrease jargon have been shown to decrease comprehension (Davison and Kantor 1982), as the style of a piece becomes more complex and thus harder to follow. Additionally, some words may be understood by themselves, but when used in a phrase, become jargon. In their experiment, Bullock *et al.* (Bullock et al. 2019) used two opposing paragraphs to test metacognition with variable jargon. Analyzing the paragraphs used in this study, we found that the jargonness of the jargon-filled paragraphs is indeed higher than that of the jargon-free paragraphs. Although the words that our script determined were jargon very nearly overlaps with the words that Bullock *et al.* saw as jargon, an interesting difference is that our code does not pick up phrases, so where Bullock *et al.*registered ”layer filaments”, our tool only flagged ”filaments” as jargon (and ”super-microsurgery”, flagging just ”microsurgery”). As discussed below, phrases can be picked out by the script, but the user must specify exact phrases for the code to flag.

Removing jargon can be problematic to scientists who are interested in effectively communicating their research to the public in a way that is understandable, while also being accurate. Some argue that jargon is necessary for clarity, concision and authority (Wilkinson 1992), and that a lack of jargon makes it easy for non-experts to draw conclusions based on prior knowledge instead of the current research (Scharrer et al. 2017). While some hurdles remain even when removing jargon, these readability measures can be helpful reminders to authors to check their work for usability within their target audience (Woodward-Kron 2008).

Further, there is more to understanding information than just comprehending jargon. Shanahan *et al.* (Shanahan et al. 2019) determine that narrative also plays a part in the understanding of risk communication. When participants listened to narratives about flood risks, they reported a more positive affective response when there was a clear ”victim to hero” narrative, and that the response was less positive when actors were portrayed as merely ”heroes” or ”victims” with no narrative arc. They concluded that story arc matters when conveying scientific information to the public, and that when presented with just scientific information, participants had a smaller overall affect than when presented with the same information framed in a narrative structure. Still, a reduction in jargon can improve metacognition and processing fluency (Bullock et al. 2019), and possibly improve a person's understanding of narratives and, thus, of the information presented. One way reduce jargon is to employ the tools of computational linguistics.

In 2014, Sharon and Baram-Tsabari (Sharon and Baram-Tsabari 2014), developed an equation that can quantify the amount of jargon in a given transcript. Naming their equation ‘jargonness’, Sharon and colleagues suggest this is an improvement over Nation's proposed 2% guideline, in part because it is a mathematical description of how much jargon is in a given transcript or paper. Sharon *et al.* chose to develop a logarithmic scale using the assumption that a word that is truly scientific jargon would be approximately 1,000 (or 10^3) times more likely to appear in the scientific corpus (or collection of words) than in the contemporary English corpus. Based on this assumption, the authors calculated the jargonness for each word and the average jargonness for the whole file. To create benchmarks for comparisons, they calculated the jargonness for TED Talks, a CERN press conference, and transcripts from the Michigan Corpus of Academic Spoken English. The authors used a British English corpus, so this code is not readily usable by American scientists. Further, the code is proprietary and thus difficult to evaluate or modify.

In 2017, Rakedzon *et al.* (Rakedzon et al. 2017) created the De-jargonizer, a tool that scrapes word data and categorizes words as "high frequency, mid-frequency, and jargon" and color codes words for easy author spotting and evaluation. Through three validation steps the De-jargonizer discovered that lay abstracts, even in their attempts to decrease jargon for a broad audience, remain above a recommended jargon level. This tool uses word frequency in the British National Corpus and the Corpus of Contemporary American English to determine the three levels of frequency and to alert authors to the frequency, and thus likelihood of comprehension, of each of their words.

By using the traditional division of words and their close derivatives, or families, into high frequency (1,000–3,000-word families), a mid-frequency group (the 3000-9000-word family level), and low frequency (above 9,000-word family level), the De-jargonizer provides a micro-level analysis of a text. Being familiar with about 5,000 words enables the comprehension of movies, newspapers, and conversations (Nation 2006). A reader who is comfortable with high frequency words and familiar with most mid-frequency words will be able to comprehend the majority of written English.

Building off of the guidelines presented by Nation, the jargonness equation presented by Sharon *et al.*, and the concept of frequency from the De-jargonizer (Rakedzon et al. 2017), we present a freely available script written in the open source language R to calculate the average scientific jargonness per word for American transcripts

(Willoughby and Johnson 2019). Nation's guideline is helpful, but does not point out which words are jargon. Sharon's approach does point out those words, but uses a British corpus for contemporary words. Rakedzon *et al.* create a map of the text word by word, though they do not provide an overall understanding of a text. The proposed model uses an open source American English Corpus and a scientific corpus of journal articles published by Elsevier. ArXiv articles could also be used to create the scientific corpus. A new series of benchmarks give general guidelines based on our analysis of these corpora for lay person, general scientific, and disciplinary specific audiences. The R script produces a list of words that have the largest jargonness values, to be flagged for consideration by the author. This tool both assesses a document as a whole and provides word level assessment for a comprehensive understanding of jargon in a document. The script is flexible and can be modified to compare documents against only works in their sub-discipline, so one can check against common jargon in their field specifically. Finally, a word cloud is created so the author can see which words are most commonly used in their text, and consider alternatives based on the target audience.

Utilizing the algorithm

Based on the work of Sharon and Baram-Tsabari (Sharon and Baram-Tsabari 2014), jargonness in this study is calculated on a base 10 log scale. For example, if the word "orbit" had a jargonness value of 2.3, it would be approximately 200 times more likely ($10^{2.3}$) to occur in scientific literature than in a newspaper article. In this study, jargonness is calculated for each word, then the mean jargonness per word is calculated for the work as a whole. This allows for a direct numerical comparison of the jargonness of multiple works. We calculated the jargonness of several types of documents, texts, and transcripts to directly compare works that are mostly free of scientific jargon (classic texts, popular podcasts), to works that are more likely to contain scientific jargon (materials safety data sheets, ArXiv manuscripts). Further details can be found in the Appendix.

For the interested user, the algorithm can be employed as follows:

1. (If needed.) Install R and R Studio, available at https://www.r-project.org/
2. Download the zip file included with the manuscript which contains the R script, pre-built corpora, and sample text files.
3. Uncompress the zip file, and place input text file into that folder.
4. Open R script and follow directions therein.
5. Output files will include:
    - ◇ A calculated value of total document jargonness (j), which can be compared to bench marked files. (See Figure 1.)
    - ◇ List of words that are likely scientific jargon.
    - ◇ A word cloud, showing the 75 most frequently used words in the input text. (See Figure 2.)

These files can be used to guide edits to the document to decrease the scientific jargon. The author can compare the average j value to the bench marked documents. For example, considering a document intended to communicate with the general public, Figure 1 indicates that a jargonness value under 0.10, preferably under 0.05, is ideal.

All words with a jargon value of 3 are also output. A j of 3 indicates that the word is not present at all in the English corpus and is scientific jargon. These words should be addressed and either replaced or defined. Authors should consider the reader's cognitive load, and reduce this jargon to aide (Bullock et al. 2019) metacongition. An example of the list of words was created from Chemistry NSF abstracts, and includes the following terms: catalyzed, spectroscopy, reactivity, ligand, marisa, mhz, eludication, atoms, quaternary, biopolymer, and xenon. Note that an author would still need to judge which terms truly are jargon, and which are not.

Finally, the R script creates a word-cloud containing the 75 most frequently used words found in the original document. This word-cloud is generated after removal of stop words from the input text file: this prevents very common words (such as 'a', 'the', 'and', etc.) from appearing so the word-cloud is useful as a frequency chart *and* as a tool for assessing jargon and word choice. The number of words created in the word cloud can be adjusted, which may be helpful for very short or very long files. An example of a word cloud can be found in figure 2. Inspection of this word cloud shows what words are used the most in this collection of published journal articles, allowing the author to define of modify certain terms when writing about this topic for a lay audience. Terms such as tensor and modified scalar will most likely need to be defined, and heavily used acronyms such as gr and qc can be fully spelled out.

This code may be used as a tool to provide flexible analysis of documents and assist authors in preparing their work for the general public. While it is not a solution to every problem of communication, the reduction of jargon can remove important barriers to understanding.

One practical application of this project could be to evaluate effective science communication skills of students. For example, in a separate study conducted by several of the authors of this paper, this jargonness code was used to evaluate the intervention in a year long NSF funded fellowship program to improve their oral communication skills within science. To gauge the fellows' reduction in jargon, each initial applicant (n=19) submitted a recording that summarized their thesis work and a transcript of that recording. We calculated the jargonness of all applications as a baseline, see Figure 1. After the year-long intervention, the fellows' utilization of jargon will be reassessed and compared to the original measure. This allows for a data-based and reproducable evaluation of the reduction of jargon in their communication.

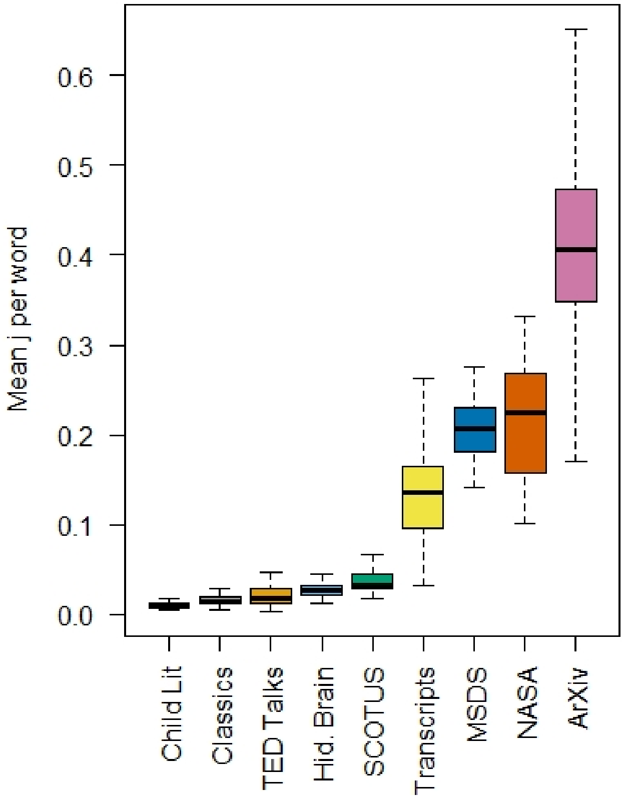
Another application of this algorithm is the assessment of stated communication objectives. We tested batches of 3,000 National Science Foundation grant abstracts of currently funded projects, based on their various grant directorates. The NSF website states, "NSF is committed to writing documents in understandable language and launched an initiative in late 2013 to make the description of our awards more transparent to all stakeholders, including the public, by improving the clarity of the titles and abstracts of awards." (NSF 2020). The amount of jargon in these abstracts was significantly higher than in any other bodies of work we examined. The values for the NSF abstracts are listed in Table 1. They were omitted from Figure 1 because they skewed the graph significantly upward. Practically, the NSF or grantees could evaluate their abstracts using this method to assist in fulfilling their stated objective of transparent communication.

Results and Discussion

Examination of Table 1 suggests an ideal range of jargonness for a given text. To present scientific information to children, content should have an average jargonness value of less than 0.015, while communications to the general public should aim for values less than 0.030. Individuals with an undergraduate-level knowledge of a scientific topic will likely comprehend information with an average jargonness value of up to 0.300, while graduate-level knowledge in the topic may extend comprehension up to 0.500. It is important to emphasize that these jargonness per word ranges are a suggestion based off of this work. Further, these numbers arise from data analyzed in the exact manner as described in this study. Should one wish to change how text is analyzed by editing our code (e.g. removing symbols) it would be best to also recalculate all document-based statistics.

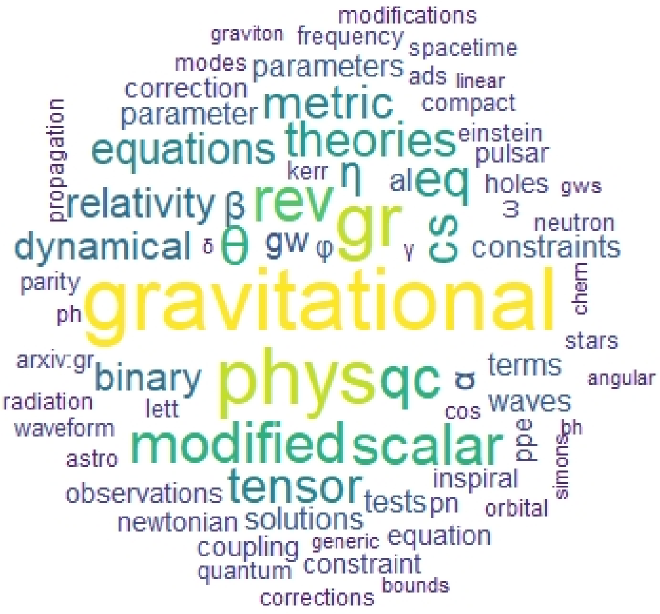
Jargonness by each benchmark group are represented in Figure 1. This box and whisker plot in shows clear differences in the means of each type of document. For example, a manuscript uploaded to ArXiv has a very high mean value for jargonness. NASA E-books, Materials Safety Data Sheets, and transcripts of STEM graduate students represent a mid-level amount of jargonness. Items that are geared toward the general public: classic texts, TED Talks, Hidden Brain episodes, etc. have a lower mean jargonness. Although Supreme Court opinions are toward the low end of jargonness levels, our corpus was developed to evaluate scientific jargon, not legal jargon and an appropriate corpus could be used with this script to calculate legal jargon. These three clusters suggest that pieces written for different target audiences do indeed differ in the mean amount of jargon present. This tool benefits authors who wish to evaluate

Figure 1. Boxplots for all benchmarked documents



their work for excessive scientific jargon (Table 1), allowing a writer to consider their audience explicitly when editing a piece. It is still up to the writer to consider flow, format, and style, though this tool may help with the goal of accurately reporting scientific findings to the public by reducing scientific jargon.

Figure 2. Word Cloud for a theoretical physicist's most highly cited journal articles



Limitations

This code calculates the mean scientific jargonness per word within a text as intended, but has limitations. Below, we discuss symbols, domains, encultration, context, and filtering.

Table 1. Minimum, mean, maximum, and standard errors for jargonness values in benchmark documents

Document	Minimum	Mean	Maximum	S.E.
Children's Literature	0.005	0.009	0.017	0.001
Classic Texts	0.006	0.019	0.064	0.003
TED Talks	0.003	0.022	0.059	0.002
Hidden Brain	0.013	0.029	0.073	0.003
SCOTUS Opinions	0.019	0.040	0.076	0.003
Transcripts	0.032	0.161	0.294	0.017
NASA E-Books	0.101	0.178	0.332	0.013
MSD Sheets	0.141	0.211	0.275	0.007
ArXiv Papers 9910	0.0716	0.425	1.034	0.005
NSF Abstracts	0.00	0.147	3	0.601

Symbols. Fundamentally, the authors aimed to write code that could handle any written English text. Unfortunately, symbols can be difficult to render from .pdf format to .txt format. For example, "cid" represents a catch-all for a symbol encoding issue when rendering the .pdf files that form our scientific corpus. "Cid" seems to be associated only with particular symbols, such as Ω or σ^2, but also may result from symbol representation issues from within the original pdfs themselves, as a result "cid" was removed from our analysis. Not all symbols fail to import, e.g. γ, β, π, etc. For those not wishing to use only some symbols, symbols may be filtered out of the corpus directly by using the command 'remove_symbols' when creating the document frequency matrix.

Domains. Our scientific corpus was drawn from four unique domains of science, meaning that in its current state it does not delineate jargon from one domain or another (e.g. "physical", "health", "physical", or "social" sciences). However, this may be overcome. When the corpus is generated it is created from each subject domain separately, in a step-wise fashion, allowing a conservative researcher to build a scientific corpus that is more specific to their domain. This approach should inflate the jargonness values of scientific words specific to their domain, and allow for a more detailed critique of their work.

Enculturation. As researchers are trained within a discipline, they are encouraged to use specific terms and measures in their communications. The practice of using scientific jargon asserts authority within a discipline, yet when communicating with a lay audience that same use of jargon may reduce comprehension (Anagnostou and Weir 2006). Disciplinary vocabulary or specialist language can often be seen as a measure of learning and comprehension (Mahon 2014), driving authors to use jargon to convey authority within their discipline. Thus, there may be resistance to alter communications, especially to remove jargon, as those lay person versions convey less scientific authority. The use of jargon can be an efficient and effective means of communication, yet the same use of jargon that is clear and concise within a discipline can be detrimental to the comprehension and engagement of a non-disciplinary trained audience.

Context. A major limitation of this work is variability of denotations based on word context. When measuring certain words, such as “vacuum,” this script does not determine the context in which the word is employed, a limitation shared with Raekdzon *et al.* (Rakedzon et al. 2017). Rakedzon *et al.* note that although their tool selects for jargon, it does not avoid cases where the same word might have specialized or jargon denotation in one context, and a separate lay understanding of the same word (Rakedzon et al. 2017). For “vacuum” it is unknown whether the word refers to a household appliance or a volume of space which does not contain any matter, without proper context. There is no automated remedy to this issue at the time of writing. One could examine correlations between the counts of discipline-specific words in a document set to assess if a particular word belongs within that discipline and should be considered jargon. For example, in a document that contains the word “vacuum” as well as the words “space” and “star”, “vacuum” probably refers to a volume of space that is absent of matter, and is likely jargon.

Filtering. Some of the filtering criteria (symbol retention, hyphenation, punctuation removal, etc.) is largely subjective. However, researchers may determine for themselves what is and is not of value to their research and proceed accordingly.

Future Work

The research team plans to use this method to analyze and assess scientific work produced by our colleagues and students. One application will be the continuation of assessment for the STEM Storytellers fellowship program. With one year completed and two future cohorts, we will apply this script as one measure the effectiveness of our intervention. We will also continue gathering different types of transcripts in order to have a full range of jargonness represented in the bench marking documents. Additional documents will increase the nuance and accuracy of our work, allowing people using the code to better understand where their writing fits relative to their discipline (Table 1). Finally, the authors suggest that this script could be used by candidates to evaluate jargon in documents used for their tenure and promotion cases. These professional documents are read by a number of colleagues in other domains and career changing decisions are made based on them being compelling and understandable. This tool could be one step in a candidate's editing process to flag language that may diminish comprehension of the value of their work.

References

Anagnostou N and Weir G (2006) From corpus-based collocation frequencies to readability measure. *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006* : 1–14.

Benjamin RG (2012) Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24(1): 63–88.

Bromme R, Nuckels M and Rambow R (1999) Adaptivity and anticipation in expert-laypeople communication. *Psychological models of communication in collaborative systems. American Association for Artificial Intelligence Technical Report FS-99-03.* : 17–24.

Bullock O, Amill D and Shulman H (2019) Jargon as a barrier to effective science communication: evidence from metacognition. *Public Understanding of Science* 28(7): 845–853.

Cook J and Jacobs P (2014) Scientists are from mars, laypeople are from venus: An evidence-based rationale for communicating the consensus on climate. *Reports of the National Center for Science Education* (34): 1–10.

Davison A and Kantor R (1982) On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly* : 187–209.

DuBay WH (2004) *The Principles of Readability*.

Mahon BM (2014) Making the invisible visible: disciplinary literacy in secondary school classrooms. *Irish Educational Studies* 33(1): 21–36.

Makoula G, Krupat E and Chang CH (2007) Measuring patient views of physician communication skills: Development and testing of the communication assessment tool. *Patient Education and Counseling* 67: 333–342.

Nation I (2006) How large a vocabulary is needed for reading and listening? *Canandian Modern Language Review* 63(1): 59–82.

NSF (2020) Nsf clarity of titles and abstracts. https://www.nsf.gov/od/transparency/clarity-titles-abstracts.jsp.

Park EW (2010) Medical jargon used in health care communication of family physician. *Korean Journal of Family Medicine* 31(6): 453–460.

Rakedzon T, Segev E, Chapnik N, Yosef R and Baram-Tsabari A (2017) Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLoS ONE* 12(8): e0181742.

Scharrer L, Rupieper Y, Stadtler M and Broome R (2017) When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts. *Public Understanding of Science* 26(8): 1003–1018.

Shanahan E, Reinhold AM and Raile E (2019) Characters matter: How narratives shape affective responses to risk communication. *PLOS One* 14(12).

Sharon A and Baram-Tsabari AB (2014) Measuring mumbo jumbo: A preliminary quantification of the use of jargon in science communication. *Public Understanding of Science* 23(5): 528–546.

Stableford S and Mettger W (2007) Plain language: A strategic response to the health literacy challenge. *Journal of Public Health Policy* 28(1): 71–93.

Stenner AJ (1999) Instructional uses of the lexile framework : 1–6.

Wilkinson AM (1992) Jargon and the passive voice: Prescriptions and proscriptions for scientific writing. *Journal of Technical Writing and Communication* 22(3): 319–325.

Willoughby S and Johnson K (2019) R script to calculate jargon. `http://www.montana.edu/stemstorytellers`.

Woodward-Kron R (2008) More than just jargon – the nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes* 7(4): 234–249.

Woodworth J (2003) National assessment of adult literacy (naal). `https://nces.ed.gov/naal/naalhistory.asp`.

Public Understanding of Science
XX(X):1–19
(c)The Author(s) 2020
DOI: 10.1177/0963662520937436

SAGE

Quantifying Scientific Jargon

S. D. Willoughby[1] K. Johnson[1] and L. Sterman[2]

Contents

This Supplemental material is available at:
Willoughby, Shannon D.; Johnson, Keith; Sterman, Leila (2020): Appendix – Supplemental material for Quantifying scientific jargon. SAGE Journals. Dataset.
https://doi.org/10.25384/SAGE.12619501.v1

[1] Department of Physics, Montana State University
[2] Renne Library, Montana State University

Corresponding author:
Shannon Willoughby, Department of Physics, 264 Barnard Hall Montana State University, Bozeman, MT
Email: shannon.willoughby@montana.edu

Supplemental Material

Methods

Benchmarking Texts

In order to create a baseline for jargonness values, we downloaded a large number of transcripts and texts from a variety of sources. From the Gutenberg project, we downloaded 20 classic texts, including *Dracula*, *Alice in Wonderland*, *Pride and Prejudice*, and others (see full list below). All were chosen using random number generation from the list of the top 100 downloads in Project Gutenberg in June 2019. One author, a subject expert in Children's literature, complied 20 of the most widely known Children's books from Project Gutenberg. For popular podcasts, we obtained transcripts from most recent 20 episodes (as of June 2019) of *Hidden Brain*, a podcast which focuses on scientific topics and is intended for a general audience. All 95 Supreme Court decisions from 2010 were downloaded. Transcripts of the top 36 most-viewed TED talks were downloaded from their host website (Wurman 2019). For a sample of transcripts we suspect contain jargon, we downloaded 30 Material Safety Data Sheets. These were also chosen randomly. All science-ebooks available from NASA, 24 in total, were obtained from their host website (Dunbar 2019). Finally, we chose 540 ArXiv publications spanning from 1999 to 2019.

Texts used in bench marking:

Classic texts

1. A Christmas Carol
2. Alice in Wonderland
3. Being Ernest

4. Doll's House
5. Dracula
6. Frankenstein
7. Hard Times
8. The Heart of Darkness
9. Meditations
10. Moby Dick
11. A Modest Proposal
12. Northanger
13. Pride and Prejudice
14. The Hounds of Baskerville
15. The Illiad
16. The Prophet
17. The Call of the Wild
18. A Tale of Two Cities
19. Ulysses
20. War of the Worlds

Children's Literature

1. A Little Princess
2. Anne of Green Gables
3. Black Beauty
4. Box Car Children
5. A Christmas Carol

6. Fairy Tales of Hans Christian Anderson
7. Fairy Tales
8. The Adventures of Huckleberry Finn
9. Little Women
10. Through the Looking Glass
11. Mother Goose
12. Peter Rabbit
13. Poems Every Child should Know
14. The Secret Garden
15. The Jungle Book
16. The Red Fairy
17. Tom Sawyer
18. Treasure Island
19. The Wind in the Willows
20. The Wizard of Oz

Materials Safety Data Sheets

1. Yttrium
2. Acetolyne
3. Acetone
4. Atrazine
5. Benzene
6. Bisphenol- A
7. Bromine

8. Chlorine
9. Copper
10. Dioxins
11. Fenchone
12. Fluorine
13. Hydrogen gas
14. Iron
15. Lead
16. Litium
17. Magnesium
18. Mercury
19. Octanol
20. Perchlorate
21. PFE
22. Phthaltes
23. Sulfur
24. Silver iodide
25. Sulfur dioxide
26. S1 Nuclease
27. T282
28. Thiazol
29. Bonide Sulfur
30. Molten Sulfur

Calculating Jargonness

In order to calculate the jargonness of a text file, the following algorithm is employed:

1. Choose a set of texts to create a scientific corpus.
2. Choose a set of texts to create a contemporary American English corpus.
3. Create both corpora, with stop words, punctuation, numbers, and hyphens removed.
4. Prepare the text file for which the jargon is to be calculated, remove punctuation, numbers, and hyphens.
5. Determine how many times each word in the cleaned text file occurs in each corpus.
6. Calculate the jargonness value for each word.
7. Calculate the mean jargonness per word for the text as a whole.

The scientific corpus is generated from a collection of scientific journal articles. We chose a 5 year compilation of all journal articles published by Elsevier from year 2011 to year 2015 in the areas of the physical, life, health, and social sciences (Kwary 2018). These articles have been converted to plain text, with the references and figures removed. The full corpus consists of 5,686,428 words collected from 895 papers. This corpus can easily be updated or made larger by the inclusion of ArXiv manuscripts. Large volumes of these manuscripts can be downloaded free of charge, and either for specific domains, or by a range of dates they were posted.

The American English corpus utilized in this study, the Open American National Corpus (Ide 2019), is a freely available and continually growing corpus of contemporary English from 1990 on. The full corpus includes over 15 million words from travel guides, switchboard conversations, government websites, fiction, online news websites, and other audio and written sources. The English corpus serves an important role in the analysis because it is being used to represent words with which the average American is familiar. For sampling consistency, we use a randomly chosen subset of this corpus in our calculations so that the number of words in this corpus has the same order of magnitude as the words in the full scientific corpus. The subset of words was chosen such that we sampled across the word lists available, and includes switchboard conversations, works of fiction, articles from Slate, travel books, letters, 911 reports, and government reports.

In order to generate relevant information regarding each corpus (word counts/frequencies), we used a package in R called Quanteda (Welbers et al. 2017). This package was created by Kenneth Benoit in order to analyze texts within R using natural language processing. Text files of various formats may be read into Quanteda (.txt, .pdf, .doc, etc.) and filtered for undesirable content (removal of stop words, punctuation, numbers, and hyphens). Following the filtering process, a document frequency matrix (dfm) is generated, which contains the word counts for every word within the input text files. Both corpora were individually read into Quanteda and a dfm was generated for each, with the aforementioned filtering criteria applied. After this work was complete, the scientific corpus contained 1,868,785 words, and the American English corpus contained 1,460,998 words. Next, the relevant text file for which jargon is to be calculated is read in. Filtering criteria are again applied, and a dfm for that text file is generated.

The filtering criteria of the corpora included removing punctuation, numbers, hyphens, and stop words. As raw numbers (e.g. 1, 7.36, etc.) and punctuation do not necessarily contain any form of jargon, they were removed. Hyphenated words such as "self-contained" are broken into their respective segments (e.g. "self" and "contained") because we assume that interpretation of the compound word as a whole requires understanding of the components of that word. Stop words are very commonly used words that are typically removed before text analysis is performed, such as 'we', 'him', 'which', and 'this'. The removal of these words reduces the overall dimensionality of the text without losing the overall meaning. We assert that stop words should have a jargonness value of zero, and are thus filtered out of the scientific and English corpora.

The stop words removed from the analysis may be found within R using the command *stopwords (source = "smart")* with more information available on the Quanteda website (Welbers et al. 2017). We also use a second source of stop words to ensure that an exhaustive list of stop words are included (Beale 2016). Note that when calculating the jargonness of a particular input file (e.g. "Moby Dick"), all the above criteria are still employed except the filtering of stop words, since this would impact metrics such as the mean jargonness per word in the file of interest.

Using the list of words associated with the text file dfm, the frequency of occurrence for the ith unique word is obtained both within the text file (f_i) and within each corpus ($f_{i,sci}$ & $f_{i,eng}$). The formula is taken directly from (Sharon and Baram-Tsabari 2014). To calculate the jargonness value j_i for the ith word, and the jargonness J for the entire document, we use the following expressions:

$$j_i = \begin{cases} \log\left(\frac{f_{i,\text{sci}}}{f_{i,\text{eng}}}\right), & 0 < f_{i,\text{sci}} < f_{i,\text{eng}} \\ 3, & f_{i,\text{eng}} = 0 \text{ and } f_{i,\text{sci}} > 0 \end{cases}$$

$$J = \sum_i j_i \times f_i$$

J is then divided by the total number of unique words in the text file, which yields a mean jargonness per word for the file as a whole.

Verifying the code

We tested our R script on *Moby Dick* and *The Adventures of Huckleberry Finn*, comparing our findings against those of several online concordancers to further check that our algorithm was working correctly. Specifically, we checked that the count of certain words within each text were similar to other published values. The words were randomly chosen using the base R function *sample*. Prior to random selection the dfm was filtered such that only the top 100 words, sorted by their jargonness contribution (jargonness of the individual word times how often it occurs in the text) were considered for random sampling. This approach was taken because those words would be most influential when determining the average jargonness per word for the text, and hence of greatest interest to this study. Tables 1 and 2 display the values we obtained compared to those available through online resources for the aforementioned texts. This check provides a base calibration of our script.

Within Tables 1 and 2 we provide two values for bench marking purposes. The first value is the frequency of occurrence for that exact word. The second value is the frequency of occurrence for both that exact word as well as any set of characters which contain the exact word, also known as a wild card. For example, the second value would not only contain counts of the exact word “light”, but also “lightning”, “skylight”, “flighty”, etc. The concordancer website only provides counts for the exact word and hence contains only a single value.

Inspection of the tables reveals that not all sources agree on the number of specific terms in the texts. Certainly there is a correct answer, given parameters as noted above, and our script finds numbers that are similar to the numbers reported by other sources. The largest discrepancy noted in the compared words comes from “mast” within Table 1, where our code extracts 28 exact occurrences of “mast” compared to the first website’s 224. This difference results primarily from our script separating hyphenated words and the website’s choice to retain them, given that when our script is run without separating hyphenated words it is found that there are 224 occurrences of “mast”, in agreement with the website. We do not know why the second website seems to miss 99 entries of the word mast, but it is likely due to words that contain mast. Additional word-by-word comparisons of the hyphenated word forms and their counts further supported this. The same behavior is seen with “oil”, as our script run without separating hyphenated words revealed 78 occurrences of “oil”, which is also in exact agreement with the website. This analysis provides an easily replicable test of the word count functionality and flexibility of our script.

The jargonness calculations and primary script used in the body of this paper were created and performed by one of the authors. Independently of this, another author generated their own methodology for calculating jargonness, as defined by this study. The only information shared between these two authors were the files from which both the scientific and English corpora were developed, the Quanteda framework, and

Table 1. Counts for various words within *Moby Dick*, as determined by the jargonness code (WJS) and sources found elsewhere. In the WJS column, the leftmost value is a count for the word only (e.g. mast), the accompanying value is for both the word and various other forms (e.g. mastless).

word	WJS script	Website (Pearce 2019)	Concordancer (Matsuoka 2003)
vessel	46, 76	81	51
species	31, 31	31	30
oil	78, 212	212	86
mast	28, 224	224	125
interval	25, 57	57	25
sperm	239, 270	270	237

Table 2. Tests of code versus word counts found elsewhere for *The Adventures of Huckleberry Finn*. The leftmost value is a count for the word only (i.e. light), the accompanying value is for the word and a wild card (*light*).

word	WJS script	Website (Li 2016)	Concordancer (Matsuoka 2003)
current	22, 23	21, 22	21
light	50, 130	50, 130	52
doan	30, 30	34, 35	35
tin	25, 327	26, 249	25
den	26, 99	27, 96	28

the filtering criteria utilized within Quanteda; the purpose of independent coding was to ensure the methodology was robust and consistent between coders. The details for this second approach to calculating jargonness were notably different, with the general algorithm as follows:

1. Choose a set of texts to create a scientific corpus.
2. Choose a set of texts to create a contemporary American English corpus.
3. Create both corpora, with stop words, punctuation, numbers, and hyphens removed.
4. Create a file containing the jargonness value for each word in the scientific corpus. Words not in the scientific corpus have, by definition (see Section), a jargonness value of 0.

5. Prepare the text file whose jargon is to be calculated, removing punctuation, numbers, and hyphens.
6. From the jargonness file, retrieve the jargonness value for each word within the text file.
7. Calculate the mean jargonness per word for the text as a whole.

For all calculations done within this paper, it was found that both methodologies and associated scripts yielded the same values.

References

Beale A (2016) List of stop words. http://wordlist.aspell.net/12dicts/. Accessed: 2019-08-02.

Dunbar B (2019) Nasa science e-books. https://www.nasa.gov/connect/ebooks/science_ebooks_archive_1.html. Accessed: 2019-05-21.

Ide N (2019) Open american corpus. http://www.anc.org/.

Kwary DA (2018) A corpus and concordance of academic journal articles. *Data in Brief* 16: 94–100.

Li S (2016) Mark twain concordancer. https://susanli2016.github.io/Mark-Twain-Novels/.

Matsuoka M (2003) Concordancer for classic texts. http://victorian-studies.net/concordance.

Pearce A (2019) Moby dick concordancer. https://roadtolarissa.com/whalewords/.

Sharon A and Baram-Tsabari AB (2014) Measuring mumbo jumbo: A preliminary quantification of the use of jargon in science communication. *Public Understanding of Science* 23(5): 528–546.

Welbers K, Atteveldt V and Benoit K (2017) Text analysis in r. *Communication Methods and Measures* 11(4): 245–265.

Wurman R (2019) Most viewed ted talks. https://www.ted.com/talks?language=en&page=1&sort=popular. Accessed: 2019-06-10.