An Overlooked Partnership:
Python Probes 1840

Computer generated-text analysis deciphers the turning point of 1840 in Black Abolitionist rhetoric, 1830-1861

Sam Strickberger
COMM 313-001: Professor O'Donnell & Jake Parelman
Final Paper
May 7, 2021

***

"What is needed is a systematic application of the scientific method to history: verbal theories should be translated into mathematical models, precise predictions derived, and then rigorously tested on empirical material. In short, history needs to become an analytical, predictive science."

Peter Turchin,
founder of Cliodynamics:
the Journal of Quantitative History and Cultural Evolution

***

INTRODUCTION

The U.S. institution of enslavement – a race, based economic system where people were treated as chattel – is inextricably linked to countless aspects of American history. The scholarship on the subject is robust, covering a diverse range of topics. Its methodology, however, is quite homogenous. After removing studies on enslavement's economic aspects, this is all the more obvious. A careful, data-minded reader will soon begin to see the glaring gap. Computerized text analysis is extremely underutilized in historical research. While it is true that quantitative analysis can only grasp and tell a certain story. This story must be told and further investigated. Historians must evaluate enslavement from as many vantage points as possible. A system so fundamental to American history necessitates nothing less.

This paper has two core aims. It will highlight an rare methodology within historical research, computerized text analysis, while also building on existing literature on enslavement. As a case study, this paper will delve into the development of Black abolitionist rhetoric, specifically from 1830 to 1861.

The literature argues that Black abolitionist rhetoric became more civic minded and nationalistic after 1840. Dexter Gordon provides three pieces of evidence, among others, to prove his point. One, Gordon argues that after 1840s the Black abolitionist discourse became increasingly focused on the topics of Black nationalism and political engagement. Black abolitionists are no longer just calling for negative freedom – freedom from bondage – but positive freedom – complete citizenship and political inclusion. Questions about integration and separatism, distinction and separation, sprung up.[1] Two, he argues that in these nationalist debates regarding Black unity, Black identity was increasingly discussed. This leap is logical.

[1] Gordon, Dexter B. "Black ideology in American Black abolitionist discourse: The materialization of a constitutive rhetoric" (1999), 185-197.

Nationalism centers around group identity. So, if Black nationalist sentiment increases, Black self-identification would, too. He explains that Black abolitionists discussed whether to identify as "colored," "black," or "n*gro." Third, he argues the following: "In black public discourse, the 1850s became a feverish decade of proposals and counterproposals relating to emigration." In other words, the tone of these abolitionists heated up after 1840.[2] This increase also makes sense. When one argues for civic inclusion of a certain group, one's antipathy towards enslavement of this same population should increase. Said differently, it is a humanitarian crime to enable forced labor of a human. It is a humanitarian and political crime to enable forced labor of a citizen.

I aim to operationalize and test these three developments through a computerized text analysis. Consequently, I will offer three hypotheses regarding the overall theory that Black abolitionist rhetoric became more civic minded and nationalistic after 1840.

1. Language referring to the United States of America increases (e.g. "this country," "the United States," etc.).
2. Language aimed at self-definition increases (e.g. "Black," "African," "colored," etc.).
3. The rhetoric becomes increasingly demanding and aggressive (i.e. negative).

METHODOLOGY REVIEW

Fortunately, my analysis does not have to start from scratch. The most cited study on Black abolitionist rhetoric is from the early 2000s. Timothy Shortell published "The Rhetoric of Black Abolitionism: An Exploratory Analysis of Antislavery Newspapers in New York State" in *Social Science History*. I will use his methodology to operationalize my three hypotheses. To summarize, he advanced two main approaches. He "generated frequencies and co-occurrences of

[2] Gordon, 191.

sixteen themes." This analysis was only partially computer generated. He used trained readers to locate themes and input them manually into the data set. He then relied engaged in computerized data analysis. He also produced a rhetorical analysis, examining tone and sentiment. In both approaches, his unit of analysis was a paragraph.[3]

I will utilize both approaches but with two modifications. His analysis was exploratory within a single corpus but mine will be a comparison between two corpora of different time periods, pre-1840 and post-1840. Additionally, my analysis will be much more computerized. In my theme analysis, for example, I will also generate themed frequency lists. I will create key word lists for different themes. I will then enable the computer to search for the words. I will then create a score for each unit of analysis and note the strength of a given theme with the different texts. I will not rely on human readers.

My rhetorical analysis will largely mirror his. However, I will rely on VADER (Valence Aware Dictionary for Sentiment Reasoning) polarity tests to operationalize my variables. Once again, no human readers are required. VADER is a sentiment analysis tool. Its pre-set lexicon has a specific rating for hundreds of words. The more positive (negative) the word, the higher (lower) the rating. Scores range between -1 and 1. For example, the word "happy" garners a rating of 0.5719 and "sad" garners a rating of -0.4767. When you feed VADER a sentence, it applies a similar approach. It provides a composite score along with a sum total of the positive, negative, or neutral scores each word in the sentence. The unit of analysis for my VADER tests will be single pages. Each is about 250 words, or large paragraphs. This aligns with Shortell's unit of analysis.

---

[3] Shortell, Timothy. "The Rhetoric of Black Abolitionism: An Exploratory Analysis of Antislavery Newspapers in New York State." *Social Science History* 28, no. 1 (2004): 75-109, 86.

There are a few possible critiques for employing VADER in my analysis. Most critically, VADER was developed to analyze the language of tweets and amazon book reviews, not antebellum writings. Its lexicon is geared toward a more modern data set, and it may not have enough of an antebellum vocabulary, so to speak, to be helpful for this project. This is a fair concern. However, I administered a few tests for external validity, which should diminish most such worries for our paper. I will discuss them further in my results section.

OVERVIEW OF DATA AND DATA COLLECION PROCESS

Now that we have a sense of the literature and my analytical approach, let us discuss my data set and how I collected it. The University of Detroit Mercy's Black Abolitionist Archive provided me with a massive, well-regarded data source. "The Black Abolitionist Digital Archive is a collection of over 800 speeches by antebellum blacks and approximately 1,000 editorials from the period," its website explains.[4] The collection portrays a critical piece of the Black abolitionist movement and the anti-slavery movement at large. While the archive hosts 1800 total texts, only about 750 are scanned PDFs that I could utilize. Most of 750 texts are speeches. The remaining 1,050 texts are not PDFS, but scanned images that I could not use for my analysis.[5]

I needed to scrape these PDFs from the online archivie to build my corpus. This task was by far the most difficult and time consuming. With the help of Jacob M. Parelman, a doctoral candidate at University Pennsylvania's Annenberg School for Communication, I succeeded in

[4] "Black Abolitionist Archive." Black Abolitionist Archive :: UDM Libraries / Instructional Design Studio. Accessed May 7, 2021. https://libraries.udmercy.edu/archives/special-collections/index.php?collectionSet=&collectionCode=baa.

[5] Since there is no substantive difference between the PDF speeches and image-based speeches, this selection does not provide any issues for our analysis.

incorporating these PDFs into my corpus. They are my raw data. My corpus includes 2,513 pages of text and 666,978 words. This is a significant size. It is nearly twenty times larger than Shortell's corpus, which included 36,00 words.[6]

My next step was cleaning my data. As aforementioned, I divided my data into two categories: texts written during and before 1840 and texts written after 1840 until 1861. For the sake of brevity, I will refer to these two sub-corpuses as pre-1840 and post-1840, respectively. The pre-1840 corpus contains about 424 pages or units of analysis. The post-1840 corpus includes about 2,513 pages or units of analysis.

EMPIRCALLY TESTING MY HYPOTHOSES

I have three main hypotheses. To empirically test them, I will aim to reject a null hypothesis for each one. They are as follows.

1. AMERICAN-CENTRIC RHETORIC: Language referring to the United States of America remains the same in frequency before and after 1840.
2. RHETORIC OF IDENTITY: There is as much language regarding self-definition before and after 1840.
3. TONE OF RHETORIC: The tone of the rhetoric remains constant from 1830-1861, with no change after 1840.

My alternative hypotheses are that instead of staying unchanged, language referring to the United States and self-identification increases after 1840. The tone, similarly, becomes more aggressive after 1840. I administered three sets of analyses to test these hypotheses.

[6] Shortell, 1.

AMERICAN-CENTRIC RHETORIC TESTS & RESULTS

First, I set about testing the change of American-centric rhetoric frequency over time. To begin with, I advanced a preliminary test. I called up the 25 most common three-word phrases, also known as trigrams, from each corpus. In the pre-1840 corpus, there were seven phrases referencing or including America: "the united states", "the people of", "of the people", "the colored American", "in this country", "of the American", and "in the united." Two phrases were in the top five; three were in the top ten. See *Table 1* for reference. In contrast, in the post-1840 corpus, the top three trigrams are all related to the United States. Six such trigrams composed the top ten. See *Table 2* for reference. This preliminary test bolsters our hunch that 1840 was a turning point Black abolitionist rhetoric.

Pre-1840 Rank	Trigram
1	people of color
2	the cause of
3	the people of
4	the united states
5	committee of vigilance
6	of the colored
7	the rights of
8	of the people
9	the colonization society
10	the colored man

Table 1: Pre-1840 Trigrams

Post-1840 Rank	Trigram
1	the united states
2	in this country
3	of the united
4	one of the
5	of the slave
6	in the united
7	the cause of
8	of the country
9	as well as
10	the people of

Table 1: Pre-1840 Trigrams

I ran a keyness analysis to confirm my preliminary results. Based on the literature, I curated a list of "political" words. These included "united", "states", "country", "nation", "citizen", "citizenship", "right", "rights", "civil", "man", and "vote." Over 80% of these political words were more likely to appear in the post-1840 corpus than the pre-1840 corpus. "States" and "vote" in particular had a 21+ log likelihood of occurring in the post-1840 corpus. I aggregated all of the political words into a single log likelihood analysis for good measure. The result was a log likelihood of 32.5. Meaning, political rhetoric was more likely to occur after 1840.

A key word in context (KWIC) analysis demonstrates that our results have external validity. To begin with, "States" denotes exactly what you would imagine. It does not refer to "states of matter," for example. It references the United States of America or its member states, according to a KWIC of a randomly generated sample. The connotations of these sentences are all politically minded. For example, we find sentences such as "…suffering in the United States, all the horrors…" and "…fundamental laws of the united states the slave was legally….". A similar analysis demonstrates that the abolitionist's use of "vote" aligns with mainstream political conceptions of enfranchisement.

As the literature noted, Black abolitionists were no longer simply fighting for negative freedom – freedom from enslavement – but positive freedom too – freedom to engage in the American political process through the most fundamental right of a citizen: enfranchisement. These preliminary and more advanced analyses confirm the overarching themes of the literature. The post-1840 abolitionist writings were more politically focused than in the preceding decade.

IDENTITY RHETORIC TEST & RESULTS

We will now test our second hypothesis. Does the frequency of Black abolitionist rhetoric of self-identification increase, decrease or remain constant after 1840?

This second analysis mimics that of the previous section. I again started with a preliminary analysis. Contrary to my hypothesis, it demonstrated that the rhetoric of self-identification may actually decrease after 1840. The trigram frequency list of pre-1840 has four phrases related to Black identity falling within its top twelve most common trigrams. These include "people of color" (first most frequent); "of the colored" (sixth most frequent); "the colored man" (tenth most frequent); and "the colored American" (twelfth most frequent). I

expected to see similar results in our post-1840 frequency list. However, in comparison, phrases related to Black identity were less common. In the post-1840 trigram frequency list, we can see that the terms "colored people", "the coloured people", "of the colored", and "the colored man" all appear within the top twenty-five most common trigrams. But they are ranked between the top fourteen and seventeen most common words. None crossed fell within the top ten.

I proceeded to test these preliminary results with a keyness analysis. Based on the literature, I assembled a list of all terms that referenced identity. This includes, "Black", "negro", "colored", "coloured" (an alternate spelling), and "African". From our keyness analysis, there are no words relating to Black identification that are more likely to occur in the pre-1840 corpus than in the post-1840 corpus. (Note that any statically significant difference has a log-likelihood with an absolute value larger than five.) More, every word except for "coloured" – the alternative spelling of "colored – is more likely to occur in the post-1840 corpus.

Word	**Post-1840 Occurrences**	**Pre-1840 Occurrences**	**Log-likelihood**
Black	456	26	67.66
Negro	390	5	118.20
Coloured	819	200	-1.81
Colored	514	17	111.72
African	251	13	40.52
Total	2430	261	109.83

Table 3: Rhetoric of Self-Identification: Log Likelihoods
Note: Any statically significant difference has a log-likelihood with an absolute value larger than five. Above 15 is equivalent to a p value of 0.0001.[1]

When I aggregated all of the identification words into a single analysis, the log-likelihood value was 145.5. Put simply, the rhetoric of self-identification intensified after 1840. The results from our preliminary analysis are thus misleading. Trigrams referring to people of color may have been more common before 1840. However, this analysis did not include trigrams of other forms of Black identification, such as "African" or "Black."

I confirmed the external validity of these results with a KWIC analysis. I found that each corpus employs the identification words in three distinct ways, all of which relate to African-American identity. The pre-1840 usage of "Black" provides a good example. Black is used in the context of the "*the* black man*"*, "*a* black man*"*, and "black *population*." In other words, it is used to describe the Black race, a member of that race, and a noun (such as "population" or "slavery") associated with that race.

POLARITY TESTS & RESULTS

Recall that our hypothesis states that after the 1840s the rhetoric of Black abolitionists became more aggressive and demanding. The underlying idea is that a movement for Black political rights will have a stronger standing to critique enslavement than one focused only on human rights.

The results of my VADER sentiment analysis concorded with my hypothesis. They were as follows. The mean polarity score of a post-1840 page was 0.38. The mean score of a pre-1840 page was 0.55. In other words, after 1840, Black abolitionist rhetoric grew increasingly negative. These results have statistical significance as their t-value was 4.15, denoting. This finding enables us to reject our null hypothesis that no change occurred in rhetoric. It supports our alternative hypothesis that abolitionist rhetoric became more demanding and aggressive.

A discussion only on average scores, however, obscures an important aspect of the results. They are bimodal. Most scores from the post-1840 (pre-1840) pages did not settle around

a rating of .38 (.55). Instead, those of both data sets hovered around -.9 and .9. See *Diagram 1* for an illustrated version of the results' bimodality. Since a greater portion of the post-1840 scores were negative, the corpus's mean rating was more negative than that of the pre-1840 corpus. Neither corpus, though, had many results that were moderately scored.

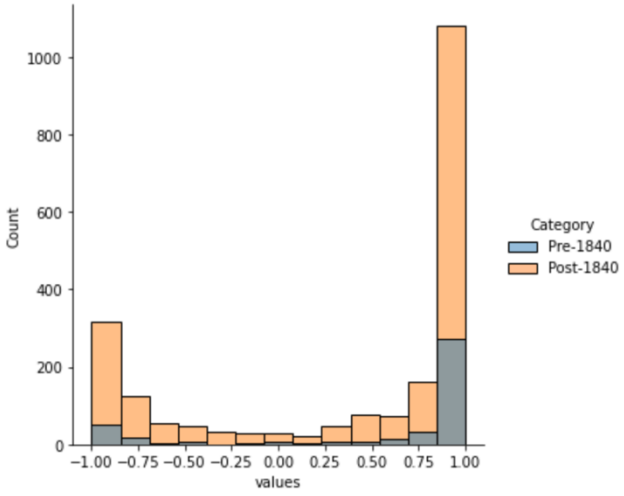

Diagram 1: Bimodal VADER polarity scores

These extreme scores may point to a lack of external validity. So, I ran four iterations of the same test to prove external validity. I assembled four sets of pages from the pre-1840 corpus: those with extreme negative scores, moderate negative scores, moderate positive scores, and extreme positive scores. I randomly selected a few pages to asses VADER's scoring. On each of these pages, I pulled up the list of words that VADER scored. (Remember that VADER does not score every word. It only scores those that it is familiar with.) This helped me understand what key words VADER relied upon. I then read through the selected page in full and gave it my own manual rating, such as highly negative or moderately positive. This approach enabled me to understand why VADER scored pages as it did and also asses whether this score was accurate.[7] I was satisfied with VADER's scoring on all but one page of the randomly selected sample, which included eight pages in total. See *Appendix 1* for an example of VADER working well. See *Appendix 2* for an example of VADER working poorly. My extern validity tests demonstrate that VADER is not a perfect tool for antebellum sentiment analysis. Perhaps, it should not be relied upon for precise results.

[7] Professor O'Donell of the University of Pennsyvian deserves credit for this approach.

However, it is successful in acquiring a broad strokes breakdown, which is all this paper requires.

CONCLUDING DISCUSSION

The data validate the existing literature's findings. In particular, the data support three hypotheses. The rhetoric becomes more civic-minded and nationalistic after 1840. It increases its emphasis on self-identification. And it becomes more negative, perhaps representing angrier and more demanding tones as the movement increases in power and advances a more civil rights-oriented approach.

The reader should also receive my findings with a healthy level of uncertainty. I say this regarding to analyses in particular. The self-identification rhetoric analysis maintains a few lingering questions. It does not deal with why the terms "colored" and "coloured" produce vastly different results. This distinction poses questions for the data set at large. Perhaps, a couple British oriented speeches tainted the findings for this specific word and others, too. The VADER analysis also deserves a closer examination. A more robust and precise study should assemble its own lexicon, scoring a set of common antebellum words. Though beyond the scope of this project, it would provide historians with an incredibly powerful tool for future analysis.

This analysis was difficult but very fruitful. Assembling a corpus of antebellum sermons was by no means easy. However, once sorted, it proved fertile for analysis.[8] It was exciting to witness the strength of python's text analysis abilities. It analyzed hundreds of pages and hundreds of thousands of words in mere seconds. This feat would require countless hours of a

[8] This project would not have succeeded without the assistance of Jacob M. Parelman and Matthew O'Donnell of the University Pennsylvania's Annenberg School for Communication. Perelman enabled me to assemble and organize the corpora. O'Donnell sharpened this studies VADER analysis test, highlighting its applicability and weaknesses.

historian's time. I appreciated supporting the existing literature with a more data driven approach. And more, I was delighted to illuminate why computer-generated text-analysis should be a more mainstream instrument in the historian's toolbox. More angles of analysis enable more complicated, multifaceted understandings of history.

Works Cited

"Black Abolitionist Archive." Black Abolitionist Archive :: UDM Libraries / Instructional Design Studio. Accessed May 7, 2021. https://libraries.udmercy.edu/archives/special-collections/index.php?collectionSet=&collectionCode=baa.

Shortell, Timothy. "The Rhetoric of Black Abolitionism: An Exploratory Analysis of Antislavery Newspapers in New York State." *Social Science History* 28, no. 1 (2004): 75-109, 86.

Gordon, Dexter B. "Black ideology in American Black abolitionist discourse: The materialization of a constitutive rhetoric" (1999), 185-197.

APPENDEX

APPENDEX 1: AN EXTERNALLY VALID VADER ANALYSIS

tion forbade the idea of their removing; and

hence in 1817, the people of color in Phila-

delphia, with James Forten at their head,Š

(and I envy them the honor they had in the

work in which they were engaged,) in an as-

sembly of three thousand, before high hea-

ven, in the presence of Almighty God, and
in the midst of a persecuting nation, resolv-
ed that they never would leave the land.Š

They resolved to cling to their oppressed

brethren. They felt that every ennobling spir-

it forbade their leaving them. They resolv-

ed to remain here, come what would, perse-

cution or death. They determined to grap-

ple themselves to their enslaved brethren as

with hooks of steel. My father, at Schenec-
tady, under great anxiety, took a journey to
Philadelphia, to investigate the subject.Š

This was the spirit to which prevailed among

the people of color, and it extended to every

considerable place in the north, and as far

south as Washington and Baltimore. They

lifted up their voice and said, this is my

country, here I was born, here I have toiled

and suffered, and here will I die. Sir, it
was a dark period. Although they were
unanimous, and expressed their opinions,

they could not gain access to the public

mind; for the press would not communicate

the facts in the caseŠit was silent. In the

city of New-York, after a large meeting,

where protests were drawn up against the

{'neg': 0.105, 'neu': 0.834, 'pos': 0.062, 'compound': -0.9086}

anxiety -0.7
death -2.9
determined 1.4
die -2.9
engaged 1.7
enslaved -1.7
envy -1.1
gain 2.4
god 1.1
great 3.1
honor 2.2
leave -0.2
oppressed -2.1
persecuting -1.5
protests -0.9
resolved 0.7
spirit 0.7
suffered -2.2

APPENDEX 2: AN EXTERNALLY INVALID VADER ANALYSIS

was to be executed on the following Friday, for having given a slave a free pass. (Shame, shame.) Mr. Thompson

had given them an account of some bad slaveholders; he

(Mr. R.) would tell them of some good ones. A master

with whom he once lived, Mr. Beveridge, in travelling from Apalache to Columbia, having to pass through the Indian nations, it was necessary for him to take arms. He

was taken exceedingly ill, and could neither stand up nor

sit down. He had a truck with him containing 20,000 dol-

lars, and he (Mr. R.) took the pistols and protected his

master and his master™s property. When he arrived at

Columbia, his master becoming embarrassed in circum-

stances, sold him on a block; that was his kindness to him

(Mr. R.) for saving his master™s life and protecting his property. Another good master was Colonel M™Gillion, a Scotchman, who held about 300 slaves, and who used to

boast that he never flogged them. His mode of punishing

them was to get a rice hogshead, into which several nails

were driven about a quarter of an inch through, and the

slaves then being fasten in, he used to roll them down a

very steep hill. (Shame, shame.) At one of the Revival

meetings, (of which he had heard so much since he came

to this country,) two ladies of colour came in and took their seats in the pew for inquirers. Holding down their heads, they were not observed; but some ladies coming in,

and noticing their colour, left the pew directly. (Hear,

hear.)

{'neg': 0.038, 'neu': 0.924, 'pos': 0.038, 'compound': -0.0387}

bad -2.5
block -1.9
embarrassed -1.5

```
free 2.3
good 1.9
ill -1.8
kindness 2.0
protected 1.9
punishing -2.6
shame -2.1
```