Scientific Research Publishing

Advances in Historical Studies, 2021, 10, 208-214
https://www.scirp.org/journal/ahs
ISSN Online: 2327-0446
ISSN Print: 2327-0438

Black Swan Years in American English, French, German, Hebrew, and Russian: Years That Reverberate in Ngram Viewer

William H. Zywiak[1], Ronald P. Bobroff[2], Gao Niu[1]

[1]Mathematics Department, Bryant University, Smithfield, USA
[2]History and Social Sciences Department, Bryant University, Smithfield, USA
Email: wzywiak@bryant.edu

How to cite this paper: 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

Received: August 7, 2021
Accepted: September 7, 2021
Published: September 10, 2021

Abstract

The Ngram Viewer database has been used to study history and culture, primarily by examining the frequencies of words, within a single or two languages (corpora). We examined numbers instead (specifically years) across five different corpora. The years were 1799, 1865, 1917, 1945, and 1948. The corpora were American English, French, German, Hebrew, and Russian. Our analyses suggest that these years reverberate in specific languages more than typical years, and therefore our approach can be used to identify years with significant historical impact, within and across languages (and by extension countries) and therefore may be useful in the field of comparative history.

Keywords

Comparative History, Ngram Viewer, Revolution, War

1. Introduction

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). Many others have used this database to conduct a variety of research. The first and third authors have been having undergraduate students in statistics courses use this database to conduct and interpret regression analyses and lagged correlation analyses. The sheer volume of the dataset allows students to conduct analyses on terms that are of interest to them.

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. Six percent of all books have been scanned. Languages (and sub-

sets) in the database include American English, British English, English, English Fiction, Chinese (simplified), French, German, Hebrew, Italian, Russian, and Spanish. Sources include over 5 million books. 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). Recently, we uncovered a startling anomalous pattern, and we present this pattern here. We used the 2019 version of the data (in July 2020).

Aiden and Michel indicate that the "forgetting" of years increases as time goes on. They present comically how "1950" was somewhat anticipated, had its hay-day in 1950, and then became "slightly passé" in the Ngram Viewer database (Please see **Figure 1**). They note that half-lives for years have been getting shorter: the year 1872 had a half-life of 24 years, while more recent years have a shorter half-life, e.g., 1973 has a half-life of 10 years (Half-life, common in the study of radioactive elements, in this case refers to the number of years that the frequency of any word decreases to half of its peak value). We found many years fit this pattern, including 1776 in American English. However, we found several years that did not fit this pattern (e.g., 1865 and 1917, see **Figure 2**). These were years that reverberated, often indefinitely in a given language. These years and languages are detailed in **Table 1**. We refer to these years that reverberate as Black Swan Years. Black swan years reflect Taleb (2007)'s perspective that history does not follow a linear process, but jumps, with quantum shifts or "earthquakes" occurring periodically. More generally, Taleb uses the term Black Swan to label an extreme, and often unexpected, outlier [the statistical definition for an outlier is a value that is extreme (being either very small or very large) compared to the general distribution of values]. While Aiden and colleagues investigated the half-lives of years, they, like most Ngram researchers, predominantly study the

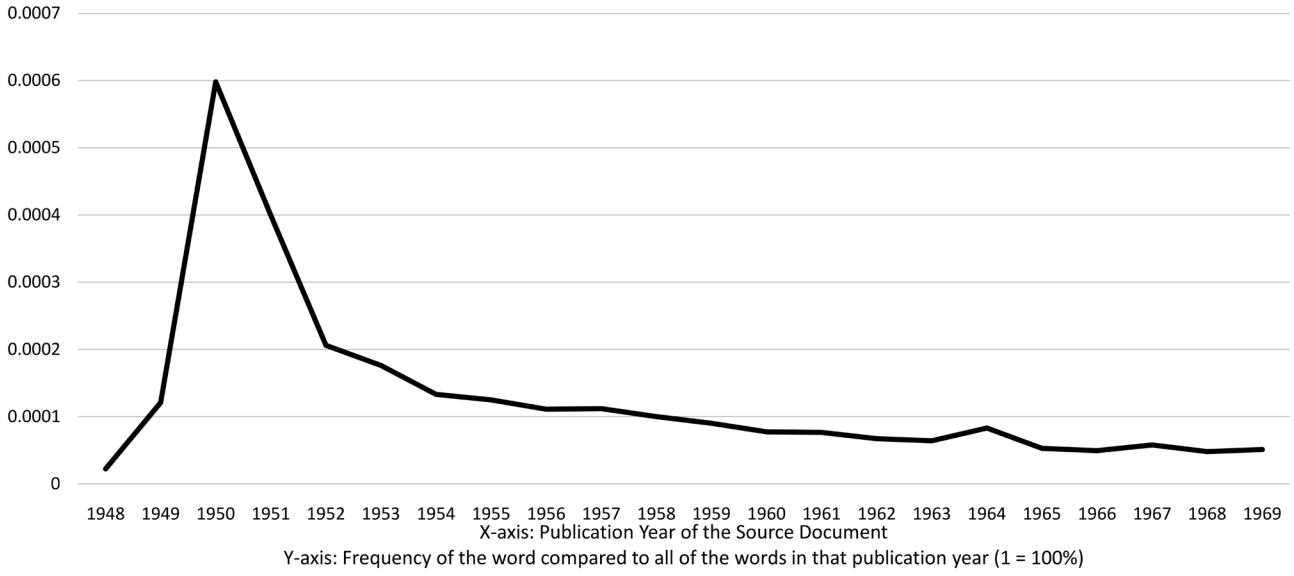

Figure 1. Frequencies for "1950" in the Russian corpus.

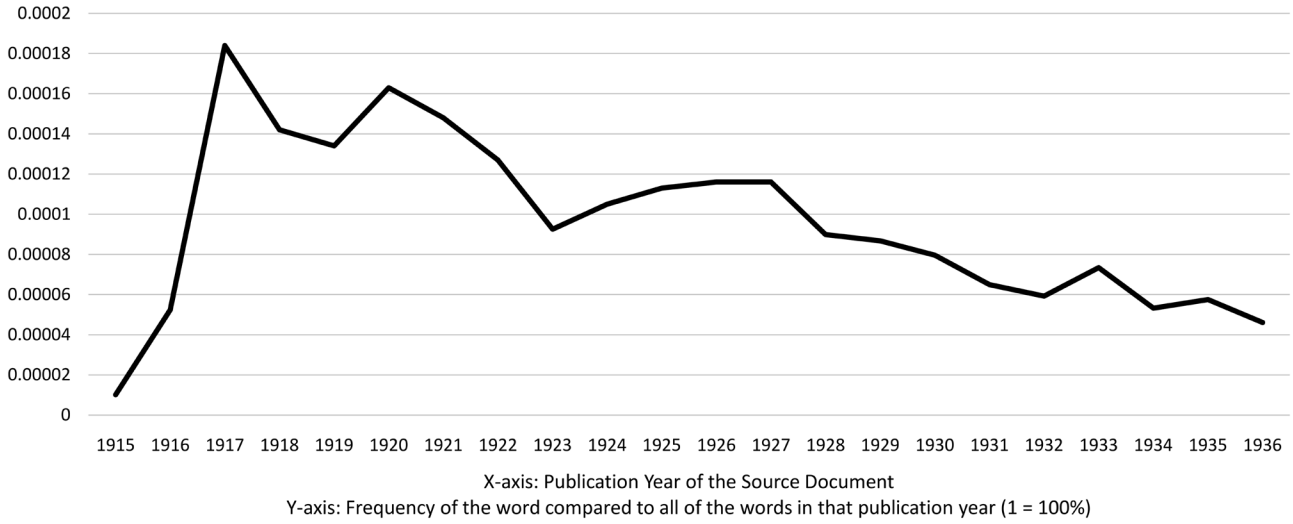

Figure 2. Frequencies for "1917" in the Russian corpus.

Table 1. Black swan years in specific corpora.

	Am. English	French	German	Hebrew	Russian
1799		X			X
1865	X			X	
1917				X	X
1945			X	X	
1948				X	

Ngram frequencies of words, not numbers. Our focus, on numbers, natural for mathematicians, complements the extant Ngram literature. Further, we hypothesize that studying years has fewer pitfalls when comparing patterns across corpora (e.g., words might have different common meanings in different corpora, and words might have to be translated).

2. Methods and Results

We conducted analyses to demonstrate that the wave form is more sustained for the Black Swan Years, than a control year (1895). We present one statistical test, but this was perhaps unnecessary given the robustness of the differences. The control year is the average value of the black swan years [(1799 + 1865 + 1917 + 1945 + 1948)/5]. We considered using 1950 as the control year, but used 1895, since Aiden and Michel indicate that the half-life of years has been getting shorter over time, so it would not be surprising if the black swan years were more persistent than 1950, so 1895 is a more conservative comparison. Below we compare the black swan years to the wave form of 1895 for all five corpora.

The nine black swan years itemized in Table 1 (counting 1799, 1865, 1917, and 1945 twice) were detected by visual inspection of the Ngram data by the first author. For each of these (e.g., 1799), frequencies of that year were recorded for

two baseline years (e.g., 1797 and 1798) and 20 subsequent years (1799 through 1818) anchored on the Black Swan year. The peak of this profile was identified (67% of the time on that year, with the remainder being one or two years later) and all frequencies were divided by the peak value, so the maximum value was 1 so that all profiles were on the same scale. The same process was used for the control year of 1895 across the five corpora. These values were averaged by year (ranging from −2 to +19) and these two wave forms plotted in Figure 3. A pairwise t-test conducted in SPSS 26 confirmed that the Black Swan profile persisted to a greater extent [M = 0.42, SD = 0.14] than the control year profile: [M = 0.34, SD = 0.22, t (19) = 3.07, $p = 0.006$] for the 20 years starting with the Black Swan Year.

3. Discussion

Historical details can help explain the persistence of some of these years in literature. Successful coups (1799 and 1917) are apparent, as is a civil war. The year that WW II ends resonated profoundly in the German and Hebrew corpora. These black swan years in general appear to be major watershed moments in countries' and/or people's political and social evolution. What requires more study is why the Ngram of 1776, certainly a crucial moment in the development of the United States, does not display a Black Swan frequency.

3.1. 1799

1799 marked the end of the French Revolution, which had included the execution of several members of the royal family and the creation of a new republic. In the wake of the 1799 coup, General Napoleon Bonaparte began the more political phase of his rise to power, bringing France to the height of its power in Europe before losing it all after the failure of the invasion of Russia (Hunt & Censer, 2017).

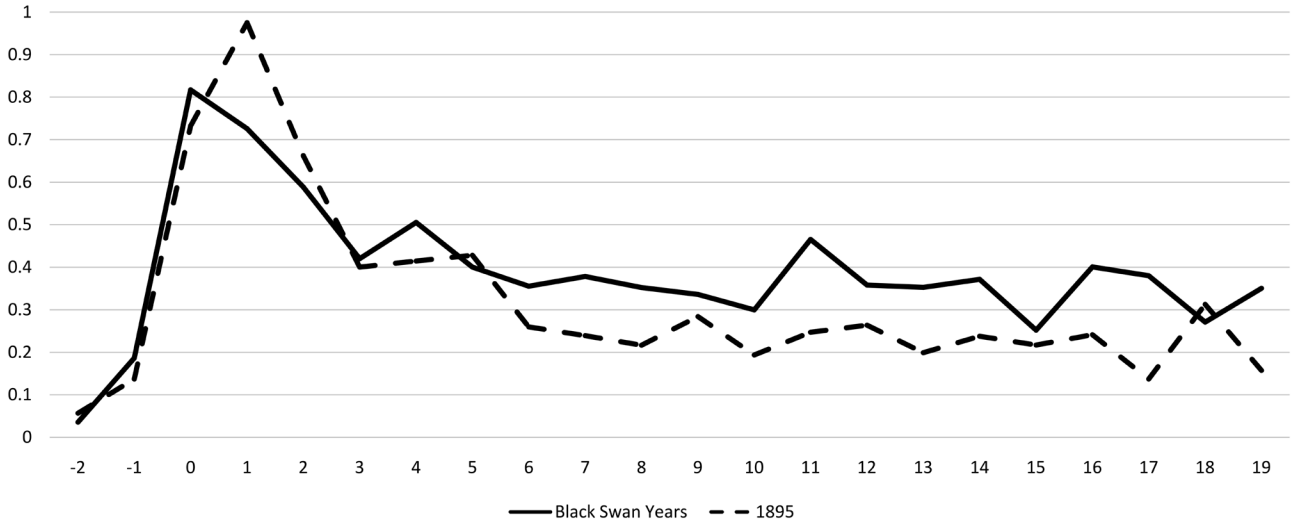

Figure 3. Frequencies for black swan years versus the control year (1895).

3.2. 1865

1865 figures strongly in the American English corpus as it marks the conclusion of the U.S. Civil War. But while the physical conflict ended at that time, the underlying problems were not resolved (Zinn, 2003). In the Hebrew corpus, this year resonates for a number of reasons. The end of slavery surely resonated with Jewish people, who have suffered a long history of being slaves by the Egyptians as well as the Assyrians, as well as being displaced more recently by the Ottoman Empire (Hanukoglu, 1988). In addition, Jewish people suffered an increase in prejudice during the U.S. Civil War, and it took more than a year for Jewish soldiers to have a Jewish army chaplain (versus a Christian army chaplain). More drastically, General Order No. 11 in 1862 expelled all Jewish men from General Grant's military department, though President Lincoln quickly had it revoked (Sarna & Golden, 2000; Karp, 1969).

3.3. 1917

1917 saw the Russian Revolution start in Petrograd in March. In November, the Bolsheviks seized power (Figes, 1996). The creation of the Soviet state resulted not only in a decisive change in the domestic affairs of its inhabitants but also introduced a new, ideological factor in European and international affairs, explaining its presence in the Russian corpus (Ward & Thompson, 2021). As for the Hebrew corpus, that same year, the Balfour Declaration announced the United Kingdom's commitment to the establishment of a Jewish homeland in Palestine, a watershed for the Zionist movement (Sachar, 2003).

3.4. 1945

1945 reverberates in the German corpus and Hebrew corpus, but surprisingly not in the English American, French, or Russian corpora. Germany surrendered in May 1945 and the Potsdam Agreement in August 1945. The Potsdam Agreement by the UK, USA, and USSR specified how Germany would be divided into four occupation zones, which soon after would evolve into the two, separate German states (Zubok & Pleshakov, 1991).

3.5. 1948

1948 marks the establishment of the modern nation-state of Israel, which reverberates in the Hebrew corpus (Sachar, 2003).

3.6. History of Science

The history of science can also be examined using the Ngram data. For example, Aiden and Michel note that while a patent for the telefax was awarded in 1843, the fax machine rises exponentially only in the 1980s. Further, they note, that while in America, Alexander Graham Bell is considered the father of the telephone; in Italy it is Antony Meucci. They also conducted analyses showing that the cultural assimilation of inventions quickens over history: time from inven-

tion to peak Ngram frequency: from 1800 to 1840: over 66 years, from 1840 to 1880: under 50 years, and from 1880 to 1920 under 27 years (Michel et al., 2011).

4. Conclusion

We have presented a statistical approach to verify which years reverberate the most in a specific Ngram corpus. We are currently using this approach to conduct an exhaustive search of black swan years in these and other corpora (i.e., Chinese, Italian, and Spanish). This new work reveals that 1895 reverberated in Russian, suggesting that this is a conservative control year in the present article. The statistical approach presented here can be used to show shared histories, as we evidenced for four of the five Black Swan Years. Similarities and differences of different black swan years can be examined across languages as we have shown. We encourage Google to digitize and add literature from other languages such as Hindi and Arabic to the Ngram Viewer database. More generally, the Ngram data can be used in statistics classes to renew interest in history, while providing some evidence of one social process contributing to an outcome (in the lagged correlation analyses for instance). For example, an escalation in "jazz" predates an escalation in "freedom" in the American English corpus in the period from 1915 through 1935.

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.

Figes, O. (1996). *A People's Tragedy: A History of the Russian Revolution.* Viking.

Hanukoglu, I. (1998). *A Brief History of Israel and the Jewish People* (pp. 53-59). Quest. https://www.science.co.il/israel-history/A-brief-history-of-Israel-and-the-Jewish-people.pdf

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

Karp, A. J. (1969). *The Jewish Experience in America.* Ktav.

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

Sachar, H. M. (2003). *A History of Israel from the Rise of Zionism to Our Time.* Knopf.

Sarna, J. D., & Golden, J. (2000). *The American Jewish Experience through the Nineteenth Century: Immigration and Acculturation.* Teacher Serve. https://nationalhumanitiescenter.org/education-material/the-american-jewish-experience-through-the-nineteenth-century-immigration-and-acculturation-by-jonathan-d-sarna/ (Accessed on July 19, 2021)

Taleb, N. M. (2007). *The Black Swan: The Impact of the Highly Improbable.* Random

House.

Ward, C., & Thompson, J. (2021). *Russia: A Historical Introduction from Kievan Rus' to the Present* (9th ed.). Routledge. https://doi.org/10.4324/9781003015512

Zinn, H. (2003). *A People's History of the United States.* Harper Collins.

Zubok, V., & Pleshakov, C. (1996). *Inside the Kremlin's Cold War: From Stalin to Khrushchev.* Harvard.