# White Paper Report

Report ID: 106099

Application Number: HD5155612

Project Director: Daniel Stowell (dstowell@papersofabrahamlincoln.org)

Institution: Abraham Lincoln Presidential Library Foundation

Reporting Period: 4/1/2012-6/30/2013

Report Due: 9/30/2013

Date Submitted: 10/1/2013

**Name of Project**: Is that You, Mr. Lincoln?: Applying Authorship Attribution to the Early Political Writings of Abraham Lincoln

**Institution**: Abraham Lincoln Presidential Library and Museum

**Project Director**: Daniel W. Stowell

**Award Amount**: $57,000

**To support**: Developing a methodology for attributing editorials and letters to the editor published anonymously and pseudonymously to a specific individual; developing a metric to report the strength of the attribution.

**Project Website**: http://www.papersofabrahamlincoln.org

**Keywords**: Authorship Attribution, Digital Editions

**Grant/Period**: NEH Digital Humanities Start-Up Grant, April 2012-June 2013

**White Paper: We Still Don't Know if that's You, Mr. Lincoln**

By Daniel W. Stowell, The Papers of Abraham Lincoln, Abraham Lincoln Presidential Library and Museum

The Papers of Abraham Lincoln applied authorship attribution methods to the problem of Lincoln's early anonymous and pseudonymous publications from the period 1833-1842. The project successfully identified 1,203 possible letters and editorials authored by Lincoln, assembled 85 known Lincoln texts from the period, and identified 219 additional texts by other authors for comparison. Our partners, computational linguist Patrick Juola and the Evaluating Variations in Language Laboratory at Duquesne University, performed the computational linguistic analysis of the questioned texts as compared to the known Lincoln texts and those created by the distractor authors. Although the results of the computational analysis were disappointing, the staff of the Papers of Abraham Lincoln learned much about the potential and limitations of computational analysis to answer these historical questions.

## Identifying Texts

A project researcher, Samuel Wheeler, prepared a list of known Lincoln documents from 1834 to 1842 for training purposes. Patrick Juola's team at the Evaluating Variations in Language (EVL) Lab at Duquesne University used these texts to try to train their Java Graphical Authorship Attribution Program (JGAAP) to recognize known Lincoln documents by identifying Lincoln's unconscious "stylome," an idiosyncratic pattern of some characteristic aspect of language. Because of potential differences in style between private correspondence and public speeches and published letters, Wheeler divided the known documents into "public" and "private" categories. Ultimately, the Papers of Abraham Lincoln was able to provide the EVL Lab with only 13 public documents and

72 private documents from 1834 to 1842. Because anonymous and pseudonymous contributions to the newspaper should be more akin to known public documents, the paucity of identified Lincoln documents posed a problem for classification from the beginning. To supplement this corpus, the Papers of Abraham Lincoln also provided an additional 36 public documents from the years 1843 to 1858, but because an author's stylome can change over time, we chose not to go beyond 1858 initially.

Wheeler also began the tedious task of reviewing each issue of the *Sangamo Journal* between 1834 and 1842 for anonymous and pseudonymous editorials and letters to the editor. He identified the date, page, column, title, and pseudonym, when available, for each article. When complete, his list included 1,203 anonymous and pseudonymous contributions to the *Sangamo Journal*.

The initial plan was to extract the text of each of these identified articles from the vendor-created OCR files for each issue. However, the project had to abandon this approach when the results of the OCR revealed that there were so many "recognition" errors that correcting the OCR text would be more time-consuming (and probably still less accurate) than typing each identified article afresh. Heroic efforts by a small group of volunteers and an intern resulted in freshly typed copies of all 1,203 articles in fewer than six weeks.

## Initial Analysis and Additional Texts

As an early attempt to test the analysis aspects of the project, the Papers of Abraham Lincoln sent the texts of ten documents that had been identified by scholars as *likely* to have been written by Abraham Lincoln to the EVL Lab. Early analysis using JGAAP identified only four of the ten documents as likely to have been authored by Lincoln. The remaining six were presumably not written by Lincoln or at least the result of the analysis was "not sure."

These mediocre results led the EVL Lab team to request known documents written by other "distractor" authors, such as fellow Whigs in Springfield who also wrote articles for the *Sangamo Journal*. These additional documents would provide known linguistic stylomes for other authors who likely contributed some of the anonymous and pseudonymous pieces to the newspaper during this same period. Wheeler identified an additional 219 documents written by known authors besides Lincoln to improve the analysis. Among the authors were Congressman and Lincoln's first law partner John Todd Stuart, Kentucky Senator and perennial presidential candidate Henry Clay, Whig Illinois Governor Joseph Duncan, local physician Anson G. Henry, and *Sangamo Journal* editor Simeon Francis. The project's volunteers also prepared transcriptions of these documents.

The EVL Lab also requested more examples of Abraham Lincoln's writings as well. The Papers of Abraham Lincoln provided the EVL Lab with an electronic version of the entire text of *The Collected Works of Abraham Lincoln*, the standard compilation of Lincoln's writings, which includes texts of more than 6,700 documents. The electronic text included neither the editorially supplied titles nor the annotations. However, the resulting text is not completely Lincoln's words. For example, the text of the Lincoln-Douglas Debates includes Douglas's words as well as Lincoln's. Lincoln signed many pieces of diplomatic correspondence during his presidency, but State Department clerks undoubtedly composed the letters. From these documents, the EVL lab selected an additional 410 Lincoln documents to add to the 121 documents (public and private from 1832-1842, and public from 1843-1858) already provided by the Papers of Abraham Lincoln.

## Final Analysis

By early January 2013, the Papers of Abraham Lincoln sent text files of 1,203 anonymous and pseudonymous texts and an additional 219 "distractor" texts by other known authors to the EVL Lab at Duquesne University for analysis.

To analyze each text, the EVL Lab team "canonized" it by eliminating uninformative or noisy variation. For example, the program removed line breaks, punctuation, and capitalization. Second, each text was broken into "events." For example, an analysis of authorial vocabulary requires that each word is an event. An analysis of strings of text requires that each set of *n* words or characters is a separate event. Third, the program classified texts based on their distance from each other in a vector space. Originally, the plan was to calculate the center of the Lincoln texts in this vector space, then classify those texts close to the central tendency of Lincoln's style as presumptively Lincoln and those farther away as presumptively by others. However, the style of the known Lincoln texts did not lend itself to this method.

The EVL Lab team then turned to a simpler classification system, comparing the centroid of the Lincoln texts with the centroid of the "distractor" texts by other known authors, then grouping unknown documents with their closer centroid. They tested a large variety of classifiers on the data with a variety of canonizers. They tested all levels of word combinations up to 5 words and all levels of character combinations from 1 to 15 inclusive. The best method was based on a specific set of canonizers and event sets of 16 characters and correctly classified 98.7 percent of the training documents. Another method with the same canonizers and event sets of 13 characters achieved 92.6 percent accuracy on the training documents.

## Analysis Results

Employing these two methods on the 1,203 documents, both tests performed by the EVL Lab identified as Lincolnian only 34 anonymous/pseudonymous contributions—a very disappointing result. Although the results do include at least one document identified as Lincolnian by other contemporary sources and others signed by several individuals including Lincoln, they do not include the four documents identified from the group of ten presumptively Lincoln documents used as an early test of method. They do include four other documents from that list.

The list of 34 documents includes some intriguing pieces, including two from Vandalia, reporting on the actions of the state general assembly, of which Lincoln was a member. However, other reports from Vandalia appear in the *Sangamo Journal* during the same sessions of the legislature; they were not on the list. Some positive results are simply inexplicable, including three sentences that introduce a speech by John Quincy Adams in Congress. Another brief article of two sentences asks the editor's opinion on "the constitutionality of an unnaturalized foreigner holding the office of Commissioner of public works."

In sum, the results of this experiment are not sufficiently consistent or authoritative for the Papers of Abraham Lincoln to withdraw from its corpus any documents not identified or to add to its corpus any documents that were identified.

Why did the computational methods fail to identify more documents as likely to have been written by Lincoln?

There are several possibilities, but the most dramatic, posited by Juola and his team, is that Abraham Lincoln may have been a "goat." In other words, Lincoln's style defies simple analysis. Some linguists have divided people into four distinct groups, including goats. Goats "are those speakers who are particularly difficult to recognize. Goats tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections." Lincoln may be a goat or "exhibit a continuum of goatish characteristics," perhaps not out of character for a rising politician who was writing under a pseudonym.[1]

Other possibilities have to do with the quality of the data. The length of the document affects classification, and some of the anonymous/pseudonymous letters and editorials are quite short, as are some of the known Lincoln writings. More problematic is the question of which documents attributed to Lincoln were actually written by him. If the additional 410 documents from *The Collected Works of Abraham Lincoln* were selected indiscriminately, they may include presidential documents authored by others but signed by Lincoln, reports of Lincoln's words mediated by reporters, and both Lincoln's and Douglas's words from the Lincoln-Douglas Debates. One solution to this problem might be to restrict the known Lincoln documents to those in Lincoln's hand that are not obvious copies of another text. Similarly, the analysis treated all "distractor" authors as an undifferentiated mass of "not Lincoln," but breaking these documents out by author may help refine the training data in helpful ways.

Two additional problems with the data stem from the complexity of the documents analyzed. First, editorials and letters to the editor, which form the preponderance of anonymous and pseudonymous documents analyzed in this project, are mediated texts. They can be altered substantially by the publisher or typesetter. Second, these documents often quoted entire documents written by others (which were usually excluded from the transcriptions in this project), and the

[1] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds, "Sheep, Goats, Lambs, and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," NIST tech report, 1998.

beginning and end of such compound documents is sometimes difficult to discern in newsprint. A careful review of the anonymous/pseudonymous texts might improve their quality as well.

## Project Results

Although this project did not achieve the goal of identifying by an objective computational linguistic standard which anonymous/pseudonymous documents Lincoln wrote, the process has been a positive one for at least three reasons.

First, the project did digitize the *Sangamo Journal* from 1834 to 1842 and make it freely available online through the Illinois Digital Newspaper Collection at the University of Illinois at Urbana-Champaign (http://www.library.illinois.edu/dnc/idnc). This project also encouraged the Abraham Lincoln Bicentennial Foundation to award a grant of $32,745 to digitize additional issues of the *Sangamo Journal / Illinois State Journal*. That effort has yielded the digitization of issues from the beginning of the newspaper's publication in 1831 through 1833 and from 1843 through 1853 thus far. Those issues will soon be available through the Illinois Digital Newspaper Collection, and that grant should allow the digitization of most of the 1850s issues. An additional grant will be necessary to complete the digitization through 1865. In addition the University of Illinois Library is considering a new interface for this collection that will enable crowd-sourced corrections to the OCR text, a feature that will improve word-searching over time, as more and more individuals correct the texts underlying the images.

Second, the project identified and created electronic texts for 1,203 anonymous and pseudonymous articles published in Lincoln's hometown Whig newspaper from a period when he was frequently contributing such pieces. Other authorship attribution techniques may be able to extract useful and authoritative classifications from these texts. The project has begun preliminary discussions with other authorship attribution specialists to see if different techniques can yield better results from this corpus.

Third, the staff of the Papers of Abraham Lincoln have learned much about the possibilities and limitations of computational linguistic analysis and even learned some of the lingo. For example, in interdisciplinary work, both disciplines must be clear on what professional norms govern their respective disciplines and what each should expect from the other. The Papers of Abraham Lincoln is much better equipped to work with linguists and computer scientists to undertake similar analyses in the future.

## Reflections

One of the principal lessons to be learned from this project is to develop clear expectations between partnering institutions and partnering disciplines. The analysis portion of the project suffered from differing expectations and poor communication. After the Papers of Abraham Lincoln turned the texts over to the EVL Lab in January 2013, a long period of silence ensued. Not until early May did the project receive any report on the results of the linguistic analysis. That report was a mere two pages long and offered little analysis of the results or discussion of why the results were so meager, other than that outlined above.

The linguistic analysis proved to be a "black box" from the standpoint of the Papers of Abraham Lincoln. Despite a request in early July for clarification of the methods and results, the EVL Lab team has yet to respond, three months later, despite repeated attempts to gain additional information. How were additional Lincoln texts selected? Were they chosen chronologically or by genre (e.g., public speeches vs. private correspondence)? Were texts by the distractor authors separated into separate groups by author or treated as undifferentiated mass? Is the order of the list of "presumptively Lincoln" documents in the report significant? Are those higher in the list more likely to have been written by Lincoln? N-grams of 13 and 16 characters proved most accurate. What 13- or 16-character strings within each document led the analysis to categorize it as Lincolnian? For known Lincoln documents, is genre more important than chronology, or vice versa? Would including only documents written in Lincoln's hand (though far more often private than public) yield a more reliable Lincoln stylome? The Papers of Abraham Lincoln never received answers to any of these questions.

It would have been far more helpful to have communications throughout the analysis phase. Having the computational linguists provide more information about their analysis—the pitfalls, the unexpected results, the alternative methods employed, the limitations of the methods, etc.—would have allowed the historians to bring their knowledge to bear on the subject. Instead, the project devolved into a black box analysis with no further input from the historians.

From this experiment, it remains difficult to tell whether the methods of authorship attribution are insufficient for this task or whether both the data and the methods applied to the data could be improved by additional refinement. Until we know the answer to that question, the Papers of Abraham Lincoln will continue to ask, "Is that you, Mr. Lincoln?"