DETECTING CONTRACT CHEATING VIA STYLOMETRIC METHODS

Patrick Juola

Abstract: Detecting plagiarism automatically (by computer) is in some cases a relatively easy string matching problem; a sufficiently long shared string is strong evidence of copy/paste plagiarism. The problem can be harder with contract plagiarism, as two independent authors are unlikely to phrase ideas in exactly the same way.

In this paper, we discuss how stylometry, the scientific study and measurement of writing style, can be used to address contract plagiarism. We first discuss the theory of stylometry: at its core, stylometry practice identifies habitual patterns of language specific to an individual and attempts to find these patterns in other, disputed, works to determine if those works are by the same author. We discuss several methods that have been empirically established to work, and several case studies where they have been applied.

Finally, we show how these methods have been instantiated into software systems that are capable of dealing with contract plagiarism. By confirming uniformity of authorship across a student's course or even entire academic career, we can identify contract plagiarism as an anomaly and thereby deal appropriately with it.

Key words: Stylometry; authorship attribution, stylometry, stylometrics, plagiarism detection; text classification; authorship analysis

1 Introduction

Plagiarism, the act of taking another's work and passing it off as your own, has almost certainly been with us since the dawn of artwork and written language. For as long as there has been art and artists, there have been people who have put their name to it incorrectly.

The previous paragraph was copied verbatim from a Web page entitled "The World's First 'Plagiarism' Case," published by Johnathan Bailey at http://www.plagiarismtoday.com/2011/10/04/the-world's-first-plagiarism-case. As this paragraph is being written, hundreds or thousands of students across the world are probably doing exactly that, reproducing verbatim the words that someone else has written as part of their papers, while hundreds or thousands of teachers may or may not know what is going on – and may or may not be able to do something about it.

Fortunately, tools like Turnitin and other automated plagiarism detectors make it relatively easy for the teachers. Simply searching Google for the phrase "the dawn of artwork and written language" turns up one and only one instance of that phrase, in that same web page. Detecting this type of "copy/paste" plagiarism is relatively easy. A more sophisticated student, however, could simply pay someone else to write an original paper, which would be turned in under the student's name. In theory, this paper is not copied, and will therefore not be found by Turnitin or by Google. This type of plagiarism is therefore much more difficult for our hapless teachers to detect.

This paper describes the science of stylometry, the detection and analysis of writing style, and discusses how the writing style of an individual person can be identified and

distinguished from others' styles. We then demonstrate practical methods to apply this to solve the so-called "contract plagiarism" problem.

We should also note that, although this paper discusses plagiarism primarily as an academic problem, it has broader implications. In 2017, Ghana's new President Nana Akufo-Addo was found (Glum, 2017) to have plagiarized the speeches of two former US presidents (Clinton and George W. Bush). In 2016, Melania Trump delivered a speech at a political convention with large sections plagiarized from an earlier speech by first lady Michelle Obama. Several German ministers have been found to have plagiarized their Ph.D. theses (and some have resigned in consequence). Accurately detecting plagiarism, then, can be important not just on a personal level, but can even have international implications.

2 Background

2.1 Types of plagiarism

Plagiarism is of sufficient interest to the educational establishment that many colleges and universities offer formal guidance about what plagiarism is (and how to avoid it). Many of these schools, have found it helpful to distinguish between different types of plagiarism. For example, the instance of plagiarism illustrated in the introduction is an example of what is sometimes called "direct plagiarism," "direct copying," or simply "copy/paste plagiarism." This is among the most obvious and the easiest to detect, because the work is obviously a word-for-word copy of the source. If the source can be found, the copying is obvious.

More subtly, a student may combine phrases from several sources or make minor changes, such as "thesaurus plagiarism", where a student will substitute synonyms for words used by the original author but retain the overall idea (and often even the syntax). This type of plagiarism can be more difficult to detect via simple pattern matching, but the use of semantic analysis (Soori, et al., 2016) can still find matches automatically. If the source can be found, the copying can be identified by more subtle similarities.

The most difficult type of plagiarism to detect, however, is one where the source cannot be found because the "original" is not a publicly available document or a document to which the teacher/judge has access. A typical example of this is so-called "contract plagiarism" or "contract cheating," where a student simply pays another person to write the assignment. This other person could be anything from another student to a professional paper mill (of which examples are left to the reader's search engine). Ideally, such a written-to-order paper will itself be original work and bear no more than typical similarity to any other paper on the same subject; the vocabulary, sentence structure, argument formation, and such, will be that of the actual author. But because this paper has never before been published, there are no sources to compare against for a suspiciously high degree of similarity.

At the same time, however, the paper will also not show the vocabulary, sentence structure, etc. of the ostensible author, the student who submitted the paper for grading. An examination of the writing style will show, not suspiciously high similarity, but a suspiciously low similarity to other work that is actually by that person. Informally, a

teacher may become suspicious, for example, when a submitted paper is substantially more sophisticated than a typical paper at the level of the class, or when the submitted paper includes material that has not been part of the class. A teacher might also become suspicious when a paper submitted to an American university contains numerous examples of Commonwealth English (for example, "colour" or "centre"). More formally, the application of stylometry, the quantitative study of writing style, may be able to show that two papers were written by different people.

2.2 Linguistic approaches to stylometry

The question of authorship has been around as long as there have been authors; disputes about authorship go back to classic times. Forensic linguists have attempted to put the process of resolving these disputes on a firmer basis by establishing sensible procedures to address such questions.

The basic theory of stylometry is fairly simple. Language does not fully constrain how any given idea can be expressed, leaving writers free to choose among many different near-synonymous ways to say what they want. McMenamin (2011) expresses it well:

> 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.

Coulthard (2013) formulates it in a similar way:

> The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.

Examples of such choices would include word-by-word choices between Commonwealth and US words (does one park the car on the pavement by the ironmonger, or on the sidewalk near the hardware store?), but also can include apparently personal choices without any obvious sociolinguistic meaning. The variation between "near" and "by" in the previous sentence is one example of such; for another example, consider the position of a fork in a typical table setting (Fig. 1).

Is the fork "to" the left of the plate, "on" the left of the plate, or "at" the left of the plate? Is it on the "left" of the plate, on the "left side" of the plate, or perhaps "on the left hand" of the plate? It should be apparent that there are many ways to describe the same fork and that individuals may make highly individual choices even on such a simple matter.

Figure 1. Typical table setting

McMenamin's report in the *Ceglia v. Zuckerberg* case (McMenamin, 2011) provides an illustration. In this report, he analyzed a set of disputed email messages and found that there were many small differences between the undisputed writings of the ostensible author (Zuckerberg) and the email in question. For example, among the features were the use of apostrophes, the expression of 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" (instead of "can not"), capitalization of the word "Internet", the use of "Sorry" as a sentence-opener, and the presence or absence of run-on sentences. While none of these differences are necessarily key to the ideas expressed (in fact, "cannot" vs. "can not" are as close to synonymous as this author can imagine), the cumulative effect is that the author of the email made different choices from Zuckerberg. As the questioned email differed significantly from other email of known Zuckerberg authorship, he concluded that the authors were "probably" different.

This observation is, in fact, the converse of a traditional rule-of-thumb regarding textual matches (Coulthard, 2004): if there are exact matches of more than five words, then the passages are likely copied (more formally, they are likely to share a common origin, possibly copied, or possibly a well-known set phrase such as a proverb or common quotation). The reason for this is simply that there are so many ways in which one might vary the expression of the same idea that it is unlikely that one could write more than five (in English) words before being provided with a choice and a chance to choose differently from other authors.

As a demonstration, consider the following two passages. These are the abridged but otherwise unmodified writings of, respectively, a 29-year old male from New England with an advanced degree ("Z") and an 11-year old female middle school student from the Pacific Northwest ("B"). These people have not met and are not even aware of each other, but both were asked (independently) to "write and email instructions about how

to make a 'PB&J'."[1] Indeed, they didn't even use the same technology to write the email (Z's was written using a smart phone, hence the increased error rate).

Z:
Take table **knife** out of drawer place on counter
Take out *two slices* **of bread** reseal bag put back in fridge
Open **peanut butter** heavily *apply to bottom slice* **bf bread** wipe off **knife** on top **slice of bread**
Open *jam apply* liberally *to bottom slice* **of bread** clean **knife on** *top slice* **of bread**
Connect bottom **and** *top slices* **of bread** to form sandwich

B:
1. *gather* all supplies – **knife**, spoon, *2 pieces* **of bread**, **peanut butter**, **and** *jelly*
2. **take knife and** *spread* **peanut butter** *on one piece* **of bread**
3. **take** spoon **and** stir *jelly*
4. *spread jelly on other piece of bread* using spoon
5. *put the 2 pieces of bread together* with the **peanut butter and** *jell* on the inside

The task was clearly understood; both B and Z produced a recognizable procedure for making such a sandwich. However, there are notable lexical differences between these accounts (e.g., B uses "jelly", or in one case "jell"; Z uses "jam") and also procedural differences (B uses a spoon to access the jam and places it on a different slice of bread from the peanut butter, both unlike Z; B numbers her steps and Z does not). B does not capitalize steps. Indeed, Z barely uses the extremely common word "and"! In the transcript above, words in **boldface** are common between the two accounts, while words in *italics* are directly analogous to one another (e.g. "spread," "apply"; "put the 2 pieces of bread together," "Connect bottom and top slices of bread") and express the same concept. (We discount here obvious typographic errors such as "bf" for "of" or "jell" for "jelly".)

Of the 54 words in B's account, only 10 are not highlighted at all; more than three-quarters of the account are conceptually duplicated between the two versions. (Of those 54 words, 18 are in fact completely duplicated, such as "of bread.") However, note that no six word sequence – in fact, no three word sequence – is exactly duplicated between the two authors, representing the different choices described by McMenamin and Coulthard.

2.3 Nontraditional approaches to stylometry

While this approach has been shown in numerous instances to be helpful, it suffers from a major problem in the modern world in that it is time-consuming and expertise-dependent. It is not practical to hire enough experts to do this kind of detailed analysis of all the papers generated at a large university. However, it may be possible to substitute statistical analysis for at least some of the literary and linguistic expertise. For example, by compiling a list of keywords and their frequencies, one can calculate frequency distributions (using ordinary statistical methods such as *t*-tests) to

[1]For readers unfamiliar with the term, a "PB&J" is an American sandwich, made with ground peanuts ("peanut butter") and fruit spread ("jelly" or "jam"). American toddlers practically live on them.

determine, first, whether two authors differ in their use of that keyword, and second, whether the word usage pattern is more typical of the first author than the second. This method was applied by Mosteller and Wallace (1961) in a now-classic study of *The Federalist Papers*. This study looked at distribution of hundreds of specific words found in undisputed writings by the various candidate authors and found, for example, that Alexander Hamilton never used the word "whilst" and that James Madison never used the word "while." They further observed that the questioned writings, the ones of less certain authorship, also never used the word "while" (and used "whilst" throughout). This, of course, suggests that the questioned writings were by Madison, not Hamilton. 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. A document with 14 instances of "by" per thousand words, then, is presumptively by Madison.

A more visual approach was used by Binongo (2003) in his studies of the Oz books. Most of the Oz books were written by one of two people: L. Frank Baum, the author of *The Wonderful Wizard of Oz* and its first eleven sequels, and Ruth Plumly Thompson, who took over the series after Baum's death. The authorship of the 13th book (*The Royal Book of Oz*) is not clear.

Binongo broke each book into reasonably small segments, and analyzed the frequencies of the fifty most common words in the undisputed works (by either author). This gave him fifty numbers for each fragment, which let him put these fragments into a fifty-dimensional vector space. Using a statistical technique called "principal component analysis" (PCA), he reduced this space to two dimensions while showing which fragments were close to which other fragments. The resulting diagram is reproduced here as Figure 1.

In Figure 2, the black dots represent (fragments of) works by Baum, while the white circles represent Thompson's work. Plotted on the same scale are white hearts, representing the fragments of the disputed work (*The Royal Book*), and black clubs, representing Baum's final work (*Glinda of Oz*). As can clearly be seen, all of Baum's work, including *Glinda*, are found on the right size of the figure, while all of Thompson's work plus the disputed *Royal Book* are on the left. Based on this, Binongo concluded (p. 13) that the first principal component "clearly separates" the two authors, and that this constituted (p. 16) "objective, independent evidence [. . .] that the book was written in Thompson's pen." He was further able to show (by inspecting factor loadings) that Thompson had (p. 14) a 'tendency to use words indicating position – "up", "down", "on", "over", "out", and "back" – more frequently than Baum. An examination of the raw data reveals that, for these words, Thompson's average rates of usage are about twice as Baum's. Baum, on the other hand, prefers "which" and "that". Moreover, he has a greater propensity for negative words: "but", "not", and "no".' If this kind of analysis were done on a student's work during a term, would anyone hold significant doubt that a significant part of the work was not the student's own?

There are many additional ways to perform this task. Indeed, in recent years, the PAN series of conferences (Juola, 2012X; Juola & Stamatatos, 2013X; Stamatatos, *et al.*, 2014X) has taken to running competitive shared-task evaluations to evaluate how newly

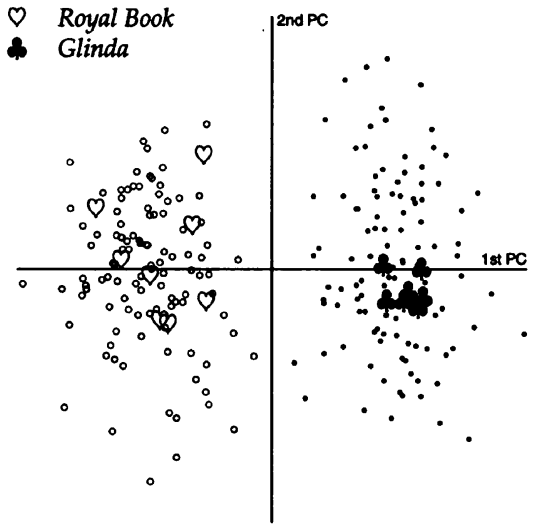

Figure 2. Differences between Baum and Thompson visualized via PCA (from Binongo, 2003)

proposed methods perform and what level of accuracy can be obtained under various conditions (e.g., which language, which genre, how much data is available).

3 Some Case Studies

Juola (2015) describes four additional case studies. Using the JGAAP system (as described in the following section), he was able to show the following:

a) For example, in 1827, an 18-year-old Edgar Allan Poe was trying to start a writing career, but was hampered by creditors. He did manage to publish two of his poems, but only under the initials of Henry (William Henry Leonard Poe), his brother. In the same newspaper and the same year, 'Henry' also published three short stories. A comparison of Henry's own writing with other writings by Edgar as well as several other contemporary authors showed the closest similarity to Edgar, strongly suggesting that these were, in fact, the earliest examples of Edgar Allan Poe's published prose.

b) In US Immigration Court, a person was seeking asylum. His claim was based on a set of anonymous newspaper articles he had (or claimed to have) written, critical of his home government. He was able to offer as supporting evidence a set of other articles, published under his own name. A comparison of writing styles showed that it was, in fact, highly likely that he was also the author of the anonymous works, and he was granted asylum.

c) In 2013, *The Cuckoo's Calling* was published, ostensibly by a first-time author named Robert Galbraith. An analysis of writing style, commissioned by the Sunday

Times, showed that Galbraith wrote very similarly to J.K. Rowling, the author of the *Harry Potter* series, similarly enough that they were probably the same author. Ms. Rowling herself later admitted that she was, in fact, the author of *Cuckoo*.

d) Bitcoin, the secure Internet currency, is based on a set of standards, protocols, and software written by a person using the name Satoshi Nakamoto. No actual person has been identified as the author of these documents, but in 2014, *Newsweek* magazine identified a certain Dorian Satoshi Nakamoto as the author, a charge Mr. Nakamoto almost immediately denied. An analysis commissioned by *Forbes* magazine confirmed that the writings of Dorian were different from those of "Satoshi". To this day, the actual identity of Satoshi remains unknown.

Clearly, the ability to determine if two documents are by the same author (a question that Koppel *et al.* (2012) has called "the fundamental question in authorship attribution") is an important question with wide-ranging application.

4 A System for Detecting Contract Plagiarism

Juola, *et al.* (2006) [see also (Juola, 2009a)] describe a system developed for general-purpose authorship analysis, focusing on selecting the most likely candidate from a set of suggested authors. This system, called JGAAP ("Java Graphical Authorship Attribution Program") proposes a pipelined architecture with several stages. The first stage is "canonicization," preprocessing the relevant documents to put them into "canonical" form. In the second stage, features or "events" are extracted from each document, and tabulated, for example into a histogram of feature frequencies. Finally, these extracted features are analyzed, comparing a questioned document to every known (training) example from every candidate author. One simple comparison, for example, would be to calculate the similarity of the unknown document's histogram as compared to each candidate author's histogram; the author with the most similar histogram is the most likely author.

As a simple example of how JGAAP might perform a replication of the Oz experiment: After preprocessing (for example, removing front and back matter, page numbers, and so forth) the system could break all documents down into "words," extract the "fifty most common" words from the entire set, and then tabulate histograms as Binongo did. Using the ordinary distance formula, the "closest" document to the *Royal Book* could easily be identified. When/if this document turns out to be authored by Thompson, we have evidence of Thompson's authorship of the *Royal Book* itself.

Juola (2015) extended this to the protocol used in the problems described in section 4. He identified three key underlying assumptions. First, he assumed that authorship attribution technology works (at least, "better than chance"), an empirical assumption nevertheless supported by numerous studies (Juola, 2009b, 2012a; Vescovi, 2011) and NIST-style competitive evaluations (Juola, 2004, 2012b; Juola & Stamatatos, 2013; Stamatatos, *et al.*, 2014). Second, he assumed that (as with the JGAAP system), a computer can be programmed to produce a rank-ordering of the authors from most likely to least. Again, many such programs exist, including JGAAP itself. A third assumption is that there are many relatively accurate analysis methods available,

and hence mixture-of-experts ensemble classification is practical; again, this is well-supported in the existing literature.

The proposed protocol uses multiple analyses of the two documents in question as well as an *ad-hoc* "distractor" corpus collected from similar writings (e.g., in the Rowling case, the distractor corpus was a collection of similar, non-Rowling-authored detective novels; the Bitcoin distractor corpus contained documents by others who had been proposed as a Bitcoin author). If the most similar document to the first questioned document was *consistently* (across multiple analyses) the second questioned document (and not one of the distractors), one can conclude that the two documents are by the same person. By contrast, if this does not appear, and particularly if the second document is not often linked closely to the first, one can conclude different authors.

This protocol has been instantiated (Juola, 2016) into a software product, called Envelope, currently available as software-as-a-service (company omitted). Envelope focuses specifically on Email written in English and has been shown to be able to successfully make this determination in controlled testing. Although Envelope focuses on English, the general approach has been proven useful in a wide variety of languages and genres, and an appropriate system could easily be built for any desired application, such as for the analysis of German-language or French-language term papers.

The application to classroom teaching is fairly straightforward. Over the course of a typical semester, students will presumably be submitting several papers to the teacher, and thus a teacher will rather quickly have a collection of multiple papers from multiple authors, all of a relatively homogenous genre. The most recent paper by a student can be compared to see whether the writing style is more similar to earlier work by that same student, or (using the rest of the class as a distractor set), more similar to that of another student. Note that finding that Thomas' paper is closer to William's work than to Thomas' other work does not indicate that William wrote this paper, or even that William has engaged in any inappropriate acts, but it does suggest that whoever wrote the paper Thomas submitted does not write in Thomas' style (and writes more like William).

5 Discussion

Envelope or an Envelope-like system could easily be used to resolve issues of suspected contract plagiarism with relatively little time or effort. In the event that plagiarism is suspected, or even as a routine precaution, student work could be compared with other work by the same student. In the event that there are two documents by two different authors, at least one of those documents must, *ipso facto*, be by a different author and hence plagiarized. A key advantage of this approach is that, unlike detection-by-search-engine, there is no need to find the source document and it will work even if the source document is not publicly available.

One potential objection to this approach is that the first document is "free," in the sense that there is nothing to compare with. While true, because it is also possible to compare across multiple courses and multiple school years, the effect of this can be expected to be minimal. Similarly, if a student consistently uses the same author to

write all of the assignments, there will only be one author, but most paper mills cannot guarantee the long-term availability of any single author.

A more serious issue is the possibility of a false positive; a document incorrectly identified (for whatever reason) as by a different author. This is a serious possibility, but it should be able to be addressed by the same notions of due process and fair procedures that characterize other suspected plagiarism cases. There is a common notion that computers are somehow error-free, but this notion must be resisted, and evidence of plagiarism generated algorithmically by a computer should be treated with the same respectful skepticism as accusations made by a human expert reader.

Unfortunately, the evidence used by a computer is often hard for humans to understand and interpret; for example, using negative words like "but" and "not" too often may not be obvious to a human reader – using character clusters like "fro" or "nd" too often are even harder to interpret. Many of the most accurate methods of analysis do not lend themselves to easy human analysis. While research continues into accurate but human readable algorithms, it is important to bear in mind that false positives exist in stylometry as in any other science.

6 Conclusions

Among the types of plagiarism, "contract plagiarism" can be the hardest to detect. Much research has focused on finding the source document, but if I have ghost-written something to order for someone else to submit under their own name, the source document may exist only on my hard drive. The science of stylometry may still be able to detect the plagiarized work, however. Like everyone else, we all have a unique authorial style. If a personal authorial style can be detected in the ghost-written work, and distinguished from the submitter's own style as evidenced in other works, the act of plagiarism becomes obvious.

This task can be performed automatically by a computer with relatively high accuracy. Computers provide several advantages; they can handle large volumes of work quickly, and are in some ways more objective and accurate than humans. At the same time, they are (as always) restricted by the limitations of their programming and a lack of common sense.

Unfortunately, as with any automatic analysis, false positive errors (as well as false negative errors) are a concern. Total elimination of errors is impractical, but also unnecessary. Action taken on the basis of a finding of plagiarism should only be taken in accordance with well-established due process and should involve human examination as well. False negatives (failure to detect actual plagiarism) is also a concern. Research continues into improved stylometric technologies, including better accuracy, better accessibilities to human decision makers, and a better understanding of the nature of writing style.

Despite these limitations, stylometry is an important and relatively mature technology that can be usefully applied to address a key problem in education as well as in the broader world.

Literature

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM,* 52(2):119–123.

Jose Nilo G. Binongo. (2003). Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance,* 16(2):9–17.

Richard Brooks. (2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed. *Sunday Times*, London: Times Newspapers Ltd. 14 July.

Richard Brooks and Cal Flyn. (2013). JK Rowling: The cuckoo in crime novel nest. *Sunday Times*, London: Times Newspapers Ltd. 14 July.

Malcolm Coulthard. (2013). On Admissible Linguistic Evidence, *Journal of Law and Policy* 21(8):441–466.

Julia Glum. (2017). Who is Nana Akufo-Addo? Ghana President's Plagiarism Scandal, Explained. *International Business Times*, January 10.

Matthew Herper. (2014). Linguist Analysis Says Newsweek Named the Wrong Man as Bitcoin's Creator. Jersey City, NJ: *Forbes Magazine*, March 10.

Matthew L. Jockers and D. Witten. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2): 215–23.

Patrick Juola. (2004). Ad-hoc Authorship Attribution Competition. ALLC/ACH 2004, Goteborg, Sweden.

Patrick Juola, John Sofko, and Patrick Brennan. (2006). A Prototype for Authorship Attribution Software. *Literary and Linguistic Computing* 21: 169–178.

Patrick Juola. (2009a). JGAAP: A System for Comparative Evaluation of Authorship Attribution. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science* 1.

Patrick Juola. (2009b). 20,000 Ways Not to Do Authorship Attribution – and a Few that Work. *2009 Biennial Conference of the International Association of Forensic Linguists*, Amsterdam, Netherlands.

Patrick Juola. (2011). Report on Authorship Attribution Subtask. *CLEF 2011* (Conference on Multilingual and Multimodal Information Access Evaluation), Amsterdam, Netherlands.

Patrick Juola. (2012a). Large-Scale Experiments in Authorship Attribution. *English Studies* 93:275–283.

Patrick Juola. (2012b). An Overview of the Traditional Authorship Attribution Subtask. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012* Evaluation Labs and Workshop – Working Notes Papers, 17–20 September, Rome, Italy.

Patrick Juola and Efstathios Stamatatos. (2013). Overview of the Author Identification Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013* Evaluation Labs and Workshop – Working Notes Papers, 23–26 September, Valencia, Spain.

Patrick Juola. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities.*

Patrick Juola. (2016). Did Aunt Prunella Really Write That Will? A Simple and Understandable Computational Assessment of Authorial Likelihood. *Workshop on Legal Text, Document, and Corpus Analytics – LTDCA 2016*, San Diego, California.

Gerald McMenamin. (2011). Declaration of Gerald McMenamin. *Ceglia v. Zuckerberg.* Available online at http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin.

Fredrick Mosteller and David L. Wallace. (1964). *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, MA.

Efstathios Stamatatos. (2013) On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy,* XXI(2):420–440.

Hans van Halteren, R. Harald Baayen, Fiona Tweedie, Marco Haverkort, Anneke Neijt. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics,* 12(1):65–77.

Koppel, M., Schler, J., Argamon, S., and Winter, Y. (2012). The 'fundamental problem' of authorship attribution. *English Studies*, 93(3): 284–91.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology,* 60(1): 9–26, (2009).

Hussein Soori, et al. Semantic and Similarity Measure Methods for Plagiarism Detection of Student's Assignments. Proc. *AECIA 2015.* 427:117–125. Basel:Springer. (2016).

Efstathios Stamatatos, Benno Stein, Walter Daelemans, Patrick Juola, Alberto Barrón-Cedeño, Ben Verhoeven, Miguel A. Sanchez-Perez. Overview of the Authorship Identification Task at PAN 2014. *Proceedings of PAN/CLEF 2014*, Sheffield, United Kingdom. (2014)

Darren M. Vescovi. (2011). Best practices in authorship attribution of English essays. Master's thesis, Pittsburgh, PA: Duquesne University.

Copyright statement

Author

Juola & Associates, 276 W. Schwab Avenue, Munhall, PA 15120 USA,
pjuola@juolaassociates.com, http://www.juolaassociates.com