Periodical Simplicity: Briefly examining modern-day newspapers' readability, sentence lengths, and word occurrence

By: Ayah Aboelela
& Gregory Minckler

May 19, 2019

Introduction

The written works published most frequently and widely at a given time and place impart many insights to the language of that environment. We choose to explore tangible aspects of the language found in three popular newspapers of the East coast of the United States from the last five years numerically, not to gain only a technical analysis but to have some characterization of the material found in today's typical written news of slightly varying styles and complexity. What we learn will be helpful both in its own right and contrasted against basic facts about typical spoken language. Some linguists might argue that newspapers do not contain a true reflection of an environment's language because of the drastic differences between the naturally spoken word and the meticulously planned details and subtleties of written words; we acknowledge the disparity and aim to use such differences as an analytical tool. Still, we focus on description and not inference, because of the limited scope of our data.

Data

While some aspects of written language are impossible to analyze numerically without great expenditure of time or expert knowledge, we have tried to choose as many useful focal points as possible. For analysis we took the Boston Globe, Washington Post, and the New York Times, and chose from them 12 articles from 2015 spread as evenly as possible highlighting international, general domestic, sports, and business news and the Lifestyle and Arts and Entertainment sections. We also took 12 editorial articles from each (from the same year). We split our 72 samples into categories based on the publisher and on whether it is an article or editorial. Our 5 main categories are: Boston Globe, New York Times, Washington Post, All Articles, and All Editorials.

We noted the length of words and of sentences in each article and took their means among articles of different newspapers and among the two genres: news and editorials. With both word and sentence length we also took the variances and contrasted the results similarly over certain collections. We took note of words that appeared in articles and not in others to find any related potential generalizations (contrast findings). We found any correlations that became obvious about the links between these findings. For a more general and meaningful overview of the complexity of the language used, we used an already-developed algorithm associated with the famous Flesch Readability Scale on all of the articles and contrasted our findings with what is known about everyday speech.

Note: We speak about 'editorials' and 'articles' as being two separate categories, even though editorials are a type of article; within the body of our analysis it is assumed that the two classifications are separate from one another.

Variables to be analyzed in our paper:

- Publisher (NYT, WP, or BG)
- Paper type (article or editorial)

- Flesch Readability Scores
- Vocabulary difficulty level
- Average frequency of most frequently occurring words
- Words that appear more frequently in some categories than others

Questions and Work Division

Ayah:

1. What is the average readability score for all articles and editorials?
2. Do the Flesch scores in our sample make up a normal distribution?
3. Is there a difference in readability between different categories?
   A. Articles VS Editorials
   B. Washington Post VS New York Times VS Boston Globe
4. Is there a correlation between vocabulary range and Flesch scores of our articles and editorials?

- Other tasks: making the graphs in R; using Python to compute Flesch scores, vocabulary ranges, sentence lengths, and word frequencies.

Greg:

5. Is there a relationship between the mean and variance of sentence lengths, and between these and the Flesch scores?
6. What is the average frequency of the most frequently used words in articles and editorials of each newspaper?
7. What words show up in the editorials and not in articles (and vice versa) and in one newspaper and not in others?

- Other tasks: writing the introduction, conclusion, and contributing to the written analyses

Methods

Background: The Flesch Reading Ease Score

The first part of the project pertains specifically to the Flesch Reading Ease scores of each of our articles. To give a little background, the Flesch Reading Ease score was first introduced in the Journal of Applied Psychology in 1948 by Rudolph Flesch. Flesch was a strong promoter of the Plain English Movement, which initially called for simpler and more accessible legal/governmental documents that can be read by average people and not just high degree holders. Eventually, his readability score became widely used in US government agencies.

The following is the Flesch Reading Ease Formula:

$$RE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Where RE = Reading Ease, ASL = average sentence length, and ASW = average number of syllables per word. RE would be a score between 0 and 100; the higher the score, the easier a text is to read.

How the Flesch Reading Ease Score was calculated

To calculate the Flesch Reading Ease Score, we first needed to calculate the ASL (average sentence length) and ASW (number of syllables per word).

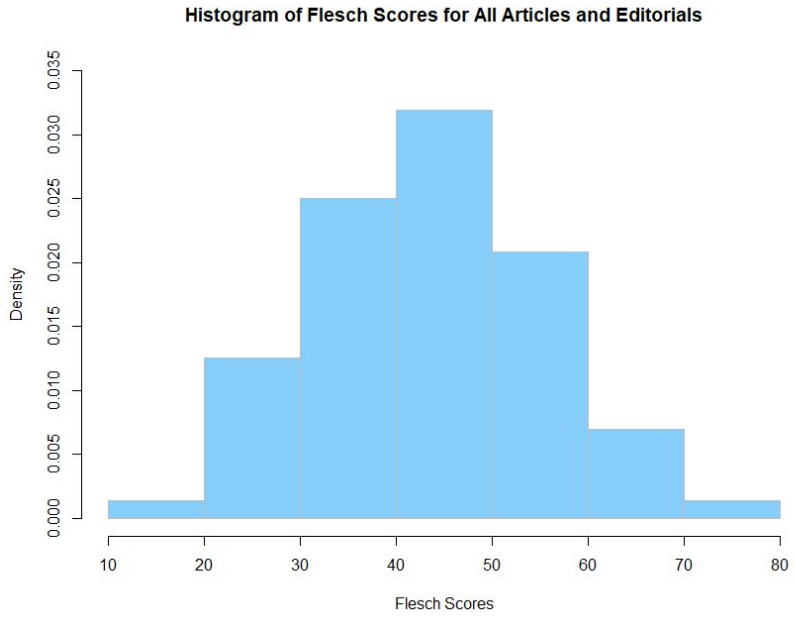
Our definition of a sentence for this project is a group of words split by a period--so colons and semicolons are not considered to be sentence delimiters in our project. To calculate ASL, we used an open-source Python function (cited below) that split a given input text into sentences, taking periods for websites and in words such as “Mr.” into consideration. We then ran each of our articles through this function, and split each of the sentences into individual words (separated by spaces) to count the sentence length.

Calculating ASW was a little tricker, since the English language does not have a systematic/mathematical rule to the number of syllables per word. For example, the letter “y” may or may not be a vowel, and sometimes pairs of vowels make 2 syllables, other times the make just one (as in the words “Leo” and “read”). We also used an open-source function to calculate the number of syllables (with some minor changes). While the code that we currently use is not 100% accurate, it is correct for most cases and considered sufficient for the Flesch Reading Ease Score.

Vocabulary Level

To calculate each paper’s vocabulary difficulty, we used an online tool developed by Mark Davies, a linguistics professor at Brigham Young University. The tool categorizes each word in a given paper into one of three possible ranges: high-frequency, medium-frequency, and low-frequency word range. The range is determined by the frequency it appears in a corpus--so words like “the” would be in the high-frequency range since it appears so often, while words like “comprehensive” would be in the low-frequency range. The tool uses data the Corpus of Contemporary American English, so the word frequencies we consider are specifically related to modern-day American English. After categorizing the words, the tool gives each paper three percentages: the percentage of words in the article that are in the high-frequency range, as well as the percentages of words in the medium- and low-frequency ranges.

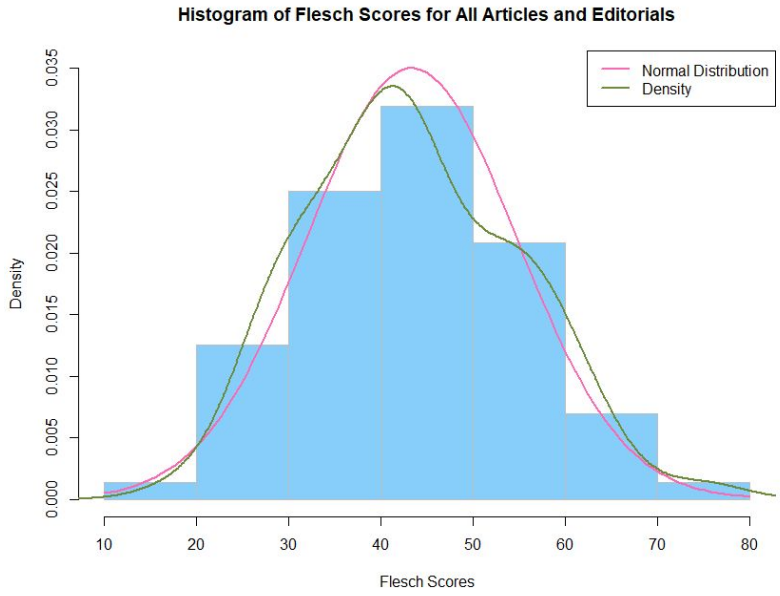
1. What is the average readability score for all the articles and editorials?



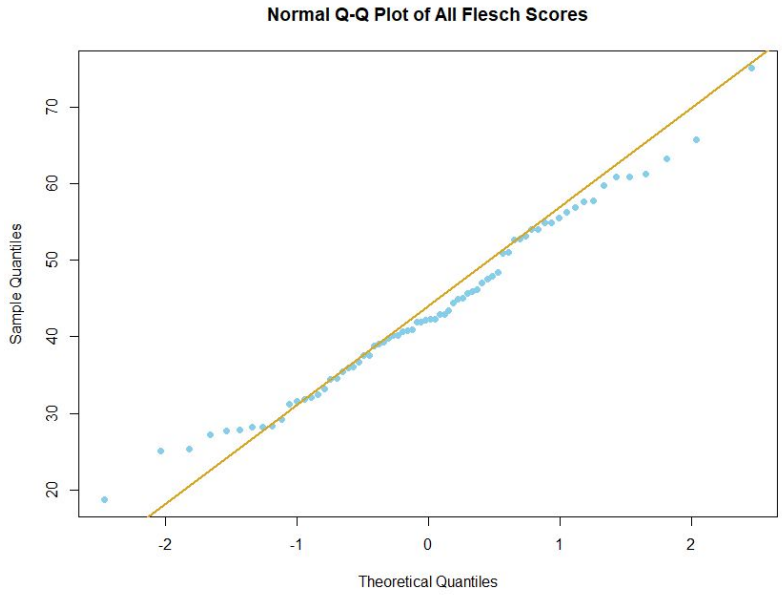
This is a histogram showing the probability density of the Flesch scores for all 72 of our articles and editorials. The average Flesch Readability Score of all 72 papers is approximately 43.3. Based on current convention of the Flesch score, this means that most of our papers (articles and editorials from all three newspapers) are on the slightly more difficult side and can be understood by 11th-12th graders or college students.

2. Do the Flesch scores in our sample make up a normal distribution?

Here, I am asking how close the distribution of our Flesch scores are to a normal distribution. To answer this, I made two graphs: a histogram with lines showing what a normal distribution would look like in comparison with our actual density function, and a normal Q-Q plot to show the difference between the theoretical ("normal") quantiles and the empirical ones.



The green line represents the distribution of our empirical data, and the pink line represents what our Flesch scores would look like if they were normally distributed. Judging visually, we see that the two lines somewhat follow the same general pattern, although there are clear differences. The density at the mean value of 43.3 is only slightly higher than the maximum (.035 vs. .033). So we can say that the distribution of Flesch scores is not fully normal although it is close to it.

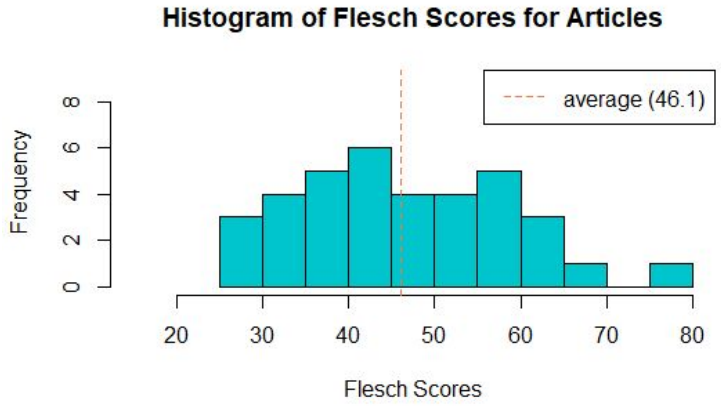


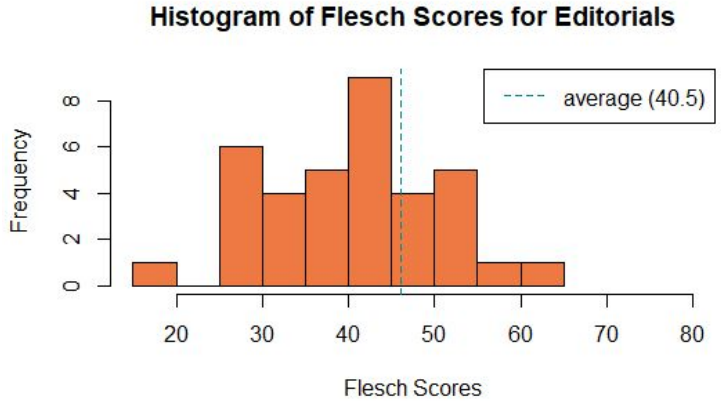
The theoretical quantiles represented on the x-axis are the quantiles for normal distribution. So the 0th quantile represents the mean. The corresponding vertical-axis value for x=0 is a little above 40, close to our mean (43.3).

The blue dots represent the actual flesch scores for each of our 72 articles. The orange line spans from the second to the third quantiles of our data, which is what our sample points would look like if they were normally distributed. Visually, we can see that the line and our empirical data seem to be close, so we can say that the distribution of Flesch scores for our articles are almost normally distributed.

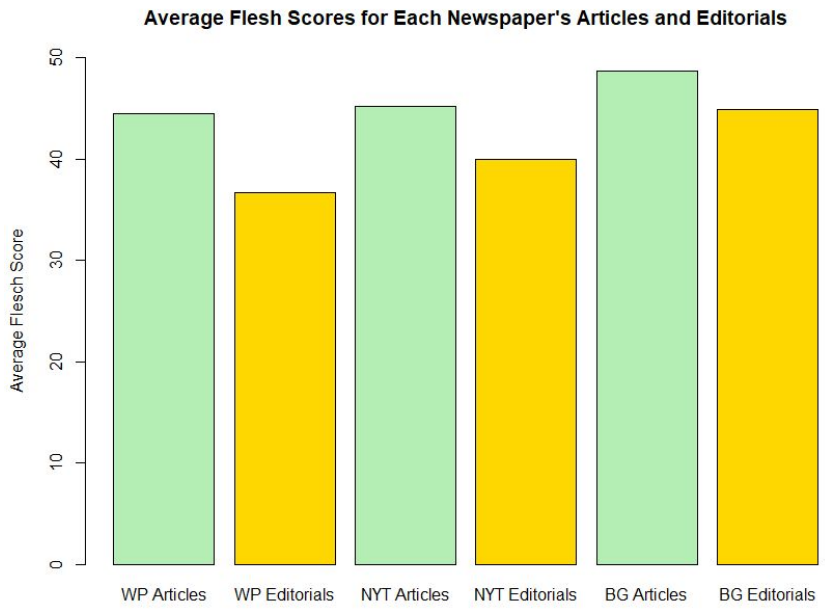
3. Is there a difference in readability between different categories?

3A. Is there a difference in readability between editorials and articles?





Our findings suggest a clear pattern: in general, articles are more readable than editorials. Our histograms above compare editorials with articles, not considering individual newspapers. Scores of less than 25 (and therefore more difficult to read) are found only in editorials, whereas scores higher than 65 correspond only to articles. The overlap is significant; the common region, again with scores between 25 and 65, accounts for a majority of the articles. Otherwise, articles dominate the 55-65 Flesch score region whereas the only other notable ‘jumps’ for editorials occur in the 25-30 and 40-45 ranges.

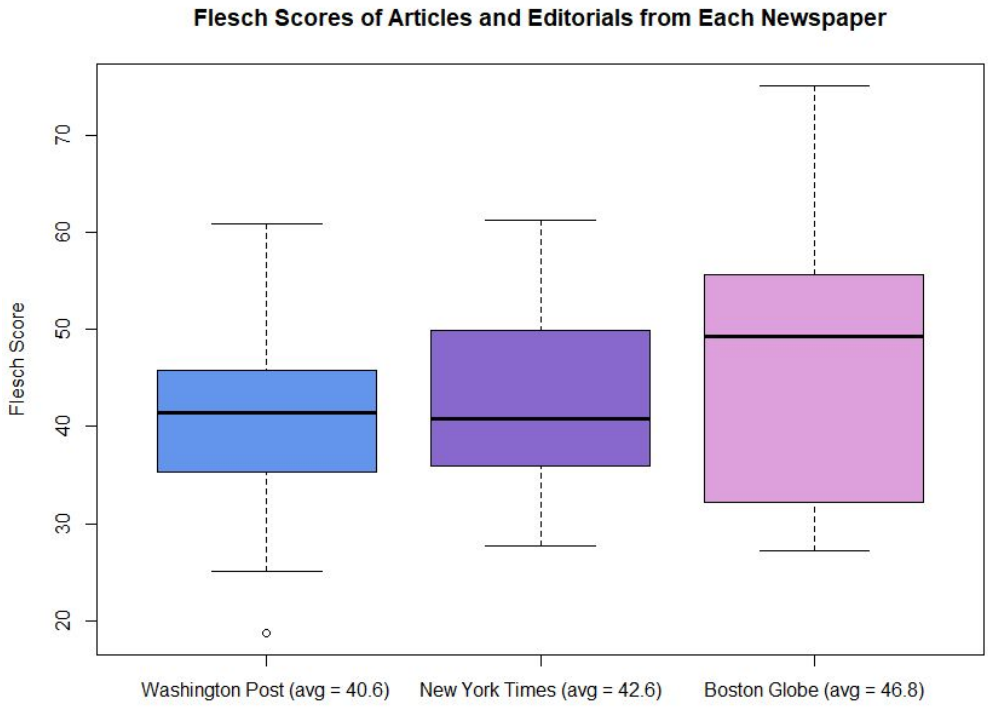


The above bar chart considers only the means of Flesch scores for articles (green) and editorials (yellow) of the three newspapers. This displays that the Washington Post has the hardest readability, and the Boston Globe has the easiest, with the New York Times in between. As we observed earlier, the articles are generally more readable than editorials. With respect to the mean Flesch scores of the Post and the Globe, which respectively score lowest and highest in both articles and editorials, the difference in articles is smaller than the difference in editorials, at 4 and 8 points respectively. This agrees with our earlier observation that the readability scores of editorials are spread somewhat more erratically than those of the general articles. This configuration gives that the highest variation of means between articles and editorials is found in the Post, followed by the Times and the Globe, respectively.

Why is it that, based on the Flesch Scores, editorials are harder to read than articles both generally and within individual publications? We know that editorials state the opinions of the newspaper's editors and are therefore biased, whereas articles are supposed to be objective reports of an event. Perhaps to more clearly state their opinions, the editors use longer sentences and more compound words when writing the editorials (both of which would contribute to a higher Flesch score). Maybe articles use shorter sentences and have an overall higher readability score because their authors have different goals than editorials' authors do--instead of detailing the view of an entire newspaper or publisher, the authors of articles (correspondents) might aim to reach as many readers as possible. If this is the case, then it would make sense that correspondents aim to make their articles as easily accessible as possible while still providing the information needed about the subject.

Does this indicate that the more opinionated a paper is, the more difficult it is to read? We would need to have a mathematical definition of "opinionated" to answer this, as well as use a much broader sample size. For this, perhaps sentiment analysis would be useful--something we did not look into for this project.

3B. Is there a difference in readability between the three different newspapers?



Without considering differences between articles and editorials we use a traditional box plot to contrast the readability of each of the three newspapers, with the Washington Post represented on the left in blue, the New York Times in the middle in purple, and the Boston Globe on the right in pink. The median Flesch scores of the Washington Post and New York Times are similar, whereas that of the Boston Globe is notably higher, by about eight points. The Post's median is similar to its mean, the Times' median is below its mean by about two points, and the Globe's median exceeds its mean by approximately two points. The first quartiles of the Post and the Times are also similar but the Globe's first quartile lies a few points below that of the other two newspapers, indicating a less regular distribution among the Globe's articles and editorials.

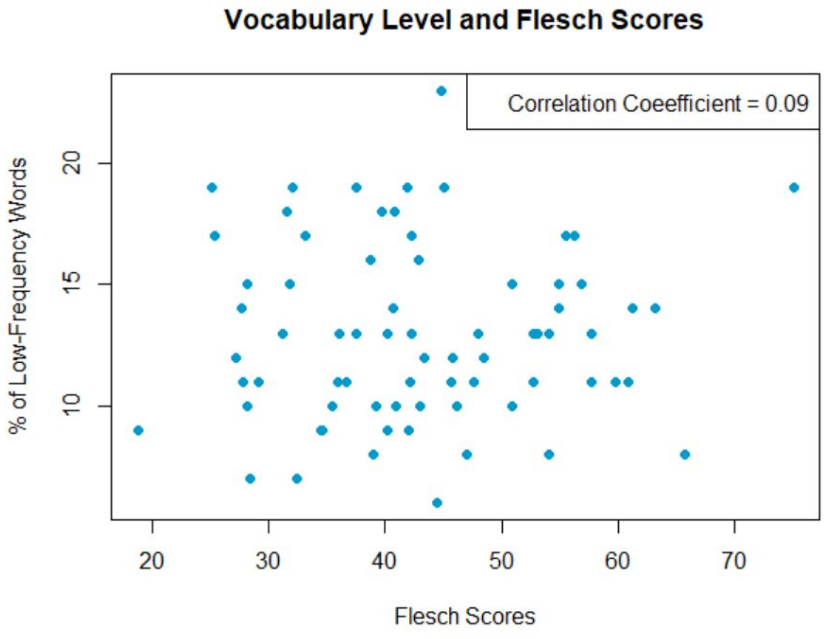
Between each other, the third quartiles of the Post, the Times, and the Globe are relatively evenly spaced at about 46, 50, and 55, respectively. Apart from an outlying article with a low readability score of approximately 19, the Post has its lower 'whisker,' representing the minimum, only slightly below that of the Times and the Globe, whose minima are similar; all three minima lie within the "very difficult to read" range by Flesch standards. With the upper whiskers/maxima, the Post and Times yield similar values of around 60 but the Globe has a maximum score about 15 points higher, at about 75, in the "fairly easy to read" range.

Overall, the score of the Washington Post and the New York Times are most similar and the Boston Globe has a wider range both between its whiskers and between its first and third quartiles than the other two papers. The standard deviation of the Flesch scores of all articles and editorials is about 2.5, with dispersions seeming to be the lowest in the Washington Post and greatest in the Boston Globe, based on the boxplot.

Based on the averages, the Boston Globe is the most easily readable of the newspapers by the highest. It is important to note that the New York Times is the only of the three newspapers that delivers papers all over the US instead of limited to a regional area. The Boston Globe and the Washington Post, in contrast, target a very local audience--BG only delivers papers to addresses in Massachusetts and neighboring zip codes while WP delivers to Washington DC, Virginia, and neighboring zip codes. Would this be the reason for the disparity in readability? While we would need a bigger sample size to say this with more certainty, examining the education levels on a regional (Massachusetts vs D.C.) and national may indicate differences in readability between the newspapers.
Another reason for the disparity may be the ages of the audience. Based on the most updated statistics we could find, the Washington Post has the lowest percentage of readers between 18-34 years old, while the New York Times has the highest (People-Press). Since BG has the highest readability, we cannot say that there is a correlation between overall readability and percentage of younger readers, although there is a correlation with WP both having the lowest readability and the lowest percentage of younger readers.

4. Is there a correlation between vocabulary difficulty and Flesch scores of our articles and editorials?

In other words, we are asking whether there is a recognizable correlation in an article or editorial between the vocabulary level of difficulty used and its readability score. The Flesch Readability Formula is not dependent upon richness of vocabulary but we are nevertheless curious about whether writing that uses a more difficult vocabulary will usually score lower on the Flesch scale.
As mentioned in the Methods section, we used an online tool to compute the percentages of easy (high-frequency), medium, and hard (low-frequency) words in each article and editorial. Since, for this question, we are interested in vocabulary difficulty, we plotted the percentage of low-frequency words in each paper against the paper's Flesch score to see if there is a correlation.

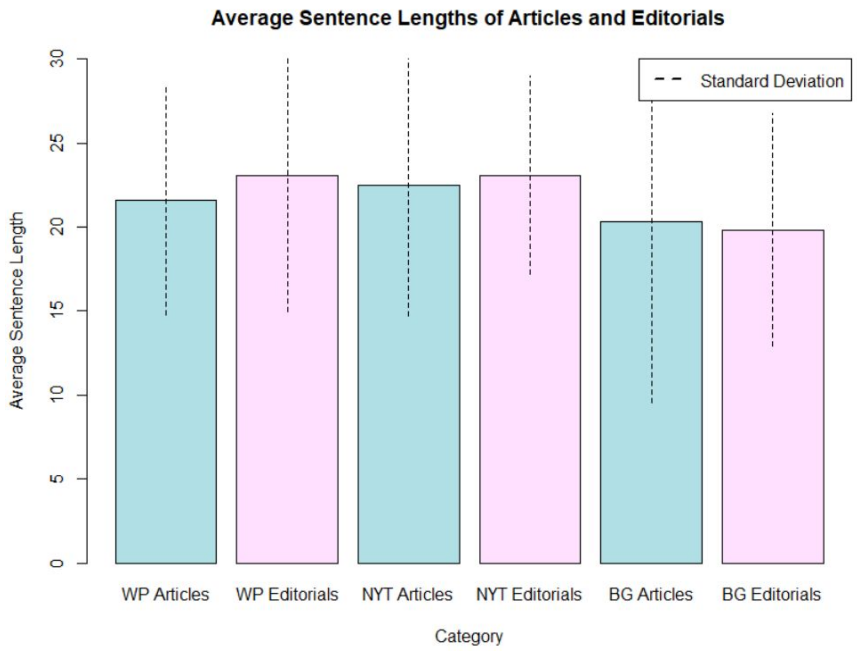


The computed correlation coefficient between vocabulary difficulty and Flesch score is approximately 0.09, so there is very little detectable linear dependence between the two variables. There is also little visible correlation based on the scatterplot. This means that, at least within our explored genre of newspaper writing, complexities are not linked regarding how broadly the vocabulary is spread in a piece of writing and how difficult it is to read according to the Flesch readability formula.

If two variables are independent, then their correlation coefficient is 0. It is important to note, however, that the converse is not always true. Therefore, we cannot say that the vocabulary level and Flesch score are independent variables, although we can say that there is little to no linear dependence between the two.

This shows one of the limitations of the Flesch Readability Score--it does not consider the use of common and uncommon words when calculating the reading-ease of a text. An uncommon word with fewer syllables may be harder to read than a common word with more syllables. As a very small example, the word “because” is very common (it is in the high-frequency range) and therefore may be more easily recognizable by readers, whereas the word “heed” (in the low-frequency range) is less common and may require more effort to understand although it only has one syllable. So reading-ease may not be determined solely by sentence length and number of syllables. Comprehensibility (and therefore vocabulary difficulty) may also play a huge role in how easily a text is read. The Flesch Score does not take comprehension or word difficulty into account, as reflected by our results. Therefore, we suggest that if a newer readability score formula were to be calculated, it should take vocabulary level into consideration. This would make the formula significantly more complicated to compute, but it would be more reflective of reading-comprehension as opposed to just the ability of sounding out words.

5. Is there a relationship between the mean and variance of sentence lengths, and between these and the Flesch scores?



We decided the standard deviation is more useful in analyzing the variation of sentence lengths than the variance itself since it can be compared with the mean values more directly. Our bar chart shows the standard deviations as dotted lines both above and below the top of each column (so the length of each dotted line is twice the standard deviation), whereas the columns represent the sentence length means for all categories.

Considering only the two categories of editorials vs. articles (with all three newspapers constant), sentence lengths are nearly the same on average, at just under 22 words. The lengths of sentences in articles varies more than in editorials, at 4.41 vs. 3.83 words, respectively.
Only considering the three newspapers while keeping the categories of articles vs. editorials constant, we see that the New York Times and Washington Post have a similar average sentence length, about 2.5 words longer than that of the Boston Globe. This pattern is inverse to the standard deviations of the three papers: the Post and Times have similar respective standard deviations of 3.81 and 3.46, while the Globe's value is higher, at 4.53, so the Globe's sentences are longer and more varied (in length) than the others.

Considering all categories as variables, there are more extreme variation levels to be found. Again, the Post and Times are similar in that they contain longer sentences in editorials than articles on average, as opposed to the Globe, but this is where patterns seem to disappear. Out of each newspaper, the standard deviation of sentence length is greater as follows: in editorials, by 0.61 words, for the Post; in articles, by 0.97 words, for the Times; and in articles, by 1.92 words, for the Globe. Clearly, given that a reader picks a random sentence from an article in the Globe, the length of a sentence is less predictable than from one of the other five categories.

As mentioned above, sentence length is a variable used to determine a written work's Flesch score, so we ignore that dependent relationship and focus on potential relationships between the standard deviations and Flesch scores. If we only contrast articles against editorials, we see that the standard deviation in sentence length is positively correlated with readability score; having a well-rounded variety of shorter and longer sentences is associated with more readable material (by the Flesch algorithm). No such generalization holds when considering our other data. Contrasting the three newspapers against each other while holding the editorials/articles category constant, the overall mean Flesch scores of the Post, the Times, and the Globe are 40.6, 42.6, and 46.8 in that order, corresponding to sentence length standard deviations of 3.81, 3.46, and 4.53. When considering the articles vs. editorials of the newspapers, there is a positive correlation between the standard deviations of sentence lengths and Flesch score of the Times and Globe, but not with the Post, and this counterexample is enough to discredit any such hypothesis.

6. What is the average frequency of the most frequently used words in articles and editorials of each newspaper?

Not unexpectedly, the most frequently occurring words within all of the articles and editorials make up common (linguistic) articles, prepositions, conjunctions, pronouns, and verbs. The first ten are as

follows: “the,” “to,” “and,” “of,” “a,” “in,” “that,” “for,” “is,” and “on,” in order starting with the most frequent. Of the 55,005 total words in our data set (including repetition), these ten account for 22.7% of the word count, with individual respective percentage figures of 5.9, 2.8, 2.7, 2.6, 2.5, 2.2, 1.3, 1.0, 0.9, and 0.8. Considering only unique words (disregarding repetition), these ten make up less than 0.1% of the total count. These ten overall frequently occurring words all appear within the top 15 within each of the six categories, and a significant drop-off of comparability occurs after (approximately) the first 30 frequent words in each.

Newspaper	year/years	country/ countries	public	health	more	law/laws
Wash. Post	77	23	8	12	65	19
N.Y. Times	66	20	12	9	41	25
Bost. Globe	58	17	45	22	67	10

Newspaper	rule/rules	system/ systems	power/powers	person/people	other/others
Wash. Post	4	6	6	43	36
N.Y. Times	17	19	13	23	48
Bost. Globe	10	14	12	35	31

We selected the eleven words in the above tables from the data set to contrast their frequencies within each newspaper. Listed are their overall frequencies within each if the three papers; these numbers are a close indication to their relative frequencies because the number of total words in each paper is comparable: the count is 17,745, 17,538, and 19,722 for the Washington Post, the New York Times, and the Boston Globe, respectively. For applicable nouns we combined figures for the singular and plural forms. Our choice was based on two factors: ‘non-triviality’ and non-specialty. We tried to select words that were not insubstantial (in that they were not simply articles or pronouns, for example); this factor is ultimately subjective but we believe it necessary to get towards insightful generalizations. From an opposing view, we wanted to avoid choosing words that were too specialized to the subject matter being discussed in the writing. Weaknesses with this ‘parameter’ include the difficulty judging and determining in how many realms the specialization lies and to what degree we should consider a word neutral enough with regard to the subject.

It is slightly haphazard to assume that the articles we chose constitute a perfectly representative sample of the writing found in each newspaper, because the numbers need to be greater to offset the specialization of writings. We will still make this general assumption warily. Our chosen word regarding time was “year,” which shows up relatively evenly throughout the three newspapers. Words like

"country," An approximately even distribution occurs with "country," a word describing a generally large scope, whereas "public," referring to a smaller, more local, scale occurs much more in the Globe than in the others. "Health," perhaps a more 'holistic' than business-related term, also occurs more in the Globe. "Law" occurs less frequently in the Globe than in the others, which could be linked to a less prevalent preoccupation with legal issues in its writings. "Rule" and "power" occur less frequently in the Post than in the others, which could be linked with less of a focus on authority in the Post's material. The Times comes highest in its use of "power," "system," "rule," and "law," which might be connected to a priority of legal and authority-related aspects. These are admittedly still rough, unsatisfactorily explored ideas.

7. What words show up in the editorials and not in articles (and vice versa) and in one newspaper and not in others?

We chose not to include data on specific words that occur in only one category (such as in articles but not editorials, or in the Boston Globe but not in the other papers) because there are so many and it would take more advanced means to determine patterns. The raw numbers are still important. The number of words that occur in articles and not in editorials is greater than its counterpart: 4,563 as opposed to 2,625. Apart from the vocabulary common to both categories, it is understandable that articles would contain more words particular to their subject matter. On the other hand, one might expect a wider use of vocabulary in editorials. In any case, there is a positive correlation between these figures and that of the higher Flesch readability scores in articles. The number of words that show up only in the Post, only in the Times, or only in the Globe are 2,188, 1,989, and 2,514, respectively. Considering the median Flesch scores, this preserves the same positive correlation between the breadth of vocabulary and readability. Potential reasons for this would indeed require further investigation.

Conclusion for report

Our findings are clear at the surface level; we know what deeper and more intricate subtopics we could explore if a future investigation were to be pursued more extensively. In general, editorials seem quantifiably more difficult to read than articles of other types and the Boston Globe is easier to read than the New York Times and the Washington Post. Within the newspapers we examined, it seems there is no discernible relationship between richness of vocabulary used and readability. Nor could we link the readability and variance of sentence lengths within articles or editorials. With the words that are exclusive to one category, all categories are nearly identical starting with the most common words but dissimilarities occur steadily. Additionally, we noted some of the flaws of using the Flesch score to calculate reading ease--one of the reasons being that it does not take more difficult (here, defined as less common) words into account, which is something significant to those learning English as a second language.

Our methods for both selecting words from a small data set and characterizing the findings were imperfect, but the size of the data set used and use of more intricate linguistic methods are large factors. Our selection method for non-editorial articles was stratified; from two given dates (January 1 and July

1) we chose 12 articles from each newspaper divided evenly into 6 subpopulations based on topic: international, general domestics, business, and sports news, and ‘arts and entertainment’ and ‘lifestyle’ writings. We chose the first articles that appeared in each category with at least 500 words based upon the classification of the database when it was available, and using the titles (and if necessary, the actual content) when pre-classification was not present. For the editorials, we chose the first samples that appeared in each newspaper with at least 450 words from each month from each paper.

The descriptive statistical methods we used mostly made numerical use of mean, variance, standard deviation, and quantiles to generalize about the data distributions.
Our project looked only at a narrow scope of data. The conclusions we reached regarding both the existence and non-existence of certain correlations lead naturally to questions whose exploration would require a much greater data set. Here are just a few obvious extensions:

- What are the patterns of readability in all newspapers throughout the United States?
- What are better ways to measure the readability of an English text than the Flesch algorithm, which only considers sentence length and number of syllables per word? (Consider also the ‘commonness’ of words, number of subordinate phrases/clauses per sentence, and the number of letters per syllable among.)
- Which variables most greatly affect differences in newspapers’ complexity of language/readability (region, specific practices of a publisher imposed on the writers regarding intended audience, etc)?

Work Cited

Boston Globe. ProQuest,

https://search-proquest-com.ezproxy.lib.umb.edu/publication/46045/citation/58BF901EA708412CPQ/3?accountid=28932. Accessed 24 April 2019.

Brown, Kevin, and Jadzia626. "Detecting Syllables in a Word." *Stack Overflow*, 3 Sept. 2018, stackoverflow.com/questions/405161/detecting-syllables-in-a-word.

Davies, Mark. "Words and Phrases." *Brigham Young University.*

https://www.wordandphrase.info/analyzeText.asp

"Difference Between Editorial and Article." *Difference Between.* 20 June 2012,

https://www.differencebetween.com/difference-between-editorial-and-vs-article/

"Find the k Most Frequent Words." *Geeks for Geeks*, 10 Dec. 2017,

www.geeksforgeeks.org/find-k-frequent-words-data-set-python/.

Greenberg, D, and Artyom. "Python Split Text on Sentences." *Stack Overflow*, 19 July 2015,

stackoverflow.com/questions/4576077/python-split-text-on-sentences.

"Home Delivery." *The New York Times*, The New York Times,

www.nytimes.com/subscription/hd/1041.

New York Times. *Academic OneFile*,

http://link.galegroup.com.ezproxy.lib.umb.edu/apps/pub/2NYT/AONE?u=mlin_b_umass&sid=AONE. Accessed 24 April 2019.

Sarkar, Tirthajyoti. "Very Simple Python Script for Extracting Most Common Words from a Story." *Towards Data Science*, 23 Sept. 2017, towardsdatascience.com/very-simple-python-script-for-extracting-most-common-words-from-a-story-1e3570d0b9d0.

"Section 4: Demographics and Political Views of News Audiences." *Pew Research Center for the People and the Press*, Pew Research Center for the People and the Press, 18 Sept. 2018, www.people-press.org/2012/09/27/section-4-demographics-and-political-views-of-news-audience.

"The Boston Globe." *Subscribe*, subscribe.bostonglobe.com/B8319/.

"The Flesch Reading Ease Readability Formula." *Readability Formulas*, www.readabilityformulas.com/flesch-reading-ease-readability-formula.php.

Washington Post. *Academic OneFile*, http://link.galegroup.com.ezproxy.lib.umb.edu/apps/pub/2PST/AONE?u=mlin_b_umass&sid=AONE. Accessed 24 April 2019.

"Washington Post Circulation and Delivery." *The Washington Post*, WP Company, live.washingtonpost.com/circulation-05052010.html.