

Chapter 1

QUESTIONED ELECTRONIC DOCUMENTS : EMPIRICAL STUDIES IN AUTHORSHIP ATTRIBUTION

Patrick Juola

Abstract Forensic analysis of questioned *electronic* documents is very difficult, because the nature of the documents eliminates many kinds of informative differences. Recent work in authorship attribution demonstrates the practicality of analyzing documents based on authorial style, but the state of the art is confusing. Analyses are difficult to apply, little is known about type or rate of errors, and no “best practices” are available. We present the results of some recent experiments and software development to address these issues, partly through the development of a systematic testbed for multilingual, multigenre authorship attribution accuracy, and partly through the development and concurrent analysis of a uniform and portable software tool that applies multiple methods to analyze electronic documents for authorship based on authorial style.

Keywords: Authorship attribution, stylometrics, software development, text forensics

1. Introduction

The forensic importance of questioned documents is well-understood — did Aunt Martha really write this disputed version of “her” will? Document examiners can look at handwriting (or typewriting) and determine authorship with near miraculous sophistication from the dot of an ‘i’ or the cross of a ‘t’. Electronic documents do not contain these clues. Any two flat-ASCII ‘A’ characters are identical. How can one determine who made a defamatory, but anonymous, post on a blog, for example? Whether the authorship of a purely electronic document can

be demonstrated to the demanding standards of a *Daubert* [7] hearing is an open, but important, research question.

2. The Problem

With the advent of modern computer technology, a substantial amount of “writing” today never involves pen, ink, or paper. This very paper is a good example — born as a PDF file, the first time these words see paper is in the bound volume. If my authorship of these words were challenged, I have no physical artifacts for specialists to examine.

Furthermore, the nature of electronic documents makes it substantially easier to “publish” or misappropriate them tracelessly or even to commit forgery with relative impunity. A network investigation will at best only reveal the specific computer on which the document was written. It is almost impossible to figure out who was at the keyboard — who wrote it.

Chaski [6] describes three incident-based scenarios where it is both necessary to pierce the GUI and impossible to do so with traditional network investigations. In all three cases, there was no question about which computer these documents came from. Instead, the question was whether the purported authorship could be validated. The key question thus can be structured in terms of the message content. Can the authorship of an electronic document be inferred reliably from the message content itself?

3. Related Work

3.1 Authorship attribution

Recent studies in authorship attribution suggest that such an inference is possible, but further research may be necessary to meet the stringent criteria of *Daubert*. As a problem, the question of determining authorship by examining style has a long history. For example, Judges 12:5–6 describes the inference of tribal identity from the pronunciation of a specific word. Such *shibboleths* could involve specific lexical or phonological items; a person who writes of sitting on a “Chesterfield” is presumptively Canadian [8]. Wellman [27] describes how an individual spelling error — an idiosyncratic spelling of “touch” — was used in court to validate a document.

At the same time, such tests cannot be relied upon. Idiosyncratic spelling or not, the word “touch” is rather rare (86 tokens in the million-word Brown corpus [22]), and it’s unlikely to be found independently in two different samples. People are also not consistent in their language,

and may (mis)spell words differently at different times; often the tests must be able to handle distributions instead of mere presence/absence judgments. The continuing discussion of methods to do this is an active research area – 70,400 hits turn up on May 4, 2006 on a Google search for “authorship attribution.” The increase from November 13, 2005 (49,500) illustrates part of the continuing activity in this area in just six months.

A key insight in recent research has suggested that statistical distribution of common patterns, such as the use of prepositions, may be universal enough to be relied upon, while still being informative. For this reason, scholars have recently focused on more sophisticated and more reliable statistical tests. Specifically, Burrows [3–5] demonstrated that a statistical analysis of common words in large samples of text could group texts by author. Since then, many additional methods [9, 11, 24, 25, 2, 13, 12, 23–1, 6] have been proposed. The current state of the art is an *ad hoc* mess of disparate methods with little cross-comparison to determine which methods work and which don’t. Or more accurately, because they all work at least reasonably well (under conditions as discussed below, 90% accuracy is fairly typical for “good” methods. See also [18]), which methods work the best.

Authorial analysis can even show more subtle aspects, such as dates of documents. Figure 1 shows such an analysis [16] within a single author (Jack London), clearly dividing works written before 1912 from works after. The apparent division is a vertical line at about 3.14 on “Dimension 1.” Finding that a newly discovered London manuscript would be placed on the left side of the diagram would be strong evidence that it was written after 1912 as well.

3.2 Test Corpus Development : The Baayen experiments

With the wide variety of techniques available, it is important and yet very difficult to compare the power and accuracy of different techniques. A fingerprint appropriate to distinguish between Jack London and Rudyard Kipling, for example, may not work to distinguish between Jane Austin and George Eliot. A proper comparison would involve standardized texts of clear provenance, known authorship, on strictly controlled topics, so that the performance of each technique can be measured in a fair and accurate way. Forsyth [10] compiled a first benchmark collection of texts for validating authorship attribution techniques. Baayen[2] has developed a more tightly controlled series of texts produced under strictly controlled conditions.

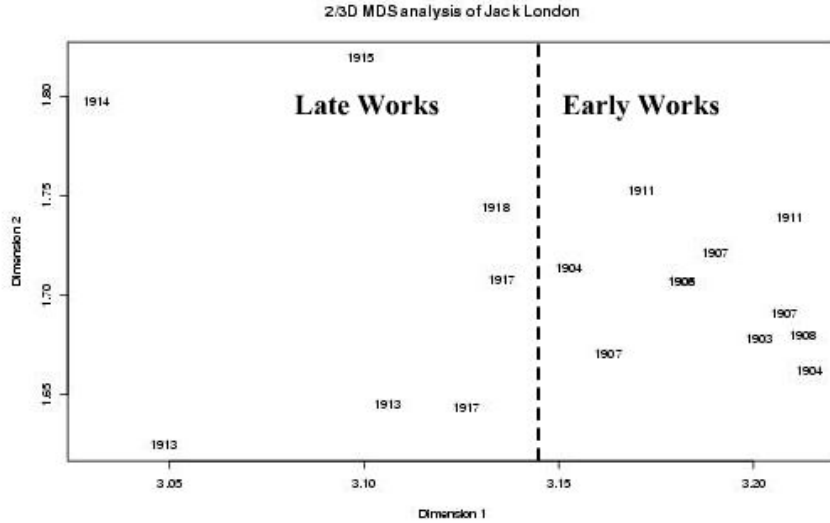


Figure 1. Spacial analysis of time development of Jack London's style

To establish this kind of clean testing material, Baayen *et al.* at the University of Nijmegen elicited writing samples in Dutch from eight U. Nijmegen students. The resulting 72 texts (8 subjects · 3 genres · 3 topics/genre) varied in length between 630 and 1341 words (3655–7587 characters), averaging 907 words (5235 characters) per text.

This corpus has been comparatively analyzed using several substantially different techniques. One of the current most well-known authorship attribution techniques, proposed in [3] and later extended, is a principle components' analysis (PCA) of the most common function words in a document. Another popular technique, linear discriminant analysis (LDA) [2], can distinguish among previously chosen classes, but as a supervised algorithm, it has so many degrees of freedom that the discriminants it infers may not be “clinically” significant. An alternative technique using measurements of cross-entropy has been independently proposed [13].

The question of which method is most accurate in this circumstance can be relatively easily answered : simply use all methods and compare. In particular, these methods have been tested [17] on the Baayen corpus.

The software was presented with repeated trials consisting of triples containing all possible author pairs and possible “disputed” documents. Using this framework, function word PCA performed at essentially chance level, while function word LDA could achieve up to about 55% to 57%, depending upon the number of function words tabulated. Cross-entropy could achieve 73% accuracy using a character-based model, and 87% accuracy across all pairwise comparisons using a word-based model.

From these results it can be concluded that *under the circumstances of this test*, cross-entropy, and in particular, word-based cross-entropy, is a more accurate technique for assessing authorship, but that the chance of a false assignment is an unacceptably high 13%.

4. Test Corpus Development : The Ad-hoc Authorship Attribution Competition

These studies raise an important followup question about the role of the test circumstances themselves. In particular, the test data was all in Dutch, the topics were very tightly controlled, and about 8,000 words of sample data per author were available. Would results have been substantially different if the authors had written in English? If there had been 800,000 words per author, as might be the case in a copyright dispute involving a prolific author? Can the results of an analysis involving expository essays be generalized across genres, for example, to personal letters?

To answer these questions, ALLC/ACH hosted an “Ad-hoc Authorship Attribution Competition” [14] as a partial response. A standardized test corpus for authorship attribution would not only allow researchers to test the ability of statistical methods to determine authors. It would also allow methods to be further distinguished between the “successful” and “very successful.” (From a forensic standpoint, this would validate the science while simultaneously establishing the standards of practice and creating information about error rates, as *Daubert* requires.)

4.1 Contest setup

Contest materials included thirteen problems, details of which are available in [14] or [18] but omitted here for brevity. These included a variety of lengths, styles, genres, and languages, mostly gathered from the Web but including some materials specifically gathered to this purpose. A dozen research groups participated (see table 1), some with several methods, by downloading the (anonymized) materials and returning their attributions to be graded and evaluated against the known correct answers.

Table 1. Partipants, affiliations, and methods

| <i>Name</i> | <i>Affiliation</i> | <i>Method</i> |
|-----------------------------------|--------------------|--|
| Andrea Baronchelli, <i>et al.</i> | Rome | Entropy-based informatic distance |
| Aaron Coburn | Middlebury | Contextual network graph |
| Hans van Haltern | Nijmegen | “Linguistic Profiling” |
| David L. Hoover | NYU | Cluster analysis of word frequencies |
| David L. Hoover | NYU | Google search for distinctive phrases |
| Patrick Juola | Duquesne | Match length within a database |
| Lana and Amisano | UNIPMN | Common N-grams (two variants) |
| Kešelj and Cercone | Dalhousie | CNG with weighted voting |
| Kešelj and Cercone | Dalhousie | CNG-wv with reject |
| O’Brien and Vogel | Trinity/Dublin | Chi by degrees of freedom |
| Lawrence M. Rudner | GMAC | Multinomial Bayesian Model/BETSY |
| Koppel and Schler | Bar-Ilan | SVM with linear kernal function |
| Efstathios Stamatatos | Patras | Meta-classifiers via feature selection |

4.2 Contest results

The contest (and results, see tables 2 through 3) were surprising at many levels; some researchers had initially refused to participate given the admittedly difficult tasks included among the corpora. (Indeed, not all groups submitted results for all test problems; problems for which no results were received are scored as $0/N$ in the tables.)

For example, Problem F consisted of a set of letters extracted from the Paston letters. Aside from the very real issue of applying methods designed/tested for the most part for modern English on documents in Middle English, the size of these documents (very few letters, today or in centuries past, exceed 1000 words) makes statistical inference difficult. Despite this apparent difficulty, almost all groups were able to achieve 90% or better on this problem.

Similarly, problem A was a realistic exercise in the analysis of student essays (gathered in a first-year writing class during the fall of 2003) – as is typical, no essay exceeded 1200 words. From a standpoint of literary analysis, this may be regarded as an unreasonably short sample, but from a standpoint both of a realistic test of *forensic* attribution, as well as a legitimately difficult problem for testing the sensitivity of techniques, these are legitimate.

Overall results from this competition were heartening. The highest scoring participant was the research group of Vlado Keselj, with an average success rate of approximately 69%. In particular, Keselj’s methods achieved 85% accuracy on problem A and 90% accuracy on problem F, both acknowledged to be difficult and considered by many to be unsolvably so. (Juola’s solutions, in the interests of fairness, averaged 65%

Table 2. Problems A–D detailed results

| Team | A | B | C | D | E | F | G |
|---------------|-------|------|-----|-----|-----|-------|-----|
| baronchelli | 3/13 | 3/13 | 8/9 | 3/4 | 1/4 | 9/10 | 2/4 |
| coburn | 5/13 | 2/13 | 8/9 | 3/4 | 4/4 | 9/10 | 1/4 |
| halteren | 9/13 | 3/13 | 9/9 | 3/4 | 3/4 | 9/10 | 2/4 |
| hoover1 | 4/13 | 1/13 | 8/9 | 2/4 | 2/4 | 9/10 | 2/4 |
| hoover2 | 4/13 | 2/13 | 9/9 | 4/4 | 4/4 | 10/10 | 2/4 |
| juola | 9/13 | 7/13 | 6/9 | 3/4 | 2/4 | 9/10 | 2/4 |
| keselj1 | 11/13 | 7/13 | 8/9 | 3/4 | 2/4 | 9/10 | 3/4 |
| keselj2 | 9/13 | 5/13 | 7/9 | 2/4 | 1/4 | 9/10 | 2/4 |
| lana-amisano1 | 0/13 | 0/13 | 3/9 | 2/4 | 0/4 | 0/10 | 0/4 |
| lana-amisano2 | 0/13 | 0/13 | 0/9 | 2/4 | 0/4 | 0/10 | 0/4 |
| obrien | 2/13 | 3/13 | 6/9 | 3/5 | 2/4 | 7/10 | 2/4 |
| rudner | 0/13 | 0/13 | 6/9 | 3/4 | 1/4 | 0/10 | 3/4 |
| schler | 7/13 | 4/13 | 9/9 | 4/4 | 4/4 | 10/10 | 2/4 |
| stamatatos | 9/13 | 2/13 | 8/9 | 2/4 | 2/4 | 9/10 | 2/4 |

Table 3. Problems H–M detailed results

| Team | H | I | J | K | L | M |
|---------------|-----|-----|-----|-----|-----|-------|
| baronchelli | 3/3 | 2/4 | 1/2 | 2/4 | 4/4 | 5/24 |
| coburn | 2/3 | 2/4 | 1/2 | 2/4 | 3/4 | 19/24 |
| halteren | 2/3 | 3/4 | 1/2 | 2/4 | 2/4 | 21/24 |
| hoover1 | 2/3 | 3/4 | 1/2 | 2/4 | 4/4 | 7/24 |
| hoover2 | 3/3 | 4/4 | 2/2 | 2/4 | 4/4 | 7/24 |
| juola | 3/3 | 2/4 | 1/2 | 2/4 | 4/4 | 11/24 |
| keselj1 | 1/3 | 3/4 | 1/2 | 2/4 | 4/4 | 17/24 |
| keselj2 | 0/3 | 2/4 | 0/2 | 1/4 | 3/4 | 15/24 |
| lana-amisano1 | 3/3 | 0/4 | 0/2 | 0/4 | 1/4 | 0/24 |
| lana-amisano2 | 0/3 | 0/4 | 0/2 | 0/4 | 3/4 | 0/24 |
| obrien | 1/3 | 1/4 | 1/2 | 3/4 | 4/4 | 5/24 |
| rudner | 3/3 | 3/4 | 1/2 | 0/4 | 1/4 | 0/24 |
| schler | 2/3 | 3/4 | 2/2 | 1/4 | 4/4 | 4/24 |
| stamatatos | 1/3 | 3/4 | 1/2 | 2/4 | 3/4 | 14/24 |

correct. As a side note, David Hoover identified a weakness in the problem structure. Since much of the data was taken from the Web, using a web search engine such as Google could identify many of the documents, and therefore the authors. Hoover himself admits that this solution does not generalize and does not address the technical questions of stylometry.)

More generally, all participants scored significantly above chance on all problems for which they submitted solutions. Perhaps because most research done focuses on English, performance on English problems tended to be higher than on other languages. Perhaps more surprisingly, the availability of large documents was not as important to accuracy as the

availability of a large number of smaller documents, perhaps because they can give a more representative sample of the range of an author's writing. Finally, methods based on simple lexical statistics tended to perform substantially worse than methods based on N-grams or similar measures of syntax in conjunction with lexical statistics.

With regard to generalization and confidence issues, the findings are very good for the field as a whole. In general, algorithms that were successful under one set of conditions tended to be successful (although not necessarily as successful numerically) under other conditions. In particular, the average performance of a method on English samples (problems A-H) correlated significantly ($r = 0.594$, $p < 0.05$) with that method's performance on non-English samples. Correlation between large-sample problems (problems with over 50,000 words per sample) and small sample problems was still good, although no longer strictly significant ($r = 0.3141$). This suggests that the problem of authorship attribution is at least somewhat a language- and data-independent problem, and one to which we may be able to expect to find wide-ranging technical solutions for the general case, instead of (as, for example, in machine translation) to have to tailor our solutions with detailed knowledge of the problem/texts/languages at hand. In particular, we offer the following challenge to all researchers in the process of developing a new forensic analysis method: *if you can't get 90% correct on the Paston letters (problem F), then your algorithm is not competitively accurate.* Every well-performing algorithm studied in this competition had no difficulty achieving this standard. Statements from researchers that their methods will not work on small training samples should be regarded with appropriate suspicion.

Unfortunately, another apparent result is that the high-performing algorithms appear to be mathematically and statistically (although not necessarily linguistically) sophisticated. The good methods have names that may appear fearsome to the uninitiated : linear discriminant analysis [2, 26], orthographic cross-entropy [17], common byte N-grams [19], SVM with a linear kernel function [20]. From a practical, courtroom setting, this may cause difficulties down the road in explaining to a jury exactly what kind of analysis is being performed, but we hope the difficulties will be no greater than explaining DNA analysis.

5. Future Developments

Because these techniques can be difficult to implement (or even to use) we cannot expect a casual user to apply these new methods without technical assistance. At the same time, the sheer number of techniques

proposed (and therefore, the number of possibilities available to confuse) has exploded, which also limits the pool of available users.

In partial mitigation of this, at the same conference where the AAAC was run, Juola [15] also proposed a *computational* framework in which the different methods could be unified, cross-compared, cross-fertilized, and evaluated to achieve a well-defined “best of breed.” We have demonstrated a proof of concept [18] during the past year.

The proposed framework postulates a three-phase division of the authorship attribution task, each of which can be independently performed. These phases are :

- **Canonicization** — No two physical realizations of events will ever be exactly identical. We choose to treat similar realizations as identical to restrict the event space to a finite set.
- **Determination of the event set** — The input stream is partitioned into individual non-overlapping “events.” At the same time, uninformative events can be eliminated from the event stream.
- **Statistical inference** — The remaining events can be subjected to a variety of inferential statistics, ranging from simple analysis of event distributions through complex pattern-based analysis. The results of this inference determine the results (and confidence) in the final report.

As an example of how this procedure works, we consider a method for identifying the language in which a document is written. We first canonicize the document by identifying each letter (an italic *e*, a bold-face **e**, or a capital E should be treated identically) and producing a transcription. We then identify each letter as a separate event, eliminating all non-letter characters such as numbers or punctuation. Finally, by compiling an event histogram and comparing it with the well-known distribution of English letters, we can determine a probability that the document was written in English. A similar process would treat each *word* as a separate event (eliminating words not found in a standard lexicon) and comparing event histograms with a standardized set such as the Brown histogram [22]. The question of the comparative accuracy of these methods can be judged empirically. This framework allows researchers both to focus on the important differences between methods and to mix and match techniques to achieve the best practical results.

The usefulness of this framework can be shown in our prototype user-level authorship attribution tool. This tool coordinates and combines (at this writing) four different technical approaches to authorship attribution [4, 5, 21, 13]. Written in Java, this program combines a GUI atop the three-phase approach defined above. Users are able to select a set of sample documents (with labels for known authors) and a set of testing

documents by unknown authors. The user is also able to select from a menu of event selection/preprocessing options and of technical inference mechanisms. Currently supported, for example, are three different choices — a vector of all the letters appearing in the sample/testing documents, a vector of all *words* so appearing, or a vector of only the fifty most common words/letters as previously selected, representing a restriction of the event model. Similarly, a variety of processing classes can be [have been] written to infer a similarity between two different vectors. Authorship of the test document can be assigned to (the author of) the most similar document.

As a specific example of application, we note that many of the AAAC methods relied on inferential statistics as applied to N-grams. But N-grams of what? Juola’s method was explicitly applied to N-grams of letters, van Halteren’s to words or word “classes,” Stamatatos’ to “common words,” and Koppel/Schler’s to “unstable words.” We can therefore in theory code Koppel’s method for identification of unstable words as a separate instance of the event set class, then calculate inferential statistics using van Halteren or Juola’s method (as an instance of the inference class) possibly resulting in an improvement over any component method.

While this program is being developed, new methods are also being developed and improved. The data from the AAAC is still available on-line to permit people to test their methods, and we hope to incorporate new practices into our continuing study of best practices. At the same time, we continue to work on extending the functionality and user-friendliness of the system with the hope of making it into more than a research-quality prototype in the near future.

The AAAC corpus itself also has some limitations that need to be addressed. As a simple example, the mere fact that the data is on the Web (in many cases, gathered from web-accessible public, archives) gives an unfair advantage to any methods that searches the Web. Similarly, the multilingual coverage is unbalanced. The coverage of different genres is spotty and there are probably important issues that have not been addressed at all. We hope to create and offer a followup contest with an improved test corpus and more stringent analysis parameters.

6. Conclusions

This paper has summarized some recent developments in authorship attribution, including large-scale empirical experiments to establish and to validate a set of “best practices” in the field of forensic analysis of questioned documents. Implicit in these experiments are an enhanced ability to create toolsets for analysis, as well as the requirement contin-

ually to push forward in creating new and more accurate experiments to test these proposed “best practices.”

References

- [1] S. Argamon and S. Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC 2005*, Victoria, BC, 2005. Association for Computing and the Humanities.
- [2] R. H. Baayen, H. Van Halteren, A. Neijt, and F.J. Tweedie. An experiment in authorship attribution. In *Proceedings of JADT 2002*, pages 29–37, St. Malo, 2002. Université de Rennes.
- [3] J. F. Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2:61–70, 1987.
- [4] J. F. Burrows. ‘an ocean where each kind...’: Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–21, 1989.
- [5] J. Burrows. Questions of authorships : Attribution and beyond. *Computers and the Humanities*, 37(1):5–32, 2003.
- [6] C. E. Chaski. Who’s at the keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1), 2005.
- [7] *Daubert v. Merrell Dow Pharmaceuticals* (92-102). 509 U.S. 579 (1993).
- [8] G. Easson. The linguistic implications of shibboleths. In *Annual Meeting of the Canadian Linguistics Association*, Toronto, Canada, 2002.
- [9] J. M. Farrington. *Analyzing for Authorship : A Guide to the Cusum Technique*. University of Wales Press, Cardiff, 1996.
- [10] R. S. Forsyth. Towards a text benchmark suite. In *Proc. 1997 Joint International Conference of the Association for Computers*

and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 1997), Kingston, ON, 1997.

- [11] D. I. Holmes. Authorship attribution. *Computers and the Humanities*, 28(2):87–106, 1994.
- [12] D. L. Hoover. Delta prime? *Literary and Linguistic Computing*, 19(4):477–495, 2004.
- [13] P. Juola. The time course of language change. *Computers and the Humanities*, 37(1):77–96, 2003.
- [14] P. Juola. Ad-hoc authorship attribution competition. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden, June 2004.
- [15] P. Juola. On composership attribution. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden, June 2004.
- [16] P. Juola. Becoming Jack London. *Journal of Quantitative Linguistics*, To appear.
- [17] Patrick Juola and Harald Baayen. A controlled-corpus experiment in authorship attribution by cross-entropy. *Literary and Linguistic Computing*, 20:59–67, 2005.
- [18] P. Juola, J. Sofko, and P. Brennan. A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 2006. Advance Access published on April 12, 2006; doi: doi:10.1093/lc/fql019.
- [19] V. Keselj and N. Cercone. CNG method with weighted voting. In Patrick Juola, editor, *Ad-hoc Authorship Attribution Contest. ACH/ALLC 2004*, 2004.
- [20] M. Koppel and J. Schler. Ad-hoc authorship attribution competition approach outline. In Patrick Juola, editor, *Ad-hoc Authorship Attribution Contest. ACH/ALLC 2004*, 2004.
- [21] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev. Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatii*, 37(2):96–198, 2000. Translated in “Problems of Information Transmission,” pp. 172–184.

- [22] H. Kučera and W. N. Francis. *Computational Analysis of Present-day American English*. Brown University Press, Providence, 1967.
- [23] T. Merriam. An application of authorship attribution by intertextual distance in English. *Corpus*, Numero 2 La distance intertextuelle, Décembre 2003.
- [24] J. Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365, 1998.
- [25] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In *Proceedings of EACL '99*, pages 158–164, 1999.
- [26] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.
- [27] F. L. Wellman. *The Art of Cross-Examination*. MacMillan, New York, 4th edition, 1936.