Foundations and Trends® in
Information Retrieval
Vol. 1, No. 3 (2006) 233–334

DOI: 10.1561/1500000005

Authorship Attribution

Patrick Juola

Department of Mathematics and Computer Science, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282, USA, juola@mathcs.duq.edu

Abstract

Authorship attribution, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and a wide range of application. Recent work in "non-traditional" authorship attribution demonstrates the practicality of automatically 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 few "best practices" are available. In part because of this confusion, the field has perhaps had less uptake and general acceptance than is its due.

This review surveys the history and present state of the discipline, presenting some comparative results when available. It shows, first, that the discipline is quite successful, even in difficult cases involving small documents in unfamiliar and less studied languages; it further analyzes the types of analysis and features used and tries to determine characteristics of well-performing systems, finally formulating these in a set of recommendations for best practices.

1

Introduction

1.1 Why "Authorship Attribution"?

In 2004, Potomac Books published *Imperial Hubris: Why the West is Losing the War on Terror*. Drawing on the author's extensive personal experience, the book described the current situation of the American-led war on terror and argued that much US policy was misguided.

Or did he? The author of the book is technically "Anonymous," although he claims (on the dust cover) to be "a senior US intelligence official with nearly two decades of experience" as well as the author of the 2003 book *Through Our Enemies*' *Eyes*. According to the July 2, 2004 edition of the Boston Phoenix, the actual author was Michael Scheuer, a senior CIA officer and head of the CIA's Osama bin Laden unit in the late 1990s. If true, this would lend substantial credibility to the author's arguments.

But on the other hand, according to some noted historians such as Hugh Trevor-Roper, the author of the 1983 *Hitler Diaries* was Hitler himself, despite the later discovery that they were written on modern paper and using ink which was unavailable in 1945. Is *Imperial Hubris* another type of sophisticated forgery? Why should we believe historians and journalists, no matter how eminent? What kind of evidence should we demand before we believe?

Determining the author of a particular piece of text has raised methodological questions for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers as in the examples above. Investigative journalism, combined with scientific (e.g., chemical) analysis of documents and simple close reading by experts has traditionally given good results. But recent developments of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option. This field has seen an explosion of scholarship, including several detailed book-length treatments [39, 41, 44, 83, 98, 103, 105, 111, 112, 150]. Papers on authorship attribution routinely appear at conference ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite — or perhaps because of — this interest, the field itself is somewhat in disarray with little overall sense of best practices and techniques.

1.2 Structure of the Review

This review therefore tries to present an overview and survey of the current state of the art. We follow the theoretical model (presented in detail in Section 3.4) of [76] in dividing the task into three major subtasks, each treated independently.

Section 2 presents a more detailed problem statement in conjunction with a historical overview of some approaches and major developments in the science of authorship attribution. Included is a discussion of some of the major issues and obstacles that authorship attribution faces as a problem, without regard to any specific approach, and the characteristics of a hypothetical "good" solution (unfortunately, as will be seen in the rest of the review, we have not yet achieved such a "good" solution).

Section 3 presents some linguistic, mathematical, and algorithmic preliminaries. Section 4 describes some of the major feature sets that have been applied to authorship attribution, while Section 5 describes the methods of analysis applied to these features. Section 6 goes on to present some results in empirical evaluation and comparative testing of authorship attribution methods, focusing mainly on the results from the 2004 *Ad-hoc Authorship Attribution Competition* [75], the largest-scale comparative test to date.

Section 7 presents some other applications of these methods and technology, that, while not (strictly speaking) "authorship" attribution, are closely related. Examples of this include gender attribution or the determination of personality and mental state of the author. Section 8 discusses the specific problems of using authorship attribution in court, in a forensic setting. Finally, for those practical souls who want only to solve problems, Section 9 presents some recommendations about the current state of the art and the best practices available today.

2

Background and History

2.1 Problem Statement

Authorship attribution, broadly defined, is one of the oldest and one of the newest problems in information retrieval. Disputes about the ownership of words have been around for as long as words themselves could be owned. Questions of authenticating documents have existed as long as the documents themselves have. And the question “what can I tell about a person from the language he uses” has been around since, literally, the days of the Old Testament:

> Judges 12:5 And the Gileadites took the passages of Jordan before the Ephraimites: and it was so, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, Art thou an Ephraimite? If he said, Nay;
>
> 6 Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand.

As will be discussed in depth, this barbaric episode illustrates in a nutshell the stylistic process. The Gileadites identified (correctly, one hopes) a single salient feature of Ephraimite speech — the inability to pronounce a particular sound — that can be used to divide the speech of one group from another. It is not clear from this passage what the success rate of this analysis was, nor how other groups such as Zebulonites were handled. But this technique, broadly interpreted, continues to be used today; "voice activated locks" are cheaply available from a wide variety of sources.

The arrival of modern statistics made it possible to investigate questions of authorship in a more sophisticated fashion, and the development of modern computers and large corpora have made it practical to investigate these questions algorithmically via information retrieval techniques. With the advent of corpus linguistics, authorship attribution (or to use another near-synonymous term, "stylometry") has become a popular and productive area of research. The principles, however, remain the same.

We can thus define "authorship attribution" broadly as any attempt to infer the characteristics of the creator of a piece of linguistic data. This is a deliberately broad definition; for example, it would include most of the research done into speech recognition. On the other hand, many of the techniques of authorship attribution could be applied to speech recognition, and may in fact be more useful in some circumstances, such as when the writer of the speech is the person of interest and different from the actual speaker. In general, most of the focus is on written text (or on aspects of spoken text that are shared with written text, such as lexical choice or sentence structure) instead of on speech-only aspects like accent or prosody.

In broad terms, there are three main problems in authorship attribution. The first is, *given a particular sample of text known to be by one of a set of authors, determine which one.* This, "closed class," version of the problem is closely related to the second, "open class," version of the problem: *given a particular sample of text believed to be by one of a set of authors, determine which one, if any*, or even *here is a document, tell me who wrote it.* The open-class version is of course much harder to solve, especially if one is required to distinguish among people outside of a small candidate set.

The third problem — for which some researchers prefer to reserve the word “stylometry” or “profiling,” reserving “authorship attribution” only for the first two — is that of determining any of the properties of the author(s) of a sample of text. For example, was the document singly or multiply authored [47]? Was the author a native speaker of English? Of US or British English? A man or a woman [5, 89]?

2.2 Theoretical Background

As will be discussed later, most researchers approach this problem as an issue of feature extraction and data mining. At an intuitive and informal level, nearly everyone is familiar with this process. Many subgroups of people have readily identifiable stereotypical habits of speech and/or writing. For example, the spelling of the word “colour” is typical in the United Kingdom and most of the British commonwealth, but very uncommon in the United States. These dialectological variations can (of course) automatically provide information about the person who wrote a given document. More abstractly, the syntactic constructions a person uses can indicate something about them, and in some cases [17] can even be used to help make (medical) diagnoses.

The assumption of most researchers, then, is that people have a characteristic pattern of language use, a sort of “authorial fingerprint” that can be detected in their writings. Van Halteren [145] has gone so far as to term this a “human stylome,” a specific set of measurable traits that can be used to uniquely identify a given author. There are good theoretical reasons for assuming that such a trait might exist. Since every person has to learn “language” by themselves, and their experiences as language learners differ, so will the “language” they learn differ in micro-aspects. On the other hand, there are also good practical reasons to believe that such fingerprints may be very complex, certainly more complex than simple univariate statistics such as average word length or vocabulary size.

A key new development in recent research has been the development of multivariate techniques tuned to distributional features instead of the mere presence or absence of features. For example, instead of looking at specific words, one might focus on properties such as average word length or vocabulary richness. Unlike specific words, these properties are always available. Similarly, use of multiple features may produce an improvement over univariate features because they provide more information overall. Furthermore, these improved features may be more reliable (as discussed in Section 8.4) because they are less susceptible to direct manipulation.

2.3 Historical Overview

2.3.1 Classical Approaches

An excellent survey of the history of stylometry is that of [56] (see also [57]). The earliest statistical publication that Holmes identifies is that of [106], who "proposed that word-length might be a distinguishing characteristic of writers." Other scholars trace the origins to a letter by De Morgan to a clergyman on the subject of Gospel authorship, suggesting that the clergyman"try to balance in your own mind the question whether the latter [text] does not deal in longer words than the former [text]. It has always run in my head that a little expenditure of money would settle questions of authorship this way.... Some of these days spurious writings will be detected by this test"[32]. This idea is at least superficially plausible, in that authors with large vocabularies may typically use longer words. Unfortunately, studies, including those of [133] have shown that average word length is neither stable within a single author, nor does it distinguish between authors. In Smith's words (cited by Holmes), "Mendenhall's method now appears to be so unreliable that any serious student of authorship should discard it."

Since that time, many other statistics have been proposed and largely discarded, including average sentence length [153], average word length [93], average number of syllables per word [45], distribution of parts of speech [134], type/token ratios [139], or other measures of "vocabulary richness" such as Simpson's D index [131] or Yule's "characteristic K" [154]. None of these methods have been demonstrated to be sufficiently distinguishing or sufficiently accurate to be reliable (see [61] for a specific discussion and strongly negative assessment).

However, failure of one specific instantiation, of course, does not invalidate the entire approach.

The underlying theory behind these approaches is that an authorial "fingerprint" can be extracted as a summary statistic — a single or small set of features — from a given text, and that different authors will vary, noticeably and consistently, along this statistic. A slightly different approach to mine is several different texts to be compared, looking for features along which they differ in a reliable and noticeable way. An example of this approach is that of Ule [143], who proposed using "Relative Vocabulary Overlap" (RVO) as a measure of the degree to which two texts draw from the same vocabulary. The advantage of this type of approach is that it can be sensitive to differences that the summary statistics hide; a major disadvantage is that the calculations of "difference" cannot be done on a per-document basis, but instead on each pair of documents, approximately squaring the amount of effort required for analysis.

A more serious problem with the Ule [143] approach in particular is that it may permit topic to dominate over authorship. Any two retellings of *Little Red Riding Hood* will probably incorporate the words "basket" and "wolf"; any two newspaper articles describing the same football game will mention the same teams and star players; indeed, any two reports of football games will probably mention "score" and "winner," and not "basket." Therefore, any measure of vocabulary overlap will find a greater match between documents of similar topic.

2.3.2 The *Federalist* Analyses

One approach that has been suggested [34] is to focus on synonym pairs. An author has, for example, almost complete freedom to choose between the words "big" and "large"; neither the structure of English grammar nor the meanings of the words place any constraints. By observing that one author consistently makes one choice and another the opposite, one has a noticeable, topic-free, and consistent way to differentiate.

Mosteller and Wallace attempted to apply this technique to the *Federalist* papers, but found that there were not enough synonym pairs to make this practical. Instead, they focused on so-called *function words*, words like conjunctions, prepositions, and articles that carry little meaning by themselves (think about what "of" means), but that define relationships of syntactic or semantic functions between other ("content") words in the sentence. These words are therefore largely topic-independent and may serve as useful indicators of an author's preferred way to express broad concepts such as "ownership."

Mosteller and Wallace therefore analyzed [112] the distribution of 30 function words extracted from the text of the various *Federalist* papers. Because this analysis has become arguably the most famous and widely cited statistical analysis of authorship, and because the *Federalist* papers themselves have become a touchstone for new methods of authorship attribution, (although later scholars have sharply criticized both the study itself and its many followups and replications) it is worth discussing the problem itself.

The *Federalist* papers are a set of newspaper essays published between 1787 and 1788 by an anonymous author named "Publius," in favor of the ratification of the newly proposed Constitution of the United States. It has since become known that "Publius" was a pseudonym for a group of three authors: John Jay, Alexander Hamilton, and James Madison. It has since become generally accepted that of the 85 essays, Jay wrote five, Madison wrote 14, and Hamilton wrote 51, with three more essays written jointly by Madison and Hamilton. The other 12 essays, the famous "disputed essays," have been claimed by both Madison and Hamilton.

Modern scholarship is almost unanimous in assigning authorship of the disputed essays to Madison on the basis of traditional historical methods. Mosteller and Wallace were able to make this determination purely on the basis of statistically inferred probabilities and Bayesian analysis.

Because of the circumstances of this problem, the *Federalist Papers* are almost a perfect test-bed for new methods of authorship attribution. First, the documents themselves are widely available (albeit with many potential corruptions, as will be discussed later), including over the Internet through sources such as *Project Gutenberg.* Second, the candidate set for authorship is well-defined; the author of the disputed papers is known to be either Hamilton or Madison. Third, the undisputed papers provide excellent samples of undisputed text written by the same authors, at the same time, on the same topic, in the same genre, for publication via the same media. A more representative training set would be hard to imagine.

For this reason, it has become almost traditional to test a new method on this problem. Studies include [58, 99, 122, 142]; Rudman [127] lists no less than nineteen studies of this particular corpus and is hardly complete. Perhaps needless to say, almost all of these studies confirm this particular assignment of authorship and the correctness of Mosteller and Wallace's results. As a result (as will be seen in later sections) the idea of mining function words for cues to authorship has become a dominant theme in modern research.

2.3.3 Controversies: Cusum and the *Elegy*

If Mosteller and Wallace are stylometry's best known success, it is equally important to discuss the best known failures. The cusum or Qsum technique, an abbreviation for "cumulative sum," [10, 11, 39, 111] is a visual method for observing similarity between sequences of measures. As applied to sequences in general, one first takes the sequence, e.g., $\{ 8, 6, 7, 5, 3, 0, 9, 2 \ldots \}$ and calculates the mean (in this case, 5). One then calculates the differences from the mean $\{ 3, 1, 2, 0, -2, -5, 4, -3 \ldots \}$ and plots their "cumulative sum" $\{ 3, 4, 6, 6, 4, -1, 3, 0 \ldots \}$. This plot measures the homogeneity or stability of a feature — in the case of cusum, the feature is traditionally something like "percentage of words with two or three letters." (but see Section 4).

This technique was rapidly adopted and used in several English court cases [including The Queen vs. Thomas McCrossen (Court of Appeal, London 1991), The Queen vs. Frank Beck (Leicester Crown Court 1992), and The Queen vs. Joseph Nelson-Wilson (London 1992)] as a forensic technique. Unfortunately, the accuracy of the technique almost immediately came under question; reports such as [24, 52, 53, 55, 60] suggested that the theory was not well-grounded and that the results were not accurate enough to be relied upon (especially given that the case mentioned were criminal cases). However, the ultimate downfall happened when “he was challenged on live British television to attribute texts that he had never seen. The result was disastrous; despite his impressive statistics and his fancy computer graphics, Morton could not distinguish between the writings of a convicted felon and the Chief Justice of England” [50].

Despite this failure variations on cusum such as WQsum (“weighted cusum”) continue to be used [15, 135, 136] and have some evidence for their validity. As will be seen, frequency analysis continues to be a major component of many successful algorithms. But the negative publicity of such a failure has cast a substantial shadow over the field as a whole.

2.3.4 Foster and the *Elegy*

Another noted episode (and noted failure) in the history of authorship attribution is the work in the late 1990s, of Don Foster [41, 44]. Grieve [50] provides a good overview of the controversy, but in summary, Foster found, by applying a battery of stylometric tests, that the relatively obscure poem “A Funeral Elegy” by “W.S.” was, in fact, the work of William Shakespeare. To suggest that this was controversial is to understate the case; the combination of a relatively well-known scholar and an iconic author propelled the finding to the front page of the New York Times. Other work by Foster, most notably in the 1995 identification of Joe Klein as the author of *Primary Colors*, was to follow, and by the mid 1990s, Foster was established as arguably the world’s best known “literary detective” [43].

Traditional Shakespearean scholars reacted with scorn and disbelief, often on the basis of traditional objections such as writing style and content (for example among many, “That the supreme master of language, at the close of his career, could have written this work of unrelieved banality of thought and expression, lacking a single memorable phrase in its 578 lines, is to me unthinkable” [140]). As Foster himself pointed out [42], this is an insistence on “a radically aestheticist ideology whereby the scholar’s literary sensibilities must overrule bibliographies and empirical evidence,” and that apparently “the elegy is not good enough to reflect the genius of a poet who never wrote a blottable line." "The Shakespeare attribution now rests on a broad and substantial foundation. What's required to dislodge it is not just the overthrow of a few minor points [...] but a systematic rebuttal."

However, such a rebuttal was in the works. Other applications of stylometry by other scholars, among them Elliot and Valenza [35, 36, 37] McDonald Jackson [68] and Brian Vickers [148] have uncovered other evidence refuting the Shakespearean attribution, and (perhaps more seriously) suggesting substantial mishandling of the *Elegy* in the course of Foster's analysis. Their analysis and debate, conducted for several years in the pages of *Computers and the Humanities*, eventually was able to establish to the satisfaction of all parties [108, 148] that the *Elegy* had not been written by Shakespeare, but was much more likely to be from the pen of John Ford. By 2002, even Foster himself accepted that.

From a purely scientific perspective, this cut-and-thrust debate can be regarded as a good (if somewhat bitter) result of the standard scholarly process of criticism. Unfortunately, for many non-specialists, this well-publicized failure was their only exposure to the discipline of stylometry, and the greater the hype that Foster managed to create, the greater and more notable the eventual fall. This public collapse of a well-known stylometric attribution may have unfortunately created a public perception of inaccuracy, hindering uptake of the results of attribution research by mainstream scholars. Indeed, given the overall accuracy of Foster's studies, the perception is probably greater than the reality, and much of the hindrance is unjust.

2.4 Methodological Issues in Authorship Attribution

2.4.1 Does it Work?

As can be seen from the previous section, a lot of work has been done on authorship attribution. The key question, however, is *Does it actually work?* In some areas (for example, in the application of authorship attribution to legal issues), the issue of accuracy is crucial. Without a well-established (and documented) reason to believe the analysis to be accurate, it is not even admissible evidence (see Section 8.3.1 for more details). A detective desperate for leads may be willing to grasp at any straw available, but few scholars would like to rely on techniques akin to guessing wildly in pursuing their studies. In light of the many public failures described above, questions of accuracy are probably the most important issue facing stylometry today.

2.4.2 Technical Accuracy

There are three major, interrelated, aspects of the accuracy question. The first is the inherent accuracy of the techniques themselves. As was touched on earlier and as will be discussed at length in later sections, there are an almost limitless number of techniques that have been proposed for doing authorship attribution. Most of these techniques have been shown to work in small-scale experiments, but there is little cross-validation or comparison. Given the tremendous number of techniques, which ones are reliable? Given the tremendous numbers of analytic conditions, which techniques are most reliable in any particular circumstance?

Issues of genre, representativeness, and corpus size combine to make these technical issues more critical. For literary studies, it may be reasonable to expect to have several hundred thousand words by the same author available in order to construct a training sample, but the sample sizes available in a typical forensic investigation may only be a few hundreds or thousands of words [26]. Perhaps more crucially, even for authors whose published works extend to millions of words, the document of interest may be in a completely different and incompatible genre. Given the well-understood [13] difference in the statistics, and therefore fingerprints, across different genres, identifying an accurate *cross-genre* technique is an even more difficult task.

The ease of writing computer programs to implement whatever techniques are found provides a seductive path to misuse and to unwarranted claims of accuracy. Any skilled programmer could write a program, for example, to calculate average word length [106] and to compare a variety of documents on the basis of that calculation. This almost invites abuse by application outside of its area of validity (a document in French was written by the same author as a document in English, because they are both far different from a set of control documents in Chinese?), but the end user may either not be aware of these limitations or may deliberately choose to ignore them. The well-known tendency to trust the output of a computer analysis as reliable — *Garbage In, Gospel Out* — adds to this problem. For this reasons, some authors, most notably Rudman, have argued against the creation of general-purpose authorship attribution programs at all, believing that their risk of misuse exceeds the benefits from their use.

Even when the techniques themselves are accurate, that still may not be adequate for a research study. "Because the computer said so" is not really a satisfactory explanation for many purposes (this is also one of the major weaknesses with artificial neural networks as decision makers, and one of the reasons that expert systems are usually required to provide explanations of the reasons underlying their decisions). A jury may be reluctant to decide based only on "the computer," while a researcher interested in gender differences in writing may be more interested in the reasons for the differences instead of the differences themselves.

2.4.3 Textual Considerations

It should be uncontroversial that a corrupt sample makes a bad basis for analysis; this is no less true for sample texts. However, determining a clean text for authorship analysis can be a very difficult, or even impossible task. In theory, only the features for which the author is directly responsible should be used for analysis. In practice, it can be very difficult to determine which features are the author's. Most published works are a product of many hands, including the author's, the editor's, the type-setter's, and possibly the printer's as well. When using machine-readable texts (such as from Google Scholar, JSTOR, or Project Gutenberg, for example), the scanning or retyping process may have introduced corruptions of its own.

Even with "pure" texts, the presence of extraneous (non-authorial) material can be a problem. Running heads, section headings and numbers, page numbers, and so forth can shift statistics substantially. A more subtle issue can arise from quotations. Although an author selects quotations to incorporate, and they can provide clues to authorship (for example, a Protestant writing a religious tract is unlikely to cite from the Catholic Apocrypha), the words used are not the author's own and will skew the statistics. In extreme cases, the author may not even mark quotations or be aware that his phrases are borrowed from a favorite author. For the cleanest possible samples, all extraneous material that did not come from the author's pen (or keyboard) should be eliminated, a task requiring extreme care and knowledge on the part of the researcher (to be discussed further in the next subsection).

Finally, the selection of materials is critical. In the words of Rudman [125], "do not include any dubitanda — a certain and stylistically pure Defoe sample must be established — all decisions must err on the side of exclusion. If there can be no certain Defoe touchstone, there can be no ... authorship attribution studies on his canon, and no wide ranging stylistic studies." Of course, this degree of conservatism carries with it its own risks and problems; by excluding questionable documents, the "range" of Defoe's style is (possibly) reduced, and in particular, documents outside of the core genres in which Defoe has written may be less accurately analyzed. Selection and preparation of control texts carry with them similar problems — which authors, which documents, and which parts of documents are to be included?

Forensic analysts such as Chaski have replicated these warnings [27]: "Known writing samples should be authenticated independently and reliably. If samples cannot be authenticated or if there is a possibility that a suspect or attorney may not be telling the truth about the authorship of a known writing sample, it is important to seek out other known samples that can be authenticated." In the event that such samples cannot be found, it is almost certainly better not to analyze at all.

2.4.4 Analyst Understanding

Rudman [126] further points out that many of these problems are consequences of the fact that "most non-traditional authorship attribution researchers do not understand what constitutes a valid study." The question of analyst competence is the second major aspect of the accuracy question. In light of the pitfalls discussed above, "does someone trained in physics know enough about Plato and all that is involved with the study of the classics to do a valid authorship attribution study of a questioned Plato work"? Rudman [127] gives the example elsewhere of the methodological problems of the Mosteller/Wallace study of "The Federalist Papers." In his words

> In the Mosteller and Wallace study, a "little book of decisions" is mentioned. This "book," originally constructed by Williams and Mosteller, contained an extensive list of items that Mosteller and Wallace unedited, de-edited, and edited before beginning the statistical analysis of the texts — items such as quotations and numerals. Unfortunately, neither Williams and Mosteller nor Mosteller and Wallace published the contents of this "little book of decisions" and only mention five of their many decisions in the published work. The little book has been lost and cannot be recovered or even reconstructed.

Rudman argues that the existence (and loss) of this book may hide critical and invalidating evidence. Certainly, without this book, no direct replication of the Mosteller/Wallace study is possible. What "decisions" have they made that might have biased their results? And perhaps most importantly, upon what basis did they make their decisions — and are their decisions supportable in light of other scholarship by historical specialists?

2.5 What Would an Ideal Authorship Attribution System Look Like?

In light of these difficulties, what should be done? Some researchers, most notably Rudman, argue that authorship attribution simply should not be automated, that the difficulties and potential pitfalls outweigh the value of any possible product. This review takes a more optimistic view, while hopefully not downplaying the difficulties.

The starting point is the observation that authorship attribution is a problem of wide and cross-disciplinary interest. The people who are interested in the results of authorship attribution are not necessarily either statistics professionals or literature experts, but may include teachers looking for signs of plagiarism, journalists confirming the validity of a document, investigators looking at a crime, or lawyers arguing over a disputed will. It is therefore necessary for the people who are statistics and language experts to make the fruits of their research as widely available as possible.

Accuracy of results is of course the primary consideration; a program that gives results that are too wrong too often is simply unacceptable. On the other hand, perfection is probably not necessary to make the program useful. What is more important to the causal user is probably a built-in confidence or reliability assessment, and to the greatest extent possible, a built-in safety net to manage and if necessary enforce the concerns and issues detailed in the previous section. For example, the ability to select the algorithms and features to use dynamically, based on the language, genre, size, of the available documents would be a nice feature to have. The ability automatically to detect areas of suspect validity (quotations, editorial insertions, phrases in foreign languages, footnotes, chapter headings and footings, etc.) would also be a nice, if almost certainly impractical, feature with current technologies.

In support of this accuracy, a good system would also have a track record; a well-validated set of results in a variety of areas and problems to help define both the expected levels of accuracy and demarcate areas where levels might be expected to drop. Implicit in this is a standardized and useful way to define the accuracy of such a program, whether by standard IR measures like precision/recall, ROC curves [138] or by application-specific measures that have yet to fully emerge.

An explanation facility — detailing the key features that allowed the determination of authorship to be made — would help make the results of the program more useful to a wide but unsophisticated audience. The standard issues of interfacing and ease of use would apply to this as to any other software package.

At least at the current stage of technology, modularity and extensibility are key features. Since the ideal feature set has not yet emerged, it must be possible to add, remove, combine, and test new proposals.

Finally, an ideal system would be theoretically well-founded. The effectiveness of the best features would be explainable in terms of linguistic, psychological, and/or neurological features of the writing process.

Few of these features are now available, and some may turn out to be impossible. However, unrealistic this wish list eventually becomes, it is nevertheless important to have a goal and to bear it in mind as one explores the current state-of-the-art.

3

Linguistic and Mathematical Background

3.1 Mathematical Linguistics

3.1.1 Language Models

Human language can be a very difficult system to study, because it combines a maddening degree of variability with surprisingly subtle regularities. In order to analyze language with computers, it is usually necessary to make simplified models of language for analysis.

In general, we can observe that a text is structured as an ordered stream of separate "events" drawn from a population of potential events. These events may be sounds, letters, words, or perhaps even phrases or sentences. Of course, language is not the only system that can be so described; researchers have studied other "paralinguistic" [76] systems (such as music). Furthermore, the relationship between different events in the same stream is not purely random, but governed by high-order regularities.

There are three major models in use to treat these regularities. The most sophisticated (and psychologically plausible) are *context-free grammars* (CFGs) and their extensions. In general a context-free grammar [65] is a set of rewrite rules that permit abstract symbols (typically representing grammatical categories) to be re-written as strings of other categories and specific words. For example, (in English) a prepositional phrase (usually symbolized/abbreviated as PP) might be rewritten (→) as a preposition ($PREP$) followed by a noun phrase (NP). A noun phrase, in turn could be rewritten as a common noun, an article followed by noun, or an article followed by one or more adjectives followed by a noun. Thus we have the following grammar in partial description of English:

- $PP \rightarrow PREP\ NP$
- $NP \rightarrow NOUN$
- $NP \rightarrow ART\ NOUN$
- $NP \rightarrow ART\ ADJ\text{-}LIST\ NOUN$
- $ADJ\text{-}LIST \rightarrow ADJ\ ADJ\text{-}LIST$

This model is often used for describing computer languages and gracefully captures some long-distance structural dependencies (such as the idea that for every opening brace, a function must also have a closing brace, or the idea that a prepositional phrase must end with a noun, no matter how many adjectives intervene). Unfortunately, this model can be computationally intensive, and fails to capture many other types of dependencies such as lexical or semantic dependencies (e.g., prepositional phrase attachment, the difference between "open the door with a key" and "open the door with a window") More psycholinguistically plausible models, such as context-sensitive grammars, are available but are still more computationally intensive yet, to the point that they are not often used.

An intermediate level of complexity is to use a model such as Markov chains, where language is modeled by each word being a probabilistic function of a fixed window of several adjacent words in context. This captures short-range dependency structures, but does not handle long-distance (beyond the fixed window) syntactic patterns.

More often, language is treated as a simple "bag of words," a collection of every word that appears in the document without regard to order (more accurately, a bag of words is a collection of every word *token* that appears; in a million-word sample, the word *type* "the" might appear sixty thousand times, and therefore show up in the bag as sixty thousand separate tokens).

3.1.2 Distributions

Another key property of language is that the distribution of words and phrases tends to be highly nonuniform and contextually irregular. Some words are much more common than others, and of course some contexts may raise or lower the probability of a given word dramatically (the overall probability of any given English word being "the" is approximately 7%, but the probability of an English word being "the" immediately following another "the" is close to 0%).

The most commonly accepted (if somewhat inaccurate) approximation to the distribution of words is the so-called Zipf distribution [158], a variation on the zeta distribution. In this distribution, the frequency of an item is inversely proportional to its rank in the distribution. For example, in the Brown corpus [18], the word "the" appears 69836 times out of slightly above one million words. The second most common word, "of," appears about half as often (36365 tokens), while third-place "and" appears about a third as often (28286 tokens). The effect is thus a small number of extremely frequent words (only 8 types appear with frequency greater than 1%, only 95 appear with frequency greater than 0.1%), and an extremely long tail of words that appear only once (43% of the total number of types) or twice (13% of types) (of course, this is only an approximation; for example, token counts are integers. More seriously, if the vocabulary of a language is assumed to be infinite, then Zipf's formulation fails to yield a distribution as it fails to converge).

In addition to simple frequency diversity, the properties of words vary dramatically with their frequencies [158]. For example, common words tend to be shorter, to be less specific in meaning, to be less decomposable into meaningful units, and so forth. Perhaps the most interesting characteristic is that different parts of speech tend to inhabit different parts of the frequency spectrum. Table 3.1 shows some samples from the Brown corpus.

Table 3.1 Samples of high frequency, medium frequency, and low frequency words from the Brown corpus.

High frequency		Medium frequency		Low frequency	
Rank	Type	Rank	Type	Rank	Type
1	the	2496	confused	39996	farnworth
2	of	2497	collected	39997	farnum
3	and	2498	climbed	39998	farneses
4	to	2499	changing	39999	farmwife
5	a	2500	burden	40000	farmlands
6	in	2501	asia	40001	farmland
7	that	2502	arranged	40002	farmington
8	is	2503	answers	40003	farmhouses
9	was	2504	amounts	40004	farmer-type
10	he	2505	admitted	40005	farmer-in-the-dell

In this table, the first ten words have token frequencies varying from about 60,000 (out of a million-token sample) to about 10,000. The second ten are taken from about the 2500th rank, and all have identical token frequency of 44. The last ten are taken from about rank 40,000 and appear only once (In this two cases, the rankings are of course somewhat misleading, as ranks 2496–2505, among others, are all tied). More strikingly, the first list is dominated by function words such as articles, prepositions, conjunctions, the second by verbs, and the third by nouns. This is a typical finding across corpora and languages, although of course the exact numbers vary.

A key feature of these common words, then, is that they are so-called “closed-class” or “function” words with relatively little meaning of their own. These words instead serve to relate other words and to provide cues [49, 110] to the grammatical and semantic structure of the rest of the sentence.

A related linguistic feature is the idea of “markedness.” In general, if the most common words are the shortest, the most common and prototypical expressions of a given concept are equally the shortest. Non-prototypical expressions are typically “marked” by the addition of extra words or morphemes. For example, the English (unmarked) word “car” usually describes a gasoline-powered vehicle. Non-prototypical power sources require extra words: “hybrid cars,” “diesel cars,” “electric cars,” “solar-powered cars” (by contrast, a “truck” is diesel-powered; there’s nothing linguistically magical about gasoline itself). The marking can involve extra words as above, additional morphology ("conventional"/"unconventional"), separate word forms ("pope"/"pontiff"), unusual phonology, stress, and many other possibilities. It can also involve nonprototypical structures ("John dropped the plate"/"The plate was dropped by John," unusually highlighting the patient instead of the agent) or in some cases simple variation from common norms; putting the verb before the agent/subject ("In the clearing stands a boxer and a fighter by his trade") or the object before the subject ("Normally I don't like spicy food, but THIS I like") illustrate some other forms of markedness.

It is precisely these areas of distributional variation that have proven fruitful in authorship research. Both function words and marked constructions represent varying ways in which the author may choose to represent a given concept, and therefore different authors may (reliably) choose different ways.

3.2 Information Theory

3.2.1 Entropy

Authorship attribution research has borrowed heavily from the field of information theory in order to measure how reliable a given cue is. The roots of the discipline lie in [129, 130], where Shannon analyzed all communications as a series of messages along a channel between an information source and a listener, and established information of such a source, measured in bits, is given by the entropy equation

$$H(P) = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{3.1}$$

where P is the distribution of messages for that source and p_i is the probability that message i is sent.

Markedness can also appear in the syntax, for fixed N, the entropy H achieves a maximum when the p_i are all equal (any message is equally probable), dropping to zero when a single probability $p_k = 1$ and all others $= 0$ (only one specific message is likely or even possible). This equation lets researchers not only measure the amount of information in a specific set of messages, but also compare different sets (as discussed below). In addition, the relevant quantity for a specific message ($p_k \log_2 p_k$) provides a measure of the unlikeliness of that specific message.

3.2.2 Cross-Entropy

Implicit in the above formulation is the idea that researchers have an estimate or measurement of the distribution for a given source. In many cases, of course, they will not have exact knowledge. The degree to which an actual distribution P differs from an inferred distribution Q is given by the cross-entropy equation

$$H(P,Q) = -\sum_{i=1}^{N} p_i \log_2 q_i. \tag{3.2}$$

This provides a metric for measuring how much event distributions P and Q differ; it achieves a minimum (the actual entropy of P) when P and Q are identical. The difference, the so-called "Kullback–Liebler divergence," or KL-distance, is thus defined as $H(P,Q) - H(P)$.

3.2.3 Kolmogorov Complexity

A major weakness with the simple formulation of entropy above is the identification of the "event" space of interest. Are events words, parts of words, phrases? Furthermore, the probabilities may disguise a hidden dependency structure unknown to the analysts. For this reason, other researchers rely on a different formulation of "information" in the form of Kolmogorov complexity [97] (also called Chaitin complexity). Kolmogorov complexity measures the informativeness of a given string (not, as in Shannon's formulation, a message source) as the length of the algorithm required to describe/generate that string. Under this formulation, a string of a thousand alternating "a"s and "b"s would be easily (and quickly) described, while a (specific) random collection of a thousand "a"s and "b"s would be very difficult to describe. For a large corpus of messages, this could be used as an operationalization of the average amount of information contained per message.

This illustrates the close relationship between Shannon entropy and Kolmogorov complexity: Shannon's entropy is an upper bound on (and under appropriate assumptions, asymptotically equal to) Kolmogorov complexity. Although the mathematics required to prove this is nontrivial, the result can be seen intuitively by observing that a decompression program *and* a compressed file can be used to (re)generate the original string. A more complex string (in the Kolmogorov complexity sense) will be less compressible, and thus require a larger program + compressed text system to reconstruct.

Unfortunately, Kolmogorov complexity is formally uncomputable, in a strict technical sense related to the Halting Problem. Although it is possible to prove that an algorithm will output a given string (by running it), it is not possible to prove that an algorithm will *not* output a given string — some "algorithms" simply run forever without stopping, and we cannot always tell beforehand which ones those will be. So one can easily show that a given (short) program will output the string of interest, but not that it is *the* shortest possible program, since another, shorter, one might do it. Despite this technical limitation, Kolmogorov complexity is of interest as an unattainable ideal. If, as argued above, Kolmogorov complexity represents the ultimate possible file compression, a good file compressor can be seen as an attempt to approximate Kolmogorov complexity within a tractable formal framework [73]. Again, this can be extended to a measurement of similarity between two strings in the degree to which they compress similarly [12, 96].

3.3 Matrices and Vector Spaces

The preceding discussion should give an idea of the huge number of ways in which differences between texts can be operationalized and measured. It is sometimes convenient to use vector representations to capture a large number of measurements in this way. For example, one could measure the percentage of one-, two-, three-, ... up to ten-letter words in a document as ten individual measurements, then capture these measurements as a vector of ten numbers. This implicitly embeds that document in a ten-dimensional vector space at a particular point. This vector space model should be a familiar construct to most IR practitioners. It also implicitly creates another metric for comparison of texts; two texts that are close in this vector space are close in the aspects of style the vector components measure. This particular vector space implicitly captures elements of style related to word length distribution.

A key problem with vector space models can be the number of dimensions and the concomitant computational burden. For this reason, vector space models are often used in conjunction with one or more methods of factor analysis to reduce the number of variables. This analysis serves two purposes: first, to identify dependencies and correlations within the different vectors, and second, to identify the most salient dimensions of variation within the vector space. Two typical types of factor analysis are principal components analysis (PCA) and linear discriminant analysis (LDA).

PCA is a method of statistical analysis that simply rotates the underlying data space in order to create a new set of axes that capture the most variance in the data. In particular, the first "principle component" is the axis on which the data has the most variance. The second principle component is the axis orthogonal to the first that captures the next greatest variance, and so forth. Two dimensions in which the original data strongly correlated would thus largely be captured by a single new dimension that is the vector sum of the original two. Algorithms for performing PCA (via calculating the singular value decomposition of the covariance matrix) are well-known [147] and relatively efficient. PCA provides an easy method for both reducing the dimensionality of a dataset (by focusing on only the first few, often the first two, principal components) and for creating diagrams for easy visualization or cluster analysis of data.

PCA is not, however, well-optimized for the task of classification. The dimensions that are the most variable are not necessarily the most salient ones. Other methods (notably Linear Discriminant Analysis and Support Vector Machines) can address this by inferring which dimensions are the most informative (information, of course, being task-specific) instead of merely the most variable. Similarly, modern machine learning methods such as support vector machines can deal directly with high-order vector spaces and do not need to perform feature analysis or other dimensionality reduction.

Of course, other classification techniques such as k-nearest neighbor, decision trees, and so forth are equally well-known and can also be applied. The key research question for authorship attribution thus becomes one of feature identification and the proper sort of classification technique to use.

3.4 A Theoretical Framework

In light of these differences among models and analytic frameworks, comparison of two different methods for authorship attribution may be problematic. Furthermore, it may be difficult to merge or hybridize two different methods. To mitigate this, Juola [76, 81] has proposed a theoretical and computational framework in which the different methods could be unified, cross-compared, cross-fertilized, and evaluated to achieve a well-defined "best of breed."

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. One can 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 boldface **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 [94]. 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 (in this case, are letter frequencies or word frequencies a better indicator of language? — a question of possible importance to both cryptographers and to designers of web search engines).

It should be apparent how classical methods mentioned in section can be described in this framework. For example, Yule's average-sentence-length [153] can be captured by canonicizing (removing page breaks and such), treating each sentence as an individual event, and determining the length of each event and taking a document-level average. We will use this framework and model to discuss the more modern approaches in the following sections.

4

Linguistic Features

4.1 Vocabulary as a Feature

The simplest way to confirm or refute authorship is simply to look for something that completely settles the authorship question. As a good example of this, the vocabulary of the Beale cipher [92, 93, 132] would seem to be conclusive proof of a hoax. The story is simple. In 1822, a man named Beale purportedly buried some treasure, leaving behind him a letter and a coded set of directions to find the treasure. This story was published in the 1860s, when the man to whom the letter was entrusted decided to go public. But did it happen that way?

One sentence of the letter of 1822 reads “Keeping well together they followed their trail for two weeks or more, securing many, and stampeding the rest.” Unfortunately, the word “stampede” is first attested by the *Oxford English Dictionary* as appearing in 1844; the earliest variant spelling (“stompado”) dates back to 1826, four years after the date on the letter. Similarly, the 1822 letter speaks of “improvised” tools, a meaning that the word “improvise” would not take until the 1850s. The word “appliance,” also appearing in the letter, had been used by Shakespeare, but was considered obsolete and outdated until it became popular again in the early 1860s [4]. The conclusion is almost inescapable that the "Beale letters" are in fact late 19th century forgeries. One would be equally suspicious of a Hitler diary that mentioned the Hiroshima bombing or the construction of the Berlin wall.

It is clear, then, that the individual words an author uses can be strong cues to his or her identity. The vocabulary labels the document as written at the time when the vocabulary existed, and most likely when it was current. Specific words can label the author by group identity — as hinted at in an earlier section, an author who writes of "colour," "honour," and "ironmongers" is likely to be British. Similarly, one who writes of sitting on a "chesterfield" is not only presumptively Canadian, but an older Canadian at that [33]. Studies such as [70] have found many examples of words that vary strongly in their usage across space and time. Idiosyncratic misspellings can even identify people as individuals. Wellman [149] gives an example of using a specific misspelling (the word "*toutch") to validate the authorship of a disputed document in court. He had noted this particular misspelling, and (under the guise of eliciting handwriting samples) was able to persuade the witness to write a phrase containing "touch" while on the witness stand. Showing that she had in fact written "*toutch," Wellman was able to satisfy the court that was how she spelled that particular word, and hence that the document with that misspelling had been written by the witness.

There are two major problems with this kind of analysis. The first is that it is relatively easy to fool, because the data can be faked easily. As an American writer, I nevertheless know all about "ironmongers" and "colours"; in fact, journals will often make demands like "standard British spelling should be used throughout," and expect their Yankee authors to be able to adjust their spelling appropriately. Editors will also routinely adjust spelling of individual words prior to publication. Thus the presence or absence of any individual word, especially in a published document, may not be compelling.

A more serious problem is that the words themselves may not be present. Idiosyncratic spelling or not, the word "touch" is rather rare. The million word Brown corpus [94] lists only 87 tokens (the same as the word "battle") or approximately one instance per 30 pages of text.

"Honor" is even rarer (66 tokens), and "stampede" gets only four. An attribution method requiring, on average, a quarter of a million words before it becomes useful is not going to see widespread application.

4.2 Vocabulary Properties as Features

A more reliable method, then, would be able to take into account a large fraction of the words in the document. One way to do this is to look at large-scale overall statistics. As discussed above, each word in the document has an age that can be used to date the document. Similarly, every word has other properties such as length, number of syllables, language of origin, part of speech, and so forth. Many classical approaches (as discussed in Section 2.3.1) have tried using this sort of superficial feature with simple summary statistics (e.g., taking the mean of the data). For example, Kruh [92, 93], following [111, 150], calculated the average sentence length in the Beale manuscripts as further evidence that they were forgeries. One particularly compelling measure is "vocabulary richness," a measure of the estimated size of the author's vocabulary. Grieve [50] cites not less than fifteen different ways of estimating this, ranging in dates from the 1940s to the 1980s. In a pre-computer age, this probably describes the limit of the feasible. Unfortunately, these methods have not been sufficiently accurate to rely upon. It remains intuitively plausible that different people have different preferred vocabulary (and different vocabulary sizes); it is not clear at this writing whether other, more sophisticated methods of inference could achieve decent accuracy based on data such as word length, and further work is obviously required.

Another approach is to limit the "features" to a reasonably sized subset of the document vocabulary. The synonym pairs of Mosteller and Wallace [112] have already been mentioned, as has the idea of "function words." In both of these cases, the intuitive idea is that the words in the chosen subset are both more common than any individual words, but more likely to vary in an interesting and informative way. Another advantage that function words have is that they are relatively easy to spot from their high frequency alone — many researchers [8, 20, 63, 145] have thus used "the most frequent N words in the corpus" as a stand-in for a more principled definition of such function words. Given the current state-of-the-art, it is probably fair to say that the accuracy baseline as currently established, as well as the most well-known techniques, derive from simple statistics such as PCA, applied to the top 50 or so words of the corpus [20, 22].

4.3 Syntactic Properties as Features

One reason that function words perform well is because they are topic-independent. But the reason for this topic-independence is itself interesting. Function words tend to be semantically "bleached," and relatively meaningless, as one can see by trying to define "the" or "of" in a satisfactory and non-circular manner. Instead, function words describe relationships between content words; i.e., syntax. For instance, the main "function" of the word "of" is simply to establish a (fairly vague) relationship between two nouns. To the extent that other prepositions do the same thing, there is a large degree of synonymity (is so-and-so a student "at" a particular university, "in" a particular university, or "of" that university?) and hence free variation and personal choice in the statistics of the specific words. They also may reflect larger-scale synonymity such as between active and passive constructions, the use of rhetorical questions against simple declaratives, or the use of conjunctions instead of a series of individual assertions.

In other words, a person's preferred syntactic constructions can also be cues to his authorship. One simple way to capture this is to tag the relevant documents for part of speech (POS) or other syntactic constructions [137] using any of the standard taggers [16, 31]. Punctuation can be another easily accessible source of such information [25]. The downside of such processing, especially for POS taggers, is the introduction of errors in the processing itself; a system that cannot distinguish between contraction apostrophes and closing single quotes or that can only tag with 95% accuracy will conflate entirely different syntactic constructs, muddying the inferential waters. An alternative approach that combines lexical and syntactic information is the use of word N-grams (bigrams, trigrams, etc.) to capture words in context. In this scheme, the bigram "to dance" is clearly different than “a dance”; distinguishing between people for whom “dance” is primarily a noun or a verb. This allows the inference mechanism to take advantage of both vocabulary and syntactic information.

Of course, nothing prevents one from combining these features, for example, by analyzing POS N-grams [7, 9, 89] as a more-informative alternative to individual POS tags. This list is of course illustrative, not exhaustive.

4.4 Miscellaneous Features

4.4.1 Orthographic properties as features

One weakness of vocabulary-based approaches is that they do not take advantage of morphologically related words. A person who writes of “dance” is also likely to write of “dancing,” “dances,” “dancer,” and so forth. Some researchers [71, 74, 117] have proposed instead to analyze documents as sequences of character. For example, the character 4-gram “danc” is shared in the examples above, something that a pure vocabulary analysis would miss (although of course this could be addressed via stemming as a pre-processing step). This information could also be obtained via a more in-depth morphological analysis.

4.4.2 Structure and Layout as Features

For many kinds of text, web pages, presentations, and WYSIWYG’ed documents among them, aspects of layout can be important clues to authorship [1]. These tools give authors control over text formatting and layout, including specific aspects such as font choice, font sizing, placement of figures and clip art, or the use of color. Similar features for typewritten documents could include the use of justification, the number of spaces that follow punctuation such as periods/full stops, the location of tab stops, and so forth.

Computer code, in particular, may lend itself to this kind of analysis, due to the importance many authors place on “style” in their programming. The difference between

```
int foo(int i, int j) {
        //  comment here
```

and

```
int
foo(int i, int j)
{
        /*  comment here */
```

or even

```
//  comment here
int foo(int i, int j) {
```

is almost entirely irrelevant to anyone except the author of the code, and yet strongly supported by most development environments.

There are, of course, two chief weaknesses of this particular feature set. The first is that they are very strongly register-bound, perhaps more so than other features. At least to a limited extent, the vocabulary of a given author is register-independent (if you do not even know the word “zymurgy,” you will not ever use it, regardless of register). The features used to lay out a Web page are not necessarily the same as one would use to lay out a PowerPoint presentations; the coding standards one follows writing C++ are not at all the same as one would use to write LISP, let alone English. The second is that layout is particularly vulnerable, much more so than vocabulary, morphology, or syntax, to editorial tinkering such as pagination, standardization of fonts, and so forth. We can be confident that the author of a scientific journal planned to put figure 1 into her article. Unless she delivered camera-ready copy, or edited the volume herself, she could not possibly have planned to put it at the top of the second column of page 138. Even just transferring a document from one format to the other may result in substantial reformatting.

4.4.3 Anomalies

An interesting approach has been taken by [26, 43, 90], focusing specifically on usage anomalies. This approach cuts across all linguistic levels and instead analyzes linguistic idiosyncracies. The example of spelling mistakes ("toutch") and cultural differences ("colour" vs. "color") have already been mentioned, but differences in grammar and usage could also qualify. To observe this, the document can be compared against a reference standard (using spelling and grammar checking tools) and the points of difference noted. Chaski, in particular, argues [26] that "markedness," as defined earlier, is a key identifier that "pervades all levels of language, from phonetics to morphology to syntax to semantics to discourse," and that markedness *per se* can be a useful feature category. She supports this with a series of experiments on her own corpus [25] (but see also [48, 104]).

4.4.3.1 Metadata

Another fruitful area for investigation is in the use of metadata to determine document authorship. Indeed, under the guise of "digital forensics," this may be one of the best-developed and most understood kinds of authorship attribution. Looking at the headers of the relevant piece of Email shows where it came from, and often other information besides. Even when deliberately masked, experts can often get some kind of useful information from the type of masking — at least enough to know where to ask the next set of questions.

Similar metadata often exists in complex documents. Microsoft Word, for example is famous for burying information in the document such as the date of creation, the program (including location) used to create it, the user who created it, the revision history, the original file name, and many other pieces of information. This information is not ordinarily visible, but is embedded in plain-text within the file itself. From a practical standpoint, finding the name "I. M. Originalauthor" in the metadata may be a clearer "smoking gun" than abstract statistical analysis when one wants to deal with a recalcitrant student-plagiarist.

Similarly, it is possible to embed specific metadata into a document in order to support a later assertion of authorship. This process goes by various names such as "digital watermarking" or "steganography" and is another well-studied and mature discipline. By their very nature, the analysis and evidence usually associated with these processes is not of the same type as the soft statistics described elsewhere in this section.

At the same time, the potential informativeness of metadata should not be overlooked (and could be considered as a generalization of formatting/layout features described earlier) [155]. That a document was written on Microsoft Word at 10 a.m. are two "features" that might tell against a theory of authorship by a notoriously nocturnal Linux advocate, and should be weighed accordingly.

4.5 Caveats

Rudman [124] has estimated that over a thousand different features have been used in authorship attribution studies. Obviously, space prevents a full discussion of each and every one. Equally obviously, no consensus has emerged about what features are the best. It seems likely that no individual feature is universally informative enough, and that the best results will come from analysis of an extremely broad set of features [1, 157] covering many different approaches. This of course raises the question about how to combine information from different feature types, a question best left to the analysis process.

It is also likely that, even with this broad-brushed approach, different languages and registers will still require different features; the discussion of "function words," for example, has been highly English-specific and may not apply directly to strongly inflected languages such as Finnish and Turkish (although see Section 6.3 for an interesting empirical take on this).

4.6 Other Domains

It is interesting to speculate about what other domains might come under this kind of analysis. Natural language, of course, is not the only creative product of the human mind, nor the only thing that is plagiarized. Cases of art forgery are well-known (Hans van Meergan is probably one of the most famous), and in many cases, the "style" of the painting is one of the key points under dispute. Music is another area ripe for plagiarism, as both Joyce Hatto and George Harrison illustrate, and of course copying of source code is the bane of computer teachers everywhere. But, as the discussion above has already illustrated, authorship disputes can be analyzed using the same methods, using an appropriate choice of features.

Features for code have already been discussed; because of the strong influence of natural languages (usually English) on their designs, many of the same lexical and syntactic features useful in English may be useful in analyzing code. In many cases, there are clear-cut examples of synonymity (such as the equivalence between `for` and `while` loops, or between `if` and `case` statements), while function and variable names provide ample opportunity for programmers to show their individual quirks.

Juola [76] has presented a similar analysis for music; focusing on individual bytes (analogous to characters) in the digital representation of songs, he was able to identify the performer of the relevant songs. He also presented several other feature-sets that could in theory be applied to the analysis of music, such as encoding monophonic melodies as a series of numbers representing the number of semitones between each note and the next, encoding each note as its representation on a scale (do, re, mi, etc.), or a Parsons' encoding of whether each note was the same, higher, or lower, than the preceding one (e.g., "* S S D U S S D" is the Parsons' code for Beethoven's 5th Symphony).

What do natural language, computer code, and music have in common? Juola identified three key characteristics. First, like text, both code and music are created from a sequence of distinct events, whether those be keypresses on a typewriter keyboard or a piano. Second, these events are drawn from a finite (or at least finitizable) set; there are only 88 possible notes on a piano keyboard. Compared to the number of characters in Chinese or possible and actual words in English, this is tiny. Third, these events occur in a time-structured sequence. Analysis along the line of "what event is likely to come next" is thus capable of distinguishing usefully in all three domains.

Juola used the term "paralinguistic" to describe such domains. It is interesting to think about what other creative enterprises are paralinguistic. Dance, for example, can certainly be described as a sequence of body shapes; it is not clear whether or not the set of these shapes are finite, but the existence of various dance notations [141] suggests that they are at least finitizable. A future researcher might examine different dancers in an effort to find key aspects of a choreographer's "style." Gestures, as in the game of Charades, is another type of sequenced set of body positions, and may prove to be paralinguistic. Film, of course, is a sequence of frames, but the potential set of frames may be (for all practical purposes) infinite and therefore too large for this sort of analysis to be fruitful. Still art, such as a painting or photograph, is not sequenced, and would therefore probably not qualify.

4.7 Summary

In the proposed framework, the first step in authorship analysis is the identification of the feature set of interest, as preparation for later analytic phase. The variety of proposed feature sets is potentially bewildering, and little consensus has yet emerged about the best features, or even the best types of features, to use. Proposals have covered all levels of language and even included salient non-linguistic features such as document layout.

The end result of this is to reduce a set of documents to a set of ordered feature vectors, which are then analyzed to produce authorship judgments. At the very grossest level, the vectors that are "most similar" are probably by the same author, and the question is simply the best and most efficient way to identify this similarity. Unfortunately, as will be seen in the next section, little consensus exists there, either.

5

Attributional Analysis

Once the relevant features have been extracted from the documents of interest, the next task is to determine by analysis of the features which document was written by which author. Again, the number of methods proposed and used is staggeringly large, and space precludes a full discussion of more than a few representative methods.

A key factor involved in the selection of an analysis method is an understanding of the requirements of the final answer. A forensic authorship report to be presented at a jury trial, for example, needs to be understandable to the general public (the jury) in a way that some specialists may not need. Graphics and visualizations may be important as a way to make authorship distinctions clear. It may be important not just to present the results, but to explain them in terms of underlying aspects of the documents; to some extent, this simply replicates the age-old debate between expert systems and neural networks — "because the computer says so" is not a particularly satisfactory explanation, even if the computer is usually right.

Other important factors involve questions such as the amount and type of training material available. For example, the distinction between "supervised" and "unsupervised" techniques applies here as elsewhere in machine learning. Supervised techniques require *a priori* knowledge of class labels, often in the form of sample documents of undisputed authorship. Unsupervised techniques are more appropriate for data exploration, with no prior information needed. These analytic methods will be addressed separately.

5.1 Unsupervised Analysis

Unsupervised analysis acts without regard to the presumptive attributions already made, and instead looks for superficial patterns in the data. At its simplest level, just making a data scatterplot and looking for groupings is a type of unsupervised analysis. And, still at this simple level, much authorship analysis can be done this way.

5.1.1 Vector Spaces and PCA

With the feature structure defined in the previous section, it should be apparent how documents can be described in terms of collections of features; quantifying the features, in turn, will implicitly create a high-dimensional “document space” with each document’s feature set defining a vector or a point in that space. For example, the token frequency of fifty well-chosen words defines a fifty-place vector for each document (some normalization would probably be necessary). Simple visual inspection of this high-dimensional space may reveal interesting clusters; in particular, if two documents are “close,” they may have similar authors.

There are two major problem with this. The first is just the difficulty of visualizing fifty-dimensional space, while the second is the problem of independence of the various dimensions. In general, the correlation between any two features (across many documents) will not be zero. For example, documents written in the first person will often have a high frequency of first person pronouns such as “I” or “me.” They are also likely to have a high frequency of first person verbs such as “am.” Therefore, these frequency measurements are not independent, since they were really just all measurements of the degree of first-person exposition in the writing (in a more technical phrasing, the correlation, or covariance, among all those frequencies will be positive). This will act to weight the single feature “personness” three times as heavily, since it is represented by three separate features.

To address this, researchers [2, 14, 20, 48, 59] often use principal component analysis (PCA), as described briefly in an earlier section. Technically, PCA is simply the eigenvectors of the covariance matrix among a set of features. Informally, PCA determines a smaller set of (orthogonal, independent) basis vectors that describe as much of the variation in the initial data set as possible. In particular, the two principal components (technically, the eigenvectors with the largest eigenvalues) describe the original data set in an easily visualizable two-dimensional space while preserving as much as possible the similarity and “closeness” between individual data items.

The most popular feature set for this kind of analysis, following the work of Burrows [20], is the token frequency of the top fifty or so most common word types in the document set. Applying this analysis method and plotting the results will often produce clear visual separation between different authors, as in Figure 5.1.

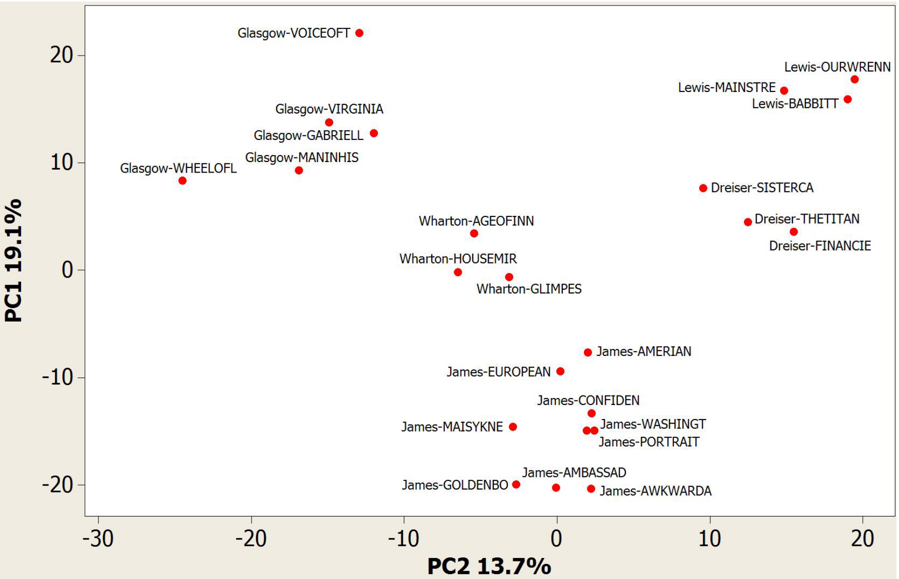

Fig. 5.1 PCA separation of various authors based on function word frequencies (data and image courtesy of David Hoover).

In this figure, Glasgow can be clearly seen to be in the upper left, Wharton in the center, Lewis above Dreiser on the right, and James in the bottom center. Learning that an unknown manuscript was also in the bottom center would be a strong evidence for its authorship by James instead of Lewis or Glasgow.

5.1.2 Multidimensional Scaling (MDS)

An alternative approach for unsupervised data exploration involves the calculation of intertextual differences as distances. A full discussion of such distance calculations would be mathematically dense, but it should be intuitively plausible that any similarity measure can be extended to a "distance." The actual mathematics are slightly more complicated, since the definition of "distance" (or more accurately "metric") involves some properties such as symmetry, the triangle inequality, and the idea that the distance between any entity and itself should be zero. For true distances, however, it is then possible to embed the distance structure in a high-dimensional abstract space while preserving (as much as possible) the distances themselves.

The primary process by which this is done is called multidimensional scaling (MDS). Essentially, MDS operates by placing objects in a space of a previously defined dimension (usually two or three dimensions) such that the total error introduced is minimized. Once this space is created, it can be examined for apparent patterns.

Two examples of this should suffice. Figure 5.2 shows a two-dimensional MDS plot of Juola's "cross-entropy" [71] distance as applied to a set of documents (Kostic, p.c., see also [75]) by three (four, counting the unknown, unrelated, and uninteresting introduction) different authors. The visual separation here corresponds largely to a separation in time; Danilo and author A are both medieval Serbians (author A, known to history as "Student," identifies himself as a student of Danilo's in this work), while B is a later interpolator; the sections written by B are clearly separable, showing that A's style is almost indistinguishable (by this test) from his mentor's, but B is clearly different (and the writing of sample 2, attributed to Danilo himself, is atypical and may stand re-examination).

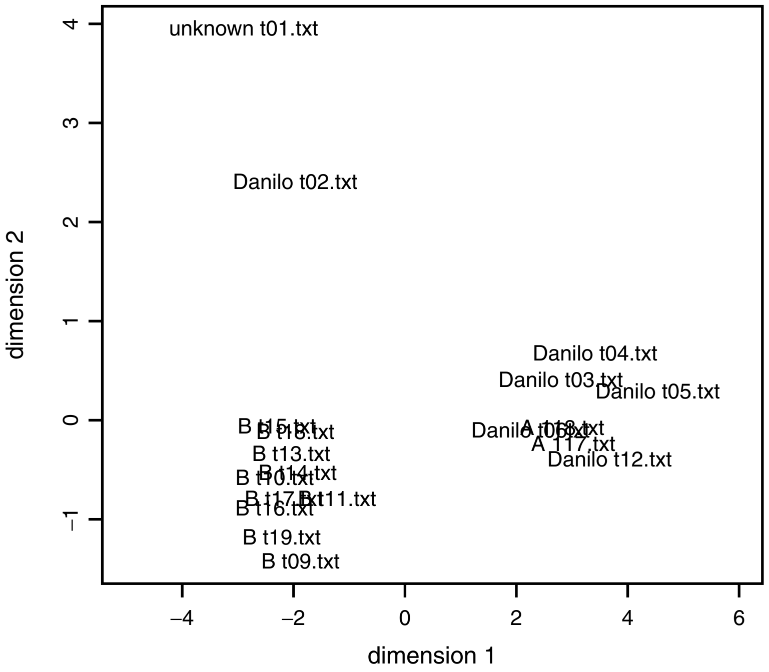

Fig. 5.2 Authorship of *Lives of Kings and Archbishops.*

Figure 5.3 shows two dimensions out of a three-dimensional MDS projection of some novel-length works by Jack London [79]. The numbers indicate date of publication. Here there is no difference in "authorship," but the pattern clearly indicates a visual separation (at the indicated vertical line) between early and late works, happening at about 1912. This, in turn, may indicate a significant stylistic change of unknown type happening at this time — a possible line of inquiry for interested literature scholars.

In both cases, the visual separation provides a strong indication of substantive differences in style.

5.1.3 Cluster Analysis

A third form of unsupervised analysis is (hierarchical) cluster analysis. As with MDS, cluster analysis presumes the existence of a distance measure between document pairs, either measured directly or inferred

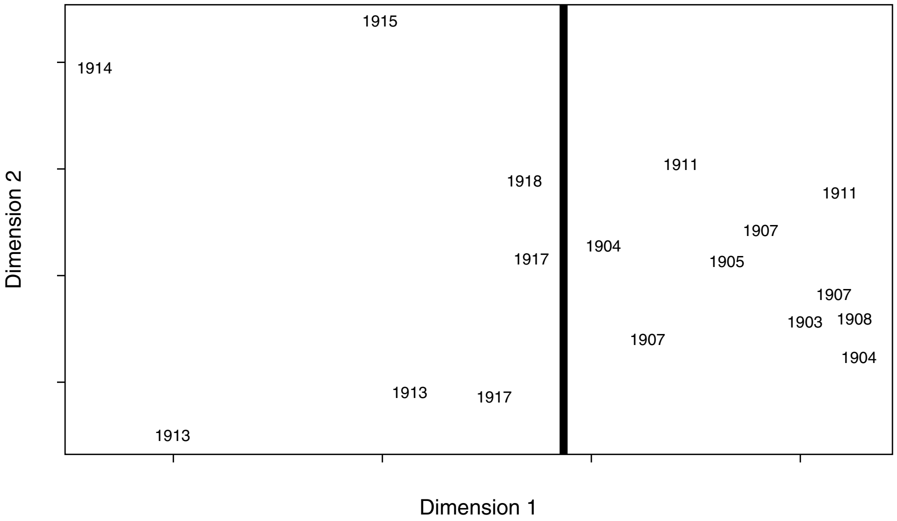

Fig. 5.3 Chronological development of style in the writings of Jack London (data and figure adapted from [79]).

from a metric applied to a vector space. In both cases, cluster analysis proceeds by grouping the closest pair of items into a cluster and then replacing that pair of items by a new item representing the cluster itself. At each step, the number of items analyzed is reduced by one, until finally all items have been joined into a single cluster.

The result of such analysis is usually displayed as a cluster diagram (sometimes also called a dendrogram), a rooted tree with binary branching. The height of each internal node represents the distance at which the closest pair was found, and the children of that node (either individual documents or subtrees) represent the two items joined at that step. An example dendrogram is provided in Figure 5.4. This, again, provides a visual indication about what kind of structure and similarities are present in the document set.

5.2 Supervised Analysis

In contrast to unsupervised analysis, supervised analysis requires that documents be categorized prior to analysis. It is perhaps unsurprising

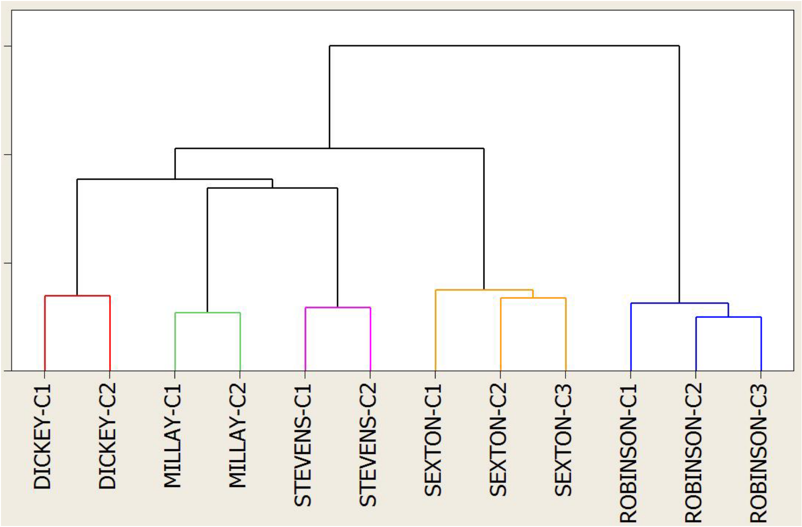

Fig. 5.4 Dendrogram of authorship for five novelists (data and figure courtesy of David Hoover).

that this additional information can be helpful in arriving at methods of categorization.

5.2.1 Simple Statistics

The simplest form of supervised analysis, used since the 1800s, is simple descriptive statistics. For example, given a set of documents from two different authors, we can easily calculate word lengths [106] and (handwaving a few statistical assumptions) apply t-tests to determine whether the two authors have different means. Once we have done that, we can apply logistic regression to estimate the authorship (and our confidence in that authorship) of a novel document (with more than two authors, we could use ANOVA to similar purposes). A similar analysis, using different features, could analyze the mean number of syllables per word, number of words per sentence, percentage of passive constructions, etc. For those unwilling to make the necessary assumptions of independence and normal distribution, non-parametric statistics such as the Wilcoxon test can be used. The mathematics of such statistics is well-known and needs no detailed explanation.

Unfortunately, these simple methods produce equally simple failures. More accurately, no single feature has been found that robustly separates different authors in a large number of cases. But these simple statistics can be and have been combined successfully.

The most notable example of this is Burrows' "Delta" method [21, 64, 138]. In its original form, Burrows analyzed the frequency of the 150 most frequent words in a collection of Restoration poets. For each of these word variables, he calculated an appropriate z-distribution (essentially, an estimate of the mean frequency of the word as well as an estimate of the variance in that frequency). Individual documents were scored on each of the 150 variables regarding how far their frequencies were above/below the norm. A positive z-score indicates a word more common than average, a negative one indicates a word less common than average, and of course a zero z-score indicates a word appearing with exactly average frequency. Burrows then defined the Delta measure as "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" [21]. The smallest Delta, representing the greatest similarity to the training text-group, is the category to which authorship of the test document is assigned.

Performance of Burrows' Delta has generally been considered to be very good among attribution specialists, and it has in many cases come to represent the baseline against which new methods are compared.

Intuitively, Delta can be viewed as creating a 150-dimensional vector space of word frequencies, scaling each dimension by the frequency variation to normalize the degree of dispersion, and embedding individual documents in this space (the analysis to this point is of course unsupervised). He then applies a simple metric to the space (the mean in this case is a simple re-scaling of the L_1-metric) to obtain average distances between the test document and the training documents (by category), selecting the appropriate category based on a variation of the nearest neighbor algorithm (choosing the nearest category instead of document) (see [138] for further discussion of this analysis). Once this general framework has been observed, variations can easily be suggested. Some simple variations would include adjusting the number of word-variables and therefore dimensions or the basis on which the word-variables are selected, adjusting the metric imposed upon vector space, or adjusting the basis on which the judgment is made.

David Hoover [63, 64] has made extensive study of such variations. Examples of the variations that he has studied include changing the number of words studied (ranging from 20 to 800 and beyond), eliminating contractions and/or personal pronouns from the set of word-variables, and "culling" the list of word-variables by eliminating words for which a single training document supplied most (70% [64]) of the words. He found the greatest accuracy occurred in a 700-dimensional space, eliminating personal pronouns but not contractions, and applying culling at the 70% level. By contrast, eliminating contractions "generally reduces the accuracy of an analysis overall," indicating perhaps that the use of contractions is an important indicator of authorship — while personal pronouns are more about the subject of the document than the author.

He also tested [63] variation in the calculation of the Delta score itself, such as replacing the z-score by percentage difference from the mean (and thus ignoring the degree of variability); this avoids the normalization step above. As might be expected this reduces performance slightly. He has experimented with different methods of treating Delta, for example, by limiting the calculations to words with relatively large z-scores (where the frequency difference is more likely to be meaningful), or words where the z-scores have opposite signs in the test and training documents (following the intuition that the difference between "more" frequent and "less" frequent may be more meaningful than the difference between "more" frequency and "a lot more" frequent). He has experimented with "alternate formulas" for Delta, for example, including words with positive difference only (words more frequent in the test document than the training set), words with negative difference only, and weighting words with positive difference more heavily than words with negative difference. The somewhat *ad hoc* modifications change the topology imposed upon the vector space, although the exact change is not clear, and in some cases the topology imposed may no longer be a true "distance" [for example, by violating symmetry conditions — Delta(A,B) may not be equal to Delta(B,A)]. In no case did he obtain substantial improvements over those cited above.

A more theoretical investigation of the underlying mathematics of Delta was given by Argamon [138], who observed that Delta (in its original version) was mathematically equivalent to the above vector space representation, and further that it effectively ranked candidate authors by their probability under the assumption that word frequencies have a Laplace distribution. He noted that the method for estimating the parameters of this distribution was nonstandard, and that an equivalent system could be built using standard maximum-likelihood estimators (MLEs) for the parameters. He further noted that this formalization leads directly to other, less *ad hoc* variations, such as assuming that word frequencies are distributed as a Gaussian, possibly a multivariate Gaussian with the possibility of covariance, or a Beta distribution (Argamon, p.c.).

5.2.2 Linear Discriminant Analysis

Of course, once a feature-based vector space has created and documents embedded in that space, many other methods can be applied to the task of assigning documents to authors. Another major approach that has been used is linear discriminant analysis (LDA). LDA is closely related to PCA in that it tries to find linear combinations of the underlying basis of the vector space that are maximally informative, and thus performs a type of dimensionality reduction. Unlike PCA, LDA works on category-labeled data (and hence can only be used as a supervised algorithm) to infer the differences among the labeled categories, and more specifically, to identify lines or (hyper)planes that best separate the categories. One advantage of LDA over PCA is that the method is free to select the most relevant dimensions (the dimensions along which the category variation is most salient), instead of the ones on which the variation is simply the largest (the principal components).

A good example of this is the Dutch experiment of [8]. This study is somewhat unusual in two respects; first, the texts studied were rather small (averaging only 908 words), and second, they were in Dutch, not a common language for study. The documents studied had been specifically elicited for this study from students at U. Nijmegen, and consisted of nine controlled-topic essays (three fiction, including a re-telling of

Little Red Riding Hood among other topics; three argumentative, and three descriptive) for each of eight students, four in their first year, four in their fourth. This thus provided a tightly controlled corpus that should produce strongly similar essays, and a very difficult test set.

Despite the relatively strong performance of Burrows' function word PCA in other studies, here analysis of "the most frequent function words in the texts shows no authorial structure." Some group structure was revealed, for example, the second principal component served to separate first year from fourth year students, and the first component provided some separation by genre (it is unsurprising that genre is more salient than authorship, and hence on the first principal component). On the other hand, LDA provided evidence of some separability, allowing about 60% of the pairwise comparisons to be made correctly, compared to a chance probability of 50%. Adjusting the weightings of the word frequencies to take into account their by-text entropy raised the accuracy to 81.5%, and adding punctuation marks to the feature set raised it to 88.1%. Baayen et al. took this to be evidence, both that authorship attribution is practical in this tightly controlled circumstances, but also that "discriminant analysis is a more appropriate technique to use than principal component analysis," as it is capable of detecting more subtle changes than simple visual inspection. This point, the appropriateness of different techniques, is of course an important one and one to which we shall return.

5.2.3 Distance-Based Methods

Alternatively, the authorship decision can be made directly, without prior embedding in a vector space (in point of fact, the vector space is really just a convenient representation, since it allows many values to be grouped under one easy-to-understand handle). For example, the distribution of words (or more generally events) can be treated as a simple probability distribution and used to define a set of pairwise interdocument distances using any of the standard, well-known probability difference, such as histogram intersection, Kullback–Liebler divergence, Kolmogorov–Smirnoff distance, and so forth. This can be the basis for a simple k-nearest neighbor attribution algorithm. In its simplest form, this method is not likely to work well because it assumes independence among the event set.

A more sophisticated application of Kullback–Liebler divergence was proposed by Juola [71, 72, 74] as "linguistic cross-entropy" based on the "match length within a database" method of estimating entropy [38, 151]. This avoids (to some extent) the problem of independence as it calculates the probability of an event occurring in the context of its immediately preceding context. Using this method, he was able to outperform [80] the LDA-based results given above on the same test data. Kukushkina et al. [96] used Markov chains to similar effect, calculating a first-order Markov model based on character bigrams for each training author, and then assigning the authorship of a test document to the chain with the highest probability of producing it. In a series of experiments, they reported between 70% and 100% accuracy in leave-one-out cross-validation. One potential weakness of both of these approaches is the combinatoric explosion as the event vocabulary increases; it is much easier to compute and estimate transition probabilities with bigrams from a 27-character alphabet than from a 50,000-word dictionary.

One of the most interesting proposed distance measures involves relative Kolmogorov complexity. In theory, the Kolmogorov complexity of a string is defined as the length of the smallest computer program whose output is that string [97]. Relative Kolmogorov complexity can be defined as the length of the smallest computer program that converts one string into another. Under this framework, authorship can be assigned to the training document that would require the least "work" to convert to the test document. Unfortunately, Kolmogorov complexity is formally uncomputable, but it can be estimated by the use of data compression techniques [73, 77]. Khmelev [96] applied this by defining relative complexity $C(A|B)$ as $|BA| - |B|$, where $|x|$ is the length of document x after compression by some standard method. Using a variety of different compression algorithms, they found that performance was generally poor, but that in some cases, they could outperform a Markov chain analysis similar to [87].

5.2.4 General Machine Learning Techniques

Given the nature of the feature sets, it is unsurprising that other general-purpose machine learning algorithms would have been deployed to solve these problems. Three specific such algorithms are neural networks (parallel distributed processing) [101], decision trees [121], and naive Bayes classifiers [156].

Neural networks are designed using the human brain as a controlling metaphor; they consist of a large number of cooperative simple arithmetic processors. In usual practice, they are arranged in a set of three or more "layers" and trained using a procedure called "backpropagation of error" [128] to minimize the error at the output layer between the desired and produced outputs. The mathematics of such networks has been well-studied [54], and, in general, is very similar to the mathematics underlying LDA; the middle layer performs a dimensionality reduction to combine the most relevant underlying dimensions of the input space, while the output layer identifies separating hyperplanes. Three examples of this application are [100, 107, 142]. One key flaw in neural networks is that, although they will often produce very accurate classifications, it is not clear on what basis they are classifying.

In contrast, decision trees [121] are a machine learning paradigm specifically designed to support descriptive classification and to explain why the classification happened as it did. A decision tree is a recursive data structure containing rules for dividing feature space into a number of smaller sub-cases, and thus implicitly defining a function mapping regions of space (usually separated by hyperplanes) to category names. A typical rule would look something like:

> (Rule 1) **if** feathers = no **and** warm-blooded = yes *then*
> type is MAMMAL **else** apply rule 2.

Rules are inferred by splitting the training data along the maximally informative (binary) division, and the recursive splitting the two resulting subsets until the problem has been solved. One key advantage of decision trees in general is that they can operate well with non-numeric data, and thus are not confined to the vector space models of other methods discussed above. Unfortunately, they appear to be outperformed [1] by other methods, most notably by support vector machines (SVM), at this writing arguably the most accurate classification method known.

Naive Bayes classifiers [156] perform similar classification but without the tree structure. Instead, they rely on a simple computational implementation of Bayes' theorem to infer the probability of a classification scheme to infer the most likely category given the observed data. They are called "naive" because they use simple and unrealistic independence assumptions (for example, they might assume that the frequency of the word "I" is independent of the frequency of the word "me," a patently false assumption), but nevertheless can perform surprisingly well and have the advantage of being extremely fast to develop and to train.

5.2.5 Support Vector Machines

SVMs [19, 146] are a relatively new classification method that manage to avoid the two classic traps of machine learning, "the computational ability to survive projecting data into a trillion dimensions and the statistical ability to survive what at first sight looks like a classic overfitting trap" [109]. They have been applied to a huge variety of problems [19], including handwriting recognition, object recognition, face detection, and, of course, text categorization [69].

The mathematics underlying SVM are unfortunately complex, and space precludes a full discussion. In general, an SVM (specifically, an LSVM, "linear support vector machine") is yet another method of inferring separating hyperplanes in a vector space model, but differs in that it is risk-sensitive, in that the inferred vector is not just a separating hyperplane, but the separating hyperplane with the greatest potential margin of error (meaning the separating hyperplane could be displaced by the greatest distance before introducing a new classification error). A more general formulation involves using nonlinear kernel function to define separating spaces other than hyperplanes.

SVMs have been widely used [1, 91, 157] for authorship attribution and "anecdotally they work very well indeed" [109] on a wide variety of problems. Researchers such as Abbasi [1] (see also [157]) have found that SVMs generally outperform other methods of classification such as decision trees, neural networks, and LDA — which in turn has been shown to outperform simple unsupervised techniques such as PCA.

Does this mean that researchers should simply use SVM and ignore the rest of this section? Unfortunately, the situation is slightly more complex than that (although a good case could be made). What has not been discussed is the degree to which different methods (and feature sets) transfer between domains and authors. Baayen et al. [8] showed, for example, that while function word PCA was adequate in some cases, it was insufficient to distinguish between authors as similar and as tightly controlled as in the Dutch experiment. We may have similar reason to question Hoover's [64] findings that eliminating personal pronouns results in a useful and substantive improvement over Burrows' Delta — not that the findings themselves are wrong, but that they may be more applicable to some forms of text than others. In scientific/technical texts, and specifically in the typical case of a researcher reporting her own results, the difference between writing "we" or "I" or "one" may be of more stylometric interest than in the case of a novel, where the dominant factor may simply be the narrator's perspective. To address this issue, what is really needed is extensive empirical testing of the system as a whole.

6

Empirical Testing

From a practical standpoint, it may not be enough to know that a given method of authorship attribution “works.” Two further questions present themselves — how accurately does the method work, and how widely/reliably? For example, as discussed in the previous section, eliminating personal pronouns from the analysis has been shown [64] to improve performance when applied to novels. Would the same improvement be seen in Web pages? LDA outperformed PCA in an analysis of small, tightly topic-controlled essays in Dutch [8]. Is this a consequence of the size, the topic control, or the fact that the documents were written in Dutch?

In many applications, most notoriously in forensics, knowledge of the expected error rate is a requirement for a technology to be useful. In any circumstance, of course, a professional would want to use the most accurate technology practical (and have a good idea of its limitations). This demands a certain degree of comparative evaluation of techniques.

6.1 Test Corpora

The need for comparative evaluation has been recognized for some time [25, 40, 63, 75, 80]. Forsyth [40] compiled a first benchmark collection of texts for validating authorship attribution techniques. Baayen's corpus [8] has also been used for comparative testing. As cited in the previous section, on this corpus, LDA was shown to outperform PCA (which performed only at chance level, as seen in the previous section). Later studies [80] have shown that cross-entropy can perform even better than LDA, and much faster. From these results it can be concluded that *under the circumstances of this test*, cross-entropy, and in particular, cross-entropy using words instead of characters for the event set, is a more accurate technique for assessing authorship.

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 eight thousand 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 eight hundred thousand 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?

6.2 The *Ad-hoc* Authorship Attribution Competition

6.2.1 Contest Structure

To answer these questions, in 2004 the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH) hosted an "*Ad-hoc* Authorship Attribution Competition" [75, 78] at the joint annual meeting. This remains the largest-scale comparative evaluation of authorship attribution technology to date. By providing a standardized test corpus for authorship attribution, not only could the mere ability of statistical methods to determine authors be demonstrated, but methods could further be distinguished between the merely "successful" and "very successful." Contest materials included 13 problems, in 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, some with several methods, by downloading the (anonymized) materials and returning their attributions to be graded and evaluated against the known correct answers.

- *Problem A* (English) Fixed-topic essays written by 13 US university students.
- *Problem B* (English) Free-topic essays written by 13 US university students.
- *Problem C* (English) Novels by 19th century American authors (Cooper, Crane, Hawthorne, Irving, Twain, and "none-of-the-above"), truncated to 100,000 characters.
- *Problem D* (English) First act of plays by Elizabethan/Jacobean playwrights (Johnson, Marlowe, Shakespeare, and "none-of-the-above").
- *Problem E* (English) Plays in their entirety by Elizabethan/Jacobean playwrights (Johnson, Marlowe, Shakespeare, and "none-of-the-above").
- *Problem F* ([Middle] English) Letters, specifically extracts from the Paston letters (by Margaret Paston, John Paston II, and John Paston III, and "none-of-the-above" [Agnes Paston]).
- *Problem G* (English) Novels, by Edgar Rice Burrows, divided into "early" (pre-1914) novels, and "late" (post-1920).
- *Problem H* (English) Transcripts of unrestricted speech gathered during committee meetings, taken from the *Corpus of Spoken Professional American-English.*
- *Problem I* (French) Novels by Hugo and Dumas (pere).
- *Problem J* (French) Training set identical to previous problem. Testing set is one *play* by each, thus testing ability to deal with cross-genre data.
- *Problem K* (Serbian-Slavonic) Short excerpts from *The Lives of Kings and Archbishops*, attributed to Archbishop Danilo and two unnamed authors (A and B). (Data obtained from Alexandar Kostic.)
- *Problem L* (Latin) Elegaic poems from classical Latin authors (Catullus, Ovid, Propertius, and Tibullus).

- *Problem M* (Dutch) Fixed-topic essays written by Dutch university students. (Data obtained from Harald Baayen, in fact, this is the data set described in the previous section.)

This data thus represents a wide variety of languages, lengths, and genres.

The contest (and results) were surprising at many levels; some researchers initially refused to participate given the admittedly difficult tasks included among the corpora. 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. Similarly, problem A was a realistic exercise in the analysis of student essays (gathered in a freshman 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.

6.3 AAAC Results

Participants in this contest are listed in Table 6.1.

Results from this competition were heartening ("unbelievable," in the words of one contest participant). The highest scoring participant was the research group of Moshe Koppel and Jonathan Schler (listed as Schler, the corresponding author, in Tables 6.2–6.4), with an average success rate of approximately 71% (Juola, the contest organizer, also submitted solutions. In the interests of full disclosure, he placed fourth averaging 65% correct). In particular, Koppel and Schler's methods achieved 53% accuracy on problem A and 100% accuracy on problem F. Kešelj, the runner-up, achieved 85% on problem A and 90% on problem F. Both problems were acknowledged, even by the organizer, to be difficult and considered by many to be unsolvably so. Actually, David Hoover identified a weakness in the problem structure. Since much of the data was taken from the Web, using a web search engine

Table 6.1 AAAC participants and method.

Name	Affiliation	Method
Andrea Baronchelli et al.	Rome "La Sapienza"	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	Piedmont Orientale	Common N-grams (two variants)
Kešelj and Cercone	Dalhousie	CNG method with weighted voting
Kešelj and Cercone	Dalhousie	CNG-wv with reject
O'Brien and Vogel	Trinity College, Dublin	Chi by degrees of freedom
Lawrence M. Rudner	GMAC	Multinomial Bayesian Model/BETSY
Koppel and Schler	Bar-Ilan	SVM with linear kernel function
Efstathios Stamatatos	Patras	Meta-classifiers via feature selection

Table 6.2 AAAC results for problems A–D.

Team[1]	A	B	C	D
Baronchelli	3/13 (23.08%)	3/13 (23.08%)	8/9 (88.89%)	3/4 (75.00%)
Coburn	5/13 (38.46%)	2/13 (15.38%)	8/9 (88.89%)	3/4 (75.00%)
Halteren	9/13 (69.23%)	3/13 (23.08%)	9/9 (100.00%)	3/4 (75.00%)
Hoover1	4/13 (30.77%)	1/13 (7.69%)	8/9 (88.89%)	2/4 (50.00%)
Hoover2	4/13 (30.77%)	2/13 (15.38%)	9/9 (100.00%)	4/4 (100.00%)
Juola	9/13 (69.23%)	7/13 (53.85%)	6/9 (66.67%)	3/4 (75.00%)
Keselj1	11/13 (84.62%)	7/13 (53.85%)	8/9 (88.89%)	3/4 (75.00%)
Keselj2	9/13 (69.23%)	5/13 (38.46%)	7/9 (77.78%)	2/4 (50.00%)
Lana-amisano1	0/13 (0.00%)	0/13 (0.00%)	3/9 (33.33%)	2/4 (50.00%)
Lana-amisano2	0/13 (0.00%)	0/13 (0.00%)	0/9 (0.00%)	2/4 (50.00%)
Obrien	2/13 (15.38%)	3/13 (23.08%)	6/9 (66.67%)	3/5 (75.00%)
Rudner	0/13 (0.00%)	0/13 (0.00%)	6/9 (66.67%)	3/4 (75.00%)
Schler	7/13 (53.85%)	4/13 (30.77%)	9/9 (100.00%)	4/4 (100.00%)
Stamatatos	9/13 (69.23%)	2/13 (15.38%)	8/9 (88.89%)	2/4 (50.00%)

[1] Not all groups submitted solutions to all problems; groups for which no solutions were received scored 0 on that problem.

such as Google could identify many of the documents, and therefore the authors. Using this method, submitted as "hoover2," Hoover was able to "cheat" and outscore Koppel/Schler. Hoover himself admits that this solution does not generalize and does not address the technical questions of stylometry; for the data not available on the Web, this method of course failed spectacularly.

As part of the AAAC procedure, participants were asked to submit brief writeups of their methods to help identify characteristics of good

Table 6.3 AAAC results for problems E–H.

Team	E	F	G	H
Baronchelli	1/4 (25.00%)	9/10 (90.00%)	2/4 (50.00%)	3/3 (100.00%)
Coburn	4/4 (100.00%)	9/10 (90.00%)	1/4 (25.00%)	2/3 (66.67%)
Halteren	3/4 (75.00%)	9/10 (90.00%)	2/4 (50.00%)	2/3 (66.67%)
Hoover1	2/4 (50.00%)	9/10 (90.00%)	2/4 (50.00%)	2/3 (66.67%)
Hoover2	4/4 (100.00%)	10/10 (100.00%)	2/4 (50.00%)	3/3 (100.00%)
Juola	2/4 (50.00%)	9/10 (90.00%)	2/4 (50.00%)	3/3 (100.00%)
Keselj1	2/4 (50.00%)	9/10 (90.00%)	3/4 (75.00%)	1/3 (33.33%)
Keselj2	1/4 (25.00%)	9/10 (90.00%)	2/4 (50.00%)	0/3 (0.00%)
Lana-amisano1	0/4 (0.00%)	0/10 (0.00%)	0/4 (0.00%)	3/3 (100.00%)
Lana-amisano2	0/4 (0.00%)	0/10 (0.00%)	0/4 (0.00%)	0/3 (0.00%)
Obrien	2/4 (50.00%)	7/10 (70.00%)	2/4 (50.00%)	1/3 (33.33%)
Rudner	1/4 (25.00%)	0/10 (0.00%)	3/4 (75.00%)	3/3 (100.00%)
Schler	4/4 (100.00%)	10/10 (100.00%)	2/4 (50.00%)	2/3 (66.67%)
Stamatatos	2/4 (50.00%)	9/10 (90.00%)	2/4 (50.00%)	1/3 (33.33%)

Table 6.4 AAAC results for problems I–M.

Team	I	J	K	L	M
Baronchelli	2/4 (50.00%)	1/2 (50.00%)	2/4 (50.00%)	4/4 (100.00%)	5/24 (20.83%)
Coburn	2/4 (50.00%)	1/2 (50.00%)	2/4 (50.00%)	3/4 (75.00%)	19/24 (79.17%)
Halteren	3/4 (75.00%)	1/2 (50.00%)	2/4 (50.00%)	2/4 (50.00%)	21/24 (87.50%)
Hoover1	3/4 (75.00%)	1/2 (50.00%)	2/4 (50.00%)	4/4 (100.00%)	7/24 (29.17%)
Hoover2	4/4 (100.00%)	2/2 (100.00%)	2/4 (50.00%)	4/4 (100.00%)	7/24 (29.17%)
Juola	2/4 (50.00%)	1/2 (50.00%)	2/4 (50.00%)	4/4 (100.00%)	11/24 (45.83%)
Keselj1	3/4 (75.00%)	1/2 (50.00%)	2/4 (50.00%)	4/4 (100.00%)	17/24 (70.83%)
Keselj2	2/4 (50.00%)	0/2 (0.00%)	1/4 (25.00%)	3/4 (75.00%)	15/24 (62.50%)
Lana-amisano1	0/4 (0.00%)	0/2 (0.00%)	0/4 (0.00%)	1/4 (25.00%)	0/24 (0.00%)
Lana-amisano2	0/4 (0.00%)	0/2 (0.00%)	0/4 (0.00%)	3/4 (75.00%)	0/24 (0.00%)
Obrien	1/4 (25.00%)	1/2 (50.00%)	3/4 (75.00%)	4/4 (100.00%)	5/24 (20.83%)
Rudner	3/4 (75.00%)	1/2 (50.00%)	0/4 (0.00%)	1/4 (25.00%)	0/24 (0.00%)
Schler	3/4 (75.00%)	2/2 (100.00%)	1/4 (25.00%)	4/4 (100.00%)	4/24 (16.67%)
Stamatatos	3/4 (75.00%)	1/2 (50.00%)	2/4 (50.00%)	3/4 (75.00%)	14/24 (58.33%)

performers. Unfortunately, another apparent result is that the high-performing algorithms appear to be mathematically and statistically (although not necessarily linguistically) sophisticated and to demand large numbers of features. The good methods have names that may appear fearsome to the uninitiated: linear discriminant analysis [8, 145], orthographic or lexical cross-entropy [74, 80], common byte N-grams [84], SVM with a linear kernel function [91]. From a practical perspective, this may cause difficulties down the road in explaining to a jury or to a frantic English teacher exactly what kind of analysis is being performed, but we hope the difficulties will be no greater than explaining DNA.

The following are extracts and summaries of the individual writeups submitted by the five top-scoring participants:

- Moshe Koppel and Jonathan Schler: SVM with unstable words
  We used a machine learning approach. For English texts, our primary feature set was a set of common "unstable" words, that is, words with commonly used substitutes (for a full description of this feature set see [88]). When necessary, we used as secondary feature sets: function words, 500 most frequent words (in training), and part-of-speech tags. In languages other than English, we used only the 500 most frequent words.

  The learning method we used is SVM with a linear kernel function. As SVM handles only binary problems, for categories with more than 2 categories, we applied "one vs. others" classifiers, and chose the authors (categories) that pass the threshold. In case of no category passing the threshold, the document was not assigned to any of the given authors in the training set. In case of collision (i.e., a document was assigned to more than one author) we used our secondary feature sets and assigned a document to the author with the most votes. With one exception (see results), this method resolved collisions (overall AAAC performance: 918% correct).
- Vlado Kešelj and Nick Cercone: CNG Method for Authorship Attribution
  The Common N-Grams (CNG) classification method for authorship attribution (AATT) was described elsewhere [85]. The method is based on extracting the most frequent byte n-grams of size n from the training data. The n-grams are sorted by their normalized frequency, and the first L most-frequent n-grams define an author profile. Given a test

document, the test profile is produced in the same way, and then the distances between the test profile and the author profiles are calculated. The test document is classified using k-nearest neighbors method with $k = 1$, i.e., the test document is attributed to the author whose profile is closest to the test profile. Given two profiles f1 and f2, which map n-grams from sets D1 and D2 to their respective frequencies, the distance measure between them is defined by [a distance formula] (overall AAAC performance: 897% correct).

- Hans van Halteren: Linguistic Profiling

  A linguistic profile for each author and text is constructed as follows:

  — For each token, determine token, token pattern, and potential word classes.

  — Take all possible uni-, bi- and trigrams, using either token or class in each position.

  — Use the token pattern instead of the token itself for low-frequency tokens.

  — Use only those n-grams which occur in more than one text.

  — Determine frequencies of all the selected n-grams.

  — Express frequencies as number of standard deviations above/below mean.

  — A text profile is the vector containing these numbers for all selected n-grams.

  — An author profile is the mean over the known texts for the author.

  To determine if a text conforms to an author profile, use [an appropriate] formula (cf. [144]). The score is then normalized to the number of standard deviations above the mean over the negative examples. Generally, a positive score indicates conformance (overall AAAC performance: 861% correct).

- Patrick Juola: Linguistic Cross-entropy
  "Cross-entropy" is a measure of the unpredictability of a given event, given a specific (but not necessarily best) model of events and expectations. This difference can be quantified [71, 74, 151] and measured as a "distance" between two samples.

  To apply this measure, we treat each document (training and test) as source of random "events," in this case letters. From each training document is taken a sample of up to 10,000 letters, and then for each position in the test document, the longest substring *starting at that position and contained in the sample* is found and its length computed.

  The mean of these [substring lengths] has been shown to be closely (and inversely) related to the cross-entropy between the two documents — the longer the length, the lower the cross-entropy.

  To complete the attribution task, the average distance between a single test document and all training documents by the same author is computed. The document is assigned to the author with the smallest average distance, or alternatively, to the authors in order of increasing distance (overall AAAC performance: 851% correct).
- Aaron Coburn: Attributing Authorship using a Contextual Network Graph
  This approach combines word-use statistics and graph theory to identify the author of an unknown text. By comparing stylistic attributes of a document to a corpus of know texts, one can make a reasonable guess about its authorship. In the first step of this method, each text is reduced to word sequences. These are typically two-word sequences, but three- and four-word phrases work well for some very large texts. Next, the indexed collection is projected onto a graph in which the documents and term sequences are represented as an interconnected network.

  For longer texts (typically greater than 3,000 words), the first stage is simple: word pairs are extracted and counted.

If, for instance, the word pair “altogether too” appears more prominently in two particular texts, one can begin to associate these documents. By extracting commonly appearing word sequences, the resulting index tends to form a fingerprint of a document’s style rather than its content. Shorter texts are more difficult; there are typically too few word pairs that appear across the collection to provide any meaningful correlations. For these texts, I apply a part-of-speech tagger, and reduce nouns, adjectives, and verbs each to a common token. Thus the phrase “I walk through the field at noon” becomes “I verb through the noun at noun.” Then, the word pair frequencies are extracted as before.

With an index of word sequence frequencies for each document, these values are applied to the connected graph described above. Document and term nodes are connected by edges, and the value of each edge is determined by the frequency of a word sequence in a document. This contextual network graph produces a network in which similar documents are closely connected and dissimilar texts are less closely connected. Once the graph is constructed, a measure of similarity can easily be determined between the document with unknown authorship and all other documents. Those documents with the highest level of similarity can then likely be identified as having the same author (overall AAAC performance: 804% correct).

6.4 Discussion

More generally, most participants scored significantly above chance on all problems for which they submitted solutions. Perhaps as should be expected, 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 (note the relatively poor performance of Hoover's first method based on word frequency analysis).

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) correlates significantly ($r = 0.594$, $\mathrm{p} < 0.05$) with that method's performance on non-English samples. This may suggest that the authorship problem is linguistically universal and that a single "best" algorithm and feature set can be found, but it may also suggest that some algorithms have simply been tuned to be better on their primary set than others, and a poor algorithm for English is unlikely to magically improve when tested on French or Serbian.

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, Juola has offered 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.

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 (as well as the more serious issue that much of the data was gathered from web-accessible archives such as *Project Gutenberg*) gives an unfair advantage to any methods (such as Hoover's second method) that rely upon searching the Web to extend or leverage their data and analysis. Similarly, the multilingual coverage is unbalanced — on the one hand, any consideration at all of languages other than English might be regarded as a waste of time by some experts who focus only on practical problems in the heartland of the United States, while at the same time, many important and widely studied languages like Arabic and Japanese are not represented. The coverage of different genres is spotty (no newspaper text, for example), and there are probably important issues that have not been addressed at all.

Research into comparative evaluation of attribution technologies continues, including active discussion of the best sort of testbeds to develop. In light of the problems mentioned above, for example, should future test corpora focus on only a single category, for example, analysis of blog or web forum messages in English? This would have some immediate and practical benefits, especially for communities such as intelligence and law enforcement that need to analyze such data on a daily basis. On the other hand, this would leave French medievalists out in the cold. A recent (2006) NSF-sponsored working group identified further evaluation, ideally as an ongoing periodic process, modeled after research groups such as TREC (Text REtrieval Conferences), with each individual evaluation focusing on a specific corpus and problem type. As authorship attribution matures, we can expect this sort of evaluation to be more common and much more detailed.

7

Other Applications of Authorship Attribution

Thus far, we have been focusing on authorship attribution primarily as a question of inferring the personal identity of the document's author. However, the definition as given in Section 2.1 is broader — "any attempt to infer the characteristics of the creator of a piece of linguistic data." As we have seen in Section 4.6, this can be broadened even further to cover other forms of data such as Web pages, music, and computer code. However, this definition also includes not just personal identity, but group identity such as gender, dialect/culture, education level, or even personality. As discussed below, researchers have found that many of the same techniques work for such evaluations as well.

Although this may seem plausible, even obvious, at least one related discipline specifically abjures such application. "Document analysis" is the term used by forensic specialists in inferring authorship from handwriting (such as ransom notes). These examiners, often certified by professional organizations such as the American Board of Forensic Document Examiners (`www.abfde.org`), try to establish "the authenticity of the contested material as well as the detection of alterations" [3]. "In their words, forensic document examination involves the analysis and comparison of questioned documents with known material in order to identify, whenever possible, the author or origin of the questioned document." They specifically do not infer the traits of the authorship, and cannot usually be used to profile the author ("the author of this document was a college-educated white woman from California"); the discipline that covers this is usually called "graphoanalysis," which ostensibly claims "to provide insight into personality traits and evaluation of a writer's personality" [67] but is usually considered a pseudoscience like astrology. The difference, simply put, is accuracy. Traditional handwriting analysis, as practiced by forensic document examiners, is not able to infer group characteristics, even simple ones like gender and handedness, to the accuracy demanded by their primary clients, the court system. It is somewhat unfortunate that there is little public documentation about the accuracy rate they are able to achieve, as it would be surprising if they could not infer gender with rates substantially better than chance, which would still be a publishable result in the authorship attribution community. Authorship attribution, having come late to the forensic community, has been happy to study group as well as individual characteristics.

7.1 Gender

An obvious place to start looking for group characteristics is in the gender of the author. In some languages, of course, gender is so marked that it is hard to avoid in first-person writing (consider guessing what gender of Francophone writes *Je suis belle*). Even in languages such as English, gender distinctions in spoken language have been studied for a long time, long enough to preclude a useful set of citations. These distinctions tend to be more obvious in spoken than in written language; in fact, some authors have suggested that male and female styles should not differ in formal writing.

Koppel [89] has studied this and found just the opposite, that there is a detectable difference, enough to enable his team to categorize documents with about 80% accuracy. His method is familiar enough in light of the previous discussion; documents from the British National Corpus (BNC) were embedded in a high-dimensional feature space, using features such as function words, POS n-grams, and punctuation. Classification was done by inferring a linear separator between male- and female-authored documents using a standard machine learning method (similar in spirit to LDA or SVM). They further found that non-fiction and fiction had very separate categorization rules, and that non-fiction was generally easier to categorize than fiction. Analysis of the specific features that were used to create the category both supported and extended prior work — for example, men used pronouns significantly less frequently than women in both fiction and non-fiction, while men used the determiners "a," "an," "the," and "no" significantly more frequently.

Other studies of this nature include [5, 29, 95] with largely similar results, although on different languages, genres, and methods. For example, [95] studied Internet chat logs (e.g., IRC, IM, ICQ, and such) in Turkish using a feature set most notable for its inclusion of smileys and a variety of classifiers including neural networks, k-nearest neighbor, and naive Bayesian analysis. The results from Baysian analysis were again in the 80% range, with other methods slightly or significantly less. An interesting variation on this type of study is that of Hota et al. [66] who studied how authors portray different genders. In particular, do the characters in a play who are supposed to be female actually speak like females? Using a corpus of character speeches from 34 Shakespearean plays, the researchers were able to classify approximately 60%–75% of the speeches correctly by the gender of the speaker. The lower accuracy may suggest that Shakespeare had some intuitive understanding of some language differences, but was unable to completely mask his own masculinity.

7.2 Document Dating and Language Change

Another aspect of authorship category is the time of authorship, both as expressed on the calendar and in terms of one's personal development as a writer. This is a surprisingly controversial area of authorship attribution research, because it reflects directly on the underlying assumption of an authorial "fingerprint." Some researchers, most notably van Halteren [145], have claimed that people should have a fixed and unchanging set of stylistic markers, but at the same time, the fact that new words routinely enter and leave the language means that there must also be continuous change. Perhaps these are other stylistic markers that can be used for other purposes?

Hota's study of Shakespearean gender is also an example of a study in document dating; Hota "somewhat arbitrarily" divided his corpus into Early and Late plays (pre- and post-1600) and analyzed the two groups separately. There was a substantial difference in accuracy between the two groups, indicating "that there is a greater stylistic difference between the genders in late Shakespeare than early Shakespeare." This is, of course, evidence of stylistic development over Shakespeare's career. Of course, establishing this development does not establish how and why this development happened; the cause may be as simple (and uninteresting) as increasing the number of female characters, or as significant as the development of a female-specific vocabulary.

Similar studies have found such development in other writers' as well; Examples include [23, 62, 79, 119]. Figure 7.1 shows an example of such development in the novels of Henry James. Unlike [66], this

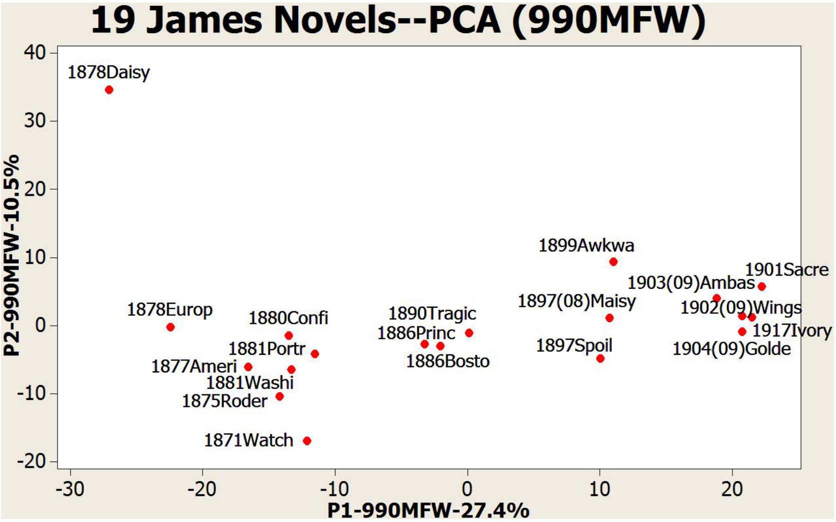

Fig. 7.1 Chronological development of style in the writings of Henry James (data and figure courtesy of David Hoover).

figure was produced by unsupervised analysis (PCA of the 990 most common words) and does not involve arbitrary categorization of texts; instead they are simply scatter-plotted by date across the first two principal components. The pattern that emerges is a clear picture of stylistic development; novels known to be earlier sort themselves to the left of the diagram. A similar picture of stylistic development can be seen in Figure 5.3, covering the work of Jack London [79]. Indeed, this may be even more striking, given the apparent lack of pattern prior to 1912 and after 1912, but the clear linear separation between early and late works; indeed, this picture argues for a non-arbitrary classification as of 1912 for the works of Jack London.

There is thus substantial work looking at the development of stylistic change within the career of a single writer. Can such stylistic change be observed for society as a whole? Obviously, lexical items will change relatively rapidly [33, 70], and other levels will also change, albeit more slowly. Juola [74] applied stylometric techniques — the linguistic cross-entropy discussed in a previous section — to infer the average distance between magazine articles covering approximately a forty-year period. He was able to infer not only that documents more distant in time are also more distant in style, but to approximate the rate at which language changed. His findings were not only that language changed, but the rate itself was non-uniform; as measured by samples from the *National Geographic*, language changed more rapidly in the 1950s than in the 1940s or 1960s (and the rate of language change in the 1940s, although measured to be positive, was not significantly different from zero).

Both of these types of analysis could in theory be applied to the task of document dating; a previously undiscovered manuscript by Henry James, for example, could be placed on the graph of Figure 7.1 to guess (perhaps via k-nearest neighbor) the approximate time period in which it was written.

7.3 Other Socioeconomic Variables

Using similar techniques, it is possible to analyze for any group identity for which data is available. Nationality, dialect, and region are almost too well-studied to deserve detailed comment. Yu [152] applied rule-based feature analysis on word *n*-grams to distinguish between native English (USA/UK) and native Chinese speakers in a study of doctoral dissertations; the methodology is an extension of that of Oakes [115] for distinguishing USA and UK English. The Baayen corpus is a good example for the identification of more subtle differences; both van Halteren [8, 145] and Juola [80] have confirmed that the educational differences between first year and fourth year college students can be detected using LDA and cross-entropy, respectively. A study by Keune [86] showed that differences in education levels could be detected in the pronunciation of the Dutch word *mogelijk* ("possible"). Further analysis showed other effects of both sex and nationality (Dutch vs. Flemish).

7.4 Personality and Mental Health

As one of the major products of the human mind, language should be expected to provide some sort of indicator of the mind that produced it. In addition to indicating identity, it may also be able to show something about process. The use of linguistic indicators as cues for medical diagnosis has a long history. A simple example is the sentential complexity measures as revised by [17], where the absence of highly complex sentences can be a sign of cognitive trouble. Similarly, other studies have shown that particular linguistic features can be a sign of specific emotional attitudes — when people are depressed, for example, they use more first-person singular pronouns [123]. Perhaps more hopefully, an increased use of causal words such as "because," as well as an increased use of cognitive words such as "realize" predicts recovery from trauma [120]. These are simple examples of lexical features that can be used to categorize people by their mental states. Other examples of such studies — and examples of intense interest to law enforcement, one expects — include the studies of lying and linguistic style performed by Newman et al. [113] as well as by Hancock [51].

Other researchers [6, 28, 114, 116, 118] have studied to what extent linguistic cues can identify personality type and profile.

Argamon [6] is another almost prototypical example of standard authorship attribution technology applied to this task. As before, the documents (student-written stream-of-consciousness essays) were embedded in a large vector space and supervised learning in the form of a linear separator was applied to determine the feature combination that best separated the categories of interest. In this case, the individual students had been tested on neuroticism ("roughly: tendency to worry") and extroversion ("roughly: preference for the company of others"). The feature set included the by-now familiar set of common words (675 in this set), but also indicators of "cohesion," "assessment," and "appraisal" measured though a Systemic Functional Grammar (SFG) approach. The analysis was able to identify neuroticism with substantially better than chance odds (58%), but did not perform nearly as well in detecting extroversion. This may reflect the inherent difficulty of the task, or it may simply reflect the fact that a more sophisticated feature set is necessary for analysis of psychological structures.

A similar experiment was conducted by Nowson and Oberlander [114]. These researchers were able to identify an "internet meme" equivalent to a personality test yielding relatively coarse scores on five personality factors (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). Applying this test to web log (blog) data, they used word-based N-grams and applied a naive Bayes classifier, typically getting above chance in classifying on each of these scales, but not by much (e.g., between 52% and 65% on two-way classification tests). Although these results are numerically disappointing, this study is remarkable for the (poor) quality of the data from which they were able to work, suggesting that good methods may have an almost phenomenal degree of resistance to noise.

7.5 Section Summary

Beyond simple questions of the identity of individual authors, the underlying technology has been applied to questions of group identity as well. Examples include gender, age, date, social class, education, nationality, and even personality type. Of course, the degree of success in these applications varies substantially.

It is interesting to compare the limited scope of document analysis with the far-ranging scope of authorship attribution technology. Document analysts deliberately avoid inferring personality, because they consider that they cannot do it with sufficient accuracy. That concern does not appear to have stopped Argamon (although Argamon presumably has not been called upon to testify in court about his personality profiles). Is 58% a useful level of accuracy for any purpose at all?

8

Special Problems of Linguistic Forensics

8.1 Forensic Issues

For many purposes, simply producing an answer is enough. A simple authoritative statement to the effect that the *Hamlet* looks like Shakespeare's work is often enough, especially when the answer is simply a confirmation of existing scholarly belief. On the other hand, in cases of genuine controversy, it is not enough to simply issue pronouncements. Before accepting the conclusions, people will usually want to see and to evaluate the evidence supporting them.

Few areas are as controversial as the law. Few subjects are as controlled by restrictive rules as the law. In few situations can one not only expect, but rely upon the presence of an intelligent, articulate person specifically dedicated to discrediting one's arguments, methodology, and evidence. It can be said with some degree of accuracy that forensic authorship attribution is among the most challenging. At the same time, questions of language and authorship can be crucial in establishing justice.

Chaski [26] presents three examples drawn from actual cases where the question of a document's authorship was key to the resolution.

In the first, an employee was fired for racist email sent from his (open) cubicle, email he denied sending. In the second, a young man was found dead by his roommate, with suicide notes typed on the man's computer. Were these notes forgeries, perhaps created to cover the roommate's murder of the man? In the third, one party claimed that their on-line journal, a crucial bit of evidence, had been tampered with by the other party before it was able to be introduced into evidence.

In each of these cases, traditional digital forensics is more or less helpless; all parties in the dispute had access to the relevant computer and could have created or modified the documents; tracing the documents back to the machine of origin was both trivial and useless. How, then, does one establish evidence sufficient to convince the investigators, judge, and eventual jury of the guilt or innocence of those involved?

There are three main problems with authorship attribution as applied to forensics — credibility, admissibility, and active malice. We will look briefly at each of these in turn.

8.2 Credibility

The first problem facing a forensic scholar, oddly enough, is not the problem of rules of evidence. Before there is a trial, there is an investigation, and even before that, there is an allegation. The first task is to make the initial finding credible enough that the investigators will take it seriously (after all, even formally inadmissible evidence can form the basis of an investigation; the Unabomber, for example, was finally caught after his brother informally noted that the Manifesto was written in his style and suggested they investigate).

In order to be credible, a statement must first be accurate, or perhaps more accurately, backed up by a track record of accuracy. This is the problem with Argamon's personality assessment with accuracy of 58%; unless an investigator were truly desperate, an assessment with barely greater accuracy than a coin flip is not a useful lead. On the other hand, the 90%+ accuracy that most AAAC participants managed to achieve on the Paston letters are probably enough to cause investigators to start focusing on Margaret instead of John. Even a suggestion that there is an 80% likelihood that the culprit was female [89, 95] is probably a useful finding.

The relevance and representativeness of the track record are also important. If the case under discussion involves Email and blog entries, a track record of accuracy derived from the study of large-scale literary documents is less convincing than one derived from shorter documents such as letters, and a study based on Email and blogs would be better yet. In the absence of such a direct study, findings that a particular method were accurate across a wide variety of genres would support the credibility of that method applied to a new one.

Another important consideration, however, cannot really be quantified. To be credible, a method should also be understandable. The idea of a shibboleth or similar direct marker of authorship is one that can be easily understood; and may not [149] even require "expert" testimony to explain. The assumption that one can infer group identity from accent and vocabulary is easily made and easily expressed. It is slightly more difficult to accept the idea of systematic variation — although everyone uses adverbs, some people use more of them than others, and do so consistently. The key difficulty here is probably consistency, and would probably need empirical support. But at least the basic idea, the varying frequency of use of a particular features, is something understandable. Quantifying and combining such variables, as with Delta and its variants, is a simple next step.

However, as shown in the previous sections, simple feature quantification tends not to be as accurate or as sensitive as more complex statistics. Unfortunately, these more complex statistics almost instantly lose the direct connection with understandable features. If two documents are separated along "the first principal component," what does that mean? What is the association between the features encoded in the study and the PCA vectors? It is possible (see [14] for an example) to plot the features themselves in terms of their representation in the principal components, to show the degree to which certain features appear in the same PCA "space" as the documents that are attributed to them, and to show (visually) that the different authors appear to cluster in this space. This permits researchers to make relatively accessible statements like "Thompson [has a] tendency to use words indicating position — 'up,' 'down,' 'on,' 'over,' 'out,' and 'back' — more frequently than Baum. An examination of the raw data reveals that, for these words, Thompson's average rates of usage are about twice as [sic] Baum's" [14] Less convincing are the diagrams (such as those in [8, 79]) that display clustering without explaining the underlying features that produce that clustering.

But even these methods are substantially more credible than those that simply report percentage accuracy, whether in the form of ROC curves, graphs, or simple statements. Unfortunately, most of the methods with the highest accuracy operate in too many dimensions to make simple representations practical; most people cannot visualize a four dimensional hypercube, let alone one with two hundred dimensions and a separating non-planar hypersurface (as would be produced by SVM with a nonlinear kernel function). Direct distance measures such as Juola's cross-entropy appear to be among the worst offenders in this regard, as the "features" are implicit and cannot even be extracted easily. Other offenders include general machine learning methods such as neural networks, from which the overall method cannot easily be extracted. There is room (and need) for much additional research in trying to find a method that is both transparent and accurate, or alternatively, in simply trying to find a way to explain methods like PCA, SVM, and LDA to a non-technical audience.

8.3 Admissibility

As discussed above, admissibility is less of an issue than one might think, because, if necessary, other evidence is likely to be found. It may also be possible to explain the methods for authorship attribution without recourse to direct expert testimony, if the methods themselves are simple enough and credible enough (as with Wellman [149]). Even with only partial admissibility, it may still be possible for an expert to take the stand, outline a set of features (such as Thompson's use of "words indicating position") and demonstrate informally that the defendant's use is more in-line with the document under discussion than other candidates, while leaving the jury to draw the necessary conclusion that the document was written by the defendant.

Nevertheless, it would be better all around if the standards for authorship attribution met the standards for admissible evidence. These latter standards, of course, are always shifting and vary not only from place to place, but also from trial to trial. We focus briefly on the evidence standards under US law.

8.3.1 US Legal Background

Expert evidence in the United States is dominated by two major cases, *Frye vs. United States* (1923), and *Daubert vs. Merrill Dow* (1993), each of which establishes separate tests for whether or not scientific evidence is admissible. Historically speaking, the *Frye* test established rules for Federal procedure that were later copied by most US states. In this case, the court recognized that, while science is a progressive enterprise, the courts have an active duty to sort out pseudoscience and avoid undue reliance on controversial or unsupported theories. The court thus held that

> Just when a scientific principle or discovery crosses the line between experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

This “general acceptance” test defines acceptable science as the uncontroversial core. In the particular case under evaluation, the proposed science (later described in *Daubert* as “a systolic blood pressure deception test, a crude precursor to the polygraph machine”) was not sufficiently accepted and therefore not admissible.

Later rewritings of the Federal Rules of Evidence (most notably FRE 702) laid down a more sophisticated epistemological framework. Rule 702 reads:

> If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

(Note that this rule does not even mention "general acceptance"!)

The Supreme Court amplified and interpreted this rule in *Daubert*, creating a new standard. In broad terms, the Court recognized that the *Frye* test was unduly restrictive and eliminated much cutting-edge but legitimate science. The role of the court was substantially reduced to evaluating whether or not a proposed bit of testimony rested on "sufficient facts or data" and was the product of "reliable principle and methods." Novel or controversial, but still scientific, findings were to be admitted, although their weight would still be for the jury to decide.

The *Daubert* decision avoided drawing a clear line or enumerating a checklist of the features that a method must have in order to be reliable and scientific, but it did note a number of factors that are ordinarily likely to be relevant. These factors include

- Whether or not the method of analysis has been tested. The Court specifically drew attention to the Popperian theory of falsifiability; a "scientific" method is one that can be subjected to empirical tests to see whether it is true or false.
- Whether or not the method of analysis has been subject to peer review and publication. Science operates largely on the basis of shared knowledge and consensus, established through

the publication process. This review process also acts as a filter, "in part because it increases the likelihood that substantive flaws in methodology will be detected." (*Daubert*)

- The known or potential rate of error — and in particular, whether the technique has a known, potential rate of error. Obviously, knowing the accuracy rate of a technique is key to knowing how much weight to give the results of its analyses.
- The existence of standards and practices governing the use of the technique, and by extension the existence of standards and practice bodies.
- Finally, "general acceptance" is still a consideration, as "it does permit [...] explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community." In particular, "a known technique that has been able to attract only minimal support within the community [...] may properly be viewed with skepticism." (*Daubert*)

As a Federal court decision (and one interpreting the Federal Rules of Evidence), *Daubert* is technically only binding at the Federal level, but many states have adjusted their rules to incorporate *Daubert*. Others have not, preferring to stay with *Frye* or extensions to *Frye*.

8.3.2 Other Jurisdictions

Other countries, of course, have their own rules and procedures with regard to expert scientific evidence. In much of the British Commonwealth, the role of the expert is somewhat different. In England, for example, expert witnesses (under the Civil Procedure Rules, Rule 35) under an overriding duty, not to the person that hired them, but to the Court. As a result, they are officially independent, overriding "any obligation to the person from whom the expert has received instructions or by whom he is paid: rule 35.3(2)."

Partly for this reason, it is (at least in theory) more difficult under English law than under US law to get an expert report written to order to support just one side of the case. English procedure focuses more on the qualifications of experts; to be admissible, a purported expert must satisfy the Court that there is "recognised expertise governed by recognised standards and rules of conduct capable of influencing the court's decision," and further, that the expert himself "has a sufficient familiarity with and knowledge of the expertise in question to render his opinion potentially of value" (*Barings Plc v Coopers & Lybrand* (*No. 2*) [2001]). This, of course, is similar in spirit to the *Frye* test, in that the key determination is whether or not an established body of practice (general acceptance) exists.

The law is similar in other common law jurisdictions [46]; for example, the wording of the Australian *Civil Procedure Act* 2005 (NSW) lays down the same overriding duty of independence. Other legal systems are beyond the scope of this report; consult your local experts.

8.3.3 Admissibility of Authorship Attribution

It should be apparent that, regardless of the actual jurisdiction, the spirit of the *Daubert* criteria are a standard that any practicing scholar should be happy to try to meet. Who would not want their methods and practices to be "reliable" and their judgments to be "based on sufficient facts and data?" It should also be apparent that the other standards, whether "general acceptance" or "recognized expertise governed by recognized standards" can only come from a body of practices that meet the *Daubert* criteria for reliability and sufficiency. The question, of course, is whether the current state of authorship attribution meets them or not.

This has been a matter of some debate; as recently as 200l, Tiersma and Solan (cited in [102]) cited disputed authorship cases as a "problem area" in forensic linguistics, while McMenamin disagreed. From a principled standpoint, there is ample reason to be concerned about many individual techniques and methods. One key problem with function word PCA, for example, is that the technique does not easily admit to the calculation of real or expected error rates. Other methods, most notably the easily explained primitive statistics, are too inaccurate (and too easily proved inaccurate) to be taken seriously by jurors, while other methods have simply not been proved against a large enough set of trials.

On the other hand, papers such as the current one should help establish both the existence of a large research community that "generally accepts" the idea of stylometric analysis, and more importantly, establishes generally accepted standards for accuracy across a number of different areas and corpora. The methods specifically recommended by such peer-reviewed papers (see the following section) can be argued to have established themselves as practices.

At any rate, the ultimate determiner of whether or not a particular method is admissible is the specific court before whom one wishes to testify. As Chaski [27] has pointed out, in many specific courts, this test has already been passed.

8.4 Malice and Deception

One final consideration in the evaluation of forensic authorship attribution is the question of active malicious alteration of writing style. In short, can I write using someone else's style in order to deceive the analyst (and by extension, the court system)? This problem is not entirely confined to the forensic realm (there are many other reasons that one could want to disguise one's true authorship), but it is a specific concern for forensics and takes on a heightened importance.

To some extent, the question of fooling the analysis obviously depends on how good the analysis is. In particular, it depends both on how robust the feature set (the "stylome") is to deliberate manipulation as well as how sensitive the analysis mechanism is. It is relatively eazy for a good speler to pritend to be a pooor one (and therefore perhaps to fool an anomaly-based scheme). It is harder for a poor speller to pretend to be a good one (without the aid of a co-author or a spelling checker). It is simple to fake the use of dialect words that are not one's own, but harder to use them in their proper context (e.g., *"voltmetre" only looks like a British word; it is not. A "metre" is a unit of length, not a measuring instrument). Harder still (one hopes) is to model one's use of abstract concepts like function words in order to match another person's patterns.

An early study [136] of this found that different methods (as might be expected) differed in their abilities to detect "pastiche." Using

Adair's *Alice Through the Needle's Eye*, a parody of Lewis Carroll, Somers and Tweedie found that measures of lexical richness distinguish the authors, but that word frequency PCA did not. A later study [82] focused on lexical variation in the *Federalist* papers. Using decision trees and SVMs, the researchers found the word tokens that were most informative of authorship and neutralized them. With an average of about 14 changes per thousand words of text, the researcher were able to essentially neutralize the SVM's ability to correctly identify the author (reducing the likelihood of a successful attribution by more than 80%). On the other hand, more sophisticated techniques were still able to identify the original author after changes.

From a research standpoint, this suggests the strong possibility of some kind of predator–prey or arms race situation; the best techniques for obfuscation will depend on what sort of analysis is expected. There is obviously great potential for further work here.

9

Recommendations

9.1 Discussion

As the previous sections should make clear, stylometry and authorship attribution are an active research area encompassing a wide variety of approaches, techniques, and sub-problems. No paper of reasonable length can attempt to do justice to the entire field.

From this confusing mess, however, a few overall themes appear to emerge. The first, quite simply, is that it works — "non-traditional" authorship attribution, attribution based on the mathematical and statistical analysis of text, can identify the author of a document with probability substantially better than chance. The second is that almost anything can be a cue to authorship under the right circumstances, and the primary task for researchers is not to find potential cues, but to evaluate and combine them in a way that produces the greatest overall accuracy. The third is that the way the problem is defined matters greatly, both since "authorship attribution" can include other aspects such as group identification, gender attribution, and so forth, and because the intended use of the results can have a strong influence on the best method to use. The needs of the forensic community, for example, are different from those of traditional English scholarship, which in turn is different from the needs of a high school teacher looking to prevent cheating.

In discussing authorship attribution, we have made extensive use of Juola's three-phase framework [76]. Separating the purely mechanical aspects of canonicization (stripping out markup, spelling regularization, case normalization, and so forth) from the process of identifying the event set or feature space, and separating that, in turn, from the analysis method make it easier to present and understand the possibly bewildering array of methods that have been proposed. Most researchers have approached this problem by embedding documents in a high-dimensional space of abstract features and then applying some sort of statistical analysis to that space. More sophisticated — but also more computationally intensive, and more difficult to understand — approaches apply machine learning techniques to the analysis, treating it as a form of document categorization problem. Different types of feature space used include vocabulary, syntax, orthography, structure/layout, or miscellaneous anomalies, or a combination of the above — the number of features used in any particular study can vary from merely one or two up to hundreds or thousands. The number of analysis methods applied is almost equally huge, ranging from simple summary statistics such as averages, through unsupervised analyses like principal component analysis or hierarchical cluster analysis, to state-of-the-art machine learning and data mining techniques such as linear discriminant analysis and support vector machines. Much additional research can be done simply in identifying new features and new methods to see whether or not they work.

However, another advantage of this three-phase framework is that it can be (and has been [81]) used to direct software development for the systematic exploration of this problem. A simple combination of twenty-five different feature sets with fifteen different analytic variations yields almost four hundred different and potentially high-performing systems to explore. Juola and his students have begun this process with the development of JGAAP, a modular framework for authorship attribution using the object-oriented capacities of the Java programming language. This program defines separate "Event" and

"Processor" classes that can be implemented in a variety of ways; for example, an Event could be defined as a letter, a word, a POS tag, a punctuation mark, or a word selected from a specific pre-defined set (as of function words or synonym sets); each document, in turn, can be represented as a vector of Events upon which the Processor operates. A Processor implementing Burrows' Delta could calculate z-distributions of any sort of Event and categorize documents appropriately; variations on Delta [63, 138] would operate similarly, but differ in the details.

A major theme implicit in the previous paragraph is the existence of a dataset for testing and cross-comparison. Many researchers [25, 40, 63, 75, 80] have been quite explicit in their call for such a framework. As a simple example, re-analysis of the AAAC data (Juola, forthcoming) has shown that a mixture of experts approach can outperform any individual AAAC participant. Absent this sort of standardized test data, this result would not have been apparent. No corpus has yet emerged to be the standard touchstone and testbed, although there are a few promising candidates such as Juola's AAAC corpus or Chaski's Writing Sample Database. Depending upon the community's needs, further development of test corpora is almost inevitable, but the exact form that it will take depends on what type of problem is held to be more important. Because of the amount of research opportunity available in this field, the next several years promise to be quite interesting and productive.

9.2 Recommendations

What, though, of the plight of a researcher who cannot afford to wait five or ten years for an answer? The AAAC and studies like it can at least provide the basis for a preliminary set of recommendations. For those interested in forensic applications, where documented "best practices" are needed before the judge will even look at your report — I can recommend the following:

- Before performing a non-standard authorship analysis, conduct a standard one, consulting with experts as necessary. Rudman, in particular, has written often and at length about the need for such an examination. The results of the analysis can only be as accurate as the data, including the quality of the document to be examined and of the training data. All data must be carefully vetted to get, as close as possible, back to the original manuscript. The set of candidate authors must be chosen as carefully and rationally as possible, and the set of candidate writings must also be chosen with equal care and rationality. For example, if the document of interest is a novel written in third person, the distribution of pronouns will [64] be radically different than that of a novel written in first person, not by virtue of an authorship difference, but simply from genre.

As Hoover showed in one of his submissions to the AAAC, in some cases, perhaps many, an analysis of this sort may render the stylometry itself irrelevant. If a search for candidate writings turns up the original document of interest as an entry in a Google search, then the attribution problem is trivial. If a close reading of the text reveals the name of the author in the metadata, there is no need for statistics.

- Methods using a large number of features seem to outperform methods using a small number of features, provided that there is some method of weighting or sorting through the feature set. In particular, simple statistics of a few simple features, such as average word or sentence length, does not usually produce accurate enough results to be used. There appears to be general agreement that both function words (or sometimes merely frequent words) and POS tags are good features to include. Methods that do not use syntax in one form or another, either through the use of word n-grams or explicit syntactic coding tend to perform poorly. Other feature categories are more controversial.
- The best choice of feature set is of course strongly dependent upon the data to be analyzed — for example, the set of function words is controlled by the language of the documents. However, the choice of analysis method appears to be largely language independent. No method has yet emerged from any

study as being particularly good within a narrow range of language, genre, size, and so forth. This lends credence to the theory that there is a best possible analysis method, and we simply need to find it.

- Simple unsupervised analysis — most notably, principal component analysis — will sometimes produce useful and easy-to-understand results. On the other hand, PCA and similar algorithms are often unable to uncover authorship structure that more powerful algorithms find.
- Understandability remains a major problem. One reason that PCA and similar algorithms remain popular, despite their modest effectiveness, is that the reasons for their decisions can easily be articulated. The same vector space that categorizes text can be used to categorize individual words (or features); one can literally superimpose on the graphic separating A from B the words used, and the words near A are the ones that A uses and B does not. For researchers more interested in the "why" than the "what," this ease of explanation is a key feature of PCA. Otherwise, you run the risk of being able to tell people apart without actually knowing anything about their styles [30].
- The real heavyweights emerging from the AAAC are the same high-performing analysis methods that have been useful elsewhere. These high-flyers include support vector machines (SVM), linear discriminant analysis (LDA), and k-nearest neighbor in a suitably chosen space (either by cross-entropy or n-gram distance). These methods were the top four performers in the AAAC and have usually been reported as the top-scoring method in other analyses. SVM, in particular, appears to be the leading contender for "best performing analysis method for any given feature set."
- Document all steps in the process. In a forensic context, this is a necessity. In a less formal context, this is still important, as the minor details of document processing will often be key to the accuracy of your results.

9.3 Conclusions and Future Work

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

This review has discussed only a few of the approaches that have been proposed over a longer-than-a-century history of statistical stylometry. Given the current research interest, it will no doubt be out of date by the time it sees print. New researchers will produce new corpora and new ways of testing to better evaluate the current state of the art. Others will invent new methods, taking advantage of advances in other disciplines such as machine learning and classification. SVM, the current front-runner, was not even dreamed of twenty years ago; twenty years from now, the front-runner will probably be a technique currently unknown. Similarly, the variety of problems — not just identification of "authorship," but of authorial gender, age, social class, education, personality, and mental state — will almost certainly increase as people find it useful to more accurately profile their correspondents.

The most serious problem facing the current discipline is disorganization. There are too many mediocre solutions, and not enough that are both good and principled. Especially if authorship is to be useful as a forensic discipline, there are a number of crucial issues that need to be addressed Fortunately, the seeds of most of the issues and developments have already been planted.

Better test data, for example, is something of a *sine qua non.* Some test corpora have already been developed, and others are on the way. A key aspect to be addressed is the development of specific corpora representing the specific needs of specific communities. For example, researchers such as NYU's David Hoover have been collecting large sets of literary text such as novels, to better aid in the literary analysis of major authors. Such corpora can easily be deployed to answer questions of literary style, such as whether or not a given (anonymous) political pamphlet was actually written by an author of recognized merit, and as such reflects his/her political and social views, to the enrichment of scholars. Such a corpus, however, would not be of much use to law enforcement; not only is 18th or 19th century text unrepresentative of the 21st century, but the idea of a 100,000 word ransom note being analyzed with an eye toward criminal prosecution borders on the ludicrous. The needs of law enforcement are much better served by the development of corpora of web log (blog) entries, email, and so forth — document styles that are used routinely in investigations. So while we can expect to see much greater development of corpora to serve community needs, we can also expect a certain degree of fragmentation as different subcommunities express (and fund) different needs.

We can expect to see the current *ad hoc* mess of methods and algorithms to be straightened out, as testing on the newly developed corpora becomes more commonplace. Programs such as JGAAP [81] will help support the idea of standardized testing of new algorithms on standardized problems, and the software programs themselves can and will be made available in standard (tested) configurations for use by non-experts. Just as tools like Flash make multimedia publishing practical, so will the next generation of authorship attribution tools make stylometry generally practical.

The new level of computer support will trigger new levels of understanding of the algorithms. Although some efforts (most notably Argamon [138]) have been made to explain not only that certain methods work, but to why they work, most research to this date has been content with finding accurate methods rather than explaining them. The need to explain one's conclusions, whether to a judge, jury, and opposing counsel, or simply to a non-technical PhD advisor, will not doubt spur research into the fundamental linguistic, psychological, and cognitive underpinnings, possibly shedding more light on the purely mental aspects of authorship.

Finally, as scholarship in these areas improves and provides new resources, the uptake and acceptance of non-traditional authorship attribution can be expected to improve. The communities that drive authorship attribution do so because they need the results. The better the results, the more badly they are needed.

Why care about authorship attribution? Ultimately, because other people do.

References

[1] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 6, pp. 67–75, 2005.

[2] A. Abbasi and H. Chen, *Visualizing Authorship for Identification,* pp. 60–71. Springer, 2006.

[3] American Board of Forensic Document Examiners, "Frequently asked questions," http://www.abfde.org/FAQs.html, accessed January 6, 2007.

[4] Anonymous, "Some anachronisms in the January 4, 1822 Beale letter," http://www.myoutbox.net/bch2.htm, accessed May 31, 2007, 1984.

[5] S. Argamon and S. Levitan, "Measuring the usefulness of function words for authorship attribution," in *Proceedings of ACH/ALLC 2005*, Association for Computing and the Humanities, Victoria, BC, 2005.

[6] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Proceedings of the Classification Society of North America Annual Meeting*, 2005.

[7] A. Argamon-Engleson, M. Koppel, and G. Avneri, "Style-based text categorization: What newspaper am I reading," in *Proceedings of the AAAI Workshop of Learning for Text Categorization*, pp. 1–4, 1998.

[8] R. H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie, "An experiment in authorship attribution," in *Proceedings of JADT 2002*, pp. 29–37, Université de Rennes, St. Malo, 2002.

[9] R. H. Baayen, H. Van Halteren, and F. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, pp. 121–131, 1996.

[10] R. E. Bee, "Some methods in the study of the Masoretic text of the Old Testament," *Journal of the Royal Statistical Society*, vol. 134, no. 4, pp. 611–622, 1971.

[11] R. E. Bee, "A statistical study of the Pinai Pericope," *Journal of the Royal Statistical Society*, vol. 135, no. 3, pp. 391–402, 1972.

[12] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, no. 4, p. 048072, 2002.

[13] D. Biber, S. Conrad, and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

[14] J. N. G. Binongo, "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9–17, 2003.

[15] A. F. Bissell, "Weighted cumulative sums for text analysis using word counts," *Journal of the Royal Statistical Society A*, vol. 158, pp. 525–545, 1995.

[16] E. Brill, "A corpus-based approach to language learning," PhD thesis, University of Pennsylvania, 1993.

[17] C. Brown, M. A. Covington, J. Semple, and J. Brown, "Reduced idea density in speech as an indicator of schizophrenia and ketamine intoxication," in *International Congress on Schizophrenia Research*, Savannah, GA, 2005.

[18] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, pp. 79–85, June 1990.

[19] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 955–974, 1998.

[20] J. F. Burrows, "'An ocean where each kind…': Statistical analysis and some major determinants of literary style," *Computers and the Humanities*, vol. 23, no. 4–5, pp. 309–21, 1989.

[21] J. F. Burrows, "Delta: A measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, pp. 267–287, 2002.

[22] J. F. Burrows, "Questions of authorship: Attribution and beyond," *Computers and the Humanities*, vol. 37, no. 1, pp. 5–32, 2003.

[23] F. Can and J. M. Patton, "Change of writing style with time," *Computers and the Humanities*, vol. 28, no. 4, pp. 61–82, 2004.

[24] D. Canter, "An evaluation of 'Cusum' stylistic analysis of confessions," *Expert Evidence*, vol. 1, no. 2, pp. 93–99, 1992.

[25] C. E. Chaski, "Empirical evaluations of language-based author identification," *Forensic Linguistics*, vol. 8, no. 1, pp. 1–65, 2001.

[26] C. E. Chaski, "Who's at the keyboard: Authorship attribution in digital evidence invesigations," *International Journal of Digital Evidence*, vol. 4, no. 1, p. n/a, Electronic-only journal: http://www.ijde.org, accessed May 31, 2007, 2005.

[27] C. E. Chaski, "The keyboard dilemma and forensic authorship attribution," *Advances in Digital Forensics III*, 2007.

[28] D. Coniam, "Concordancing oneself: Constructing individual textual profiles," *International Journal of Corpus Linguistics*, vol. 9, no. 2, pp. 271–298, 2004.

[29] M. Corney, O. de Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *Proceedings of Computer Security Applications Conference*, pp. 282–289, 2002.

[30] H. Craig "Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?" *Literary and Linguistic Computing*, vol. 14, no. 1, pp. 103–113, 1999.

[31] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the Third Conference on Applied Natural Lanugage Processing*, Association for Computational Linguistics, Trento, Italy, April 1992. Also available as Xerox PARC technical report SSL-92-01.

[32] A. de Morgan, "Letter to Rev. Heald 18/08/1851," in *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*, (S. Elizabeth and D. Morgan, eds.), London: Longman's Green and Co., 1851/1882.

[33] G. Easson, "The linguistic implications of shibboleths," in *Annual Meeting of the Canadian Linguistics Association*, Toronto, Canada, 2002.

[34] A. Ellegard, *A Statistical Method for Determining Authorship: The Junius Leters 1769–1772*. Gothenburg, Sweden: University of Gothenburg Press, 1962.

[35] W. Elliot and R. J. Valenza, "And then there were none: Winnowing the Shakespeare claimants," *Computers and the Humanities*, vol. 30, pp. 191–245, 1996.

[36] W. Elliot and R. J. Valenza, "The professor doth protest too much, methinks," *Computers and the Humanities*, vol. 32, pp. 425–490, 1998.

[37] W. Elliot and R. J. Valenza, "So many hardballs so few over the plate," *Computers and the Humanities*, vol. 36, pp. 455–460, 2002.

[38] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence," in *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 48–57, San Francisco, California, January 22–24, 1995.

[39] J. M. Farringdon, *Analyzing for Authorship: A Guide to the Cusum Technique*. Cardiff: University of Wales Press, 1996.

[40] R. S. Forsyth, "Towards a text benchmark suite," in *Proceedings of 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.

[41] D. Foster, *An Elegy by W.S.: A Study in Attribution*. Newark: University of Delaware Press, 1989.

[42] D. Foster, "Attributing a funeral elegy," *PMLA*, vol. 112, no. 3, pp. 432–434, 1997.

[43] D. Foster, *Author Unknown: Adventures of a Literary Detective*. London: Owl Books, 2000.

[44] D. Foster, *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company, 2001.

[45] W. Fucks, "On the mathematical analysis of style," *Biometrika*, vol. 39, pp. 122–129, 1952.

[46] J. Gibbons, *Forensic Linguistics: An Introduction to Language in the Justice System*. Oxford: Blackwell, 2003.

[47] N. Graham, G. Hirst, and B. Marthi, "Segmenting documents by stylistic character," *Natural Language Engineering*, vol. 11, pp. 397–415, 2005.

[48] T. Grant and K. Baker, "Identifying reliable, valid markers of authorship: A reponse to Chaski," *Forensic Linguistics*, vol. 8, no. 1, pp. 66–79, 2001.

[49] T. R. G. Green, "The necessity of syntax markers: Two experiments with artificial languages," *Journal of Verbal Learning and Verbal Behavior*, vol. 18, pp. 481–96, 1979.

[50] J. W. Grieve, "Quantitative authorship attribution: A history and an evaluation of techniques". Master's thesis, Simon Fraser University, 2005. URL: http://hdl.handle.net/1892/2055, accessed May 31, 2007.

[51] J. Hancock, "Digital deception: When, where and how people lie online," in *Oxford Handbook of Internet Psychology*, (K. McKenna, T. Postmes, U. Reips, and A. Joinson, eds.), pp. 287–301, Oxford: Oxford University Press, 2007.

[52] R. A. Hardcastle, "Forensic linguistics: An assessment of the Cusum method for the determination of authorship," *Journal of the Forensic Science Society*, vol. 33, no. 2, pp. 95–106, 1993.

[53] R. A. Hardcastle, "Cusum: A credible method for the determination of authorship?," *Science and Justice*, vol. 37, no. 2, pp. 129–138, 1997.

[54] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison Wesley, 1991.

[55] M. L. Hilton and D. I. Holmes, "An assessment of cumulative sum charts for authorship attribution," *Literary and Linguistic Computing*, vol. 8, pp. 73–80, 1993.

[56] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[57] D. I. Holmes, "The evolution of stylometry in humanities computing," *Literary and Linguistic Computing*, vol. 13, no. 3, pp. 111–117, 1998.

[58] D. I. Holmes and R. S. Forsyth, "The Federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111–127, 1995.

[59] D. I. Holmes, "Stylometry and the civil war: The case of the Pickett letters," *Chance*, vol. 16, no. 2, pp. 18–26, 2003.

[60] D. I. Holmes and F. J. Tweedie, "Forensic stylometry: A review of the CUSUM controversy," in *Revue Informatique et Statistique dans les Science Humaines*, pp. 19–47, University of Liege, Liege, Belgium, 1995.

[61] D. Hoover, "Another perspective on vocabulary richness," *Computers and the Humanities*, vol. 37, no. 2, pp. 151–178, 2003.

[62] D. Hoover, "Stylometry, chronology, and the styles of Henry James," in *Proceedings of Digital Humanities 2006*, pp. 78–80, Paris, 2006.

[63] D. L. Hoover, "Delta prime?," *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 477–495, 2004.

[64] D. L. Hoover, "Testing Burrows's Delta," *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 453–475, 2004.

[65] J. Hopcroft and J. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley, 1979.

[66] S. R. Hota, S. Argamon, M. Koppel, and I. Zigdon, "Performing gender: Automatic stylistic analysis of Shakespeare's characters," in *Proceedings of Digital Humanities 2006*, pp. 100–104, Paris, 2006.

[67] IGAS, "IGAS — Our Company," http://www.igas.com/company.asp, accessed May 31, 2007.

[68] M. P. Jackson, "Function words in the 'funeral elegy'," *The Shakespeare Newsletter*, vol. 45, no. 4, p. 74, 1995.

[69] T. Joachims, *Learning to Classify Text Using Support Vector Machines*. Kluwer, 2002.

[70] E. Johnson, *Lexical Change and Variation in the Southeastern United States 1930–1990*. Tuscaloosa, AL: University of Alabama Press, 1996.

[71] P. Juola, "What can we do with small corpora? Document categorization via cross-entropy," in *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK, 1997.

[72] P. Juola, "Cross-entropy and linguistic typology," in *Proceedings of New Methods in Language Processing and Computational Natural Language Learning*, (D. M. W. Powers, ed.), Sydney, Australia: ACL, 1998.

[73] P. Juola, "Measuring linguistic complexity: The morphological tier," *Journal of Quantitative Linguistics*, vol. 5, no. 3, pp. 206–213, 1998.

[74] P. Juola, "The time course of language change," *Computers and the Humanities*, vol. 37, no. 1, pp. 77–96, 2003.

[75] P. Juola, "*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)*, Göteborg, Sweden, June 2004.

[76] P. Juola, "On composership attribution," 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)*, Göteborg, Sweden, June 2004.

[77] P. Juola, "Compression-based analysis of language complexity," Presented at *Approaches to Complexity in Language*, 2005.

[78] P. Juola, "Authorship attribution for electronic documents," in *Advances in Digital Forensics II*, (M. Olivier and S. Shenoi, eds.), pp. 119–130, Boston: Springer, 2006.

[79] P. Juola, "Becoming Jack London," *Journal of Quantitative Linguistics*, vol. 14, no. 2, pp. 145–147, 2007.

[80] P. Juola and H. Baayen, "A controlled-corpus experiment in authorship attribution by cross-entropy," *Literary and Linguistic Computing*, vol. 20, pp. 59–67, 2005.

[81] P. Juola, J. Sofko, and P. Brennan, "A prototype for authorship attribution studies," *Literary and Linguistic Computing*, vol. 21, no. 2, pp. 169–178, Advance Access published on April 12, 2006; doi: doi:10.1093/llc/fql019, 2006.

[82] G. Kacmarcik and M. Gamon, "Obfuscating document stylometry to preserve author anonymity," in *Proceedings of ACL 2006*, 2006.

[83] A. Kenny, *The Computation of Style*. Oxford: Pergamon Press, 1982.

[84] V. Keselj and N. Cercone, "CNG method with weighted voting," in *Ad-hoc Authorship Attribution Contest*, (P. Juola, ed.), ACH/ALLC 2004, 2004.

[85] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING03*, pp. 255–264, Dalhousie University, Halifax, NS, August 2003.

[86] K. Keune, M. Ernestus, R. van Hout, and H. Baayen, "Social, geographical, and register variation in Dutch: From written MOGELIJK to spoken MOK," in *Proceedings of ACH/ALLC 2005*, Victoria, BC, Canada, 2005.

[87] D. V. Khmelev and F. J. Tweedie, "Using markov chains for identification of writers," *Literary and Linguistic Computing*, vol. 16, no. 3, pp. 299–307, 2001.

[88] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1519–1525, 2006.

[89] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, doi:10.1093/llc/17.4.401, 2002.

[90] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003.

[91] M. Koppel and J. Schler, "*Ad-hoc* authorship attribution competition approach outline," in *Ad-hoc Authorship Attribution Contest*, (P. Juola, ed.), ACH/ALLC 2004, 2004.

[92] L. Kruh, "A basic probe of the Beale cipher as a bamboozlement: Part I," *Cryptologia*, vol. 6, no. 4, pp. 378–382, 1982.

[93] L. Kruh, "The Beale cipher as a bamboozlement: Part II," *Cryptologia*, vol. 12, no. 4, pp. 241–246, 1988.

[94] H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English*. Providence: Brown University Press, 1967.

[95] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," *Lecture Notes in Computer Science*, vol. 4243, p. 274283, 2006.

[96] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, "Using literal and grammatical statistics for authorship attribution," *Problemy Peredachi Informatii*, vol. 37, no. 2, pp. 96–198, Translated in "Problems of Information Transmission," pp. 172–184, 2000.

[97] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications. Graduate Texts in Computer Science*, New York: Springer, second ed., 1997.

[98] H. Love, *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press, 2002.

[99] C. Martindale and D. McKenzie, "On the utility of content analysis in authorship attribution: The Federalist Papers," *Computers and the Humanities*, vol. 29, pp. 259–70, 1995.

[100] R. A. J. Matthews and T. V. N. Merriam, “Neural computation in stylometry I: An application to the works of Shakespeare and Marlowe,” *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 203–209, 1993.

[101] J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1987.

[102] G. McMenamin, “Disputed authorship in US law,” *International Journal of Speech, Language and the Law*, vol. 11, no. 1, pp. 73–82, 2004.

[103] G. R. McMenamin, *Forensic Stylistics*. London: Elsevier, 1993.

[104] G. R. McMenamin, “Style markers in authorship studies,” *Forensic Linguistics*, vol. 8, no. 2, pp. 93–97, 2001.

[105] G. R. McMenamin, *Forensic Linguistics — Advances in Forensic Stylistics*. Boca Raton, FL: CRC Press, 2002.

[106] T. C. Mendenhall, “The characteristic curves of composition,” *Science*, vol. IX, pp. 237–249, 1887.

[107] T. V. N. Merriam and R. A. J. Matthews, “Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe,” *Literary and Linguistic Computing*, vol. 9, no. 1, pp. 1–6, 1994.

[108] G. Monsarrat, “A funeral elegy: Ford, W.S., and Shakespeare,” *Review of English Studies*, vol. 53, p. 186, 2002.

[109] A. W. Moore, “Support Vector Machines,” Online tutorial: http://jmvidal.cse.sc.edu/csce883/svm14.pdf, accessed May 31, 2007, 2001.

[110] J. L. Morgan, *From Simple Input to Complex Grammar*. Cambridge, MA: MIT Press, 1986.

[111] A. Q. Morton, *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribner’s, 1978.

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

[113] M. Newman, J. Pennebaker, D. Berry, and J. Richards, “Lying words: Predicting deception from linguistic style,” *Personality and Social Psychology Bulletin*, vol. 29, pp. 665–675, 2003.

[114] S. Nowson and J. Oberlander, “Identifying more bloggers: Towards large scale personality classifiation of personal weblogs,” in *International Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[115] M. Oakes, “Text categorization: Automatic discrimination between US and UK English using the chi-square text and high ratio pairs,” *Research in Language*, vol. 1, pp. 143–156, 2003.

[116] J. Oberlander and S. Nowson, “Whose thumb is it anyway? classifying author personality from weblog text,” in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, pp. 627–634, Sydney, Australia, 2006.

[117] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, “Language independent authorship attribution using character level language models,” in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 267–274, Budapest: ACL, 2003.

[118] J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference,” *Journal of Personality and Social Psychology*, vol. 77, pp. 1296–1312, 1999.

[119] J. W. Pennebaker and L. D. Stone, “Words of wisdom: Language use over the life span,” *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 291–301, 2003.

[120] J. Pennebaker, M. Mehl, and K. Niederhoffer, “Psychological aspects of natural language use: Our words, ourselves,” *Annual Review of Psychology*, vol. 54, pp. 547–577, 2003.

[121] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1993.

[122] M. Rockeach, R. Homant, and L. Penner, “A value analysis of the disputed Federalist Papers,” *Journal of Personality and Social Psychology*, vol. 16, pp. 245–250, 1970.

[123] S. Rude, E. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students,” *Cognition and Emotion*, vol. 18, pp. 1121–1133, 2004.

[124] J. Rudman, “The state of authorship attribution studies: Some problems and solutions,” *Computers and the Humanities*, vol. 31, pp. 351–365, 1998.

[125] J. Rudman, “Non-traditional authorship attribution studies in eighteenth century literature: Stylistics, statistics and the computer,” URL: http://computerphilologie.uni-muenchen.de/jg02/rudman.html, accessed May 31, 2007.

[126] J. Rudman, “The State of Authorship Attribution Studies: (1) The History and the Scope; (2) The Problems — Towards Credibility and Validity,” Panel session from ACH/ALLC 1997, 1997.

[127] J. Rudman, “The non-traditional case for the authorship of the twelve disputed Federalist Papers: A monument built on sand,” in *Proceedings of ACH/ALLC 2005*, Association for Computing and the Humanities, Victoria, BC, 2005.

[128] D. Rumelhart, G. Hinton, and R. Williams, “Learning internal representations by error propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 318–362, The MIT Press, 1986.

[129] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, no. 4, pp. 379–423, 1948.

[130] C. E. Shannon, “Prediction and entropy of printed English,” *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[131] E. H. Simpson, “Measurement of diversity,” *Nature*, vol. 163, p. 688, 1949.

[132] S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*. Anchor, 2000.

[133] M. Smith, “Recent experiences and new developments of methods for the determination of authorship,” *Association of Literary and Linguistic Computing Bulletin*, vol. 11, pp. 73–82, 1983.

[134] H. H. Somers, “Statistical methods in literary analysis,” in *The Computer and Literary Style*, (J. Leed, ed.), Kent, OH: Kent State University Press, 1972.

[135] H. Somers, "An attempt to use weighted cusums to identify sublanguages," in *Proceedings of New Methods in Language Processing 3 and Computational Natural Langauge Learning*, (D. M. W. Powers, ed.), Sydney, Australia: ACL, 1998.

[136] H. Somers and F. Tweedie, "Authorship attribution and pastiche," *Computers and the Humanities*, vol. 37, pp. 407–429, 2003.

[137] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," *Computers and the Humanities*, vol. 35, no. 2, pp. 193–214, 2001.

[138] S. Stein and S. Argamon, "A mathematical explanation of Burrows' Delta," in *Proceedings of Digital Humanities 2006*, Paris, France, July 2006.

[139] D. R. Tallentire, "Towards an archive of lexical norms — a proposal," in *The Computer and Literary Studies*, Cardiff: Unversity of Wales Press, 1976.

[140] S. Thomas, "Attributing *a funeral elegy*," *PMLA*, vol. 112, no. 3, p. 431, 1997.

[141] E. Tufte, *Envisioning Information*. Graphics Press, 1990.

[142] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The Federalist Papers," *Computers and the Humanities*, vol. 30, no. 1, pp. 1–10, 1996.

[143] L. Ule, "Recent progress in computer methods of authorship determination," *Association for Literary and Linguistic Computing Bulletin*, vol. 10, pp. 73–89, 1982.

[144] H. van Halteren, "Author verification by linguistic profiling: An exploration of the parameter space," *ACM Transactions on Speech and Language Processing*, vol. 4, 2007.

[145] 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*, vol. 12, no. 1, pp. 65–77, 2005.

[146] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.

[147] W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 2002.

[148] B. Vickers, *Counterfeiting Shakespeare*. Cambridge: Cambridge University Press, 2002.

[149] F. L. Wellman, *The Art of Cross-Examination*. New York: MacMillan, fourth ed., 1936.

[150] C. B. Williams, *Style and Vocabulary: Numerical Studies*. London: Griffin, 1970.

[151] A. J. Wyner, "Entropy estimation and patterns," in *Proceedings of the 1996 Workshop on Information Theory*, 1996.

[152] B. Yu, Q. Mei, and C. Zhai, "English usage comparison between native and non-native english speakers in academic writing," in *Proceedings of ACH/ALLC 2005*, Victoria, BC, Canada, 2005.

[153] G. U. Yule, "On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship," *Biometrika*, vol. 30, pp. 363–90, 1938.

[154] G. U. Yule, *The Statistical Study of Literary Vocabulary.* Cambridge: Cambridge University Press, 1944.

[155] P. M. Zatko, “Alternative routes for data acquisition and system compromise,” in *3rd Annual IFIP Working Group 11.9 International Conference on Digital Forensics*, Orlando, FL, 2007.

[156] H. Zhang, “The optimality of naive bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, (V. Barr and Z. Markov, eds.), Miami Beach, FL: AAAI Press, 2004.

[157] R. Zheng, J. Li, H. Chen, and Z. Huang, “A framework for authorship identification of online messages: Writing-style features and classification techniques,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[158] G. K. Zipf, *Human Behavior and the Principle of Least Effort.* New York: Hafner Publishing Company, 1949. Reprinted 1965.