

CRI: Collaborative Research: Planning Proposal: Community Resources for Research in Automated Authorship Attribution

1 Overview

Homeland security and the criminal and civil justice systems increasingly require reliable and valid methods to identify the authors of anonymous documents. Further demand arises from fields as diverse as computer forensics and literary studies. However, despite the clear and growing need for methodological research in this arena, there are as yet no standard test suites for authorship attribution, and hence no agreed upon ways to compare research results and validate techniques. This situation, combined with the highly interdisciplinary nature of the field, has led to much redundant and sometimes unsound research. We therefore propose a one-year NSF-supported project to develop a detailed plan for a test corpus and associated resources for research on automated authorship attribution. We envisage enabling future proposals that will seek funding from multiple sources to implement the plan.

Individuals have distinctive ways of speaking and writing, and there is a long history of linguistic and stylistic investigation into authorship attribution. In recent years, practical applications for authorship attribution have grown in areas such as intelligence (linking intercepted messages to each other and to known terrorists), criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (tracking authors of computer virus source code). This activity is part of a broader growth within compute science of identification technologies, including biometrics (retinal scanning, speaker recognition, etc.), cryptographic signatures, intrusion detection systems, and others. As with these other technologies, there is increasing interest in the accuracy and scientific validity of determinations of authorship by both human experts and automated methods.

Automating authorship attribution promises more accurate results and objective measures of reliability, both of which are critical for legal and security applications. Recent research has used techniques from machine learning [3, 16, 21, 41, 57], multivariate and cluster analysis [35, 36, 12, 30], and natural language processing [5, 56] in authorship attribution. These techniques have also been applied to related problems such as genre analysis [4, 1, 8, 23, 34, 56] and author profiling (such as by gender [2, 20] or personality [52]). This work is scattered among the literatures of several areas of computer science (machine learning, computational linguistics, computer forensics), as well as statistics, linguistics, and literary analysis. Methodological inconsistencies, unsettled controversies, and redundancy are rife.

We propose to advance authorship attribution research by constructing a detailed plan for developing community resources. The plan will be developed in concert with leading authorship researchers from a range of fields and reviewed at a workshop held towards the end of the project period. The plan will form the basis of one or more funding proposals by ourselves and others to construct these community resources. The process of creating the plan will itself draw together the authorship attribution community and aid research progress.

Our focus will be the design of a collection or *corpus* of texts of known authorship, for use as a research/development testbed, along with the definition of standard attribution tasks for comparing techniques. Use of a standardized corpus and standardized tasks will enable objective comparisons between different techniques and provide for validated estimates of reliability (as required by legal

standards of evidence, for example).

Designing a corpus adequate for testing hypotheses and drawing supportable conclusions in authorship attribution is a complex and subtle task. How language is used in a text is affected by many factors besides the identity of the author, including the text's topic, genre (news, business report, fiction), purpose (information, persuasion, entertainment), time of composition, intended audience (broad/specific, young/old, etc.), medium (speech, book, email), and others. Past research has sometimes ignored these factors, leading to misleading conclusions. Properly controlling for these factors, and drawing correct conclusions about their impact, requires a large and carefully selected range of texts. Negotiating access to materials will be a major issue, as will formatting data for ease of use and to omit unrealistic cues to authorship.

Another important area which we will address is determining exactly which tasks should be defined and supported by the corpus. Different identification tasks include (i) identifying which of a (small or large) set of authors wrote a text, (ii) determining whether two texts were written by different authors, and (iii) determining if a text was written by none of a set of defined authors. A variety of 'profiling' tasks may also be considered, such as determining the age, dialect, or gender of an author. Some such tasks may require additional annotation of the corpus, and all will affect the choice of what sorts of texts to include.

Finally, we will consider what activities beyond corpus design and construction should be associated with the project. Formal evaluations (with or without associated meetings), periodic releases of new data, updating and improving content and annotations, and support of hosting and maintenance are all issues that will be discussed.

A good precedent for our proposal is Sparck Jones and van Rijsbergen's *Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection*, funded by the British Library in 1975. This report had a major influence on the design of the DARPA/ARPA Tipster [19] and NIST TREC (Text REtrieval Conference) [27] test collections in the early 1990s. These collections and associated evaluations led to new text retrieval techniques usable on millions of heterogeneous documents. Later TREC and CLEF¹ collections have unified and stimulated research communities working on text filtering, speech and video retrieval, cross-language retrieval, and more. The TREC question answering track has been particularly successful at bringing together a diverse set of researchers in natural language processing, information retrieval, databases, machine learning, and statistics, significantly boosting research progress on the problem.

Corpus and testbed building efforts in other fields have had similar positive impacts. The data sets produced by the Linguistic Data Consortium have been used in hundreds of studies and have had a major impact in the explosion of high-quality empirical and statistical work in computational linguistics [15]. The UCI repositories for machine learning and data mining have had a similarly significant impact on their fields [10, 7]. The Reuters collections [39] have helped turn text categorization from a niche application into a thriving research and operational field bringing together machine learning and information retrieval. Our goal is to similarly unify and galvanize the authorship attribution research community and enable rapid research progress.

Proposed Activities: To ensure that the proposed planning project addresses the needs of the authorship attribution community, we have recruited an advisory Working Group of eminent computer scientists, statisticians, linguists, and literary scholars who actively work on authorship

¹Cross-Language Evaluation Forum, sponsored by an international consortium of national organizations and institutes in the US and Europe.

attribution and related problems. In consultation with this Working Group, we will undertake the following activities during the project period to gather information and devise a draft plan for development of community resources. We will survey both what resources already exist and what practitioners feel they need. We will also assess the costs of obtaining, cleaning, annotating, distributing, and maintaining corpus data, as well as legal issues such as copyright. The plan we develop will be discussed and revised at a workshop held towards the end of the project period, with the Working Group and other invited researchers. The revised plan will then be disseminated, and will serve as the basis for funding proposals to develop the planned resources. The workshop will be hosted by the Center for Discrete Mathematics and Theoretical Computer Science at Rutgers University (DIMACS), which has nearly two decades of experience organizing and hosting similar workshops.

Intellectual Merit and Broader Impacts: The proposed planning project will help unify researchers working on authorship attribution, encouraging greater communication between researchers coming from different fields. The corpus itself, once developed, will facilitate both more work in authorship attribution, and better work, with deeper understanding resulting from more rigorous and quantitative results. This work will both contribute to and benefit from an increasingly sophisticated understanding of other identification technologies, such as biometrics, authentication protocols, and computer intrusion detection.

More systematic research on authorship attribution is likely to produce results of broad interest in information retrieval, natural language processing, machine learning, statistics, and related fields. Authorship attribution work forces researchers to confront a range of challenges very different from those in, for instance, content-oriented text classification or information extraction. This synergy will be enhanced by the highly interdisciplinary nature of the field, which already involves computer scientists, statisticians, linguists, and literary theorists. Creating a common testbed will encourage discussion and cross-fertilization of ideas from these disparate fields.

The community resources we propose to design will also have a substantial impacts on applications of authorship attribution. Intelligence analysis, criminal investigations, commercial dispute resolution, and literary scholarship will all benefit from more effective and more rigorously evaluated authorship attribution technology. The resources may also stimulate the development of entirely new technologies based on linguistic analysis of style (e.g. in marketing).

2 Background

2.1 Computational stylistics and forensic authorship analysis

Our long-range goal is to develop community resources to support research on automated authorship attribution and related problems. To place this proposal in context, we review briefly the relevant history and methods of research on computational and forensic stylistics.

Scientific investigation into measuring style and authorship of texts goes back to the late nineteenth century, with the pioneering studies of Mendenhall [47] and Mascol [43, 44] on distributions of sentence and word lengths in works of literature and the gospels of the New Testament.

The underlying notion was that works by different authors are strongly distinguished by quantifiable features of the text. By the mid-twentieth century, this line of research had grown into what became known as “stylometrics”, and a variety of textual statistics had been proposed to quantify

textual style. The style of early work was characterized by a search for invariant properties of textual statistics, such as Zipf’s distribution [60] and Yule’s K statistic [59].

Modern work in authorship attribution (often referred to in the humanities as “nontraditional authorship attribution”) was ushered in by Mosteller and Wallace in the 1960s, in their seminal study *The Federalist Papers* [49]. The study examined 146 political essays from the late eighteenth century, of which most are of acknowledged authorship by John Jay, Alexander Hamilton, and James Madison, though twelve are claimed by both Hamilton and Madison. Mosteller and Wallace showed statistically significant discrimination results by applying Bayesian statistical analysis to the frequencies of a small set of ‘function words’ (such as ‘the’, ‘of’, or ‘about’), as stylistic features of the text. Function words, and other similar classes of words, remain the most popular stylistic features used for authorship discrimination.

Other stylometric features that have been applied include various measures of vocabulary richness and lexical repetition, based on Zipf’s [60] studies on word frequency distributions. Most such measures, however, are strongly dependent on the length of the text being studied, and so are difficult to apply reliably. Many other types of features have been applied, including word class frequencies [2, 24], syntactic analysis [5, 56], word collocations [30, 55], grammatical errors [38], and word, sentence, clause, and paragraph lengths [3, 42]. Many studies combine features of different types using multivariate analysis techniques.

One effective technique, pioneered for authorship studies by Burrows [12, 13], is to use principal components analysis (PCA) to find combinations of style markers that can discriminate between a particular pair (or small set) of authors. This method has been used to good effect in several studies, including [29] and [5]. Another related class of techniques that have been applied are machine learning algorithms (such as Winnow [40] or Support Vector Machines [17]) which can construct discrimination models over large numbers of documents and features. Such techniques have been applied widely in topic-based text categorization (see the excellent survey [54]) and other stylistic discrimination tasks (e.g. [2, 37, 56]), as well as for authorship discrimination [3, 21]. Often, studies have relied on intuitive evaluation of results, based on visual inspection of scatterplots and cluster-analysis trees, though recent work (e.g. [3, 20, 21]) has begun to apply somewhat more rigorous tests of statistical significance and cross-validation accuracy. The proposed project will encourage the development of more rigorous and widely-accepted evaluation standards for authorship studies.

A somewhat controversial technique is that of Cumulative Sums (CUSUM), developed by Morton and Farrington [48, 22]. This method seeks to analyze whether or not a given document is of homogeneous authorship, or was written by multiple authors (to do authorship attribution, then, the test document is simply concatenated with documents of known authorship). A graph is constructed showing the cumulative deviations of different marker features (typically sentence length, and frequency of short words per sentence) from their overall means in the document. The claim is that if different authors wrote different parts of the document, the different curves will diverge. This is somewhat dependent, however, on the analyst’s reading of the graph, and the method has been criticized as unsound and ineffective in the literature [6, 28], though it has been accepted as expert evidence in British courts. A standardized corpus will help resolve such controversies.

Indeed, in forensic text analysis, where quantifiable effectiveness is now a *sine qua non* (due to the *Daubert* decision—see below), disputes have become heated (see [14, 26, 45]). These disputes concern the use of different feature sets, analysis techniques, and evaluation methodologies. The standard forensic approach (see [46]) for determining if a candidate author wrote a test document

consists of (a) choosing an appropriate ‘confusion set’ of comparison documents by other authors (controlling for dialect, genre, and register), (b) examining a variety of textual features to find those that consistently differentiate between the candidate and other authors, (c) determining which of those occur in the test document, and (d) performing statistical tests to see if the characteristics of the test document can be attributed to chance. There is a great deal of expert judgement involved in this sort of procedure, which would be acceptable if there were generally accepted standards in the field. However, there is very little agreement in the field over what level of control is necessary in the confusion set, what sorts of features are most reliable, what statistical tests to use, and most critically, how to determine the reliability of a proposed attribution. This state of affairs underscores the necessity both for a standard corpus and tasks, as well as for forums that bring together researchers to share results.

2.2 Need for reliable authorship attribution

Accuracy in both attribution, and in estimating the reliability of attribution, are essential for automated authorship attribution to meet its promise in real-world tasks. However, estimating reliability will be possible only after large-scale systematic research on attribution efficacy in different scenarios. The community resources we propose to design will enable such research to be undertaken.

A key application is homeland security, where authorship attribution by human experts (say of intercepted communications), can be prohibitively costly. Informal language-based authorship analysis has in the past proven important for identifying terrorists; for example, Ted Kaczinsky, the Unabomber, was found in part because a relative recognized his writing style. On a larger scale, automated authorship analysis might be combined with other data mining techniques in analyzing document collections such as those seized in Iraq and Afghanistan. To combine information from authorship analysis with other, perhaps conflicting, data, it will be important for to have a clear measure of attribution reliability,

Human authorship analysis has been important in a number of criminal cases (for example the ransom note in the still unsolved JonBenét Ramsey case [31]). Such cases are currently addressed by expert forensic linguists, whose field which has been growing in prominence [46]. A fundamental difficulty in the forensic application of authorship attribution, however, is determining and describing reliability of the evidence. In contrast to, for instance, DNA analysis, accurate statistical determinations of attribution reliability are not currently possible for human linguistic analysis. This concern was deepened by the decision of the United States Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993), which established a test for expert testimony which includes (among others), “whether the theory or technique in question can be (and has been) tested, ... its known or potential error rate, and the existence and maintenance of standards controlling its operation”. A carefully designed corpus holds the promise that methods can be validated to satisfy this test.

Potential commercial applications of automated authorship attribution include automated detection of copyright infringement and plagiarism (a problem that has grown with the Internet), and analysis of customer communications. Sophisticated authorship analysis tools may help improve the readability of collaborative texts, by ‘smoothing out’ their style [25], and may also contribute to development of more sophisticated information retrieval systems.

Finally, as the history of statistical authorship analysis shows, reliable techniques in this area

have the potential to contribute to scholarly debates about the authorship of historical documents, and in turn to improve our understanding of past events.

3 Methodological Issues

As we hope is now clear, authorship attribution is a complex interdisciplinary problem, to which a great variety of approaches have been applied. The nature of the problem has led to fragmentation in the field, as linguists, literary scholars, statisticians, computer scientists, and historians have all worked on authorship attribution, using different methods, making different assumptions, and publishing in a large number of different journals. As Rudman notes in his recent critique of the field [53], such fragmentation has led to much methodological irregularity.

There have been previous small-scale attempts to share data sets and encourage common methodology among authorship attribution researchers. Most notable is a recent authorship attribution “competition” organized by Patrick Juola and Stephen Ramsay [33, 32]. It followed good practice in defining separate training and test sets, and included several types of material (essays, novels, poetry, plays, and speech transcripts). However, the amount of data was tiny, with only 98 test documents spread across 13 subtasks and 5 languages. Only one of the subtasks allowed for textual controls (see below) and that one only for topic. While shared data sets and small evaluations such as this have given some insight into the merits of different authorship attribution methods, they have not led to significant progress in authorship attribution technology, nor allowed reliability estimates to be developed suitable for legal testimony and critical applications.

The main goal of this planning proposal is working towards developing research community resources to help remedy such irregularities. Methodological difficulties in current authorship research can be roughly divided into four categories:

Corpus design: Standard authorship attribution studies begin with a corpus of documents of known authorship. This is complicated, however, by the possibility of collaboration and editing, which are not always controlled for. Indeed, Rudman [53] provides several examples of published authorship attribution studies where the provenance of the documents is questionable. One issue concerns multiple editions of the same document. For example, the word “further” appears only once in the first edition of Daniel Defoe’s “Memoirs of an English Officer” but appears 27 times in a later definition. Other issues include edited documents and multi-authored documents. When dealing with historical documents, orthographic conventions can differ between different editions, such as the common interchange of ‘u’ and ‘v’ in Elizabethan English—standardizing all documents used in a study is time-consuming and expensive, but crucial. Unfortunately, many studies have used just the most easily available editions, for convenience, making their results suspect.

A large, publicly available corpus containing verifiably single-authored documents of known provenance, standardized for formatting and orthography would advance the field of authorship attribution considerably, by enabling comparison of properly controlled studies.

Proper and complete use of textual controls: Picking appropriate control documents presents a significant challenge in authorship attribution studies. For example consider documents by a sports journalist, say John, writing about the exploits of Babe Ruth in the 1920’s. Suppose we have a challenge document, also about Babe Ruth in the 1920’s, but actually written

by a journalist other than John. If the training corpus contains no articles about baseball or about Babe Ruth or written in the 1920's, then a "John" attribution may be inevitable, no matter how sophisticated the attribution methodology. On the other hand, if the corpus contains articles by many different sports journalists of the period, a correct identification may emerge. A substantial corpus with sub-corpora that control for period, topic, and genre does not currently exist and represents a significant obstacle.

Feature selection: Traditional authorship attribution studies represent documents using so-called function words (such as "and," "of," "the," etc.), as well as numerical statistics such as average word, sentence, and paragraph length. The idea is that such features are topic independent; should an author write about a new topic in the future, the relative frequencies of the various function words would remain approximately constant. However, different researchers have used different function words and the notion of "topic-free" does not admit a simple definition. Argamon et al. [2], for example, have used a list of 450 function words that include "art" and "hopeless." Mosteller and Wallace [49] used a smaller list but included such words as "violence" and "rapid." Novak et al. [50] used *all* words in the document. Defining topic-free author attribution features is an open problem. In the context of disputed authorship, Rudman [53] warns against picking the features that best align a set of documents with a particular author, referring to this as "cherry picking."

One approach to better understanding the topic-dependence of textual features is systematic experimentation examining statistics and discrimination utility of various kinds of features in a variety attribution tasks controlled for topic at different levels of specificity. Such experimentation will be enabled by a large and varied corpus, stratified for topic, as we propose to work towards here.

Proper use of algorithms and statistics: According to Rudman, the authorship attribution literature presents many examples of inappropriate statistical practice such as selecting the statistical test that gives the desired result and, a more subtle problem, inappropriate accounting for and reporting of, uncertainty. The literature rarely provides estimates of predictive error. A large corpus with specific guidelines concerning test-training splits and cross-validation procedures can help, as will the development of a more integrated research community.

The issues we have just outlined have complex causes. First, construction of properly controlled and balanced corpus presents a significant challenge. Development of a communal corpus and associated set of tasks will directly help researchers overcome this difficulty. Second, and more profoundly, authorship attribution research teams rarely include expertise in all the relevant areas of computer science, statistics, linguistics, and textual scholarship. Hence it is only too natural for shortcuts to be taken in one area or another. Furthermore, the fact that authorship attribution is pursued in several different research communities in parallel makes it far too easy for work that is suspect in some way to get published and accepted by some subset of the field. Here too we believe that our proposed program for development of communal resources will help, by providing a mechanism for researchers to compare and communicate their results based on a common testbed. In the following sections, we describe in detail the issues that we will explore in the proposed planning project, and how development of such a corpus will aid research in authorship attribution.

4 Constructing an Author Identification Corpus

As discussed in the previous section, a variety of problems have plagued work in authorship attribution. We believe that many of these problems result largely from a lack of a critical piece of research infrastructure, namely open-access, high quality corpora. Corpora are to text analysis what sequence databases are to genetics: resources that provide a multiplier effect far in excess of the initial investment.

A well-designed corpus is a boon to any data intensive field. Barriers to entering the fields are reduced and a larger proportion of time and money can be spent on research questions. Experimental results have improved comparability (since data has been prepared the same way) and quality (since the preparation can more easily be done to high standards). The methodological issues discussed in the previous section suggest authorship attribution will benefit even more than other fields.

This section outlines the main issues that we will explore during the proposed activity, and how we propose to resolve them, with the active involvement of the project Working Group. These issues include:

- The needs of authorship attribution researchers from different fields.
- What kinds of texts are needed to properly enable comprehensive and controlled experiments—while easily-available texts such as those downloadable from the Web are appropriate for some purposes, they do not come near to covering the range of kinds of texts used in authorship analysis applications;
- How to preprocess and format texts in the corpus—raw texts are more useful for some purposes, while more standardized texts (for spelling, say) may be more appropriate for others;
- What sort of auxiliary annotations to include, such as word tagging, demographic author information, and the like;
- Whether and what software should be included with the corpus;
- What (if any) auxiliary activities should be established in conjunction with the corpus (such as public evaluations and other dissemination activities);
- How to deal with possible legal issues such as copyright and licensing;
- Design of a workable maintenance plan for managing the resources that are constructed over the long term.

4.1 Target user communities

There are many research communities whose needs must be considered in designing an authorship attribution corpus. They include:

- *Forensic linguistics*: A primary concern here is improving and validating (ideally to courtroom standards) the reliability of authorship judgments under conditions encountered in practice. A major question is how reliable can judgments be when the texts available on suspect authors are few, short, or of different character than texts to be identified. These researchers will be particularly interested in a wide diversity of material in a corpus.
- *Literary stylistics, stylometrics, corpus linguistics, etc.*: These are the researchers with the most direct interest in documents as writings. The pedigree, literary merit, and choice of

versions of texts are of interest. These researchers may prefer genres (e.g. fiction, poetry, historical materials) different from those of interest to more applied fields.

- *Machine learning, statistics, pattern recognition, etc.:* Researchers in these fields usually view author identification as just one of many benchmark tasks on which to test algorithms. They may be unfamiliar with text processing, and so providing them corpus texts in the form of extracted feature values will be of interest. Documentation that makes clear the particular concerns of authorship attribution and presents standard experimental designs may encourage them to produce results comparable with those of other authorship attribution researchers.
- *Information retrieval, computational linguistics, etc.:* Researchers here have similar interests to those in machine learning, though they are more likely to do their own text processing. They may find uses for corpus data besides author identification experiments, and so are more likely to be interested in the particular genre and origins of texts included.
- *Data mining, text mining, link analysis, collaborative filtering, etc.:* These fields are similar to the previous two, but with more emphasis on combining multiple data sources. Research here is particularly relevant to the application areas of business analytics, law enforcement, and intelligence. Researchers in these fields are particularly interested in data sources beyond the authored document, such as demographic data on authors, texts they may have read, etc. Data sets that include the complete message traffic within some community (e.g. a mailing list) may be of great interest.
- *Computer security, software forensics, etc.:* The interest here is identifying authors of texts in computer-related security and law enforcement: electronic mail, chat, source code, internal documents from large organizations, and others. Researchers may desire to use nontextual data such as message headers, web server logs, and packet traces, some of which may not be practical to provide in a general authorship attribution corpus.

The above descriptions are necessarily oversimplifications, and many researchers in these fields have overlapping interests. However, the above indicates the range of user concerns that will be taken into account in the proposed planning activity, as described in the following section.

4.2 Choice of Material and Textual Control

As discussed in Section 3, many factors besides authorship affect the character of a text. The ability to systematically control for these factors in experiments is the single greatest benefit a well-designed corpus will provide. Only through such systematic experiments can the accuracy of authorship attribution techniques in complex, data poor situations be validated and improved.

Genre is perhaps the most critical factor researchers would like to control, since it influences many of the stylistic clues commonly used in determining authorship. Further, forensic authorship attribution tasks often involves samples of authored material from one genre and material to be identified from another. (This is almost axiomatic in some cases, e.g. anonymous threats, or suicide notes.) The range of genres that might be included is immense: technical articles, short stories, email discussion list postings, home pages, homework assignments, press releases, wills, diaries, interviews, phone messages, source code comments, and scores of others. Members of the working

group will draw on an extensive literature on taxonomies and characterizations of genres [9, 51, 18] in proposing genres to include in a corpus.

Other factors to control include whether a text had multiple authors, whether the text has been edited, the topic of the text, the year in which it was produced, and the medium of production (handwriting, typing, audio). The choice of particular languages or dialects to include is important, though for simplicity we will likely focus on English only. The number of authors, and number and length of documents, will be a major factor in balancing issues of corpus expense, diversity, and usefulness.

Demographic characteristics of authors such as age, country of origin, native language, educational history, and profession are of interest for two reasons. First, they influence style and so controlling for them is desirable. Second, in some applications it may not be possible to identify a particular person as author of a text, but determining the likely demographic characteristics of the author would still be of value. Producing corpus materials and metadata to allow studies of both sorts is desirable.

Finally, an author may alter their natural style in an attempt to foil identification. Anonymous harrassers are an important law enforcement example as are, to an unknown extent, communications within criminal and terrorist groups. Author identification cases of popular interest (e.g. *Primary Colors*) may have this characteristic as well. An intriguing possibility in corpus building would be to solicit deliberately deceptive texts from cooperative writers of various skill levels.

4.3 Preprocessing and Formatting

A surprisingly subtle question is how to preprocess and format the texts. Two issues predominate here. First, researchers want formats that make explicit the information they care about, while removing extraneous information that complicates processing. However, one researcher's nuisance may be another's vital clue. A computer forensics expert may be interested in formatting codes in raw word processor files, while a literary stylist may want normalized ASCII text. One researcher may want quoted, non-author material removed from e-mail messages, while another may want to specifically study it.

Second, a corpus is constructed from available materials of known authorship, and used to *simulate* situations of unknown, disputed, and concealed authorship. A realistic simulation is not always straightforward, however. Using e-mail messages, for instance, may require the removal not only of headers, but of signatures, signature blocks, and possibly other self-references by the author in running text. If material is taken from electronic conversations (mailing lists, chat rooms) then the name of an author may appear not only in their own texts, but in those of other authors. These issues are a particularly concern when applying machine learning techniques, which may implicitly use such cues without a researcher realizing it. Different researchers will also have different perspectives on what information should and shouldn't be available.

During the proposed activity, we will obtain samples of material from a variety of sources and genres, and estimate what cleanup and annotation will be necessary for different purposes. Together with the Working Group, we will produce a plan for dealing with these issues, taking into account the needs of the various research communities. Various within-file and auxiliary file markup strategies will be considered, as will providing several versions of the same data.

4.4 Auxiliary Material

In addition to texts in appropriate formats, an authorship attribution corpus might contain other materials. Some of these have been mentioned already, such as vectors of extracted features and demographic information on authors. Good documentation on the structure, origin, and licensing restrictions of texts will of course be important. Documentation on standard experimental designs will also be useful in increasing comparability of results between fields.

Two types of software are also of possible interest. Programs that extract features from text in a standard way may be useful for improving comparability of results, particularly when researchers have access to additional texts of interest. In cases of linguistically complex features, providing such software could also greatly reduce the effort to produce meaningful results.

Also of interest may be software for computing standard evaluation measures on authorship attribution tasks. It can be surprisingly difficult to get two implementations of evaluation software to agree exactly on the value of a measure in all circumstances, particularly for ranking and clustering problems. Standard software, such as *trec_eval* [58], that embodies agreed upon strategies for averaging across data, handling boundary cases, and so on can be helpful.

We will investigate the relative expense and desirability of these various resources, in consultation with the Working Group.

4.5 Auxiliary Activities

There are a variety of other activities that might or might not be associated with a corpus creation project. The biggest decision here is whether to run one or more evaluations based on the corpus, as TREC and other programs have. It is important to note that TREC-style evaluations incorporate several different activities: definition of a standard task, time-delayed release of test materials, manual annotation of data, consistent computation of effectiveness measures, and a conference where users of the corpus present results. The Working Group will consider whether none, all, or some subset of these activities would be useful to associate with the corpus.²

Other activities we will consider include creation of a website with various resources on authorship attribution, development of software and documentation to aid others in building corpora, generation of baseline results on various authorship attribution tasks, and various outreach activities.

4.6 Legal Issues

Efforts to produce corpora must contend with intellectual property law, particularly copyright law. In choosing texts to include, the ability to negotiate licensing terms with copyright owners, and the impact of those terms on researchers using the data, are a major factor. Possible tax benefits to donors are another legal issue of importance.

Some material of interest was originally distributed in a fashion that would have allowed a researcher (or anyone) to access it without cost or license (e.g. web pages without robot exclusions, blogs, public mailing lists). Some previous corpus building efforts have used such material without licensing it [11] while others have policies³ that forbid this.

²The eventual funders of any corpus effort would be a major influence on this choice. In the U.S., DoD and intelligence agency funders have had a particular interest in organized evaluations.

³<http://www ldc.upenn.edu/Providing/>

If message traffic is used, then privacy laws may be relevant as well, again potentially even for broadcast material. A corpus intended for international use must take into privacy regulations in, for instance, the European Union.

We will give close attention to these issues, and to the novel aspects that arise for authorship attribution in particular. For instance, if demographic information on authors is desired, then the cooperation of those authors may be critical. Such cooperation could also lead to the availability of additional unpublished texts from authors that could be of great value.

4.7 Resource Management

Creating a corpus is a substantial effort involving many people, lengthy and unpredictable negotiations with content providers, problems with data formats, conflicts over annotation and metadata choice, and a variety of other issues. Members of the Working Group will draw on their own expertise, published literature on corpus building and revision efforts, and interviews with managers from past and ongoing corpora projects to help us develop appropriate strategies for project management, recruiting and training of personnel, and technical infrastructure.

A major consideration will be whether to build the project from scratch, or work with an existing organization with expertise in corpus creation, such as the Linguistic Data Consortium. There are arguments for both approaches, and influence from funders is likely to be a major consideration as well.

5 Workplan and Project Management

5.1 Workplan

We have subdivided our proposed activities for the proposed one-year project into four quarters, listing the activities to be performed and goals to be reached in each.

Quarter One. The first quarter will focus on information gathering for corpus construction.

Working with members of our Working Group, we will conduct a detailed survey of the author identification literature. We will examine existing corpora such as the various literary sources, the Enron e-mail data, Reuters RCV1 stories with identified journalists, and the recently acquired DIMACS newsgroup dataset. Via a mailing list, we will initiate a discussion with the Working Group members about the key requirements for an author identification corpus and associated materials. We will also develop a list of candidate data sources.

Quarter Two. The second quarter will focus on the author identification experimentation process and address the issues we discussed in Section 3. In particular the Working Group will develop design standards for author attribution test corpora, develop a catalog of author identification features along with their advantages and disadvantages, discuss guidelines for appropriate use of algorithms and statistical methods, and prescribe desiderata for appropriate controls in author attribution experimentation. We will also examine the legal issues associated with the various data sources under consideration.

Quarter Three. In the third quarter we will prepare for the workshop. We will develop a detailed plan for corpus construction along with standards for author identification experimentation

and evaluation. We will also look at the processes and costs associated with data cleaning. This quarter will culminate in a draft design document ready for discussion by the Workshop participants.

Quarter Four. We will hold a two-day workshop near the start of the fourth quarter, including the Working Group members and interested parties by them. We envisage providing financial support to around 20 participants but the total number of participants may be closer to 30. Following the workshop, we will refine the existing design documents and summarize the workshop proceedings in a comprehensive report.

At the end of year we will publish our conclusions as a technical report. We will also submit a version to a high-impact journal. This report will serve as the core of one or more proposals to different funding sources for the development of high impact community resources for author identification.

5.2 Project management

Shlomo Argamon, the PI, will coordinate the overall planning effort, via frequent emails and phone meetings with the coPI, David Madigan, and the consultant, David Lewis. An email list including project members and the Working Group will be set up, to enable continuous feedback on the planning process. Two project meetings of all three principals will be held, one at IIT, and one at Rutgers, in addition to the DIMACS-hosted workshop which will be held with the Working Group towards the end of the project year. Also, Argamon and Lewis will also meet on approximately a monthly basis at IIT; Lewis and Madigan have several ongoing collaborations and meet regularly.

A research assistant at IIT will help with gathering and collating information about corpora and other resources, as well as working on any small-scale feasibility studies (for example, of software resources or annotation schemes) required in the course of the planning activity.

5.3 Working group

The Working Group will play a key role in the proposed project, providing ideas and feedback, as well as, in some cases, resources that may be useable for those we intend to design. The following individuals have enthusiastically agreed to participate:

Dr. David Banks is a Professor of the Practice of Statistics at Duke University. Previously he was on the faculty at Carnegie Mellon University. He has also worked for the FDA, NIST, and the US Department of Transportation. He has broad research interests in Bayesian statistics and multivariate analysis. He was co-editor of the recent *Chance* issue on author identification and has published a number of articles on information retrieval and text mining.

Dr. W. Bruce Croft is a Distinguished Professor in the Department of Computer Science at the University of Massachusetts, Amherst. He is currently Director of the NSF Collaborative Research Center for Intelligent Information Retrieval (CIIR), as well as Chair of the Department of Computer Science. He has published more than 120 articles in information retrieval and related areas, has consulted for a wide range of companies and government agencies, and has developed technologies used in a number of operational information retrieval systems. Dr. Croft was a member of the National Research Council Computer Science and Telecommunications Board, 2000-2003, and was Editor-in-Chief of *ACM Transactions on Information Systems*, 1995-2002. He was elected

a Fellow of ACM in 1997, received the Research Award from the American Society for Information Science and Technology in 2000, and received the Gerard Salton Award from the ACM Special Interest Group in Information Retrieval (SIGIR) in 2003.

Dr. Haym Hirsh received his PhD in Computer Science from Stanford University in 1989. He is Professor and Department Chair of the Computer Science Department at Rutgers University. He has published more than 100 papers in artificial intelligence, machine learning, information retrieval, data mining, and related areas. Hirsh is a member of the Executive Council of the American Association for Artificial Intelligence, the Board of Directors of Institute for the Study of Learning and Expertise, and the editorial board of various journals.

Dr. David Holmes is Professor of Mathematics and Statistics at the College of New Jersey and a prominent stylometrist. He has been involved in many disputed authorship identifications and has a long publication record in this area. His most recent work has involved the civil war-era Pickett letters.

Dr. Moshe Koppel is Associate Professor of Computer Science at Bar-Ilan University. He has held visiting faculty positions at the Institute for Advanced Study in Princeton, CUNY and Cornell University. His main area of research concerns machine learning, especially text categorization. In particular, he has published on demographic profiling of text authors and on attribution of anonymous and pseudonymous texts. Dr. Koppel's research has been presented in leading journals and conferences and has been featured in Nature News, New Scientist, The New York Times and many other publications. In addition to his academic work, Dr. Koppel consults for leading hi-tech firms and has published two books on Talmudic Law.

Dr. Gerald McMEnamin received his doctorate in Spanish linguistics from El Colegio de Mexico in 1978. He is currently Professor of Linguistics and Director of the Forensic Linguistics Institute at California State University, Fresno. He has taught several courses and training seminars on linguistic stylistics, and has worked on more than 250 civil and criminal cases of questioned authorship. Dr. McMEnamin is the author of a number of publications in forensic linguistics, including the book *Forensic Stylistics* (1993). He recently edited the book *Forensic Linguistics: Advances in Forensic Stylistics*, and authored 12 of its 15 chapters.

Dr. Joseph Rudman has been working on non-traditional authorship attribution since the mid 1970s. He has published widely on the topic and lectured throughout the United States and in seven foreign countries on the subject of non-traditional attribution. Dr. Rudman regularly referees papers on stylistics and non-traditional authorship studies for Literary and Linguistic Computing, Computers and the Humanities, Science, and other journals. He has acted as an Assessor for the Australian Research Council, and is a member of many professional societies that both act as gatekeepers and publish attribution studies, such as The Association for Computers and Humanities, The Association for Literary and Linguistic Computing, The Classification Society of North America, and the International Association of Forensic Linguists.

Dr. Ian Soboroff is a Computer Scientist with the Information Access Division of the National Institute of Standards and Technology. Each year, NIST hosts the Text REtrieval Conference (TREC), which is the world's largest coordinated research effort in information retrieval. Dr. Soboroff has organized TREC evaluations in web search, text filtering, novelty detection, and searching large-scale collections. He has a number of research publications in text filtering, collaborative filtering, information retrieval evaluation, authorship attribution, and intelligent software agents. He received his Ph.D. in Computer Science from the University of Maryland, Baltimore County, in 2001.

6 Significance and Broader Impacts

Authorship attribution has deep roots in the humanities, addressing sometimes long-standing disputed authorship issues. More recently, new applications focusing on intelligence and security have emerged and propelled an old field to new prominence and importance. Researchers in author identification currently face significant challenges in terms of test corpora to evaluate novel methodology as well as evaluation methodologies. Our proposed project, if it proceeds beyond the planning phase we propose here, could have a profound unifying effect on the now-dispersed author identification research community.

The possible broader impacts of this work are legion. A well developed and rigorous methodology for assessing the reliability of authorship attribution claims will have a significant impact in the criminal justice system. Homeland security could come to rely heavily on well-understood and validated author identification methodologies to find and track known terrorists. With development of a standard test corpus and associated experimentation standards, research progress is likely to be highly accelerated.