Authorship Attribution and Optical Character Recognition Errors

Patrick Juola[+][*] — John I. Noecker Jr[*] — Michael V. Ryan[*]

[+] *Evaluating Variations in Language Laboratory*
Duquesne University, Pittsburgh, Pennsylvania USA
[*] *Juola & Associates, Pittsburgh, Pennsylvania USA*

juola@mathcs.duq.edu, pjuola@juolaassociates.com
jnoecker@juolaassociates.com, mryan@juolaassociates.com

ABSTRACT. *Stylometric authorship attribution is a fundamental problem. The basic idea behind the research is that one can determine the authorship of a document on the basis of cognitive and linguistic quirks that uniquely identify a person. In many cases, however, noise in the original documents can make this analysis more difficult and less reliable. We investigate the errors introduced by a typical optical character recognition (OCR) process. Using simulated (random) errors in a standard benchmark corpus, we test to see how sensitive the authorship attribution process is to character mis-recognition. Our results indicate that, while accuracy decreases measurably with noise, the decrease is not substantial.*

RÉSUMÉ. *Le problème de l'attribution stylométrique d'auteur est un problème fondamental. L'idée fondamentale derrière cette recherche est que l'on peut déterminer la paternité d'un document sur la base d'un ensemble de trait cognitifs et linguistiques qui permettent d'identifier de manière unique le style d'écriture d'une personne. Dans de nombreux cas, cependant, le bruit présent dans les documents originaux peut rendre cette analyse plus difficile et moins fiable. Nous étudions les erreurs introduites par un processus typique de reconnaissance optique de caractères (OCR). En utilisant des erreurs simulées (aléatoirement) dans un corpus de référence standard, nous évaluons la sensibilité au bruit du processus d'attribution d'auteur. Nos résultats indiquent que, bien que la précision diminue avec un niveau de bruit, cette baisse n'est pas substantielle.*

KEYWORDS: *authorship, stylometry, optical character recognition, OCR errors*

MOTS-CLÉS : *paternité des documents, stylometrie, reconnaissante optique de caractèdes, erreurs d'OCR*

1. Introduction

Authorship attribution, the problem of inferring authorship from the writing style of a document, is an important problem not just in computer science, but also in education, journalism, history, and law. Given the stakes involved in some legal cases (in one case discussed here, several hundred million dollars), it is critical that the technology be as accurate as possible. But what of the language samples themselves? Can authorship attribution be performed accurately on inaccurate texts?

This paper discusses a series of experiments on authorship attribution in the presence of simulated errors typical of those produced when physical documents are scanned and processed via optical character recognition. This process, which is quite common in the production of large volumes of historical or contemporary texts, is known to produce high error rates, in some cases nearly 30% of any given document. We perform an analysis using a standard benchmark corpus and several different analysis methods to determine whether these errors substantially decrease accuracy and under what conditions.

The remainder of the paper is structured as follows. After a substantial background (section 2), section 3 describes the experimental corpus (the Ad-hoc Authorship Attribution Competition corpus), the instrument (the Java Graphical Authorship Attribution Program) and its settings for this set of experiments, and then the details of the experiments themselves. Results are largely in the form of tables with a discussion provided in section 4; this discussion includes a followup experiment suggested by the referees. Finally section 5 provides our conclusions.

2. Background

We summarize here the basic background in authorship attribution, both tradition and non-, as well as the relationship of OCR technology (and the errors it introduces) into authorship attribution.

2.1. *Authorship attribution*

Authorship attribution is an increasingly important and popular problem in NLP research. The problem itself can be stated very simply: *Given a document, who wrote it?* Koppel *et al.* (2012) have suggested an alternate variation of what they call the fundamental problem of authorship attribution : *Given a pair of documents, are they by the same person?* Either formulation requires the identification of the characteristic writing style of a person and a determination of whether or not a specific questioned document fits that style.

In theory, every author is capable of making almost an infinite number of choices from the available set of language, but many of these choices are habituated and form part of an individual style, or what van Halteren *et al.* (2005) call the “stylome”.

Other researchers have preferred the metaphor of an “authorial fingerprint”. In either case, the idea is that one’s language is clearly different from any other and that this difference can be detected and used to identify the writer of a given document.

2.1.1. *Traditional authorship attribution*

Traditional linguists, historians, and literature scholars have of course been studying questions of authorship for centuries; as a simple example, the authorship of the Biblical *Second Epistle to the Thessalonians* has been in dispute since at least 1798 (Best, 1972), and the authorship of the *Book of Revelation* has been disputed since the 2nd century CE (Cross, 1997). More recently, the *Federalist Papers* were a set of American political essays written in 1787–8, written to argue for the adoption of the newly proposed US Constitution, and published pseudonymously in newspapers under the name “Publius”; as recently as 1937 (Earle, 1937), scholars were still disputing who had written them. More recently still, in 2010, Paul Ceglia sued Mark Zuckerberg claiming ownership of a major stake in Facebook, citing email (from Zuckerberg to Ceglia) from 2004 that (Ceglia claimed) demonstrated Ceglia’s early investment in the then-startup that gave him control of up to 80% of the company. A noted forensic linguist (McMenamin, 2011) was called in to give evidence and concluded that it was “probable that Mr. Zuckerberg was not the author of the QUESTIONED [caps in original] writings”.

As McMenamin noted in his report, “At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer’s ‘choice’ of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer’s own unique set of habitual linguistic choices. Identification and analysis of a writer’s choices, i.e., of his or her style markers, constitute stylistic analysis, which is well established as a generally accepted and peer-reviewed method of author identification in both literary and forensic contexts.” He identified in particular eleven different style-markers, including the use of apostrophes, suspension points (aka ellipsis markers), the spelling of “backend” as a single word (as opposed to “back-end” or “back end”), the use of the single word “cannot,” capitalization of the word “Internet,” the use of “Sorry” as a sentence-opener, and the presence or absence of run-on sentences. As the questioned email differed significantly from other email of known Zuckerberg authorship, he concluded that the authors were “probably” different.

There are several epistemological and methodological issues with this approach, of which analysis bias is one of the more problematic. It is possible, especially in the context of an expert report commissioned by the lawyers of one party in the dispute, for the expert to be aware of “which side his bread is buttered on” and (consciously or unconsciously) select features that support one side of the argument while neglecting other features that would support the other side. (Consider as a thought experiment: why apostrophes specifically and not other punctuation marks such as single vs. double quotations, use of semicolons, or the Oxford comma?) It’s even possible under US law for a law firm to commission five reports from five different experts and submit to the court only the ones supporting their client's interests.

Another issue is evaluating the reliability of the report; if little data is available, the odds of a spurious finding go up. In the case of the McMenamin report, he found two uses of "can not" in the questioned documents and six instances of "cannot" in the known Zuckerberg documents, a total of eight word types upon which to base judgement. Is this really enough instances for a reliable judgment? What if McMenamin had missed one instance of "can not," possibly from a typographical error that made it "cna not" or from simple fatigue? McMenamin made no attempt in his report either to formalize the statistical models underlying his analysis, to calculate p-values associated with his finding, or to perform error analysis to assess the reliability and robustness of his methods.

2.1.2. *Nontraditional authorship attribution*

Some of these difficulties can be addressed by the use of more formal statistics, in what has been called (Rudman, 2005) "nontraditional" authorship attribution. Although the underlying concepts date back to the 19th century (de Morgan, 1851/1882; Mendenhall, 1887), this approach has received serious scholarly attention only in the past fifty years. Mosteller and Wallace (1964) re-analyzed the *Federalist Papers*, paying particular attention to the distribution of hundreds of specific words found in undisputed writings by the various candidate authors. They found, for example, that Alexander Hamilton never used the word "whilst", that James Madison never used the word "while", and that the questioned writings, the ones of less certain authorship, also never used the word "while". Similarly, they found that Hamilton never used the word "by" more frequently than 13 words per thousand, while Madison never used it less than 5 per thousand and often as much as 19 per thousand. By itself, these findings might suffice for a convincing forensic analysis, but Mosteller and Wallace went further, applying formal statistics to the data, including an early example of Bayesian analysis, and derived formal probabilities for the likelihood that any given document had been written by Madison as opposed to Hamilton. This classic study has become a model for many studies, and the *Federalist Papers* have become the classic touchstone for testing a proposed new method of attribution (Martindale and McKenzie, 1995; Rockeach *et al.*, 1970; Holmes and Forsyth, 1995; Rudman, 2005).

Since this study, work in authorship attribution has exploded (Juola, 2006a; Koppel *et al.*, 2009; Stamatatos, 2009; Jockers and Witten, 2010). A particularly good example is Binongo's study of the *Oz* books (Binongo, 2003). The backstory is fairly simple: the series was started with L. Frank Baum's publication of *The Wonderful Wizard of Oz* and continued until his death in 1919. After his death, the publishers asked Ruth Plumly Thompson to finish "notes and a fragmentary draft" of what would become *The Royal Book of Oz*, the 15th in the series, and then Thompson herself continued the series until 1939, writing nearly twenty more books. The underlying question is the degree to which this "fragmentary draft" influenced Thompson's writing; indeed, scholars have no evidence that the draft ever existed. Binongo collected frequency statistics on the fifty most frequent function words across the undisputed samples and analyzed them using principal component analysis (PCA). Reducing these fifty variables down to their first two principal components produced an easily graphable distribution that showed clear visual separation between the two authors. When the *Royal Book* was plotted on the same scale, it was shown clearly to lie on Thompson's side of the graph, confirming that "from a statistical standpoint, [the *Royal Book*] is more likely to have been written in Thompson's hand".

2.1.3. *An ur-study in authorship*

Both the Mosteller/Wallace and Binongo studies share many characteristics common to modern authorship attribution studies. A prototypical authorship study includes most or all of the following steps:

– collect a training set of undisputed documents covering the set of candidate authors (for example, Baum and Thompson) and as similar as practical to the questioned documents in dating, genre, theme, and so forth;

– extract features from the various documents that show little intra-author variation, but much inter-author variation;

– apply standard classification technologies to determine which author is more likely to have written the various questioned documents.

Of course, the details of this process will vary from study to study. In the Mosteller/Wallace analysis, the features selected were individual words hand-chosen from a lengthy study of the undisputed documents. Binongo's study used a more objective set of fifty function words chosen purely on the basis of a frequency cutoff. De Morgan suggested (de Morgan, 1851/1882) that (average) word lengths could be used to distinguish authorship. In a recent presentation, Stamatatos has argued (Stamatatos, 2012) for the use of character n-grams (sets of n consecutive characters drawn from the document) instead of words. Chaski (Chaski, 2005) has argued for the use of more sophisticated linguistic features based on the theory of "markedness". Indeed, Rudman wrote more than a decade ago (Rudman, 1997) that more than 1,000 features had been proposed to solve this problem.

Similarly, the specific type of classifier has been the subject of much research. Mosteller/Wallace used naive Bayesian analysis; Binongo used PCA; Tweedie, Singh, and Holmes used artificial neural networks (Tweedie *et al.*, 1996), and many researchers have used support vector machines. Noecker and Juola noted (Noecker Jr. and Juola, 2009) that simple nearest-neighbor algorithms could achieve comparable results to support vector machines at a fraction of the computational effort, and specifically recommended dot-product or normalized cosine distance; other work (Grant, 2012) recommends the use of Jaccard or intersection distance in this framework. One of the primary findings, in fact, is that there are many methods that tend to work (Juola, 2012) and that there does not appear to be a "magic bullet" that produces vastly improved performance in controlled testing.

2.2. *Optical character recognition (OCR)*

One area of authorship attribution that has not received much attention is the relationship between text quality and authorship attribution. We use the word “quality” here in the sense of “authenticity”; if an author is a poor speller, then their characteristic spelling errors may be a reliable indicator of authorship (Wellman, 1936), and similarly an author prone to grammatical errors (or using a non-standard dialect or idiolect) may produce notably unusual texts (Chaski, 2005). In either case, if an editor proofreads and corrects the text, even if “improving” it, this distinctiveness will be lost and the attribution process made more difficult. Rudman, in particular, has written about this, going so far as to issue the directive (Rudman, n.d.) “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.” We have cited this elsewhere (Juola and Baayen, 2005) as Rudman’s Law (Rudman, 2003): “the closest text to the holograph should be found and used.”

2.2.1. *The need for scanning and OCR*

While this argument is great in theory, in practice one must often resort to texts far removed from the holograph. Both the Binongo and Mosteller/Wallace experiments showed this; in both cases, the holographs were long gone, and the published versions were accessible only in physical print instead of computer readable form. Mosteller and Wallace, in the largely pre-computer days, were forced to resort to retyping and hand-counting; Binongo was able to download some needed works from Project Gutenberg, works that themselves had been retyped or machine scanned, but had to purchase and scan others using character-recognition software. In this analysis, “proofreading both the downloaded and scanned texts was... the most time-consuming... part of the study”. For large-scale experiments, proofreading OCR’ed documents costs a substantial amount of money, takes a huge amount of time, and is monotonous and prone to error itself.

In some ways, this problem is likely to get worse instead of better as time and technology progress. Many documents submitted to courts, and thus subject to forensic authorship analysis when appropriate, are submitted as physical artifacts but only available for analysis as scanned images; many other documents are submitted purely as electronic scans. Cases of e-discovery can result in millions of documents to analyze, requiring thousands of person-hours to review and proofread. Google Books has made available the texts of millions of books that can serve as baselines for inquiry, but again these are largely the product of scanning and using character recognition software.

2.2.2. *OCR errors*

OCR software inevitably introduces errors. Even when the documents are clearly printed, errors still happen, and with historical documents, print quality will itself become an issue. Especially with large-scale digitization projects forming the base for many linguistic and computational studies, how much the OCR process will impact the final result is an important question, and one that has not received enough attention.

Holley provides a good description (Holley, 2009) of the impact of OCR errors as well as a good overview of the state-of-the-art. For example, the letter "h" is often mis-read as the letter pair "li". Holley found that the mistranslated word "tlie" occurred in one of 8 articles in the Australian Newspaper Digitization Project database. More significant mistranslations occurred on likely search terms — she found 30,000 articles with "Sydnoy" instead of "Sydney". As of 2009, her project had decided "that 'acceptable' OCR was still not good enough", and was looking for long-term improvements. But for an authorship study, exactly how bad is "acceptable" and how does that impact the overall accuracy of the study?

To answer the first question, there are several studies of newspaper digitization available. Powell and Paynter (2009) found in 2009 that she could average an accuracy of 97.53% on bitonal images and 94.10% on 400 dpi (dots per inch) greyscale. (Greyscale at 300 dpi, by contrast, scored only 83.88%). Klijn (2008) cites rates varying from 68% to 99.8%. Holley's findings were similar: 71% to 98.02% accuracy on a per-character basis. Her summary statistics are as follows:

– Good OCR accuracy: 98–99% accurate;

– Average OCR accuracy: 90-98% accurate;

– Poor OCR accuracy: Below 90% accurate.

From a practical standpoint, then, the question becomes whether the errors introduced in this process are likely to have a substantial effect on authorship classification. Authorship attribution itself carries its own baseline error rate, but if the error rate for documents with 2% or fewer errors is the same as the accuracy rate for perfect documents, then authorship attribution can be performed without a need for expensive proofreading if the OCR is "good"; if adding 10% errors does not substantially lower the overall judgment accuracy, proofreading is unnecessary even with average OCR.

2.3. ***Why it matters***

Since statistical authorship attribution hinges on detectable difference in data, noisy data will create problems. Consider the word "the", one of the fifty words studied by Binongo and also a candidate word for Mosteller/Wallace. (Madison used "the" between 8–13% of the time, Hamilton 7–11%. This word, then, can help distinguish but is not as informative as the thirty words eventually chosen.) If one assumes that 10% of the instances of "the" are thus corrupted, the absolute frequency of the word will be reduced, as will the differences in average frequency between the authors. Furthermore, with smaller samples, the statistical power of an analysis will be reduced. The combination of larger error bars and smaller differences will make it harder to analyze correctly. It is, however, not clear by how much.

Previous work (Noecker *et al.*, 2009) suggests that small error rates can be dealt with. In a sample (described later) of 98 unknown documents, introducing small error rates (less than 5%) reduced performance by less than two additional incorrect answers above baseline, while larger error rates (up to 20%) still only introduced five more incorrect answers. These preliminary results suggest a high degree of robustness.

However, these results were obtained using one specific feature set (words), one specific form of classification (nearest-neighbor), and one particular distance (normalized cosine distance). It is possible and indeed even likely that some forms of analysis are substantially more robust than others. In particular, we note that in a word-based analysis, any error creates an entirely new word that is not comparable to the original. Since long words create more opportunities for error, they are likely both to create more errors and more new words. Analyzing by character n-grams is likely to preserve similarity among erroneous long words and might be expected to be more robust to OCR errors.

Similarly, analysis of vocabulary overlap or richness would be dominated by new words created via error. Grant (2012) argues in favor of using Jaccard or intersection distance (defined below) for authorship studies, but distances may be artificially inflated by the creation of new words. Noecker Jr. and Juola (2009) argue for cosine similarity distance, which may perform better. An empirical comparison is in order.

We therefore describe in the following section an enlarged study to replicate and extend these findings, with an eye to determining to what extent various levels of OCR errors affect accuracy of statistical authorship attribution, whether proofreading can reasonably be left out of a set of “best practices”, and to whether this varies with the type of analysis done.

3. Materials and methods

3.1. *The Ad-hoc Authorship Attribution Competition corpus*

Key to any comparative study is a normative baseline and benchmark. One of the weaknesses, historically, of authorship studies is the lack of an accepted benchmark corpus. Instead, most researchers relied on a corpus of personal interest (a Defoe scholar, for example would focus on disputed works of Defoe and close contemporaries) and validate a proposed method against the baseline of random guessing. This is one reason that more than 1,000 different feature sets have been proposed, most of which work better than random chance, but with little comparison between them. The Ad-hoc Authorship Attribution Competition (Juola, 2004; Juola, 2006a; Juola, 2006b) represented the first TREC-style comparative evaluation on a set of documents specifically designed for this task.

The AAAC corpus is divided into thirteen problems, designed to cover a wide (but ad-hoc) collection of languages, genres, themes, styles, and document sizes. Languages incorporated include English, Middle English, French, Serbian-Slavonic, Latin, Dutch; genres include essays, personal letters, novels, plays, and transcribed speech. The corpus contains 264 documents of known authorship, and 98 "unknown" documents whose authorship was withheld at the time of competition and which act as test documents. Details of the corpus are listed below:

– *Problem A* (English) Fixed-topic essays written by thirteen US university students;

– *Problem B* (English) Free-topic essays written by thirteen 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 (père);

– *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).

– *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.

3.2. *Java Graphical Authorship Attribution Program*

In addition to developing the AAAC corpus, we have also developed a modular, Java-based program (Juola *et al.*, 2006; Juola, 2006a; Juola *et al.*, 2009) to permit and encourage comparison between methods as well as cross-fertilization of techniques. This program, the Java Graphical Authorship Attribution Program (JGAAP) is freely distributed at `www.jgaap.com`. As an example, JGAAP supports a replication of the Mosteller/Wallace experiment, or a variation of the Mosteller/Wallace experiment using instead the fifty most common words from Binongo, a change that would help analyze whether the elaborate hand-tuning of Mosteller and Wallace is in general a good practice.

JGAAP approaches the problem via a modular pipelined architecture. The pipeline stages include:

– Document selection — JGAAP supports a wide variety of formats but ultimately converts them to raw UTF text for analysis;

– Canonicization — Documents are converted to a "canonical" form, and uninformative variations can be eliminated at the option of the analyst. For example, document headlines and page numbers, which are often the product of the editor and not the author, can in theory be removed. This also supports a certain amount of normalization — for example, punctuation (often also a product of an editor) may be stripped out or case variation neutralized to make sure that "The" and "the" are treated as the same word;

– Determination of the event or feature set — The input stream is partitioned into individual non-overlapping "events". (This term is used instead of "feature" to emphasize that, in text documents, there is an implicit ordering to the features that may or may not be of interest.) At the same time, uninformative events can be eliminated from the event stream. "The fifty most common words" would be an example of such "events," as would "all part-of-speech 3-grams" or "all character 4-grams";

– Event culling — the overall set of events can be culled, for example, to make sure that only frequent, only infrequent, only widely distributed, or only events with high information gain are included;

– 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. JGAAP supports several different distance-based nearest-neighbor analyses (with more than 20 separate distances) as well as other classifiers such as LDA, SVM, and the WEKA suite of classifiers. The results of this inference determine the results (and confidence) in the final report.

Each of these phases is implemented as a generic Java class which can be instantiated by any of several different classes. For example, the event set (EventSet) factory class is defined as EventDriver, which takes in a document and returns the set of events from that document. Specific EventDrivers include words, word n-grams, word lengths, parts of speech, characters, character n-grams, part of speech tags, word stems, and so forth. Specific AnalysisDrivers include support vector machines, linear discriminant analysis, and so forth. It is not difficult to add a new module, and module selection is performed at run-time through an automatic GUI that searches the codebase for appropriate class files and adds them as options to the various menus.

With combinatorics, JGAAP supports more than one million different analyses and can thus be used for large-scale experiments in search of best practices (Juola and Vescovi, 2010; Juola and Vescovi, 2011; Vescovi, 2011) or to create individual elements for ensemble classification such as mixture-of-experts (Juola, 2008; Juola, 2011). We use it here as a testbed to compare different approaches to authorship attribution in the presence of errors as detailed below.

3.3. *JGAAP settings*

Because of the modular nature of JGAAP, there are a nearly limitless set of analysis variations that can be used inside that framework. We used a fairly standard set of parameters that have generally performed well (as will be seen later in the comparison to prior work). These parameters are as follows:

– Canonicizers: We used the "Normalize Whitespace" and "Unify Case" canonicizers. The first converts all instances of adjacent whitespace characters in to a single space character, thus neutralizing variant spacings, kerning, paragraph indentation, and so forth; the second converts all characters to lower case, neutralizing case variations;

– Event Sets: All analysis was performed using "Character Ngrams," groups of adjacent characters. Stamatatos (2012) provides a detailed recommendation for this particular Event Set;

– Event Culling: No event culling was performed;

– Analysis Method: We used an author-centroid-based nearest neighbor driver, with a variety of distances as detailed below. This method is based on the idea of a nearest neighbor in some abstract stylistic space (as defined by the distance), but calculates an author-based "centroid" as an average representation of the authorial style, removing a potential source of variation when two or more authors have similar style.

3.3.1. *Cosine distance*

The results described below incorporate author-centroid nearest-neighbor analysis using three different distance functions. Cosine or dot-product distance (Noecker Jr. and Juola, 2009) has been suggested as an efficient and well-performing substitute for more computationally complex methods such as support vector machines or linear discriminant analysis. As with the two other distance functions, cosine distance does not use the ordering of the events (hence events are treated as true features), instead creating a histogram or vector of event frequencies. Each document is converted to a vector of frequencies $\mathbf{u} =< u_1, u_2, u_3, \ldots u_n >$ where u_i is the relative frequency of the ith event (which might be zero, if that event does not appear in that document). In this specific set of experiments, all documents/vectors associated with a single author are averaged to produce a single vector representing that author's "typical" or average style, as discussed above.

The distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ is computed as:

$$CosDis(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{\parallel \mathbf{u} \parallel \parallel \mathbf{v} \parallel} = \frac{\sum\limits_{i=1}^{n} u_i v_i}{\sqrt{\sum\limits_{i=1}^{n} {u_i}^2}\sqrt{\sum\limits_{i=1}^{n} {v_i}^2}} \qquad [1]$$

This of course returns a value between -1 and 1 depending upon the angle between the vectors; to convert this to a true distance, JGAAP uses the value 1-$CosDis(\mathbf{u},\mathbf{v})$; thus identical documents have distance 0 as required.

3.3.2. *Manhattan distance*

The second distance used in these experiments is Manhattan distance, more formally the Minkowski or Lebesque L_1 distance, and intuitively the city-block distance between two points on a grid. As before, documents are converted to frequency histograms/vectors; the distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ is computed as:

$$ManhattanDis(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} |u_i - v_i| \qquad [2]$$

3.3.3. *Jaccard or intersection distance*

The final distance used is intersection distance, more formally known as Jaccard distance, which measures essentially the degree of event/feature overlap. Unlike the previous distances, which rely on frequency histograms, intersection distance instead calculates the set of events associated with each document, and thus a single appearance of a word is treated identically as the appearance of a word occurring thousands of times. For two documents (sets) U and V, we calculate the intersection distance as:

$$IntersectionDis(U,V) = 1 - \frac{|U \cap V|}{|U \cup V|} \qquad [3]$$

Intuitively, if U and $V$are identical, then $U \cap V = U \cup V$ and the distance is 0 as required; if U and V are disjoint, $|U \cap V| = 0$ and the distance is maximized at 1.

3.4. ***Simulated OCR errors***

To simulate OCR errors, we created a canonicizer (AddErrors) that inserts substitution errors at random into a document. Specifically, it accepts a parameter defining the desired per-character error rate p and then for each character token, it independently with probability p decides to replace it (creating an effective error) or not. If the character is replaced, the replacement is again chosen randomly. In particular, whitespace is always transformed into whitespace, and printable characters are always transformed into printable characters, but no attempt is made to model specific OCR confusion matrices and replacements are always random and uniform within these broad categories. This of course is unrealistic for several reasons; not only are certain errors more likely (e.g. “n” becoming “r” instead of “q”), but it also ignores insertion or deletion errors (e.g. “h” becoming “li”).

3.5. *Experimental details and scoring*

The experiment was repeated at all lengths of character n-gram from 1 (individual characters) to 15, for each of 9 percentage error levels [0 (perfect accuracy), 1, 2, 3, 4, 5,10, 15, and 20%]. This creates one sample of “perfect” OCR, two of “good,” four of “average,” and two of “poor” as defined earlier by Holley’s criteria.

Following previous work, we report total percentage correct for each of the 13 AAAC problems, hence scores range in theory from 0 to 1,300 (100 points on thirteen separate problems). In this framework, scores in the 700s or 800s are generally considered “good”.

Selected results are presented in graphical format for brevity in the following pages.

4. Discussion

4.1. *Overall results*

The preliminary observations are simply confirmations of some previous results: first, authorship attribution is possible at substantially better than chance levels. Secondly, confirming the results of Noecker *et al.* (2009), cosine distance performs well under nearly all conditions (including at all error levels as will be discussed later), supporting the idea that it may be a “best practice” for authorship attribution in general.

Intersection distance, another candidate for best practice, strongly underperforms cosine distance and Manhattan distance at small values of n, reflecting the fact that most reasonable-sized documents (in alphabetic languages) are likely to use the entire character set (and come close to exhausting the set of commonly-used small n-grams up to a value of about 7). By n=7, intersection distance has closed most of the performance gap and will sometimes outperform cosine distance.

The performance of Manhattan distance is more problematic; while it initially performs well, performance drops off with increasing n-gram length. And as will be seen in the following subsection, it is much more sensitive to errors than the other two distance measures.

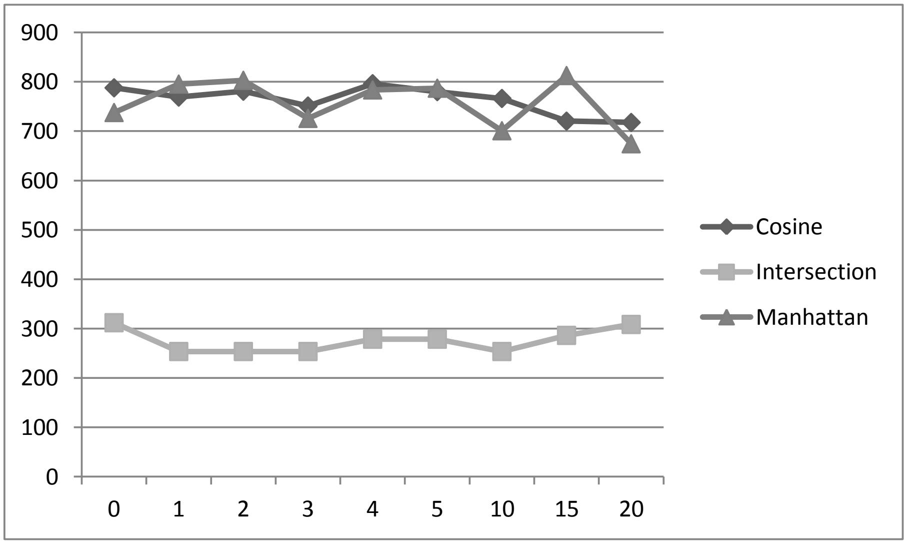

Figure 1. *AAAC score for character 1-grams at differing error levels*

4.2. *The effects of errors*

The graphs make it clear that "good" OCR, in particular, does not have a substantial effect on overall authorship accuracy. The exact data for character 4- and 5-grams (a typical length used in many experiments) is attached as table 1. From this table it can be seen that, while the overall trend may be down slightly, the trend is dominated by random variation (in particular, note that at least one instance of each distance actually scored *higher* at the 10% to 20% error range than did it at the 0% "perfect document" baseline. The performance loss, even at "poor" OCR error rates, is almost always less than 10% of baseline. Performance loss at 2% error rates tends to be at most 2% of AAAC score.

At the same time, it also appears that Manhattan distance tends to be much more sensitive to OCR errors. Even at short lengths, when Manhattan is performing well overall (see the table of 2-grams for an example), adding 10% or more error consistently produced a loss of more than 10% of baseline accuracy. At longer lengths (e.g.,

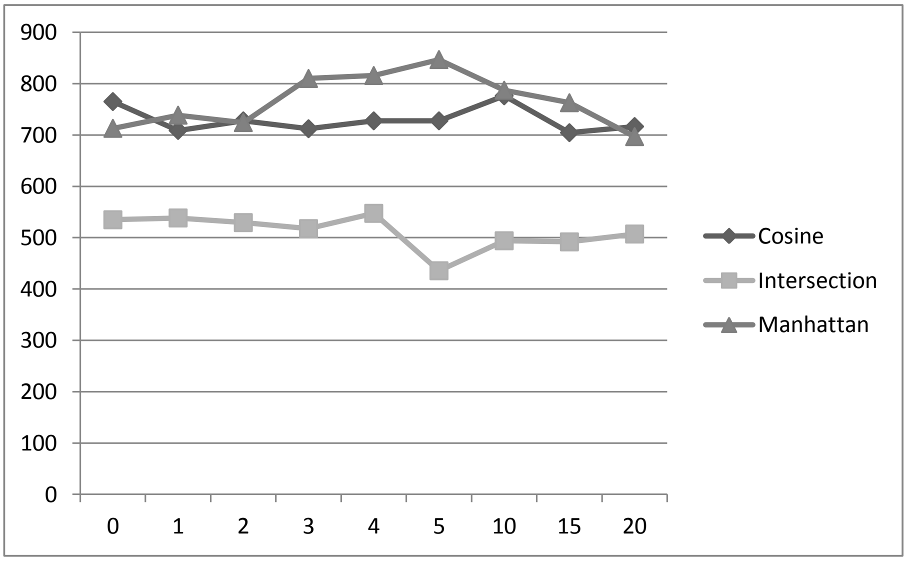

Figure 2. *AAAC score for character 2-grams at differing error levels*

Character 4-grams				Character 5-grams			
% Error	Cosine	Intersection	Manhattan	% Error	Cosine	Intersection	Manhattan
0	758.7607	641.2393	820.1923	0	686.6453	543.3761	757.4786
1	755.3419	504.3803	811.7521	1	711.0043	538.5684	735.8974
2	754.594	504.3803	775.641	2	674.7863	430.8761	694.9786
3	766.453	544.765	731.0897	3	675.5342	527.3504	676.0684
4	731.6239	533.547	820.0855	4	703.312	455.8761	690.7051
5	737.2863	529.3803	759.5085	5	663.5684	523.1838	754.594
10	717.735	537.7137	805.9829	10	680.8761	519.6581	804.594
15	689.3162	526.6026	719.1239	15	687.1795	548.1838	673.9316
20	673.2906	511.2179	691.8803	20	635.6838	540.4915	744.871

Table 1. *Exact results for character 4- and 5-grams (AAAC Score)*

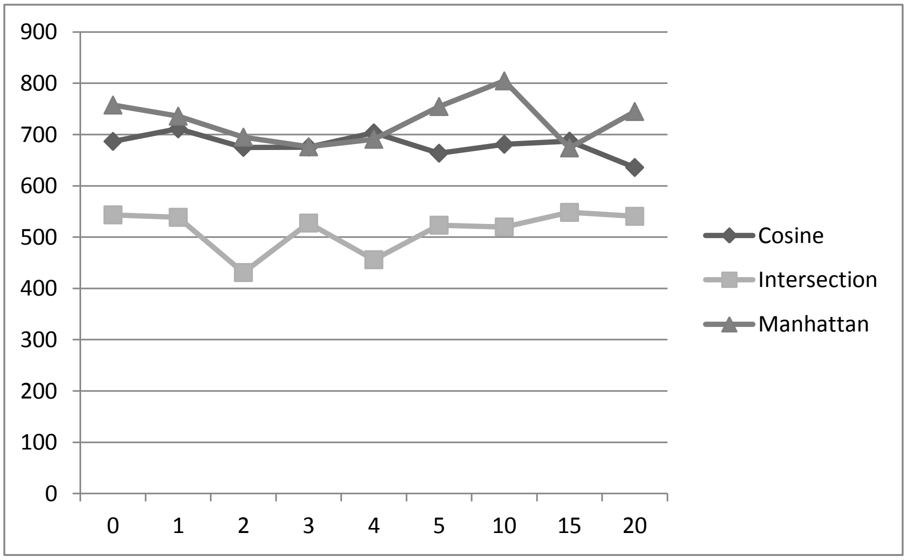

Figure 3. *AAAC score for character 5-grams at differing error levels*

n=10) poor quality OCR will halve an already low baseline performance. Even at n=6, poor OCR (more than 10% errors) produces this strong drop-off. This strongly suggests that where documents have been processed via OCR, Manhattan distance should be avoided not simply for performance reasons (Manhattan actually outperforms intersection at this length range) but for reasons of error sensitivity.

4.3. *Comparison to previous work*

A comparison to previous work (Noecker *et al.*, 2009) is also instructive. Noecker *et al.* found that “we probably don’t have to worry about small abnormalities in the text” because: “If we trust the current estimated OCR error rate (∼2% error), it is probably not necessary to proofread OCR’d documents thoroughly to perform meaningful authorship attribution on them. Most OCR errors will ‘come out in the wash’.”

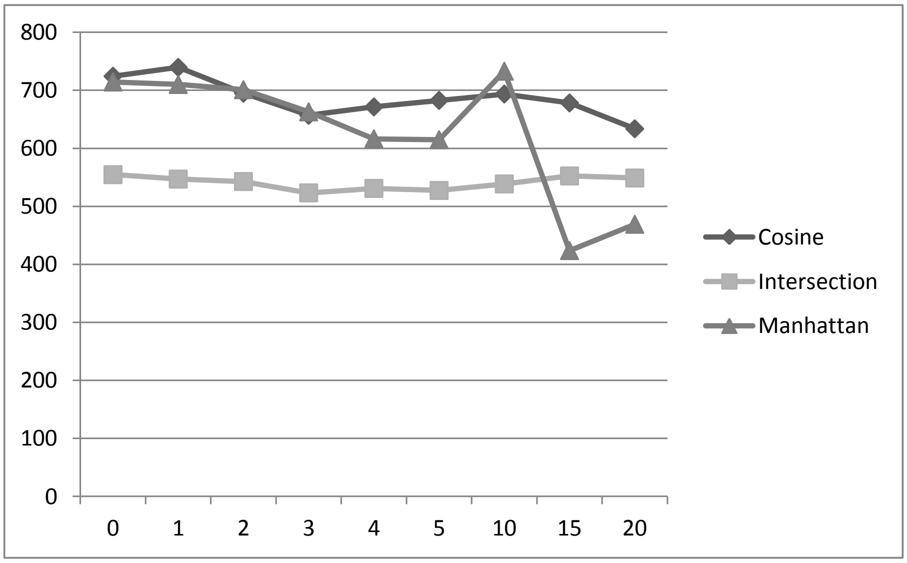

Figure 4. *AAAC score for character 6-grams at differing error levels*

These results were obtained by analyzing not character n-grams (as in the present study) but words, and using document-based instead of author-based nearest neighbor. We present comparative data in table 2. Aside from the clear superiority of character 4-grams to words, and the lower variance due to the use of repeated measures in Noecker *et al.* (2009), we are able largely to confirm their original results.

4.4. *Modeling limitations*

One issue with this work is its transfer to the real world. One of the more serious limitations on this work is the unrealism of the error transformations. In a realistic setting, some types of errors are much more likely than others; an “c”, for example, is more likely to be converted into an “c” (or vice versa) than into a “k”, especially with older documents where a worn “o” key is likely to have a break on the typeface. Similarly, our approach does not allow for insertion or deletion errors, such as converting

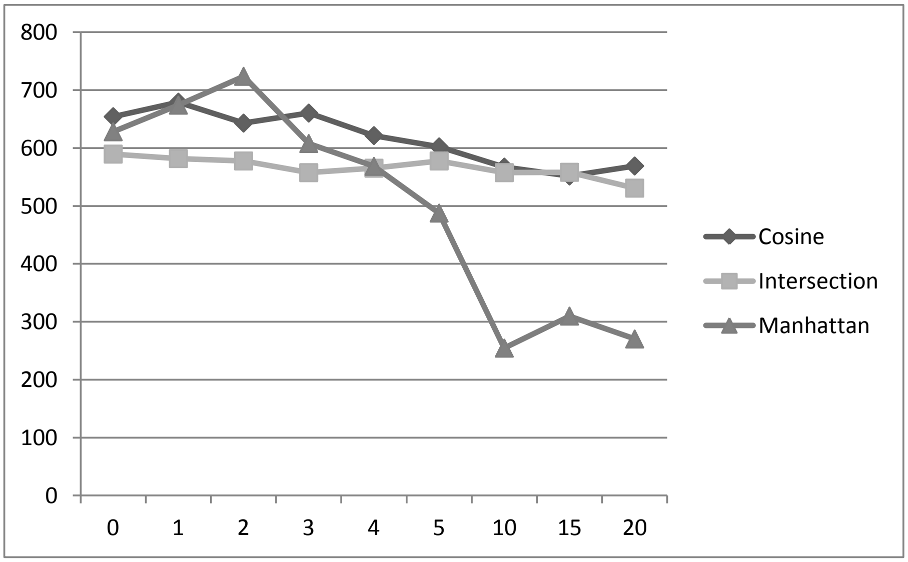

Figure 5. *AAAC score for character 7-grams at differing error levels*

% Error	4-grams	Words
0	758.7607	645.0
1	755.3419	639.9
2	754.594	634.8
3	766.453	632.9
4	731.6239	629.0
5	737.2863	624.9
10	717.735	597.7
15	689.3162	577.7
20	673.2906	577.7

Table 2. *Comparison of character 4-grams and words (AAAC Score) using cosine distance*

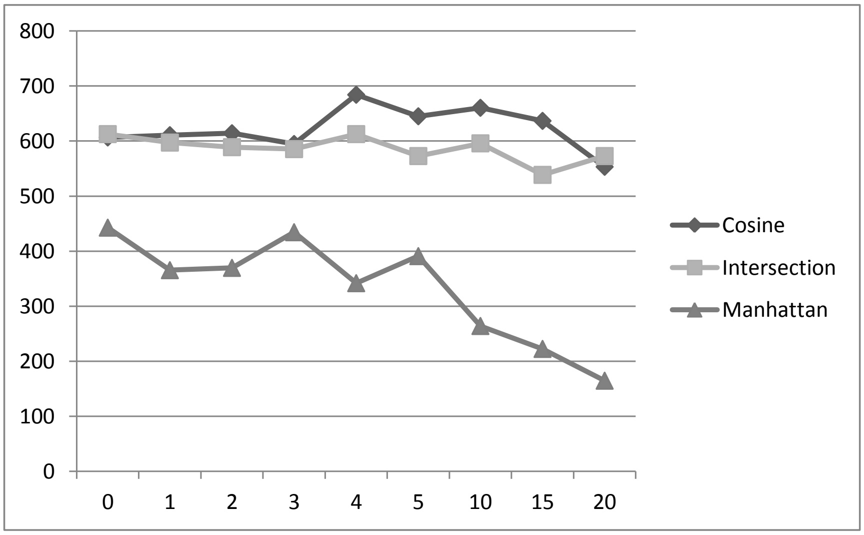

Figure 6. *AAAC score for character 8-grams at differing error levels*

an “h” to the two letters “li” as discussed above. We found it difficult to find published conversion matrices from which to develop a more sophisticated model.

More to the point, though, the probability of confusion is context-dependent; most modern OCR systems apply modeling not only at the level of the character but also at the level of the language (or at least word). If a specific glyph were ambiguous between “h” and “n”, for example, but that glyph appeared between the letters “t” and “e”, it is overwhelming more probable to be an “h” than an “n” (at least according to the statistics of standard English). On the other hand, that same glyph occurring after the letter sequence “the” and before whitespace is likely to be an “n”. Thus producing properly plausible OCR errors is challenging and in many cases would need to be engine-specific, as each program has its own model. Furthermore, in many cases the models themselves are language-specific and an engine tuned for English text might fail horribly on Spanish text precisely because the expectations would not fit.

Another approach would, of course, be to generate OCR errors manually. We have run preliminary experiments to this end using lossy image compression to create

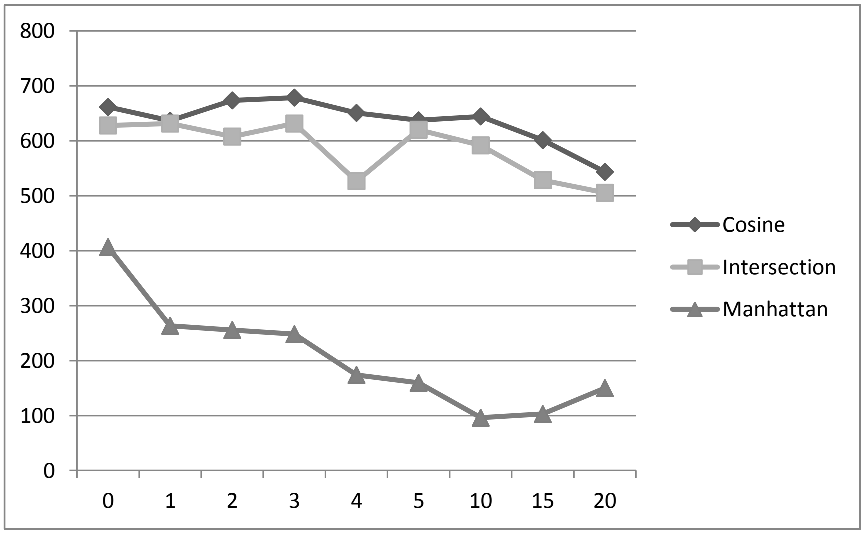

Figure 7. *AAAC score for character 9-grams at differing error levels*

errors. Specifically, the documents in the AAAC were converted (using imagemagick) to a set of (approximately 13,500) JPEG page images at various levels of quality from 100% (lossless compression) down to 10%. (Experiments at 3% and 1% compression did not produce usable images; the OCR software failed to recognize any text in these images.)

These documents were then assembled into multipage PDFs for each document and re-OCR'ed using commercial off-the-shelf software (ABBYY FineReader 11 Corporate Edition). As seen from the samples below, this produced relatively good text (1% to 3% errors) even at high compression levels (down to about quality-level 25%).

– Original: Today, people young and old, male and female, short and tall work. Employment in today's fast moving economy almost seems essential. Children start to work at a young age at little jobs like being a newspaper carrier and con\tinue to work into their sixties or later. Work has become an embedded idea that you need to do just like sleeping or eating. Money and the strive for wants and necessities is the

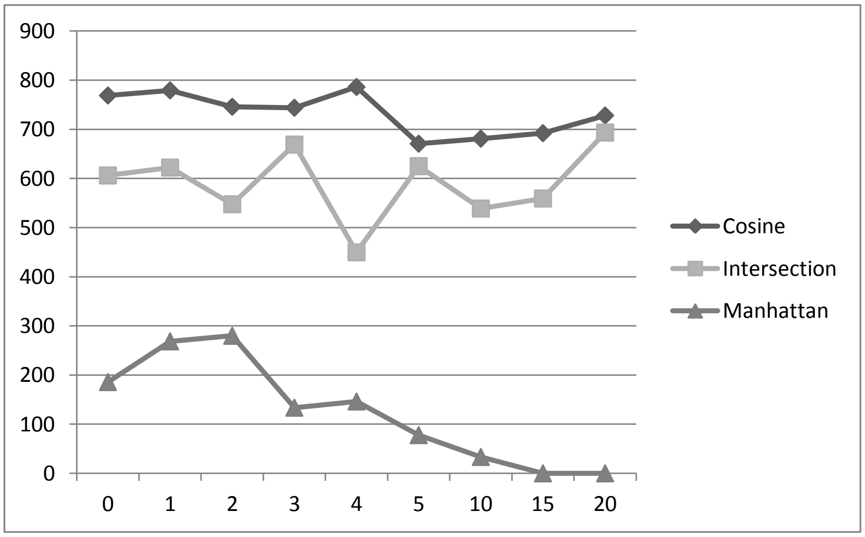

Figure 8. *AAAC score for character 12-grams at differing error levels*

key reason why people around the world work. In my family work is essential. My parents work not only as adults, but also as high school students. My family believes that you have to work for what you want. That meant for me to have a car and money to spend, I had to go \to work and earn money just like my parents did when they were my age. I really enjoy my job working at Rita's Italian Ice and I feel that it is important that you enjoy what you are doing and appreciate the rewards you receive from \your hard work and dedication.

– 100%: ?Today, people young and old, male and female, short and tall work. ?Employment in today?s fast moving economy almost seems essential. ?Children start to work at a young age at little jobs like being ?a newspaper carrier and continue to \work into their sixties ?or later. Work has become an embedded idea that you need to do ?just like sleeping or eating. Money and the strive for wants ?and necessities is the key reason why people around the world ?work. In my family work is essential.?My parents work not only ?as adults, but also as high school students. My family believes ?that you have to work for what you want. That meant for me ?to have a car and money to spend, I had to go to work \and earn ?money just like my parents did when

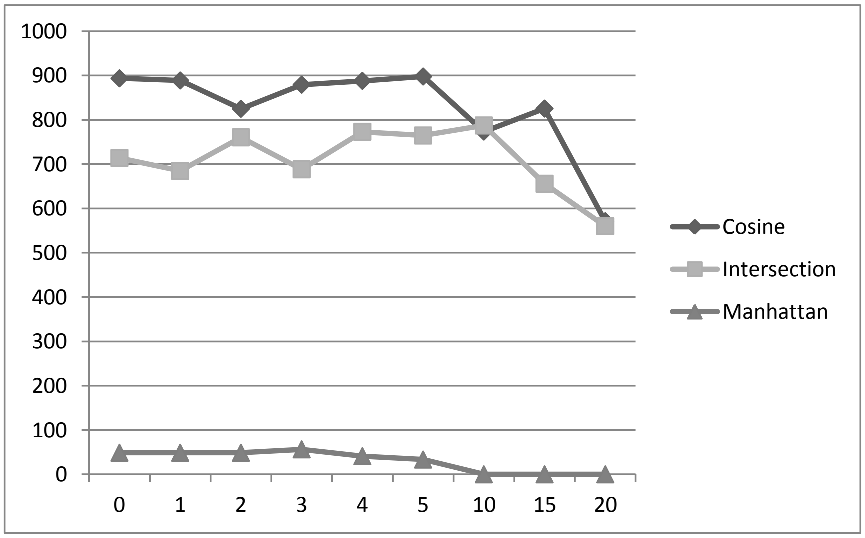

Figure 9. *AAAC score for character 15-grams at differing error levels*

they were my age. I really ?enjoy my job working at Rita?s Italian Ice and I feel that it ?is important that you enjoy what you are doing and appreciate ?the rewards you receive from your har\d work and dedication.

– 75%: ?Today, people young and old, male and female, short and tall work. ?Employment in today?s fast moving economy almost seems essential. ?Children start to work at a young age at little jobs like being ?a newspaper carrier and continue to \work into their sixties ?or later. Work has become an embedded idea that you need to do ?just like sleeping or eating. Money and the strive for wants ?and necessities is the key reason why people around the world ?work. In my family work is essential.?My parents work not only ?as adults, but also as high school students. My family believes ?that you have to work for what you want. That meant for me ?to have a car and money to spend, I had to go to work \and earn ?money just like my parents did when they were my age. I really ?enjoy my job working at Rita?s Italian Ice and I feel that it ?is important that you enjoy what you are doing and appreciate ?the rewards you receive from your har\d work and dedication.

– 25%: ?Today, people young and old, male and female, short and tall work. ?Em-

Character 4-grams				Character 5-grams			
Quality	Cosine	Int.	Man.	Quality	Cosine	Int.	Man.
Original	758	641	820	Original	686	543	757
100	772	597	799	100	733	567	764
75	747	602	799	75	733	579	764
50	772	608	803	50	733	570	768
30	747	609	807	30	733	563	785
25	768	593	774	25	751	545	768
10	685	548	737	10	662	542	718

Table 3. *Real OCR results for character 4- and 5-grams (AAAC Score)*

ployment in today?s fast moving economy almost seems essential. ?Children start to work at a young age at little jobs like being ?a newspaper carrier and continue to \work into their sixties ?or later. Work has become an embedded idea that you need to do ?just like sleeping or eating. Money and the strive for wants ?and necessities is the key reason why people around the world ?work. In my family work Is essential.?My parents work not only ?as adults, but also as high school students. My family believes ?that you have to work for what you want. That meant lor me ?to have a car and money to spend, I had to go to work \and earn ?money just like my parents did when they were my age. I really ?enjoy my job working at Rita?s Italian Ice and I feel that it ?is important that you enjoy what you are doing and appreciate ?the rewards you receive from your har\d work and dedication

– 10%: ’Today, paòplè young and old, male and female, short and tall, work ?Employment in today’s fast moving economy almost seems essential ?Chiidren stari lo work at a young age al little jobs like being ´a newspaper carrier and continue to w\ork into their sixties ´or later. Work has become an embedded idea that you need to do ´just like sleeping or eating. Money and the strive for wants-?and necessities Is the key reason why people around the world Õwork In my lamlly work is essential.´My parents work hot only ?as adults, bulˆlso as high school students. My family believes ´that you have to work for what you want. That meant for me ´to have a car and money to spend, I had to go to work a\nd earn ´money just like my parents did when they were my age. I really ´enjoy my job working at Ritn?s Italian ice and I feel that il ?is Important that you enjoy what you are doing and appreciate ´the rewards you.receive from your hard\work and dedication

These documents were used as a set of natural OCR errors for a similar set of experiments of different event sets and distances. Selected numerical results of the experiment are presented in table 3; a comparison with table 1 is appropriate and instructive.

We should note that this specific OCR engine is English-based; the AAAC includes multilingual content that may have a negative effect on some problems, but that despite this, overall results are in keeping with the simulation; “good” OCR (quality 25 or above) is typically only marginally lower if any; in some cases, the errors introduced by the artificial OCR process actually increased accuracy.

4.5. *Representativeness*

Of course, there are other issues in blindly taking these results at face value; one, for example, is the question of how well the AAAC represents any particular authorship problem, such as Binongo's work on the (scanned) *Oz* books. Previous work (Juola, 2006a) has shown that there is at least reasonable grounds for supposing that good authorship attribution technologies transfer, as there is a strong correlation between any given method's performance on different subproblems within the AAAC. (This makes sense; a poorly performing method is unlikely to magically turn into a superstar when one takes it into a random new environment, one for which it was probably not designed.)

4.6. *OCR engine effects*

A more serious challenge would be the effects of any specific OCR engine, and whether the use of two or more engines would produce specific bias. If an engine introduces specific errors — for example, is substantially more likely to turn "the" into "tlie" — then all documents produced by that engine will share a substantial amount of vocabulary introduced by the engine, which in turn will create substantial spurious similarity. Again, Binongo provides an example. All of his Baum samples were out of copyright and available from Project Gutenberg, but only some of the Thompson samples were. Assuming that Gutenberg's OCR system differed from Binongo's, one might expect an engine-related difference between the public domain Thompson and the copyright Thompson. That no such difference was observed may lend support to the idea that OCR errors are not substantial in this area.

5. Conclusions

Authorship analysis is an important problem, and controlling for errors has been argued to be a key aspect of doing a proper analysis. However, modern OCR technology is quite good (capable of less than 2% character error rate), and proofreading hundreds, let alone hundreds of thousands, of documents is time-consuming, expensive, and introduces its own errors.

Our conclusions are that modern authorship attribution methods are largely robust to the type of error produced by the OCR process. We single out, in particular, nearest neighbor classification systems using normalized cosine distance as being particularly well-performing in a large variety of conditions. Jaccard or intersection distance can also perform well in some conditions, and we specifically recommend against the use of Manhattan distance both for performance reasons and because it has been shown to be much more sensitive to errors and to degrade much more. These results, obtained in simulation, have been confirmed in experiments using artificially degraded images (via lossy compression).

However, in a practical application of authorship attribution, researchers (or lawyers, for that matter) should not be concerned about typical examples of OCR, producing less than about 5% to 10% character error rates.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. OCI-1032683 and by DARPA under BAA-12-06. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or DARPA. The errors ours, the thanks theirs.

6. References

Best E., *The First and Second Epistles to the Thessalonians*, Harper and Row, New York, 1972.

Binongo J. N. G., "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution", *Chance*, vol. 16, n° 2, p. 9-17, 2003.

Chaski C. E., "Who's at the Keyboard: Authorship Attribution in Digital Evidence Invesigations", *International Journal of Digital Evidence*, vol. 4, n° 1, p. n/a, 2005. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Cross F. L., *The Oxford Dictionary of the Christian Church*, Oxford University Press, New York, 1997.

de Morgan A., "Letter to Rev. Heald 18/08/1851", *in* S. E. de Morgan (ed.), *Memoirs of Augustus de Morgan by His Wife Sophia Elizabeth de Morgan with Selections from His Letters*, Longman's Green and Co., London, 1851/1882.

Earle E. M., *The Federalist*, Modern Library, Washington, D.C., 1937.

Grant T., "TXT 4N6: Describing and Measuring Consistency and Distinctiveness in the Analysis of SMS Text Messages", *Proceedings of the Brooklyn Law School Authorship Attribution Workshop*, 2012.

Holley R., "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", *D-Lib Magazine*, March/April, 2009.

Holmes D. I., Forsyth R. S., "The Federalist Revisited : New Directions in Authorship Attribution", *Literary and Linguistic Computing*, vol. 10, n° 2, p. 111-27, 1995.

Jockers M. L., Witten D., "A Comparative Study of Machine Learning Methods for Authorship Attribution", *LLC*, vol. 25, n° 2, p. 215-23, 2010.

Juola P., "Ad-hoc Authorship Attribution Competition", *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden, June, 2004.

Juola P., "Authorship Attribution", *Foundations and Trends in Information Retrieval*, 2006a.

Juola P., "Authorship Attribution for Electronic Documents", *in* M. Olivier, S. Shenoi (eds), *Advances in Digital Forensics II*, vol. 222 of *International Federal for Information Processing*, Springer, Boston, p. 119-130, 2006b.

Juola P., "Authorship Attribution : What Mixture-of-Experts Says We Don't Yet Know", *Proceedings of American Association for Corpus Linguistics 2008*, Provo, UT USA, March, 2008.

Juola P., "Fishing the Ocean: Lessons from Large-Scale Experiments in Styometry", 2011, Talk given at School for Advanced Studies (29 March, 2011).

Juola P., "An Overview of the Traditional Authorship Attribution Subtask", *Proceedings of PAN/CLEF 2012*, Rome, Italy, 2012.

Juola P., Baayen H., "A Controlled-Corpus Experiment in Authorship Attribution by Cross-Entropy", *Literary and Linguistic Computing*, vol. 20, p. 59-67, 2005.

Juola P., Noecker Jr. J., Ryan M., Speer S., "JGAAP 4.0 — A Revised Authorship Attribution Tool", *Proceedings of Digital Humanities 2009*, College Park, MD, 2009.

Juola P., Sofko J., Brennan P., "A Prototype for Authorship Attribution Studies", *Literary and Linguistic Computing*, vol. 21, n° 2, p. 169-178, 2006. Advance Access published on April 12, 2006; doi:10.1093/llc/fql019.

Juola P., Vescovi D., "Empirical Evaluation of Authorship Obfuscation using JGAAP", *Proceedings of the Third Workshop on Artificial Intelligence and Security*, Chicago, IL USA, October, 2010.

Juola P., Vescovi D., "Authorship Attribution for Electronic Documents", *in* G. Petersen, S. Shenoi (eds), *Advances in Digital Forensics VII*, International Federal for Information Processing, Springer, Boston, chapter 9, p. 115-129, 2011.

Klijn E., "The Current State-of-art in Newspaper Digitization: A Market Perspective", *D-Lib Magazine*, January/February, 2008.

Koppel M., Schler J., Argamon S., "Computational Methods in Authorship Attribution", *Journal of the American Society for Information Science and Technology*, vol. 60, n° 1, p. 9-26, 2009.

Koppel M., Schler J., Argamon S., Winter Y., "The "Fundamental Problem" of Authorship Attribution", *English Studies*, vol. 93, n° 3, p. 284-291, 2012.

Martindale C., McKenzie D., "On the Utility of Content Analysis in Authorship Attribution: The Federalist Papers", *Computers and the Humanities*, vol. 29, p. 259-70, 1995.

McMenamin G., "Declaration of Gerald McMenamin", , Available online at `http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin`, 2011.

Mendenhall T. C., "The Characteristic Curves of Composition", *Science*, vol. IX, p. 237-49, 1887.

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

Noecker Jr. J., Juola P., "Cosine Distance Nearest-Neighbor Classification for Authorship Attribution", *Proceedings of Digital Humanities 2009*, College Park, MD, June, 2009.

Noecker Jr. J., Ryan M., Juola P., Sgroi A., Levine S., Wells B., "Close Only Counts in Horseshoes and. . . Authorship Attribution?", *Proceedings of Digital Humanities 2009*, College Park, MD, 2009.

Powell T., Paynter G., "Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images", *D-Lib Magazine*, March/April, 2009.

Rockeach M., Homant R., Penner L., "A Value Analysis of the Disputed Federalist Papers", *Journal of Personality and Social Psychology*, vol. 16, p. 245-50, 1970.

Rudman J., "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.

Rudman J., "On Determining a Valid Text for Non-Traditional Authorship Attribution Studies : Editing, Unediting, and De-Editing", *Proc. 2003 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 2003)*, Athens, GA, May, 2003.

Rudman J., "The Non-Traditional Case for The Authorship of the Twelve Disputed Federalist Papers : A Monument Built on Sand", *Proceedings of ACH/ALLC 2005*, Association for Computing and the Humanities, Victoria, BC, 2005.

Rudman J., "Non-Traditional Authorship Attribution Studies in Eighteenth Century Literature: Stylistics, Statistics and the Computer", , URL: http://computerphilologie.uni-muenchen.de/jg02/rudman.html, accessed 5.31.2007, n.d.

Stamatatos E., "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, vol. 60, n° 3, p. 538-56, 2009.

Stamatatos E., "On the Robustness of Authorship Attribution Based on Character n-gram Features", *Proceedings of the Brooklyn Law School Authorship Attribution Workshop*, 2012.

Tweedie F. J., Singh S., Holmes D. I., "Neural Network Applications in Stylometry : The Federalist Papers", *Computers and the Humanities*, vol. 30, n° 1, p. 1-10, 1996.

van Halteren H., Baayen R. H., Tweedie F., Haverkort M., Neijt A., "New Machine Learning Methods Demonstrate the Existence of a Human Stylome", *Journal of Quantitative Linguistics*, vol. 12, n° 1, p. 65-77, 2005.

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

Wellman F. L., *The Art of Cross-Examination*, 4th edn, MacMillan, New York, 1936.