COMPUTATIONAL ANALYSIS OF AUTHORSHIP AND IDENTITY FOR IMMIGRATION

Patrick Juola

Juola & Associates

256 W. Schwab Avenue

Munhall, PA 15210

Email : pjuola@juolaassoc.com

**ABSTRACT** : An asylum-seeker sought to establish his case on the basis of anonymously-published news articles which he claimed to have written. Unfortunately, original manuscripts of these documents were not available for traditional (physical-document-based) analysis. Using statistical linguistics, we were able to analyze the writing style against an ad-hoc collection of distractor authors and to establish using non-parametric rank-order statistics that the documents had indeed been written by the seeker.

Keywords : Forensic science, information science, authorship, stylometry, attribution, computational linguistics

 Michel Souris[[1]](#footnote-0), a native of Ruritania, was facing immigration procedures that might have led to his removal from the United States. He claimed in immigration court that deportation was inappropriate because he was a noted and published activist against the Ruritanian government and he feared negative consequences if forcibly repatriated. As evidence for this claim, he offered a number of articles he had written for a Ruritanian-language newspaper, as well as a set of newer (anti-government) articles he claimed to have written, but that had been published anonymously while outside Ruritania.

 On the surface, this appears to be an ordinary questioned documents case: did the author of the older and authenticated documents also write the newer and anonymous/disputed ones? However, the nature of the documents precluded the application of traditional document examination technology.

 Traditional document analysis typically focuses on physical aspects of physical documents; for example, Green (1) analyzes variance in paper brightness with an eye to determining whether one page of a multi-page document has been surreptitiously replaced. Others (2) look at variances in the physical instantiation of characters in a manuscript. These approaches, however, were inappropriate for this situation for several reasons:

* The original manuscripts were no longer available; the documents to be analyzed were photocopies or prints of professionally produced journalism.
* At least the more recent documents were “born-digital,” having been created on a word processor and published to the Web via HTML. As a result, the “characters” were artifacts of pure information, and one 12-point Times font capital A is identical to any other. Not just similar, but identical.
* Similarly, no paper or other physical artifacts were involved in the production of these born-digital documents.

Did Mr. Souris write the newspaper articles in question? More accurately, was the author of known documents (named as Mr. Souris) the same as the author of the questioned one (also assumed to be a single person, but not necessarily Souris)?

**BACKGROUND**

Unfortunately, this kind of information-only document is becoming increasingly common in the computer age. Consider the following situations [some taken from (3,4)]:

* a suicide note typed into a word processor instead of scribbled as a note. Is it genuine?
* an anonymous blog posting containing confidential or defamatory information. If the Internet service provider is not cooperative, how can you determine who wrote it?
* a piece of inappropriate email, sent from a public access terminal or from the keyboard of an employee away from their desk. Who actually wrote the email?

The alternative to analyzing the physical document is, of course, analyzing the abstract document, the writing style and the words it contains. If the writing in the disputed newspaper articles could be shown to be in the style of Souris, then it would confirm that the disputed newspaper articles had been written by him.

Close reading, hopefully by an expert in language, may be able to provide cues to the authorship, and for centuries this was the method that scholars and experts used to solve cases of disputed authorship. Modern computational linguistic technologies have produced a faster, more objective, and more “scientific” method of answering such questions on the basis of document statistics. If some notion of writing style can be objectively measured, it becomes possible to quantify exactly the idea that “this document is likely to have come from this person.”

Just as every person has their own fingerprint or DNA, so every person has their own writing style, an extension of their personality and cognitive habits. Some examples of this are clear even to laymen. Vocabulary choice is an aspect of style, as the reader can confirm by asking what is the name of the huge piece of furniture in front of the TV, the one that seats at least three people? Depending upon where you are from (and how old you are), that might be a *sofa*, a *couch*, a *settee*, a *divan,* a *chesterfield*, or a *davenport*. The rubber-soled shoes you wear to work out might be *sneakers*, *tennis shoes*, or *trainers*; Wikipedia (5) lists at least eighteen regional synonyms for these things. Something even bigger than *large* could be *huge*, *gigantic*, *colossal*, *enormous*, and so forth. And, of course, the readership is well cognizant of idiosyncratic individuals with a pronounced tendency toward sesquipedalian lexical items—or, alternatively, we all know some people who use a lot of big words.



Figure 1. A place setting: where is the fork?

A more interesting example comes from figure 1. Look at the picture and tell yourself where the fork is. The important word in your answer is not what you think it is -- it’s the preposition you used. Is the fork *to* the left, *on* the left, *at* the left, or something else?

Prepositions, like pronouns, conjunctions, and determiners, are examples of ``function words.’’ Unlike content words such as nouns, verbs, and adjectives, function words tend to be common, short, and not very meaningful. (Consider the difficulties in defining the word “of” or “the.”) Instead, they describe relationships between other words – structural information, not content. In part because of this, people tend not to pay much attention to them. In one famous experiment (6), subjects were largely unable to notice or remember whether a sentence they had just memorized was “Three turtles rested on a floating log, and a fish swam beneath them” or “Three turtles rested on a floating log and a fish swam beneath it.” The difference between the pronoun *them* and the pronoun *it* just isn’t that salient, especially as, in the situation described, the fish would have necessarily been underneath both the log and the turtles.

For these reasons, these function words are very useful for forensic stylometry. Since the words themselves are so common, a single page of text will have dozens of examples, unlike words like *huge* which may come one to a page if that. People’s characteristic patterns of these words are stable, often over a period of decades (7). More importantly, because people pay so little attention to them, they are unlikely to be aware of their own characteristic patterns, and would find it difficult to change their patterns or imitate someone else’s even if so inclined, an important consideration when considering the adversarial nature of much forensics.

However, using this kind of subtle evidence introduces its own issues. It’s simple to say of a word like *chesterfield* that a person either uses it or not. Since literally every English-speaker uses the word *to*, analysts need to dig deeper and ask harder questions, such as how often and in what contexts they use that word, and what other words might be used instead?

 Standard stylometric practice (8,9,10,11) involves collecting distributional statistics on a well-defined collection of potential authors and then applying some sort of categorization technology to the statistics. Examples of the type of statistics collected include words (possibly with part-of-speech or frequency restrictions), part-of-speech tags, character substrings, or (as in the case presented here) multi-word phrases. Example analysis techniques include Bayesian statistics, principle component analysis, linear discriminant analysis, or (as in the case presented here) nearest neighbor algorithms. (12,13) These studies have confirmed the high accuracy of this type of analysis [one well-publicized set of experiments (8) shows better than 90% accuracy at analyzing personal letters].

 Unfortunately, Mr. Souris’ case presents unusual aspects that make standard practice inappropriate. The least important aspect is that the documents in question were in Ruritanian, and none of the lab staff were familiar with that language. Because of the nature of computer processing, this is not a major handicap. Ordinary string processing techniques are language-agnostic; words and phrases can be collected using the same procedure in any space-separated language, whether English, French, Irish, Maltese, or Polish. More important is the lack of a well-defined distractor set among whom the computer is asked to choose.

 Normal stylometric practice asks the analyst (or computer) to select between a small, closed group of authors. For example, (14) analyzes the authorship of political pamphlets known to be either by Hamilton or Madison, while (15) does the same for novels known to be by Baum or Thompson. Even when the number of potential authors is large (16), “none-of-the-above” is not typically an answer. In Souris’s case, there was only one potential author (Souris himself) or an unbounded universe of unnamed “none-of-the-above” authors from whom samples could not be gathered.

**MATERIALS AND METHODS**

 Michel Souris himself supplied us with ten copies of newspaper articles published under his name approximately ten years before the date of the case, articles that comprised a set of known documents. These documents (photocopies of clippings) were hand-transcribed by Ruritanian-speaking typists into a machine-readable corpus. In addition, he supplied us with eleven web page images from a recent news site, published anonymously, as the set of questioned documents. (Of these eleven documents, one was in English and unsuitable for study, so the actual questioned documents comprised ten web pages from which text was extracted. No typists were needed to extract text from these pages as they were in standard HTML.)

 The JGAAP software package (8,17,18) provided the necessary technology for this text analysis. All relevant files were pre-processed to convert them into plain text (Unicode) format. All case distinctions were neutralized, and all whitespace (inter-word spacing, line breaks, paragraphing, etc.) was normalized to avoid any spurious findings of dissimilarity caused by simple formatting and editing issues. All documents were converted into word trigrams (phrases of three adjacent words, as in the English phrase “in the English”), a unit of processing known (8) to give good results in authorship queries.

 To establish with reasonable certainty that Souris had or had not written the document, it was necessary for us to create our own distractor set, which we did by gathering a collection of Ruritanian-language newspaper articles on political issues from another online newspaper. This corpus consisted of 160 news articles by five different named authors, none of whom were Souris. This provided us with five separate comparison “baseline corpora,” each containing at least thirty articles known to be by a distractor author.

 The word trigram distributions of the ten documents in the known document set were averaged to produce a central or typical example of Souris’ writings. Each individual document in the questioned corpus as well as the five baseline corpora was individually compared against this ``typical’’ Souris style to determine a stylistic “distance.” Two identical documents would be at distance zero, and in general, the smaller the distance (the “closer” the document pair), the more likely two documents were to share authorship. These distances were averaged to produce a per-author average distance from the known documents.

 These distance comparisons were actually performed twice, using two different distance formulae and hence two different analyses. The first analysis was performed using normalized dot product or cosine distance (19), in which the frequency of each individual word trigram is taken into account. The second was done with Jaccard or intersection distance between the sets of word trigrams, which does not take into account frequency but simply measures whether or not a particular author used a particular three word phrase at any point in the samples. As will be discussed later, this provides a certain degree of independence between the two samples, a key point in evaluating our findings.

**RESULTS**

 As a result of this analysis, we obtained a set of distances comparing the known documents both to the questioned documents as well as to five other distractor authors. On the assumption that Souris had written the questioned documents, we would expect to see the average distance between the known and questioned documents (documents by the same author) to be smaller than the distance between any of the distractor sets and the known documents.

 The analytic results can be summarized in tables 1 and 2 (data has been

rescaled for clarity of presentation):

Cosine distance:

|  |  |  |
| --- | --- | --- |
| Corpus | Average (mean) distance (normalized) | Variance (s.d.) (normalized) |
| BD-1 -> KD | *0.9437975* | 0.082 |
| BD-2 -> KD | 0.9517967 | 0.064 |
| BD-3 -> KD | 0.9576155 | 0.071 |
| BD-4 -> KD | 0.9530338 | 0.059 |
| BD-5 -> KD | 0.9534134 | 0.052 |
| QD -> KD | **0.8840330** | 0.158 |

 Table 1 : analytic results for cosine distance

Intersection distance:

|  |  |  |
| --- | --- | --- |
| Corpus | Average (mean) distance (normalized) | Variance (s.d.) (normalized) |
| BD-1 -> KD | 0.806731 | 0.276 |
| BD-2 -> KD | *0.739381* | 0.258 |
| BD-3 -> KD | 0.852844 | 0.251 |
| BD-4 -> KD | 0.747444 | 0.273 |
| BD-5 -> KD | 0.777530 | 0.218 |
| QD -> KD | **0.735449** | 0.353 |

 Table 2: analytic results for Jaccard or intersection distance

 As can be seen, the distance between the questioned documents (QD) and the known documents (KD) is less than any distractor/baseline author (BDs 1—5). Unfortunately, these differences are not statistically significant according to standard *t* tests.

**DISCUSSION**

 As acknowledged in the previous section, standard *t* tests could not find significant differences among the authors, although the results were in the expected direction. From a non-parametric point of view, however, the results were extremely interesting. We examined instead the rank-order statistics (i.e. ordered the corpora by closeness to the known documents without regard to the magnitude of the differences). In both analyses, the questioned documents were ranked first (closest to the known documents). With six different authors, none of whom were Souris (under the null hypothesis), we have no *a priori* reason to expect baseline author #2 to be any closer (or any further) than baseline author #4, or than the author of the questioned documents. Thus, if Souris were not the author of the questioned documents, there would only be one chance in six (16.7%) of his being closest to the QD author, and hence once chance in six of a false positive. (The issue of a false negative does not arise in this case.)

 A similar argument applies to each analysis run; there is no *a priori* reason to assume that (given the null hypothesis) the closest author under two different analyses will be the same person. We support this by observing that the closest distractor author differs between the two analyses, being BD-1 in the first and BD-2 in the second. Thus the chance of the QD author being closest to Souris in two independent analyses would be approximately one in 36 (six times six) or 2.78%.

 In theory, this could be repeated against as many analyses as necessary. The analysis software we used incorporates literally thousands of different methods, and using only ten of them would yield a naïve probability of being first on all ten analyses (given six authors and the null hypothesis) of approximately 0.00000165%, one chance in slightly over sixty million.

 It’s not clear, however, that the assumption of complete independence can be supported in this case, since the analyses are of the same underlying dataset and share correlations. In the case of the present analysis, distances measured according to the normalized cosine formula and distances measured according to the Jaccard formula are substantially different as they weight different parts of the frequency spectrum differently. Common phrases (e.g. “instead of the”) will be present in both corpora and hence will contribute zero to the Jaccard distance; the contribution is made exclusively by uncommon phrases present in one corpus but not the other. By contrast, cosine distance focuses on differences in frequency and hence only a common phrase has the possibility of generating a large difference. However, if two documents are completely different—if, for example, we had allowed a document in English to be analyzed against the Ruritanian corpora—the English document would have shared almost nothing, and would have scored furthest of all documents under any reasonable analysis.

 For this reason, we can strictly claim only the chance of Souris’s style being closest to QD twice by mere chance is less than 17% and could be as low as 2.78%. This number was considered acceptable by counsel.

**CONCLUSIONS**

Analysis of questioned electronic documents creates difficulties for traditional document analysis. We describe here a case in which the known documents were photocopies of newspaper clippings and the questioned documents were born-digital web pages, a cross-media comparison that permits no analysis of the physical artifacts such as handwriting or paper. By analyzing the language used in the documents, we were able to produce an opinion on the authorship of the questioned documents.

 The Souris case presented several unusual aspects in stylistic investigations; the standard stylometric analysis paradigm selects among others rather than giving a simple “yes/no” answer. Using non-parametric rank order statistics and an ad-hoc set of distractor authors, we could still get an answer and validate it statistically.

 Oh, and Michel Souris? The judge permitted him to remain in the United States.

**ACKNOWLEDGEMENT**

This material is based upon work supported by the National Science Foundation under Grant No. OCI-1032683 to Duquesne University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

**REFERENCES**

1. Green JA. Reliability of Paper Brightness in Authenticating Documents. J Forensic Sci 2012; 57(4):1003—7.
2. Saunders CP, Davis LJ, Buscaglia J. Using Automated Comparisons to Quantify Handwriting Individuality. J Forensic Sci 2011; 56(3):683—9.
3. Chaski CE. Who’s at the keyboard: Authorship attribution in digital evidence investigations. Int J Digital Evidence 2005; 4(1):n/a (Electronic-only journal: http://www.ijde.org, accessed May 31, 2007)
4. Chaski CE. The keyboard dilemma and forensic authorship attribution. In Craiger P, Shenoi S, editors. Advances in Digital Forensics III 2007. New York:Springer.
5. Wikipedia. Athletic shoe. http://en.wikipedia.org/wiki/Athletic\_shoe. Retrieved 9 July 2012.
6. Bransford JD, Barlay JR, Franks JJ. Sentence memory: A constructive vs. interpretive approach. Cog Psy 1972;3:193–209.
7. Ryan M, Juola P. Age or identity: A comparison of stylistic similarity. In: Exploring the Boundaries and Applications of Corpus Linguistics 2011. Tuscaloosa, AL: University of Alabama.
8. Juola P. Authorship attribution. Foundations and Trends in Information Retrieval 2006. 1(3):233—334.
9. Koppel M., Schler J. and Argamon S. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 2009. 60(1):9—26.
10. Stamatatos E. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 2009. 60(4):538—56.
11. Calle-Martin J, Miranda-Garcia A. Stylometry and authorship attribution: Introduction to the Special Issue. English Studies 2012; 93(3):251—58.
12. Juola P. 20,000 ways not to do authorship attribution – and a few that work. In Proc. 2009 Biennial Conference of the International Association of Forensic Linguists. Amsterdam, Netherlands.
13. Juola P. Large-scale experiments in authorship attribution. English Studies 2012;93(3):275—83.
14. Mosteller F, Wallace D. Inference and Disputed Authorship: The Federalist Papers. Chicago: University of Chicago Press. 1964 (reprinted 2007).
15. Binongo, JNG. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance 2003. 16(2):9—17.
16. Koppel M,  Schler J, Argamon S, Messeri E. Authorship attribution with thousands of candidate authors. In SIGIR 2006: 659-660.
17. Juola P. JGAAP 5.3.0: A system for general purpose text classification experiments. In Proc. EACL Workshop on Computational Approaches to Deception Detection 2012. Avignon, France:EACL
18. Java Graphical Authorship Attribution Program [computer program]. Pittsburgh:Duquesne University, 2012. http://www.jgaap.com
19. Noecker Jr J. and Juola P. Cosine distance nearest-neighbor classification for authorship attribution. In Proc. Digital Humanities 2009. College Park, MD: Alliance of Digital Humanities Organizations.
1. All identifying information has been changed to protect the privacy of the individuals. [↑](#footnote-ref-0)