Scientific Research Publishing

Advances in Historical Studies, 2022, 11, 188-195
https://www.scirp.org/journal/ahs
ISSN Online: 2327-0446
ISSN Print: 2327-0438

War and Money in Ngram Viewer

Robert H. McFadden[1,2], William H. Zywiak[3], Ronald P. Bobroff[4], Gao Niu[3]

[1]Finance Department, Bryant University, Smithfield, USA
[2]Veterans Services Office, Community College of RI, Lincoln, USA
[3]Mathematics and Economics Department, Bryant University, Smithfield, USA
[4]History, Literature, Arts, and Cultural Studies Department, Bryant University, Smithfield, USA
Email: rmcfadden@bryant.edu

How to cite this paper: McFadden, R. H., Zywiak, W. H., Bobroff, R. P., & Niu, G. (2022). War and Money in Ngram Viewer. *Advances in Historical Studies, 11,* 188-195.
https://doi.org/10.4236/ahs.2022.114016

Received: August 26, 2022
Accepted: October 29, 2022
Published: November 1, 2022

Abstract

The second and fourth authors have been inviting Intro to Applied Analytics and Statistics 1 students to use the Ngram Database to explore historical topics of their choosing. This is the first article derived from this exercise. The first author examined the historical relationship between war and money from 1775 to 2005 in the American English corpus. This is followed by an examination of the 3-gram "cost of war" in the American English and British English corpora. Specific to the analyses presented here several military and economic events are discussed. More specifically, both economies and wars are somewhat unpredictable, with wars being more so. Through this exercise, more experience with statistics and a greater appreciation of history are achieved.

Keywords

Ngrams, War, Money, Cost of War, Statistics

1. Introduction

The ngram database allows statistics and analytics students to test for statistical relationships among terms interesting to them. This is in contrast to giving the student a database and telling them what variables to conduct what analyses on. It is hoped that analyses on the ngram database lead to a renewed interest in history. The second and fourth authors have been having students in statistics and applied analytics courses use the ngram database using terms of their choice for the last few semesters. This allows students to examine constructs that are specifically of interest to them. This is the first paper we have written based on a student's work. In this case, the first author: a finance major graduating in 2023.

Aiden and Michel (2013) introduced the world to Google's Ngram Viewer through a best-selling book and a Science article (Michel et al., 2011). The ngram database includes words and phrases, up to 5 words, from books created between 1500 and 2019. These words and phrases are anchored to year of publication. The ngram software plots the frequency of a given word across a selected time span. The ngram software includes a smoothing function which we recommend setting to zero, so the actual pattern of the data can be observed (the default setting is smoothing in 3-year segments).

In statistics and analytics courses, we ask students to examine ngram patterns. If the plots of two ngrams intersect or look like they will eventually intersect (see Figure 1) the student is invited to use the intersecting regression line method, which will be described to a greater extent later in the Methods section. In contrast, if two plots share the same slope (typically positive), and one increases before the other, the student is invited to conduct lagged correlations. An example of lagged correlations conducted after noticing that the increase in "jazz" preceded the increase in "freedom" is presented in Figure 2.

For his Ngram Project, the first author researched the relationship between "war" and "money" in the ngram database. As an Iraq War veteran, he has often reflected on the reason why countries go to war. Knowing that money is a driver of many global decisions, he thought that researching if there were any statistical,

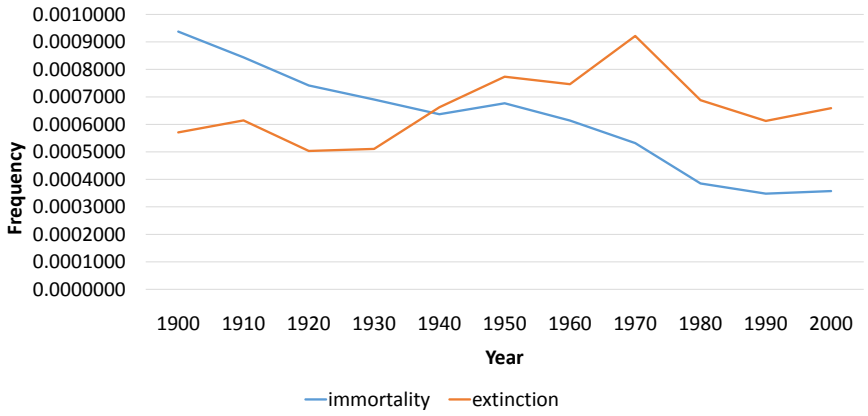

Figure 1. Plots from Data Spreadsheet (immortality and extinction).

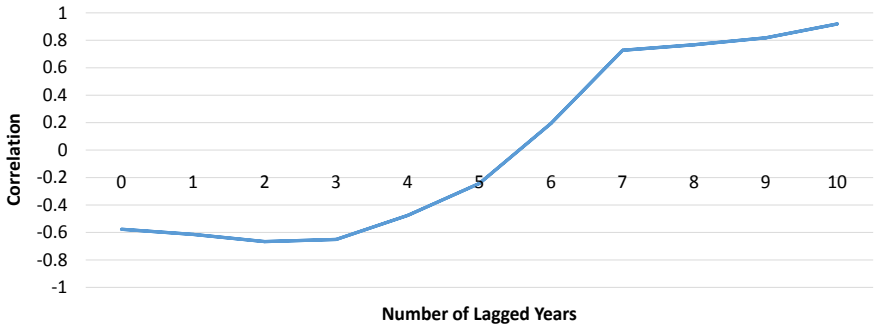

Figure 2. Correlation between "jazz" and "freedom" based on lag in years.

literary, trends between war and money would be interesting. The economy has been an antecedent of several wars. Examples include, "taxation without representation" was an American colonist rallying cry, and plantation owners did not want to give up slave labor during the Civil War. French peasants suffered from starvation prior to the French Revolution (Hunt & Censer, 2017). Germany suffered extreme unemployment during the 1930s, prior to WW2.

The rationale for this paper is underscored by curriculum development at Bryant University. For the Class of 2027 and thereafter all juniors will take a Junior Capstone class which will focus on one of the 17 UN Sustainability Goals while partnering with an external community (broadly defined). Financing is helpful in achieving these goals, and war is antithetical to many of these goals. More specifically, financing is helpful in achieving no poverty, zero hunger, good health and well-being, and quality education. War and the displacement of persons that often ensues are antithetical to no poverty, zero hunger, good health and well-being, clean water and sanitation, sustainable cities and communities, life on land, and peace. We understand that war is nuanced. Wars exist on a continuum of justified wars and wars that are less justified, and we posit that when oppressed persons (or allies of oppressed persons) win a war the results may be more positive (e.g., the American Revolution and the US Civil War) than wars which are based on an acquisition of natural resources and/or land [the American Indian Wars (Echo-Hawk, 2010)].

The purpose of this paper is to provide an example of the intersecting regression line approach in the ngram data, to interpret these results, to interpret the descriptive statistics of the terms searched, to highlight how valence is important in ngram analyses, to present an approach on comparing frequencies in two corpora, and to interpret the results based on the 3-gram of "cost of war".

2. Method

When conducting the intersecting regression line approach students are instructed to use at least 11 data points per term. These frequencies are multiplied by 10^6 to reduce rounding errors in the calculation of the slope and intercept by SAS Enterprise Guide. In the first regression line, year was used to predict the frequency of war ($\times 10^6$). In the second regression line, year was used to predict the frequency of money ($\times 10^6$). The student then sets these to regression lines equal to each other, and uses a calculator to solve when (what year) the values are equal: when the lines intersect. If this month and year have already occurred, the student can google what happened in that month and year (e.g., what happened in December 1943) and report events that resonate with the two terms selected. More specifically, the first author examined the frequencies of war and money, in the American English 2019 corpus, from 1785 through 2005 in 20-year intervals (i.e., 12 data points). The span of 220 years was chosen to see how these two words have interacted with each other throughout our country's history. This raw data is plotted in **Figure 3**.

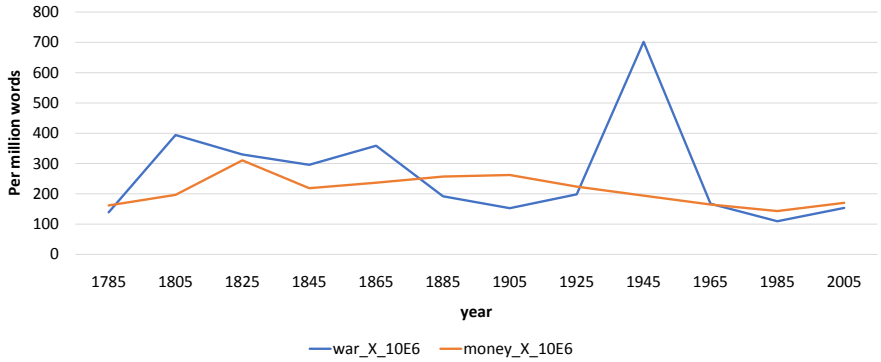

Figure 3. War and money raw data (times 1,000,000).

3. Results

3.1. Initial Statistical Results

Pertinent to the Ngram Project instructions the results of the two regression analyses are as follows:

Dependent	Independent	R^2	*p*-value	Slope	Intercept
War × 10^6	Year	0.0235	0.6343	−35.45898	93,800
Money × 10^6	Year	0.1583	0.2002	−27.29919	72,899

The two regression lines were set equal to each other, and then solved for the year = 2561. While the R^2s are low and the *p*-values high, this still demonstrates that the student has learned how to conduct regression analyses using SAS Enterprise Guide.

3.2. Descriptive and Contextual Analyses

The first author also plotted the data from the ngram database and added in the timing of specific wars, conflicts, and financial periods in **Figure 4**.

3.3. Results for "Cost of War" in Two Corpora

An issue with the ngram analyses can be that the valence of the term is not always certain. For example, Aiden and Michel (2013) identified the person whose name appeared at the highest frequency for each annual birth cohort. This maximum value might be a result of fame, or infamy, or both. On this list are great leaders [e.g., Winston Churchill, Mother Teresa, and MLK Jr. (see also Zywiak & Niu, 2021)] as well as some infamous persons: Saddam Hussein and Lee Harvey Oswald. To shed light on the war and money analyses, we examined the frequency of the 3-gram "cost of war" from 1901 through 2019 in the American English corpus and the British English corpus for comparison. "Cost of war" has an unambiguous negative valence, while "money" is more ambiguous and depends on the context (i.e., a credit versus a debit). We chose a more recent period and up to the latest available ngram data (i.e., 2019). Results are plotted in

Figure 5. Over this period, we also calculated the frequency of the 3-gram in American English minus the frequency of the 3-gram in British English for each year, and plotted this in Figure 6. The last line we plotted was the cumulative value of this difference (or the sum of this difference). For 1901, this was just the value of the difference. For 1902, it was the sum of the difference for 1901 and the difference for 1902. Over the 119 years, the British English frequency value exceeded the American English frequency value.

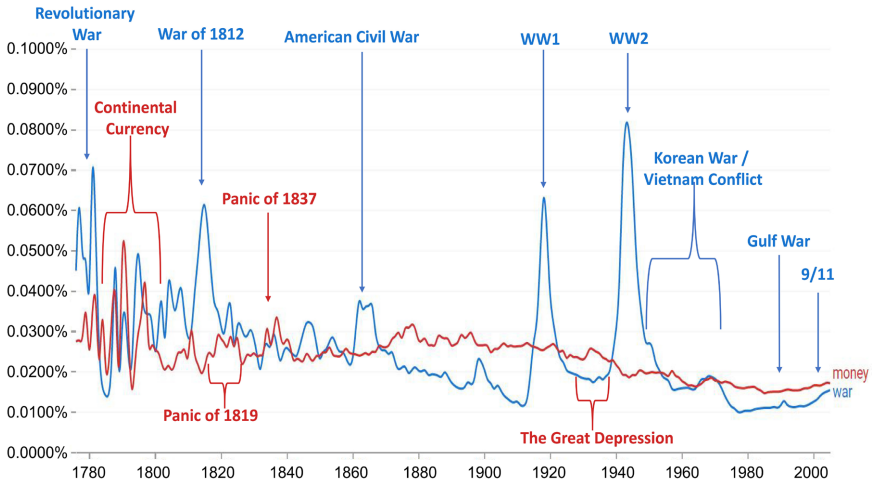

Figure 4. Ngram Viewer: War and Money with events tagged.

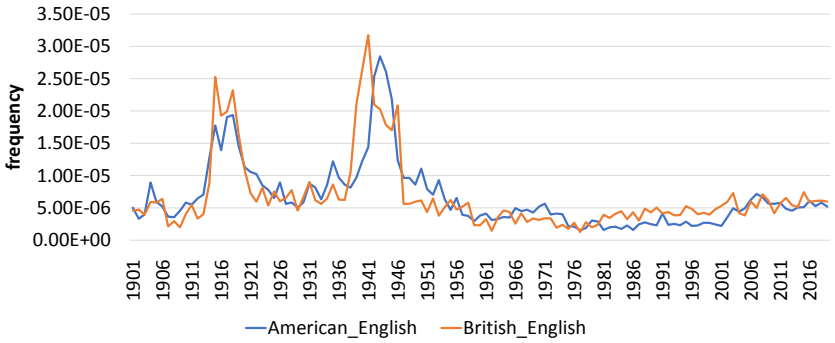

Figure 5. "Cost of war" in American and British English.

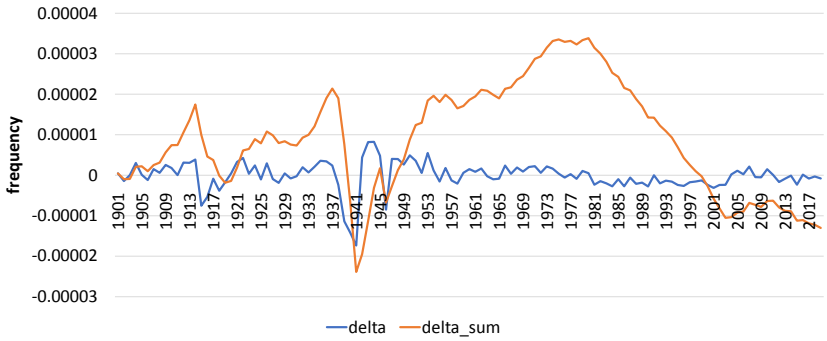

Figure 6. Delta (American English value minus British English Value) and Cumulative Delta.

4. Discussion

4.1. Discussion of the War and Money Analyses

When studying the Ngram Viewer chart of "war" and "money" (see **Figure 4**) significant increases in the frequency of the use of the words "money" and "war" are noticeable. Spikes in the frequency of the use of the word "money" occur in the late 1700's, when the U.S. government issued its first currency, as well as during the Panic of 1819, and during the Panic of 1837 (Perkins, 1984). Surprisingly, during the time of the Great Depression, there was not a large spike in the frequency of the word "money". Prominent increases in the frequency of the word "war" are seen during time periods that coincide with wars involving the United States (Williams, 1981). Interestingly, after World War 2, the frequency of the use of the word "war" did not increase as dramatically during times of war when compared to pre-WW2 frequency increases.

Regression lines for "money" and "war" both have negative slopes. This implies that they are both expected to be used less frequently as time goes on. This may be due to "substitute" words. Examples of a substitute word(s) are "the Vietnam Conflict" instead of "the Vietnam War", and "cash" instead of "money." The literary frequency of words could also decrease during times of censorship, but this probably doesn't apply to these terms.

The R^2 of both models is low, and the *p*-values are not significant suggesting the data does not fit a linear function. (Similarly, humor has a low R-square in this type of analysis.) This is consonant with economies and wars being multi-determined and somewhat unpredictable. The R^2 is lower and the p-value is higher for "war" compared to "money" suggesting that wars are more unpredictable than economies. There are several examples of how wars are unpredictable in past and recent history. WW1 which involved more than 30 countries was precipitated by an assassination in Sarajevo. More recently explosions of the Nord Stream 1 and 2 pipelines were quite unexpected. Additionally, the element of surprise can be strategic [e.g., General Washington and his men attacking the Hessians (fighting for the British) the day after Christmas]. In **Figure 4**, the huge peaks for "war" during WW1 and WW2 both explain why the R^2 is so low, and also evidence the distinction between wartime and peacetime.

The sampled data points from the Ngram Viewer were plotted on a line chart in **Figure 2**. The highest frequency of "money" is seen in 1819, which coincides with the Panic of 1819. The Panic of 1819 was centered around a banking crisis that brought an economic downturn to the US. It has been regarded "as the first nationwide economic depression in American history" (Lehman, 2011). The highest frequency of the word "war" is seen in 1945. This peak is historically significant because during 1945, a number of major "war" events happened regarding World War 2. The first was Germany's surrender in May which was preceded by the suicide of Adolf Hitler. Also in 1945, the U.S. dropped atomic bombs on Japan in August, which caused the Japanese to surrender in September (Mowat, 1968). When examining the graph of the sampled data, it is clear that both lines are trending in a downward direction, lending credence to the

negative slopes found in the regression analysis.

This project shows that by analyzing the frequency of use of certain words in literary texts, we can model a timeline of past historical events that are relevant to the words that we are researching. These models can be made without the researcher having prior knowledge of any significant events in history. For example, before this project, three of the authors had never heard of the "Panic of 1819". It was only by looking at the frequency of the word "money" that this was "discovered". Theoretically, by looking at the historical frequency of other words, we would notice spikes in their usage that would correspond to important historical events.

4.2. Discussion of the "Cost of War" Analyses

In **Figure 5** during the onset of WWII, the increase in "cost of war" is apparent in the British English corpus before the American English corpus. This is consistent with Great Britian's earlier declaration of war against an Axis member (i.e., Germany in September 1939) than that of the USA declaring war on Axis member (i.e., Japan in December 1941, Mowat, 1968). In **Figure 6**, the difference score or delta is above zero when the frequency of "cost of war" is greater in the American corpus than the British corpus. The sum or cumulative value of the difference score is also plotted. This last plot is quite striking in that it peaks in 1980, generally descends through 2003, and crosses zero between 1998 and 1999. This peak after the 1970s is consistent with war protests during the 1970s in America, which would emphasize the cost of war. The US withdrew from Saigon in 1975. The crossing of zero between 1998 and 1999 suggests that "the cost of war" didn't matter as much to US citizens after 9-11 compared to British citizens. The Afghanistan War (U.S. led coalition) started in 2001 and the Iraq War (U.S. led coalition) started in 2003.

5. Conclusion

This paper demonstrates that college students can conduct statistical analyses on the ngram data to explore topics in history. This paper builds on previous work by the second, third, and fourth authors (Zywiak, Bobroff, & Niu, 2021; Zywiak & Niu, 2021) by testing a 3-gram for the first time and shows that a useful comparison can be made between the American and British English corpora. One limitation is that the timeline for the cumulative delta in **Figure 6** ends in 2019. It would be interesting to know how this changes with the withdrawal of the US from Afghanistan in August 2021, which in some ways paralleled the withdrawal of the US from Vietnam in 1975. Future studies could examine other ngram corpora (e.g., German, Italian, and Russian). Additionally, other sources of data could be examined (e.g., economic, military, and UN data).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Aiden, E., & Michel, J.-B. (2013). *Uncharted: Big Data as a Lens on Human Culture.* (Riverhead) Penguin.

Echo-Hawk, W. R. (2010). *In the Courts of the Conqueror: The 10 Worst Indian Law Cases Ever Decided.* Fulcrum Publishing.

Hunt, L., & Censer, J. R. (2017). *The French Revolution and Napoleon: Crucible of the-Modern World.* Bloomsbury.

Lehman, J. D. (2011). "The Most Disastrous and Never-to-Be-Forgotten Year": The Panic of 1819 in Philadelphia. *Pennsylvania Legacies, 11,* 6-11. https://doi.org/10.5215/pennlega.11.1.0006

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). *Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331,* 176-182. https://doi.org/10.1126/science.1199644

Mowat, C. L. (1968). *The New Cambridge Modern History. Vol. 12: The Shifting Balance of World Forces, 1898-1945.* Cambridge University Press. https://doi.org/10.1017/CHOL9780521045513

Perkins, E. T. (1984). Langdon Cheves and the Panic of 1819: A Reassessment. *Journal of Economic History, 44,* 455-461. https://doi.org/10.1017/S0022050700032058

Williams, T. H. (1981). *The History of American Wars from 1745 to 1918.* Louisiana State University Press.

Zywiak, W. H. & Niu, G. (2021). Love, Hope, Perspective, and Leadership in the Ngram Database: Solace for Modern Times. *Open Journal of Social Sciences, 9,* 159-166. https://doi.org/10.4236/jss.2021.911013

Zywiak, W. H., Bobroff, R. P., & Niu, G. (2021). *Black Swan Years in American English, French, German, Hebrew, and Russian: Years That Reverberate in Ngram Viewer. Advances in Historical Studies, 10,* 208-214. https://doi.org/10.4236/ahs.2021.103013