Summer 2009

An Inquiry into Authorial Attribution

Sarah Potvin
Center for Digital Research in the Humanities, University of Nebraska-Lincoln, spotvin@library.tamu.edu

Sarah Potvin
Center for Digital Research in the Humanities
Summer 2009

An Inquiry into Authorial Attribution.

Executive Summary.
At attempt to establish the authorship of several unsigned pieces that appeared in the Pittsburgh magazine *Home Monthly* circa 1896-1897 using computational attribution techniques did not deliver robust results. The report that follows examines the computational methods advocated by professors David Hoover, Patrick Juola, and Matthew Jockers. It emphasizes the importance of contextualizing any statistical attribution results with traditional scholarship, as demonstrated by both Hoover and Jockers, and calls for collaboration between humanists and statisticians in interpreting and integrating statistical data.

Goal.
Survey current techniques used by authorial attribution specialists and assess how such techniques might be employed on this or similar projects.

Relevance of authorial attribution to CDRH.

- General interest in exploring authorial attribution.
- May be helpful to include in Cather Journalism Project.
- Use of electronic tools to supplement or challenge more traditional research methods in the humanities.

"Why care about authorship attribution? And, especially, why care about statistical methods for doing authorship attribution? Because 'style,' and the identity underlying style, has been a major focus of humanistic inquiry since time immemorial. Just as corpus studies have produced a revolution in linguistics, both by challenging long-held beliefs and by making new methods of study practicable, 'non-traditional' stylometry can force researchers to re-evaluate long-held beliefs about the individualization of language. The practical benefits are also immense; applications include not only literature scholarship, but teaching, history, journalism, computer security, civil and criminal law, and intelligence investigation. Automatic authorship attribution holds the promise both of greater each of use and improved accuracy."

-Patrick Juola, *Authorship Attribution*, 322.

"The immense power and ultimate limitation of statistical analysis are one and indisseverable: statistical analysis deals in probabilities and not in certainties. If certainty is indeed our goal, we must look for it elsewhere."

-John F. Burrows, "Questions of Authorship: Attribution and Beyond," 26.

Scholars of note.

- David L. Hoover (Department of English, NYU).
- Patrick Juola (Department of Math & Computer Science, Duquesne University).
- Matthew L. Jockers (Department of English, Stanford University).

David Hoover: Delta.
david.hoover@nyu.edu
Homepage: https://files.nyu.edu/dh3/public/
Facts about Delta:

- Developed by John F. Burrows.
- Delta promises to be accessible to "anyone interested in textual difference."
- Designed for open problems, to attribute anonymous texts to one in a set of possible authors.

Hoover uses Excel spreadsheets to perform text analysis. Spreadsheets are available on his web site at https://files.nyu.edu/dh3/public/The%20Excel%20Text-Analysis%20Pages.html; recently-updated (circa 2009) instructions for using the spreadsheets may be found at: https://files.nyu.edu/dh3/public/UsingtheDeltaCalculationSpreadsheets.html.

Hoover has written extensively on the use of Delta to analyze authorship. Delta was introduced by John F. Burrows as "a measure of textual difference that can be used effectively in authorship attribution investigations and stylistic studies. It is especially useful in investigations involving relatively large numbers of texts and authors, situations where other methods would be unwieldy." Hoover writes that the "user can now enter a raw word frequency list for a whole corpus of texts and raw word frequency lists for each single text, and the spreadsheet can prepare the word lists for processing and perform a series of Delta analyses based on different numbers of frequent words, collect and analyze the results, and group the data for graphing. Putting all the functions except creating the word lists themselves into one self-contained package that can operate automatically should allow anyone interested in textual difference to perform multiple experiments easily, even if they are not particularly proficient in computer programming or computational stylistics."[1]

What Delta does:
Delta was designed to suggest which of a set of primary authors is likely to have written an anonymous text known to be authored by one of them. The user must input a set of primary samples of text by known authors and a secondary set of texts

[1] David L. Hoover, "Using the Delta Calculation Spreadsheets," 2009. Accessed at https://files.nyu.edu/dh3/public/UsingtheDeltaCalculationSpreadsheets.html, July 21, 2009.

or samples, some by authors in the primary set (as test texts) and others of unknown/disputed authorship. Burrows describes it as a "primitive form of cluster analysis, for use in those open inquiries that have always given us most difficulty" (26).[2]

Hoover notes that, when test running a set of texts, if Delta is successful in attributing the known texts to their correct authors, the user might confidently accept the results of the attributions.

Available parameters for Delta: the user may choose to remove personal pronouns or other words from analysis; enter a percentage for an automated culling process, which removes words that are very frequent in the entire corpus only because they appear frequently in a single text; select number of words to process; select number of words to include in analysis.

Hoover provides some research findings to guide setting of parameters, suggesting that "increasing the size of the word list almost invariably improves the accuracy of authorship attribution methods based on frequent words (Hoover, 2001, 2002, 2003a, 2003b, 2004a, 2004b, 2007, 2008; Hoover and Corns 2004). Because of these studies, the Delta Spreadsheets allow the processing of up to the 4000 most frequent words. In spite of the fact that nearly all of these are content words, the accuracy of analyses of large texts like novels often improves steadily up through the 2000 most frequent words, and sometimes higher, and often remain at their most accurate levels all the way up to the 4000 most frequent words (at which point the selected words typically account for more than 90% of all the words in the texts)."

Hoover has done interesting work with stylometry to chart authors' careers. He specifically observed of Cather: "A comparison with Charles Dickens and Willa Cather shows that Dickens's early and late novels tend to separate, but do not fall into such neat groups as [Henry] James's do, and that Cather's novels form consistent groupings that are not chronological. These authors seem not to have experienced the kind of progressive development seen in James."[3]

---

[2] John Burrows, "Questions of Authorship: Attribution and Beyond," A Lecture Delivered on the Occasion of the Roberto Busa Award, ACH-ALLC 2001, New York, *Computers and the Humanities* 37 (2003): 5-23.

[3] David L. Hoover, "Stylometry, Chronology and the Styles of Henry James," Digital Humanities 2006, Single Sessions, p. 79.

Patrick Juola: JGAAP 4.0.
juola@mathcs.duq.edu
Homepage: http://www.mathcs.duq.edu/~juola/
Facts about JGAAP:

- JGAAP is a "Java-based, modular, program for textual analysis, text categorization, and authorship attribution."[4]
- Advocates for support vector machines, linear discriminant analysis, and *k*-nearest neighbor in a suitably chosen space (either by cross-entropy or *n*-gram distance) as methods of analysis.
- JGAAP incorporates three phases:
    - Purely mechanical canonicization (normalize case, spelling, etc.)
    - Identification of event set or feature space (such as letters, words, parts of speech, function words, etc.)
    - Selection of analytic technique and analysis of event/feature set[5]
- JGAAP 4.0 incorporates "nearly 20 different analytic methods (including eight different distance-based nearest-neighbor algorithms), more than 20 different event set and models ranging from character- and word-based N-grams to reaction times, and several different preprocessors incorporating a wide variety of different document types including remote (Web-accessible) files and text extraction from different formats. We estimate that JGAAP is capable of performing more than 20,000 different types of analysis for authorship attribution or similar text classification tasks, with more being added as development continues."[6]

JGAAP 4.0 is available for download at www.jgaap.com.

Juola, notably, is the developer of JGAAP and the author of *Authorship Attribution*, a Foundation and Trends in Information Retrieval monograph dedicated to explaining attribution methods broadly and his own work and philosophy in particular. The following text includes frequent reference to the monograph, as it provides the underpinning for Juola's approach to attribution via JGAAP.

---

[4] "Main Page," www.jgaap.com.
[5] Patrick Juola, *Authorship Attribution*, Foundation and Trends in Information Retrieval, v. 1, no. 3 (2006): 319.
[6] Patrick Juola, John Noecker, Jr., Mike Ryan, Sandy Speer, "JGAAP 4.0—A Revised Authorship Attribution Tool," Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park; June 22-25, 2009 (Maryland Institute for Technology in the Humanities), www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf, 357.

Juola acknowledges the various pitfalls of automating authorial attribution, as he has sought to do with JGAAP.[7] Among the methodological issues associated with doing so, he counts the coincidence of a proliferation of attribution techniques and a reluctance to cross-validate results: "The ease of writing computer programs to implement whatever techniques are found provides a seductive path to misuse and to unwarranted claims of accuracy."[8] Further, he recognizes that authorship is itself a complex issue, as published works depend on the input and alterations made by editors, typesetters, and printers, as well, who incorporate non-authorial material into any given text.

Authorial attribution is a theoretically well-founded practice, as the authorial fingerprint might be attributed to linguistic, psychological, and/or neurological features of the writing process. Some ideas—such as the one that authors' "fingerprints" change as they mature (as advanced by Hoover's examination of various authors)—remain theoretically controversial. Juola claims of computational attribution: "it works." That is, "attribution based on the mathematical and statistical analysis of text, can identify the author of a document with probability statistically better than chance." But researchers must approach the data they generate and the problem they define carefully.[9] Juola suggests that texts inputted to JGAAP should resemble, as much as possible, the original texts produced by the author's hand, and "the set of candidate authors must be chosen as carefully and rationally as possible, and the set of candidate writings must also be chosen with equal care and rationality," in order to control for other differences not attributable to authorship, such as genre or period.[10]

JGAAP offers many options for selecting event vectors and performing statistical analysis, allowing the user to experiment with approaches and combinations rather than prescribing a particular path.

This flexibility is a strength of the tool in the hands of the well-informed user but requires that the neophyte user thoroughly and separately investigate the available techniques (and, indeed, triangulate results!). At this time, JGAAP has no users manual, though Juola responds quickly to e-mails soliciting advice or help with the program. In *Authorship Attribution*, he offers an overview of the efficacy of some combinations of features and analytic techniques:

> Methods using a large number of features seem to outperform methods using a small number of features, provided that there is some method of weighting or sorting through the feature set. In particular, simple statistics of a few simple features, such as average

[7] Patrick Juola, *Authorship Attribution*, Foundation and Trends in Information Retrieval, v. 1, no. 3 (2006): 233-334.
[8] Juola, *Authorship Attribution*, 246.
[9] Juola, *Authorship Attribution*, 317.
[10] Juola, *Authorship Attribution*, 319-20.

> word or sentence length, does not usually produce accurate enough results to be used. There appears to be general agreement that both function words (or sometimes merely frequent words) and POS tags are good features to include. Methods that do not use syntax in one form or another, either through the use of word *n*-grams or explicit syntactic coding tend to perform poorly. Other feature categories are more controversial."[11]

In a Digital Humanities 2009 abstract, Juola et al. indicate that:

- "Symmetric ('commutative') distance-based methods tend to outperform asymmetric ones."
- "Linear classifiers such as LDA tend to outperform nonlinear classifiers despite the apartment oversimplicity of the underlying model."
- "Character-based methods tend to outperform word-based ones for authorship attribution in Chinese."
- "Both cosine distance (normalized dot product) and simple event-based Kullback-Leibler divergence tend to be the best-performing methods for distance-based nearest-neighbor methods."
- "The seminal word list of Mosteller and Wallace does not generally perform well for texts other than the Federalist Papers."[12]

Juola (2006) reports that support vector machines, linear discriminant analysis (LDA), and *k*-nearest neighbor in a suitably chosen space (either by cross-entropy or *n*-gram distance) methods of analysis were heavyweights in a recent Ad-hoc Authorship Attribution Competition, with support vector machines emerging as "the leading contender for 'best performing analysis method for any given feature set."[13]

By 2009, Noecker, Jr., et al. found that Cosine Distance, the process of using "a normalized dot product as a nearest neighbor algorithm to assign authorship tags to unknown documents… outperforms many more complicated techniques and is especially well suited to the feature set we chose."[14] Cosine distance classification was found to rival—and even outperform—support vector machines.[15]

---

[11] Juola, *Authorship Attribution*, 320.
[12] Juola, Noecker, Jr., Ryan, Speer, "JGAAP 4.0—A Revised Authorship Attribution Tool," 358.
[13] Juola, *Authorship Attribution*, 321.
[14] John Noecker, Jr., Mike Ryan, Patrick Juola, Amanda Sgroi, Stacey Levine, Benjamin Wells, "Close Only Counts in Horseshoes and… Authorship Attribution?" Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park; June 22-25, 2009 (Maryland Institute for Technology in the Humanities), www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf, 380-1.

Despite the seeming efficacy of methods such as Cosine Distance classification and support vector machines, other approaches are used for distinct reasons. Juola recognizes that "Understandability remains a major problem. One reason that PCA and similar algorithms remain popular, despite their modest effectiveness, is that the reasons for their decisions can easily be articulated."[16]

Juola concludes his report on the field of authorship attribution with a caveat about the disorganization of the field and returns to the issue of understandability that, in my inquiry into attribution, rings especially true and important:

> The new level of computer support will trigger new levels of understanding of the algorithms. Although some efforts … have been made to explain not only that certain models work, but to why they work, most research to this date has been content with finding accurate methods rather than explaining them. The need to explain one's conclusions, whether to a judge, jury, and opposing counsel, or simply to a non-technical PhD advisor, will not [sic] doubt spur research into the fundamental linguistic, psychological, and cognitive underpinnings, possibly shedding more light on the purely mental aspects of authorship.[17]

In short, understandability and the theories that underpin successful statistical methods remain elusive goals in computational attribution. My own difficulties with JGAAP point to the potentially haphazard manner in which the tool might be used, a misuse directly linked to the lack of explanation for *correct* or explicable usage of attribution tools and techniques. Juola has expressed his concern that the tool be

---

[15] See John Noecker, Jr., and Patrick Juola, "Cosine Distance Nearest-Neighbor Classification for Authorship Attribution," Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park; June 22-25, 2009 (Maryland Institute for Technology in the Humanities), www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf, 208.

[16] Juola, *Authorship Attribution*, 321. Burrows writes of PCA as "the first port of call in computer-assisted studies of authorship [citing David Holmes]. … put to good use… by a growing number of scholars. …The fact that pca displays the phenomena most responsible for a given outcome means that the evidence is more transparent than in the methods mentioned above [e.g. artificial neural networks, discriminant analysis, cluster analysis]. And, especially with the shrewd use of control-specimens evident in some recent studies, pca can yield extremely accurate inferences" (8). He notes that pca is "not intrinsically a test of authorship but only of comparative resemblance," best suited for "the middle stages of the game" (8).

[17] Juola, *Authorship Attribution*, 322.

misused, referring in 2006 to the "Faustian bargain" incurred in making the application widely available.[18]

JGAAP stands as an intriguing possibility and prototype, but its claims of being well-suited to amateur users are unrealistic, given the disorganization of the field of attribution and disputes over the efficacy and understandability of methods. Juola sees the importance of JGAAP in serving a transitional role, claiming that it and similar programs "will help support the idea of standardized testing of new algorithms on standardized problems, and the software programs themselves can and will be made available in standard (tested) configurations for use by non-experts."[19]

[18] Juola, et. al., "A Prototype for Authorship Attribution Studies," *Literary and Linguistic Computing* 2006.
[19] Juola, *Authorship Attribution*, 322.

Matthew Jockers
mjockers@stanford.edu
Homepage: http://www.stanford.edu/~mjockers/cgi-bin/drupal/

Jockers et al.'s study of the *Book of Mormon* is notable for its extensive documentation of the authorship controversy that has attended the work since soon after its publication in 1830.[20] The study documents a whole range of efforts to establish authorship of the *Book*, including 1830s theories of authorship pinned on witnesses' testimonies of similarities to other authors' works, stylometric studies deploying multivariate, cluster, and classification analysis in the 1980s, and 1990s-era investigations relying on multivariate measurement of vocabulary richness. This documentation provides an impressive account of a longstanding interest in the authorship of the *Book*, of the various methods employed in service of attribution, and of the potential for error encountered by researchers who subjectively chose and grouped passages from the work. Jockers et al. criticize and frame the whole history of attribution attempts for the *Book*.

The *Book* study claims to incorporate "a new approach that differs from past work both in source selection and methodology," using the entire corpus of the text, grouped by chapter and compared with texts by five authors with "known or alleged connections to the *Book of Mormon*" and two period authors as controls.[21] The study uses both delta and nearest shrunken centroids (NSC) as attribution methods. NSC, a "statistical technique for classification in high-dimensional settings," was "developed to assist cancer diagnosis by classifying patient tumor samples into cancer subtypes based on gene expression measurements."[22] Applied to the *Book of Mormon*, "NSC assigns a probably that each potential author wrote each…chapter, just as it would assign a probably that a tissue sample manifests a particular cancer sub-type."[23] The study finds that NSC is a more effective method of analysis that delta,[24] offering "a lower cross-validation error rate… a lower rate of false positive assignments, and a probability-based output that enabled in-depth interpretation of the results, including speculation regarding possible connections between candidate authors."[25]

Based on the results of the analysis, the study finds "signals" of authorship by each of the five potential authors and charts relative and significant authorship. Statistical results are reported alongside literary inquiry; for example, the study reports that signals of a particular author were discovered in a particular section

---

[20] Matthew I. Jockers, Daniela M. Witten, Craig S. Criddle, "Reassessing authorship of the *Book of Mormon* using delta and nearest shrunken centroid classification," *Literary and Linguistic Computing* 23:4 (2008): 465-91.
[21] Jockers et al., 469-70.
[22] Jockers et al., 470.
[23] Jockers et al., 470.
[24] Jockers et al., 475.
[25] Jockers et al., 482.

but cautions that, if that author "had a direct hand in the authorship of the *Book of Mormon* it was likely a lesser one. It is more likely that his primary role was editorial given both the historical and stylometric data."[26] The use of NSC allows for a more nuanced approach to authorship and influence, as befits an analysis of a work as complicated and contested as the *Book of Mormon*.

[26] Jockers et al., 479.

Reflections on using JGAAP to evaluate unsigned articles from Home Monthly*.*

In his 2004 doctoral dissertation for the English Department at Duquesne University, Timothy W. Bintrim describes the contributions of Willa Cather scholar John P. Hinz, who, working to unearth Cather's writings from her years in Pittsburgh, "did much reading in period magazines and newspapers and identified some fifty articles and reviews that Cather wrote under thirteen pseudonyms in addition to three variations of her own name," discoveries which he published in a 1950 article in *The New Colophon*.[27] Bintrim writes: "Hinz predicted that identifying Cather's *unsigned* writings for the Home Monthly and other Pittsburgh periodicals 'will offer someone an interesting, if vigorous exercise in literary detection'—a prediction the present researchers find accurate even fifty years later."[28] Bernice Slote, whom Bintrim calls "the foremost authority on Willa Cather's early writing,"[29] and William Curtin have, in compiling bibliographies of Cather's writings, implemented methods for attributing her anonymous or pseudonymous published works. Bintrim adopted their method of cross-referencing to link rewritten pieces publishing in the "sibling" journals *Home Monthly* and *National Stockman and Farmer*.[30]

Initially, it seemed that the unsigned, unattributed articles Cather Archive Editor Andy Jewell had in XML format were the perfect material for JGAAP. The articles were important to Cather scholarship, in that they included biographical material and other insight into Cather's editorship and were clearly of interest to scholars such as Bintrim. Working with CDRH assistant professor Brian Pytlik Zillig, I applied a simple XSLT script to strip encoding from the XML files of the unsigned articles as well as pieces authored by Cather and saved them as text files. Here was the moment of decision! I carefully uploaded the files into JGAAP, selected my event set and method of analysis, and recorded the results in a spreadsheet. I changed the event set and method several times and continued to record the results.

It soon became clear that JGAAP was attributing all of the files submitted with unknown authors to Cather. I added texts that were clearly not Cather's work, such as Lewis Carroll's *Jabberwocky*, Jane Austen's *Mansfield Park*, and Edgar Allan Poe's *The Fall of the House of Usher* and obtained identical results: according to JGAAP, these texts were all Cather's.

What had gone wrong? I e-mailed Patrick Juola, the developer of JGAAP, to ask his advice. JGAAP 4.0 has no real users manual, just a bare bones wiki and a largely empty "help" section. But Juola proved to be very helpful. He responded quickly and pointed to the problem:

---

[27] Timothy Bintrim, "Recovering the Extra-Literary: The Pittsburgh Writings of Willa Cather," PhD diss., Duquesne University 2004, 12.
[28] Bintrim, 67-8.
[29] Bintrim, 1.
[30] Bintrim, 68.

> You need a set of distractor authors as well. The way JGAAP currently works is to identify the most likely author of a document from among the group of testing authors represented. If there is only one testing author (Cather, in this case), then there is only one candidate for most likely author -- Cather is, as it were, running unopposed.
>
> On the other hand, if you listed Jabberwocky *et al*. with their correct attributions, then it would be making decisions about whether or not Carroll was a more likely author than Cather. This, again, may or may not be meaningful given the tremendous stylistic dissimilarity between the two.[31]

Now we were tasked with finding and applying an appropriate "distractor author" to pair with Cather. The task itself required us to gather more information on Cather's editorial process and her work as an author and editor of *Home Monthly* (Cather edited the magazine from August 1896 to June 1897, for eleven issues). I spoke with Andy Jewell about the problem. Andy showed me images of the magazine and pointed to the short sections he was interested in attributing to (or ruling out as the work of) Cather. Actually looking at these short, editorial pieces (Andy referred to them as "blogs") raised the question of what Cather's role as an editor or author was for these pieces. Andy pointed out that there was a blurry line between these roles, as the owners of the paper might have insisted or dictated that Cather write or include certain pieces.

Who would be most appropriate as distractor authors? Andy suggested that authored pieces that appeared in the magazine, such as fashion columns, might be well suited to the role. Alternately, we might draw on similarly situated short pieces that appeared under the direction of the editor who replaced Cather, who, seemingly, would have worked with similar authors. But what authority did these women wield as editors? Did their involvement extend to rewriting and "authoring" pieces submitted by others?[32] It was at this point that, seeking to provide more information about the role that Cather occupied as an editor, Andy suggested and provided Bintrim's dissertation, "Recovering the Extra-Literary: The Pittsburgh Writings of Willa Cather." Bintrim's dissertation indicated that the managing editor who succeeded Cather was Mary Beers (Mrs. T.E.) Orr.

I decided to include other authors whose pieces appeared under their bylines in *Home Monthly* during Cather's editorship as distractors, in an attempt to reframe the authorship question as a problem digestible to JGAAP. In order to obtain text

---

[31] Patrick Juola to Sarah Potvin, e-mail, "JGAAP Advice?" July 6, 2009.
[32] For a broader look at the role of women editors in this period, see Ellen Gruber Garvey, Foreword to *Blue Pencils & Hidden Hands: Women Editing Periodicals, 1830-1910*, ed. Sharon M. Harris and Ellen Gruber Garvey (Boston: Northeastern University Press, 2004): xi-xxiii.

samples from other authors whose work appeared under bylines in *Home Monthly*, I found TIFF files of the magazine, which I then OCRed using OmniPage, corrected, and saved as text files.

In the course of selecting distractor authors and laboriously preparing text files of their work, however, I lost faith in the usefulness of this exercise. It appeared that JGAAP was simply not the right tool for this project, as it required the use of distractor texts that, in this case, seemed irrelevant to the original question. Asking JGAAP to determine, essentially as a yes-or-no question, whether an unsigned text that appeared in a magazine edited by Cather was in fact the work of Willa Cather (which, given her editorial role, it most likely was—to some degree) or that of *Home Monthly* fashion columnist, Mildred Beardsley (which, given her limited role as a contributor with a byline, it most certainly was not) seemed like at best a deeply flawed and at worst a misleading and false inquiry.

An appropriate distractor author remained elusive, as no contributor appeared to be a real candidate for authoring the unsigned articles. An inherent difficulty of subjecting the unsigned *Home Monthly* texts to computational analysis was the problem's failure to conform to the ideal closed set required by JGAAP. As Burrows writes, the current role of computational techniques is "strictly ancillary to the traditional work of scholarship, corroborating or casting doubt on the product of other sorts of evidence but rarely opening fresh ground." As such, those involved with computational attribution "should confine [themselves] … to cases where the choice lies within a narrow range of well-matched sets and we should proceed with only two authors' texts at a time."[33] The question of authorship surrounding the unsigned *Home Monthly* texts simply did not conform to this strict standard.

Further, JGAAP's provision of a yes/no answer to attribution queries, while posed as a user-friendly feature, belied the hidden, complicated process of computation. The tool presents computation attribution as almost a non-technical question, but unease accompanies use by the inexpert. Ultimately, however, JGAAP was incompatible with our Cather attribution question because the question itself was inexact. Burrows cautions that "the appropriate moment to revert to the literary and historical evidence must always be considered."[34] In this case, the appropriate moment had not yet been reached.

[33] Burrows, 10.
[34] Burrows, 26.

Appendix.

Other Attribution Resources.

Hugh Craig, "Stylistic Analysis and Authorship Studies," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Oxford: Blackwell, 2004).

- Contrasts stylistics and authorship studies, "Yet stylistic analysis needs finally to pass the same tests of rigor, repeatability, and impartiality as authorship analysis if it is to offer new knowledge. And the measures and techniques of authorship studies must ultimately be explained in stylistic terms if they are to command assent."
- Explains the increasing dominance of the cognitive function of the author over the post-structuralist "author function": "If the one long-term memory store of human beings works not systematically but by 'casting out a line for any things directly or indirectly associated with the object of our search,' then 'the organisation of memories…reflects the person's own past experience and thought rather than a shared resource of cultural knowledge' and this would imply a 'unique idiolect' for each individual's speech or writing" (Lancashire 1997:178).
- "Statistics depend on structured variation—on finding patterns in the changes of items along measurable scales. It is easy to see that language samples must vary in all sorts of ways as different messages are composed and in different modes and styles. The claims of statistical authorship attribution rest on the idea that this variation is constrained by the cognitive faculties of the writer. …The approach to authorial idiolect through cognitive science and neurology offers it own reinforcement of the notion that different genres are treated differently, and that word-patterns can be acquired and also lost over a lifetime."
- When performing authorship study, more than two author candidates is ideal and multiple approaches allow for triangulation and cross-checking of results.

David V. Erdman and Ephim G. Fogel, *Evidence for Authorship: Essays on Problems of Attribution* (Ithaca: Cornell University Press, 1966).

Harold Love, *Attributing Authorship: An Introduction* (Cambridge: Cambridge University Press, 2002).

A.Q. Morton, *Literary Detection: How to prove authorship and fraud in literature and documents* (New York: Charles Scribner's Sons, 1978).

Frederick Mosteller and David L. Wallace, *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers*, Springer Series in Statistics (New York: Springer-Verlag, 1964).