The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions

Patrick Juola
Evaluating Variations in Language Lab, Duquesne University, Pittsburgh, PA, USA and Juola & Associates, Pittsburgh, PA, USA

Abstract

We propose a possible solution to one of the major weaknesses in the application of authorship attribution—the absence of clear-cut standards for accurate analytic practice. To address this, we propose a specific practice as a possible standard and present four recent cases applying this standard. The key elements of this protocol are the use of an ad hoc distractor set in conjunction with multiple analyses structured as a set of elimination tests. This protocol (or close variants of it) has been used in at least four separate cases across a wide variety of documents and consumers. It is mathematically supported while still being easy to understand. We are confident that the proposed protocol will provide a relatively straightforward and understandable way to reduce controversy regarding stylometric authorship attribution, and thereby increase its uptake and credibility.

Correspondence:
Patrick Juola,
Math/CS Department,
Duquesne University,
Pittsburgh PA 15282
USA.
E-mail:
juola@mathcs.duq.edu

1 Introduction

Authorship is one of the key baseline questions in the humanities. It is through the body of their writings that we know most of the great thinkers of the past. In light of the importance of authorship, how can questions about it be resolved? What are the appropriate standards for decision-making?

For example, in 1827, an 18-year-old Edgar Allan Poe was trying to start a writing career, but was hampered by creditors. He did manage to publish two of his poems, 'The Happiest Day' and 'Dreams', in a weekly newspaper called the North American, but under the initials of Henry (William Henry Leonard Poe), his brother.

In the same newspaper and the same year, 'Henry' also published three short stories. Were these stories actually written by Henry, or were they also works by Edgar Allan, hidden from creditors for the same reason and using the same technique? And how could this question be addressed?

As Collins (2013) put it, '[i]n a past era, any suspicions about Edgar's authorship of these pieces would be dutifully wrapped in supporting quotes and biographical context—and short of finding other documentation, that would be the end of it'. Recent advances in text analysis have allowed a different approach using a statistical analysis of individual quirks of language, the emerging science of stylometry (Juola, 2006; Koppel *et al.*, 2009; Stamatatos, 2009; Jockers and Witten, 2010). Collins applied the Java Graphical Authorship Attribution Program (JGAAP) stylometry tool (Juola *et al.*, 2006, 2009) and obtained a very robust finding: out of eight possible authors including both Henry and Edgar, Edgar was picked as the most likely author of those short stories fifteen times out of fifteen different analyses. Edgar, therefore, was

Digital Scholarship in the Humanities, Vol. 30, Supplement 1, 2015.

doi:10.1093/llc/fqv040

not just a consensus pick, but a unanimous #1 choice.

Based on this, he concluded 'the results are suggestive—but perhaps they are only that'. A degree of scholarly conservatism is admirable, but at what point do suggestions turn into actionable findings, or findings turn into practical certainties? Put bluntly, should these three works not be added to the curated canon of Poe's work? If not, why not?

One of the barriers to the development of new scholarly methods is uptake among the relevant community. In pharmacology, a new drug requires, among other things, a rigorous set of tests to find out not only if it works, but also appropriate ways to use it (e.g. how large a dose, how often, and under what conditions?). The admission of science into law (as evidence) follows similar guidelines, including the need for a relatively well-understood protocol for performing the science. In medicine, the Cochrane Collaboration (Higgins and Green, 2009) defines standards for high-quality studies to be incorporated into systematic reviews to support evidence-based medicine.

Digital scholarship lacks these explicit standards. In this article, we argue that authorship attribution, in particular, is a mature enough subfield that it has implicit standards. We argue further that formalizing these implicit standards into explicit ones can improve both uptake as well as scholarly rigor.

One possible reason for this lack may be neglect by the larger universe. 'The community as a whole tends not to be aware of the tools developed by Digital Humanities(DH) practitioners [. . .], and tends not to take seriously many of the results of scholarship obtained by DH methods and tools.' (Juola, 2008). Part of the reason for this neglect may be epistemological; the types of evidence developed by digital methods and the arguments employed are new and unfamiliar to many humanists. This does not necessarily make them unreliable, but traditional scholars may feel themselves challenged to assess the reliability of any specific argument. I can presumably rely on the scholar who prepared the scholarly edition of the work I am studying (or the biography of its author) without tracking down the original manuscript myself—but upon what can a non-stylometrist rely in assessing a statistical argument about authorship? Prevalidating methods—that is, the establishment of appropriate formal protocols—will help this assessment.

2 Background

2.1 Paradigms and Mature Science

One of the key transitions in the development of a new field, according to Kuhn (1996), is the transition to 'normal science',[1] a state characterized by 'some accepted examples of actual . . . practice—examples which include law, theory, application, and instrumentation together—[that] provide models from which spring particular coherent traditions of. . . research', which in turn enables scholars to build on each other's work instead of reinventing wheels. Contrast this with Kuhn's description of pre-Newtonian optics: 'being able to take no common body of belief for granted, each writer on physical optics felt forced to build his field anew from its foundations' (Kuhn, 1996).

The increasingly tool-rich environment of digital humanities shows this process; while digital humanities is a 'new set of practices, using new sets of technologies' (Borgman, 2009), there is an increasing need for broadly useful tools. '[W]hat use are the digital libraries, if all they do is put digitally unusable information on the web? The digital libraries don't offer a platform for traditional note taking, much less for larger scale analysis, either quantitative or qualitative.' (Borgman, 2009). Borgman, implicitly, is expressing the need and desire for shared 'theory, application, and instrumentation'.

This transition is perhaps particularly evident in authorship studies over the past few decades. The idea that some sort of statistical process can infer authorship dates at least to the 19th century (de Morgan, 1851/1882; see Juola, 2006, for a fuller history) and achieved prominence in the 1960s (Mosteller & Wallace, 1964). However, only recently has research emphasis shifted from 'can authorship be inferred?' to 'what is the best way of inferring authorship?' and the search for best practices. Much of the research, for example, focused on specific documents (often documents of specific interest to the authors) and the development of a new

method tuned to work on those documents, without regard to generalization or comparative accuracy.

One weakness of this approach is the multiplicity of methods proposed in the literature (and hence implicitly validated by peer review); with more than 1,000 feature sets (Rudman, 1998) and similarly large numbers of analysis methods proposed, it is possible to run dozens or hundreds of analyses and select/publish the one that supports a desired point of view. This of course, opens the door to cherry-picking and means that the credibility of a single analysis is low, especially in a situation where there is a strong incentive present.

What is needed, of course, is an understanding of what methods can consistently be relied upon to produce accurate results. Among the first explorations of the question of best practices was the 2004 Ad-hoc Authorship Attribution Competition (Juola, 2004), which presented a standardized set of test problems in a (Text REtrieval Conference) TREC-style evaluation to permit scholars to compare performance across different methods, implicitly moving the discussion to questions of accuracy and applicability instead of capacity.

Since then, research (Juola, 2012b; Juola & Stamatatos, 2013; Stamatatos *et al.*, 2014) has continued to work on this issue, including the use of larger corpora, many more languages, many more genres, and a resulting trend of continuous improvement in accuracy. The current state of the art (PAN 2014) shows empirically that authorship attribution will work with substantial accuracy under a wide variety of conditions, and that it can be applied to a wide variety of documents. Perhaps more importantly, the community has also developed a huge variety of different methods, all of which are known to work with substantial accuracy, but do not rely on each other, and can therefore cross-check each other.

2.2 Communities of practice

In his 2014 Zampolli lecture (Siemens, 2014), Ray Siemens discussed the digital humanities as a 'community of practice', a collection of people drawn together by a common interest and expertise, but also implicitly sharing and learning both core knowledge and specific practices in handling those knowledge. Wikipedia specifically identifies the purpose of such a community as 'to provide a way for practitioners to share tips and best practices'.

This is of particular importance as the DH community is among the best-placed to interpret evidence of authorship, in comparison to a judge, jury, or newspaper reporter that may not be familiar with the nuances of practice. We, collectively, already validate scholarly practices by what we accept into our journals and adopt for our own work. Some of us, however, may not understand our role in this process and may not be attentive to the broader societal implications of what we do. On one hand, this may prevent legitimate new scholarship from being accepted, and on the other hand, the veneer of 'scholarship' and 'expertise' can also produce unwarranted acceptance of shaky assumptions, shoddy methods, or false conclusions—simple reliance on what-has-been-published may permit 'a short-cut decision rule that allow[s] judges to avoid having fully to understand the proffered scientific evidence' (Odgers & Richardson, 1995), substituting instead reliance on a plausible-seeming title page in a reputable journal. It is important to understand, and to express clearly and forcefully, that exploratory scholarship is the search for best practices, and that not everything published becomes such a practice.

It is therefore both necessary and appropriate that the community of practice should publicly acknowledge a collective opinion about the reliability of opinions and, more accurately, of the methods used to get to those opinions. This both enhances uptake of these methods and opinions outside the narrow confines of the digital scholarship specialists, and also provides a shared basis for understanding. This understanding will be useful for helping new members to participate, supporting and sustaining the long-term growth and improvement of the community. The question then becomes—what kind of analyses would we as a community of practice like to see used for significant decisions, decisions that may directly impact the lives and happiness of real people?

2.3 Rules of evidence

We are thus in a position to discuss seriously the idea of best practices and even to provide a rubric

for gauging the reliability of a proposed analysis. Based partly on some of the areas of application for this technology (which, as will be seen, include legal disputes), we propose that the rules of legal evidence are a good basis to use for such a rubric.

Rules of evidence, of course, are specific to a given legal system, but the USA provides a good starting point. Expert evidence in Federal courts is controlled by the Federal Rules of Evidence, and specifically Rule 702, which holds that:

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

(a) *the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;*
(b) *the testimony is based on sufficient facts or data;*
(c) *the testimony is the product of reliable principles and methods; and*
(d) *the expert has reliably applied the principles and methods to the facts of the case.*

Stripped of the legalese, this roughly means that expertise is admissible if it is both useful and reliable. The question of what constitutes a 'reliable principle and method', then, is exactly the question set we have been discussing in the previous subsection: does authorship attribution work, and what are the best practices to maximize its usefulness? Similarly, the question of 'sufficiency' of data is based, ultimately, on questions of sample size and representativeness (Eder, 2010), and the case-specific application hinges in part upon whether the analysis has been cherry-picked or represents a practice that has been shown to be, in general, good.

The controlling case for this kind of testimony is generally considered to be *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993). In this case, the Supreme Court of the United States identified 'scientific knowledge' on a Popperian (Popper, 2002) basis: 'generating hypotheses and testing them to see if they can be falsified' (*Daubert*, 1993).

Daubert, along with various follow-up cases, established a specific set of tests to determine whether or not such evidence should be admissible. It superceded the earlier Frye standard [*Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923)], which held that scientific (and by extension, scholarly) evidence was admissible only when it is 'sufficiently established to have gained general acceptance in the particular field in which it belongs'. This, perhaps obviously, both opens the door to generally accepted pseudoscholarship and also places a substantial burden on new and controversional, but obviously relevant, theories.

By contrast, *Daubert* notes that this is not how scholars themselves operate, and that instead 'science [...] represents a *process* for proposing and refining theoretical explanations about the world that are subject to further testing and refinement' and that therefore 'proposed testimony must be supported by appropriate validation—i.e., ''good grounds,'' based on what is known'. The court therefore suggested criteria—a 'checklist' and not, explicitly, a definitive test—to establish what constituted these good grounds.

Daubert established a five-point checklist, as follows:

- Whether the theory can be (and has been) tested.
- Whether the theory has been subject to peer review and publication.
- The known or potential error rate of the technique.
- Whether there exist standards controlling the technique's operation.
- Whether or not the theory/technique is widely accepted.

Daubert therefore includes *Frye* as one element on the checklist ('a known technique that has been able to attract only minimal support within the community' [cites omitted], may properly be viewed with skepticism) but allows for the possibility of new and controversial theories with a solid body of evidence behind them. While the direct application of this case is confined to the USA, other jurisdictions including both Canada and the UK have suggested adopting *Daubert*-like standards.[2]

Again, we try to focus on the essence instead of the legalese. This standard is essentially a restatement in more formal terms of the anti-cherry-picking argument of the previous section; one cannot simply make up an analysis technique and

present it. However, a theory that has been shown in independent testing (backed up by peer review and publication) to be a reliable method of analyzing data, should be taken seriously, a technique that has been studied to a degree that standards of practice are available even more so. This argues that authorship attribution, with a long history of empirical testing, should be able to create evidence to convince both courts and scholars. At the same time, this also argues more strongly for the discussion and development of such standards by the relevant DH sub-community.

3 Four Case Studies

3.1 The cases

The case of the Poe short stories, described in an earlier section, is an almost classic example of an academic dispute. The documents in question are published documents (and no access is available to the unpublished drafts), the candidate author and his contemporaries are all long-dead, and the stakes of a wrong decision are mercifully low. The second case we wish to discuss (Juola, 2013b) has much higher stakes and is a good example of how authorship attribution can have a real impact on a real person.

In this case (Juola, 2013b), a person (identified only pseudonymously as 'Bilbo Baggins') was seeking asylum in the USA. His claim was based on a set of anonymous newspaper articles he had (or claimed to have) written for an online publisher, articles critical of his home government. If he were returned to this country, he feared persecution, possibly amounting to arrest and torture, for these articles. He was able to offer as supporting evidence a set of other articles, on other subjects, that had been published under his own name. A key question for the immigration court, and the key question for authorial analysis, is whether the author of the undisputed documents was the same as the author of the anonymous, critical, political essays.

A third case, that of *The Cuckoo's Calling*, is not of such immediate personal impact, but is of much greater public interest (Brooks, 2013; Brooks & Flyn, 2013; Juola, 2013a). *The Cuckoo's Calling* is a detective novel, published in the spring of 2013 under the pen name of 'Robert Galbraith'. In July, an anonymous Twitter user announced that Galbraith was really J.K. Rowling, the famous author of the Harry Potter series. *The Sunday Times* was interested enough to approach the author to resolve this issue as a matter of public interest: Was *Cuckoo* written by Rowling?

The fourth problem is that of the Bitcoin documents (Herper, 2014). Bitcoin is the name of an increasingly popular cryptocurrency, an electronic payment system that can be used in peer-to-peer transactions and hence free of the need for government or large corporate networks to facilitate. The original design for the Bitcoin protocol and the original versions of the reference software were written in 2009 by a person using the name Satoshi Nakamoto. However, no actual person has been identified as the author of these documents, despite rampant speculation.

One of the more high-profile suggestions was the 2014 *Newsweek* article that identified a certain Dorian Satoshi Nakamoto as the Bitcoin author. Dorian almost immediately denied this, and Forbes magazine inquired about the possibility of using stylometric analysis to validate or disprove the Newsweek article. Again, this is a matter of substantial public interest, but also could in theory be a matter of litigation, if Dorian chose to sue *Newsweek* (Cohen, 2014; Volokh, 2014).

3.2 Characteristics of the cases

These cases, while very different in detail, share a number of aspects. In fact, there are enough similarities that these could be regarded as a specific subclass of the general authorship attribution problem. Furthermore, this is demonstrably a rather common subclass. Designation of a protocol for looking at this subclass, a protocol approved by the DH community of practice (as discussed above), would be useful.

The first element we note: there is a substantial amount of text available in each case. This applies both to the questioned documents (QD) to be analyzed and to the known documents to be used as an analytic baseline or training documents. Obviously, Rowling not only has seven lengthy Harry Potter novels to her credit, but also another adult crime

novel published under her own name (*The Casual Vacancy*). Poe, of course, was a prolific short-story author with more than fifty stories to his credit,[3] plus poetic works, essays, one novel, and a partial play. Dorian Nakamoto was a professional engineer with a number of technical documents to his credit, and Baggins was a professional newspaper columnist, and so had an extensive back catalog. Similarly, the questioned document was relatively large, and in the case of *The Cuckoo's Calling*, an entire novel. The amount of data necessary is of course an area of active research (Eder, 2010), but was easily met in all these instances.

The second element is a methodological assumption: with the exception of the Bitcoin case, no one seriously discussed or considered the idea of multiple or co-authorship. Even in the Bitcoin case, the theory proposed (by *Newsweek*) was a specific single author. In any of the cases described, collective authorship would fall under the broad group of 'other', and count against the hypothesis (as opposed to a hypothetical question asking, for example, 'was this person involved in writing this document?')

The third, related, element, is that there is a specific single candidate author proposed for the questioned document, whether Poe, Baggins, Rowling, or Dorian Nakamoto. In each case, this author was designated prior to the start of the analysis.

Because of the third element, these problems are formally structured as 'verification' problems, where the question is related to what Koppel *et al.* (2012) have called 'the fundamental question of authorship attribution', to wit, 'are these two documents by the same author?' The verification problem can be contrasted both with the closed-class attribution problem (where the author is known to be one of a small group of people, but there is no primary candidate), and the open-class attribution problem (where the author is believed to be one of a small group, but 'none of the above' is also a plausible candidate).

Finally, one of the most important characteristics is the need for a 'yes' or 'no' answer. Neither the Sunday Times nor US Immigration courts are particularly interested in 'problematizing' the discussion of what authorship means or challenging the notion of the creation of a text; instead there is a factual dispute to be resolved, often at substantial stakes.[4] While the courts are receptive to a certain degree of hedging (the difference, for example, between 'practically certain', 'likely', and 'possible'), it is still ultimately necessary for the answer to 'help' in resolution.

4 A Proposed Protocol

In this section, we describe in general terms the protocol used (with minor variations) to resolve each case described above. This serves two purposes: first, to show how this protocol has been used, and second, to offer the protocol itself to the community for commentary and ultimately (we hope) validation.

4.1 Underlying assumptions

Our first assumption in constructing the protocol is that authorship attribution itself works; if an analysis undertaken by a 'person having ordinary skill in the art'[5] identifies someone as the author of a document, that identification is more likely to be right than wrong. This is of course an empirical assumption, but has been supported by numerous studies (Juola, 2009a, 2012a; Vescovi, 2011) as well as by the published results of many competitive evaluations (Juola, 2004, 2012b; Juola and Stamatatos, 2013; Stamatatos *et al.*, 2014). In the 2014 (Plagiarism Action Network/Conference and Labs of the Evaluation Forum) PAN/CLEF conference, for example, six of the thirteen participants were able to achieve 80% accuracy on a corpus of Dutch essays, using training samples of less than 1,000 words from a single distractor author. Based on this, we assume (for purposes of calculation) a baseline 80% accuracy of authorship attribution methods generally.[6]

A second assumption is that the analysis will produce a rank-ordering of the authors by likelihood (e.g. A is the most likely author, B the next most likely after A, and C still less likely), but not necessarily provide specific probability judgments. Standard analysis methods such as Burrow's Delta (Burrows, 2002; Hoover, 2004; Stein & Argamon, 2006) or other nearest-neighbor analyses (Noecker & Juola, 2009) will produce this sort of result, but direct probability measurements are, at this writing, still an open research problem (DeCarlo, 2013, 2014).

A third assumption is that multiple independent analyses are available. The JGAAP program, for example, provides several tens of thousands of different analyses (Juola, 2009a, 2012a), and the various conferences (Juola, 2012b; Juola & Stamatatos, 2013) similarly show a wide variety of approaches and feature sets. [Rudman's (1998) estimate of more than 1,000 feature sets is relevant here as well.] However, not all of these analyses would be independent. An analysis of the frequency of the fifty most common words using random forests would not be independent of an analysis of that same set of words using nearest neighbor or support vector machines. Similarly, an analysis of the fifty most common words would be very similar to an analysis of the sixty or even 100 most common words. On the other hand, there is no reason a priori to believe that an analysis of common word frequencies would correlate strongly with an analysis of word lengths.

As will be seen, this assumption allows performance to be boosted substantially over the 80% baseline described above. We are specifically agnostic to the particular analysis methods used, as best practices are a moving target, and since 'best' is unique, any collection of several analyses will of necessity use a variety of good-but-not-best practices.

Finally, we assume (perhaps controversially) that the basic attributes such as age, gender, and nationality of the presumptive author are known to the analyst in order to enable rough matching. As will be discussed, this assumption may not be necessary, and further investigation is appropriate.

4.2 Protocol elements

4.2.1 *Ad hoc distractor corpus*

The overall cases are structured as verification problems ('did this person write that document?'), but the current state-of-the-art obtains best results on closed-class attribution problems ('which of these people was most likely to have written that document?'). To address this gap, we follow Koppel *et al.* (2012) and propose the collection of an ad hoc distractor corpus of different works by comparable authors.

Koppel specifically proposed gathering samples from 10,000 separate authors based on a single and largely unrelated genre, in this case, collected from blogs. In practical terms, this may go too far. For certain situations, writers, and genres, the sort of documents gathered by Koppel-like harvesting may be systematically different, and the relevant documents may not be adequately represented by the Koppel harvest. Consider, for example, how a collection of blog posts in modern English would not provide an adequate control sample for Elizabethan. No matter how large the control sample, Marlowe would still probably have written Shakespeare's plays if the alternatives were Maureen Dowd, Paul Krugman, and Mark Bittman (all bloggers for the *New York Times*). But Koppel's point should still hold if gross discrepancies are controlled for.

We therefore propose collecting a corpus of distractor authors matched for time period, language, region, genre, and gender. The exact number of distractor authors at this point is open, but three to seven seems a reasonable range. Perhaps obviously, the discriminative power is greater with more distractors, but the harvesting is more challenging (How many female crime novelists from the 1930s from New Zealand are there? Other than Ngaio Marsh, what distractors are available?).

For the Rowling case, in particular, we worked together with the *Sunday Times* to produce a set of female-authored contemporary British crime novels, consisting of Rowling's own *The Casual Vacancy*, Ruth Rendell's *The St. Zita Society*, P.D. James' *The Private Patient*, and Val McDermid's *The Wire in the Blood*. (One advantage of working with contemporary documents is that clean e-copies are often available at relatively low prices without requiring arduous digitization.) For the Bitcoin analysis, Noecker used a collection of documents of various types and genres written by different people who had been proposed as the real 'Satoshi Nakamoto' (Michael Clear, Neal King, Shinichi Mochizuki, Vili Lehdonvirta, Dorian Nakamoto, Hal Finney and Nick Szabo). The Poe analysis used the two primary candidates (Edgar Allan as well as his brother Henry, in separate analyses) plus six contemporary authors (James Fenimore Cooper, Nathaniel Hawthorn, Washington Irving, George Lippard, John Neal, and William Gilmore Simms). The Baggins case involved Baggins himself

plus five contemporary online political newspaper columnists.

The interpretation of such analyses is fairly simple. If the predesignated author (e.g. Rowling) is suggested as the author of the QD, then this is strong evidence that she is more likely to be the actual author than any of the distractors. On the other hand, if a distractor author, such as Ruth Rendell, is selected, this does not mean that the selected author is particularly likely to be the actual author, since the world is full of people who are not J.K. Rowling, and Ruth Rendell is merely one among many. The interpretation is not symmetric; settling on Rendell rejects Rowling authorship, but does not prove Rendell's.

4.2.2 *Multiple independent elimination tests*

The key insight here is that, quoting Koppel, any given wrong author 'is highly unlikely to be consistently [similar] over many different feature sets'. An author who consistently uses Rowling's preferred set of prepositions, for example, may not use her grammar, and vice versa. Similarly, an evaluation based on the fifty most common words (Binongo, 2003) is likely to produce different results than an evaluation based on the 1,000th through 1,500th most common words, and both are likely to be different than an evaluation based on distribution of sentence lengths. Indeed, even an analysis of the same feature set can be substantially different if the features are weighted differently.

How do we interpret a finding that an author uses Rowling's prepositions but not her verbs? Intuitively, this is no different from learning that a person shares a suspect's eye color but not his height. If the height assessments are accurate, this person is not the suspect despite similarities. To ignore the dissimilarity would likely be to commit a false acceptance error. We can therefore use multiple analyses as a multistage filter, where the proposed author is required to pass each stage in order to be confirmed as a plausible candidate.

This insight can be formalized mathematically as follows:

- If a technique is X% accurate, the chance of it being wrong is $(1-X)$ (e.g. an 80% chance of being right yields a 20% chance of being wrong).
- If two independent techniques are X% accurate, the chance of them both being wrong is $(1-X)\times(1-X)$ or $(1-X)^2$.
- This generalizes. If K different and independent techniques are each X% accurate, the chance of them all being wrong is $(1-X)^K$, which becomes arbitrarily small as K increases.

Thus, using multiple independent analyses will reduce the chance of false acceptance error to as small a value as desired.

Of course, using too acceptance criteria that are too strong can result in false rejection errors. We propose handling these errors by using a relaxed acceptance criterion, and essentially treating the top few candidates as 'successful'. In other words, we accept the possibility that a given test might deliver an erroneous result, but we expect that if the true author is not the most likely candidate, then the true author is more likely to be the next-most-likely than one of the others. This can be continued for as long as needed to establish a high probability of acceptance in the event that the true author is among the candidates.

This again can be demonstrated rigorously. Assume that the correct author is indeed among our candidate set. If our technique is 80% accurate among this set of distractor authors, there is a 20% chance that the most similar author will not be the correct one. However, consider the modified candidate set where the (erroneous) most similar author is no longer considered. In this case (and with suitable independence assumptions), there will also be an 80% chance that the most similar author in this second set will be the correct one (by assumption). The chance of the correct author not being first among either set is 20% times 20%, due to independence. There is thus only a 4% chance that the correct author will not be among the top two in the original set. (This chance drops to 0.8% for the top 3.) Thus we can say with high probability that any author not among the top few most similar candidates has been eliminated as a plausible author.

4.2.3 *The proposed protocol formalized*

We can thus formalize the proposed authorship analysis protocol as follows: Gather an ad hoc collection of authors (our preliminary recommendation is of

three to five candidates) other than the author of interest. Based on the confidence levels desired, run a prespecified number of independent tests of different feature sets to determine which author is most similar to the questioned document on that specific feature set. Any author not in the top few most likely candidate authors (our preliminary recommendation is for the top two) is eliminated as a potential author. If, after enough experiments have been run, the only author not eliminated is the author of interest, his or her authorship of the QD is deemed confirmed.

4.2.4 *Detailed examples*

4.2.4.1 ***The Rowling example enlarged upon, with numbers.*** The Galbraith/Rowling case is instructive. In this case, I was provided a distractor set of three authors, all contemporary female British crime writers, so their writings would be comparable to 'Galbraith's' (Juola, 2013a). Tests were run on four separate feature sets: word lengths, character 4-grams, word pairs, and the 100 most frequent words. For each of these tests, an author was 'eliminated' if she did not appear in one of the top two positions in the ranked candidate set. Of the four authors, Rowling, and only Rowling, was not eliminated by at least one analysis.

We can determine the likelihood of a false negative as follows: Dismissing for a moment the possibility that we had inadvertently hit on the true author as one of our ad hoc distractors, if Rowling was not the author, then any of the four could have been most similar in her use of the most frequent words. Any of the remaining three could have been the second most similar. Rowling herself could have been, with equal likelihood, in first, second, third, or fourth place, and hence would have been in the first two places 50% (half) of the time.

Since the four analyses are assumed independent, the chance of her placing in the first two places on all four analyses is $(50\%)\times(50\%)\times(50\%)\times(50\%)$ or one in sixteen (6.25%). While above the standard 5% *P*-value cutoff for 'significant' results, this is still a relatively low number, enough to provide useful information to the client or consumer. (If textbook 'significance' were needed, a fifth test would reduce the likelihood of a false acceptance to one in thirty-two.)

Conversely, if Rowling (or any other distractor author) were indeed the true author, the analysis below indicates a 96% chance of not being eliminated on any single test. (This number of course relies on the assumed 80% accuracy described above.) The chance of the true author not being eliminated is thus $(0.96)\times(0.96)\times(0.96)\times(0.96)$ or about 85%. In other words, there is about a 15% likelihood of a false rejection error. In summary, we estimate that this analysis had about one chance in sixteen of a false acceptance error, and one chance in seven of a false rejection error.

4.2.4.2 ***The Poe example.*** Collins did not formally define acceptance or rejection criteria in his study. Instead, he used an informal test based on the average rank order of the candidate authors. With six distractor authors, Edgar could have been first, second, third, fourth, fifth, sixth, or seventh, with equal probability. The 'average' rank across multiple studies would be 4.0. (Indeed, Henry's average rank was 3.86, almost exactly what would be expected if Henry were not the author of the works in question.) Edgar's average rank was 1.0. This average-rank test could be formalized using standard tests for rank-order distributions, but the math involved is somewhat complex (Dunn, 1964) and it may be easier to understand the multistage filter interpretation proposed in this article.

Post hoc, we will conservatively allow the top two candidates of the seven to 'pass' each of the analysis filters. With seven potential authors, this yields a pass rate of two in seven (about 28%) for each of fifteen analyses. The chance of Poe, or anyone else, passing all seven filters by chance alone is approximately seven in a billion. Conversely, the chance of Poe managing to survive all fifteen filters (e.g. Poe being rejected at least once) is remote if our estimates of 80% are accurate, but fortunately Victorian fiction is one of the more well-studied domains in authorship, and accuracy is typically much higher. But even a 99% accuracy rate would create as much as a 14% false reject rate. This, however, does not appear to have happened in this case.

4.2.4.3 ***The 'Baggins' example.*** In the Baggins case, five distractors were provided, and only the

top candidate passed, creating a chance pass rate of one in six (roughly 17%). Two analyses were performed (although the independence of these analyses is/was questionable), creating an overall false acceptance rate of one in thirty-six (about 2.8%). Conversely, the chance of a false rejection on one of these tests (assuming 80% accuracy) would be 36%, approximately one in three. This chance of false rejection is probably too high to be relied upon, and a better framework would use more tests and a less stringent acceptance criteria.

One key issue in the Baggins case was the language barrier. Unlike the other cases described here, the Baggins case was not in English, and in fact, not in a language for which extensive analysis of authorship attribution accuracy has been done. (To protect Mr. Baggins' identity and in the interests of his personal safety, the actual language of the documents will not be revealed here.) In the case of the Poe documents, we have a huge body of work to draw upon both to determine the best practices and to estimate reliability. One robust finding (Juola, 2009b; Hasanaj *et al.*, 2014) in lesser-studied languages is that there is a high correlation in performance across languages. Therefore, of two methods, the one that works better in English is also likely to work better in Russian, Hebrew, or Tagalog. (Given that the bulk of the research is done by English-speaking researchers on English corpora, this is little more than common sense. It would be surprising if a method that worked badly on a language familiar to the designer suddenly started working well on a new and unfamiliar language.) Unfortunately, 'better' does not establish standards of performance, and our estimate of 80% accuracy in the Baggins case is little more than a guess.

4.2.4.4 *The Nakamoto example.* The Nakamoto case provided an interesting variation, in that Dorian Nakamoto was not found to be a plausible candidate author, and in fact, one of the distractor authors (Neal J. King) was found to be a better match to Satoshi Nakamoto than any other distractor or than Dorian. 'No method ever identified Dorian as being a more likely author than King.'(Herper, 2014). Despite this, King is not necessarily the author of the Bitcoin documents either. The chances of selecting the true author of a document in an ad hoc distractor set is low simply because there are more than six billion people in the world who were not studied.

5 Discussion and Conclusions

Perhaps obviously, there are some caveats to the proposed protocol. The most key is, of course, the implicit assumption of independence. Is it reasonable to believe that the distribution of word lengths is independent of the use of common function words? (An argument could be made, and indeed one reviewer made it, that since the same mind produced all of the features in a given document, then of course no features are independent.) More importantly, can this belief be validated empirically and justified theoretically? Alternatively, can we use develop and use knowledge of joint distributions (without the independence assumption) to deliver improved authorship judgments, and/or can the independence assumption be justified as a simplifying approximation (that does not reduce the overall accuracy significantly)? This is obviously a point requiring substantial research effort, but we believe that this type of research will justify itself in improving the overall quality of this type of analysis.

Similarly, there are some numbers in the protocol that may need tightening—is three to five distractor authors enough? Are five better than three? How many analyses should be run? Can these numbers be justified?

To some extent, the specific numbers in the protocol are dependent on an empirical factor, the accuracy level of the underlying analysis or analyses. The presentation above has assumed a level 80% accuracy rate irrespective of the number of distractor authors, but in practical terms, more distractors will invariably lower the accuracy. More importantly, changes in our understanding are likely to improve expected accuracy measures. If 99% accuracy becomes the norm, then the chance of a false acceptance surviving three independent screenings drops to 1 in 10,000, while the chance of a false rejection is less than 3%. The accuracy of various measures will strongly affect the parameters such as

the recommended number of analyses and the size of the distractor set—and in particular, a more formal and detailed analysis would note that the accuracy rate of different methods is likely to be different from each other, which in and of itself makes the calculations more difficult.

Beyond this, however, is a question of the needs of the various users of this kind of analysis. The legal profession, for example, recognizes several different levels of 'burden of proof', ranging from a simple 'some credible evidence', up to the much more demanding 'beyond a reasonable doubt' required of criminal convictions in the Anglo-American court systems. Less formal areas, such as journalism or historical studies, do not have formal levels of evidence, but the notion (attributed to Carl Sagan) that 'extraordinary claims require extraordinary evidence' is generally accepted as a method for evaluating highly controversial or unlikely claims. What degree of evidence would be necessary for a Shakespeare scholar to believe that Hamlet was not, in fact, written by William Shakespeare? This is a question that can only be answered by disciplinary scholars and not by stylometrists; the cumulative judgment of centuries of scholarship should not lightly be cast aside, but questions like this must not become a matter of dogma. Alternatively, it is not the strength but the type of evidence that is key, an important matter for discussion. What types of text analysis can produce clear and compelling evidence of the authorship of an unknown work?

At a more practical level, is it ever possible for this kind of statistical-stylistic evidence alone to rise to the level of 'beyond a reasonable doubt'? In this light, how would one regard stylistic evidence as different from, for example, fingerprint evidence or DNA evidence? If 'my' DNA is found at a crime scene, it is at least possible that it was left by someone else, possibly even an identical twin whom I do not know about (Bouchard *et al.*, 1990; Boomsma, 2012; Bidgood, 2014). Does that argument by itself constitute 'reasonable doubt'? In practice, this question is less important than it appears at first glance, precisely because criminal cases are usually brought on the basis of multiple pieces of evidence—my DNA (or stylometry alone) probably should not be enough to convict me, but as one piece of evidence along with others (such as a showing that I had the means, the motive, and the opportunity) a generic argument such as an unknown and possibly non-existent twin is less 'reasonable'.

One serious potential flaw in the proposed protocol is that no specific analysis methods are incorporated into it. The sheer number of proposed methods makes it practical for an unscrupulous scholar to 'game' the protocol by running not three to five analyses, but thirty to fifty, and selecting/publishing only those results that show the desired conclusions. This could be avoided in a rather heavy-handed way by providing an approved list of ten or so analytic methods and insisting that only the approved methods count. From a legal standpoint, this might greatly enhance the credibility and admissibility of the findings. From a scholarly standpoint, however, any proposed list would be obsolete as soon as it were published. We suggest instead that questions of the appropriateness of the methods used should be dealt with as part of any project, and that researchers should be encouraged to justify their choices by reference to appropriate empirical studies such as the PAN series of conferences or appropriate journal references.

Considering the broader question, one must still ask 'did Rowling really write *Cuckoo*?' Stylistic evidence aside, this is an ordinary question for journalists or literature scholars that can be investigated in the ordinary course of practice. The analyses presented here show clearly that there are stylistic similarities between Cuckoo and other writings by Rowling, lending credibility to the idea that she is the author. Similarly, the Poe analysis strongly suggests that Edgar Allan was the author of the questioned stories and that Henry was not. This of course does not compel agreement, but in the absence of any other evidence, we believe that a scholar should grant at least tentative belief. In the case where a decision is mandatory (such as a court case), we feel it is not reasonable simply to withhold assent on the grounds that perfect evidence is not available. The perfect should not be allowed to be the enemy of the good.

It should be clear that this article does not *ipso facto* establish a mandatory standard for authorship

studies. We invite discussion and even competing proposals, in addition to further studies to establish not only what other protocols might be more accurate, but also which ones are easier to apply, or even more likely to generate useful information (beyond simple authorship). One key aspect of this proposal is that it relies primarily on rank-order statistics and does not take into account the degree of variation; a more sophisticated protocol might use parametric statistics for greater power, at the possible cost of increased complexity. Alternatively, a simpler protocol may produce more compelling evidence simply because the argument itself is easier to follow, even if, in a technical sense, it is less powerful.

From a practical standpoint, however, this protocol may represent a substantial maturation of the field. Not only have we used it ourselves, but it has also been used by third parties. The results have been validated by reference to independent ground truths (Rowling acknowledged authorship on 12 July 2013.) The results have even been accepted in courts of law, as when Baggins was permitted to remain in the USA. We are thus confident that the proposed protocol will provide a relatively clear-cut way to reduce controversy regarding stylometric authorship attribution and increase its uptake and credibility. More importantly, this protocol and the field generally should now be in a position to help resolve long-standing questions of authorship.

What, then, is the next step? Formalizing standards of practice in the digital humanities has been done before, but it is a substantial undertaking requiring a long time and much discussion. For the specific problem of authorship attribution, the stakeholders who would need to be involved are not necessarily part of the traditional digital humanities community, and include people like lawyers, journalists, forensic scientists, traditional linguists, and psychologists. It is not even clear which organization or organizations would be a logical candidate to spearhead formalization efforts. At the same time, if the digital humanities community can achieve a robust understanding of this particular 'community practice', this would provide a firm basis for discussion with other groups as well as improving the practices themselves. Authorship attribution, as a subfield, is rapidly 'maturing', in the Kuhnian sense. It may be time to acknowledge we are growing up, and to engage in shaping the long-term use of our field.

Funding

This material is based in part upon work supported by the National Science Foundation [OCI-1032683]; and by the Defense Advanced Research Projects Agency [Active Authentication, Phases I and II]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or DARPA.

References

Bidgood, J. (2014). New DNA test sought in identical twin's rape case. *New York Times*.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16(2): 9–17.

Boomsma, D. I. (2012). Similarities despite separation. *Science*, 337(6091): 157.

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, 3(4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Bouchard T. J. Jr., Lykken, D. T., McGue, M., Segal, N. L., and Tellegen, A. (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science*, 250(4978): 223–8.

Brooks, R. (2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed. *Sunday Times*, 14 July. London, England: Times Newspapers Ltd.

Brooks, R. and Flyn, C. (2013). JK Rowling: The cuckoo in crime novel nest. *Sunday Times*, I, 14 July. London, England: Times Newspapers Ltd.

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17: 267–287.

Cohen, N. (2014). Putative Bitcoin author categorically denies it. *New York Times*, March 17, 2014. New York City: Arthur Ochs Sulzberger, Jr.

Collins, P. (2013). Poe's debut, hidden in plain sight? *New Yorker Blog. 7 October*, 2013. New York City: Condé Nast.

DeCarlo, E. (2013). Inferring authorship through Myers-Briggs Type Inventory. In *Proceedings of DHCS 2013*, Chicago.

DeCarlo, E. (2014). Inferring authorship through Myers-Briggs Type Inventory and Naive Bayes. In *Nebraska Conference for Undergraduate Women in Mathematics*, Lincoln, NE.

de Morgan, A. (1851/ 1882). Letter to Rev. Heald 18/08/1851. In Elizabeth, S. and Morgan, D. (eds), *Memoirs of Augustus de Morgan by His Wife Sophia Elizabeth de Morgan with Selections from His Letters*. London: Longmans, Green, and Co.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3): 241–52.

Eder, M. (2010). Does size matter? authorship attribution, small samples, big problem. In *Proceedings of Digital Humanities 2010*, London.

Hasanaj, B., Purnell, E., and Juola, P. (2014). Cross-linguistic transference of authorship attribution. In *Proceedings of the International Quantitative Linguistic Conference* (QUALICO). Olomouc, Czech Republic: Philosophical Faculty of Palacký University.

Herper, M. (2014). Linguist analysis says Newsweek named the wrong man as Bitcoin's creator. *Forbes Magazine*, March 10. Jersey City: New Jersey.

Higgins, J. P. T. and Green, S. (eds) (2009). *Cochran Handbook for Systematic Reviews of Interventions. The Cochrane Collaborative*, 5.0.2 edn. www.cochrane-handbook.org.

Hoover, D. L. (2004). Testing Burrows's delta. *Literary and Linguistic Computing*, 19(4): 453–75.

Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2): 215–23.

Juola, P. (2004). Ad-hoc authorship attribution competition. In *Proceedings of 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, G öteborg, Sweden.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Juola, P. (2008). Killer applications in digital humanities. *Literary and Linguistic Computing*, 23(1): 73–83.

Juola, P. (2009a). 20,000 ways not to do authorship attribution and a few that work. In *Proceedings of 2009 Biennial Conference of the International Association of Forensic Linguists (IAFL-09)*, Amsterdam.

Juola, P. (2009b). Cross-linguistic transferrence of authorship attribution, or why English-only prototypes are acceptable. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P. (2012a). Large-scale experiments in authorship attribution. *English Studies*, 93(3): 275–83.

Juola, P. (2012b). An overview of the traditional authorship attribution subtask. In *Proceedings of PAN/CLEF 2012*, Rome, Italy.

Juola, P. (2013a). How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. *Scientific American*, August.

Juola, P. (2013b). Stylometry and immigration: A case study. *Journal of Law and Policy*, 21(2): 287–98.

Juola, P., Noecker, Jr., J., Ryan, M., and Speer, S. (2009). Jgaap 4.0—a revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P., Sofko, J., and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2): 169–78.

Juola, P. and Stamatatos, E. (2013). Overview of the authorship identification task. In *Proceedings of PAN/CLEF 2013*, Valencia, Spain.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9–26.

Koppel, M., Schler, J., Argamon, S., and Winter, Y. (2012). The 'fundamental problem' of authorship attribution. *English Studies*, 93(3): 284–91.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*, 3rd edn. Chicago, IL: University of Chicago Press.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Noecker J. Jr. and Juola, P. (2009). Cosine distance nearest-neighbor classification for authorship attribution. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Odgers, S. J. and Richardson, J. T. (1995). Keeping bad science out of the court-room—changes in American and Australian expert evidence law. *University of New South Wales Law Journal*, 18(1): 108–29.

Popper, K. R. (2002). *The Logic of Scientific Discovery, reprint edition*. Hove: Psychology Press.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31: 351–65.

Siemens, R. (2014). Communities of practice, the methodological commons and digital self-determination in the humanities. Zampolli Prize Lecture at Digital Humanities 2014, U. Lausanne.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538–56.

Stamatatos, E., Stein, B., Daelemans, W., Juola, P., no, A. B.-C., Verhoeven, B., and Sanchez-Perez, M. A. (2014). Overview of the authorship identification task at PAN 2014. In *Proceedings of PAN/CLEF 2014, Sheffield*, UK.

Stein, S. and Argamon, S. (2006). A mathematical explanation of Burrows' Delta. In *Proc. Digital Humanities 2006*, Paris, France.

Vescovi, D. M. (2011). *Best Practices in Authorship Attribution of English Essays.* Master's thesis, Duquesne University.

Volokh, E. (2014). Reverse Defamation, the Newsweek Bitcoin Story, and Satoshi Nakamoto. *Washington Post*, March 18, 2014.

Notes

1 Although Kuhn writes specifically about 'science', this description also applies to any scholarly or creative field that builds upon the work of others; see, for instance, the development of the fugue or of sonata form in music composition.

2 For example, see R. v. Mohan 1994 CanLII 80, [1994] 2 SCR 9 (5 May 1994) or the House of Commons Science and Technology Committee, (2005) Forensic Science on Trial, London: The Stationery Office Limited, HC96-I, para. 173.

3 The Web site *www.online-literature/poe* lists sixty-one separate short stories.

4 For the Baggins case, the stakes could literally be life and death for the asylum claimant.

5 U.S. Patent Act, 35 U.S.C. § 103(A)

6 More formally, under the specific assumption of a closed-class problem with a relatively small number of distractor authors (certainly fewer than ten, but more than one), we assume that if the actual author is among the candidates, an analysis is 80% likely to identify that author as the most likely one. We also, less plausibly, assume that this probability is the same irrespective of the number of distractor authors.