Detecting Stylistic Deception

Patrick Juola
Evaluating Variation in Language Laboratory
Duquesne University
Pittsburgh, PA 15282 USA
juola@mathcs.duq.edu

Abstract

Whistleblowers and activists need the ability to communicate without disclosing their identity, as of course do kidnappers and terrorists. Recent advances in the technology of stylometry (the study of authorial style) or "authorship attribution" have made it possible to identify the author with high reliability in a non-confrontational setting. In a confrontational setting, where the author is deliberately masking their identity (i.e. attempting to deceive), the results are much less promising. In this paper, we show that although the specific author may not be identifiable, the intent to deceive and to hide his identity can be. We show this by a reanalysis of the Brennan and Greenstadt (2009) deception corpus and discuss some of the implications of this surprising finding.

1 Introduction

Deception can occur in many different ways; it is possible to deceive not only about the content of a message, but about its background or origin. For example, a friendly invitation can become sexual harassment when sent from the wrong person, and very few ransom notes are signed by their authors. Recent research into stylometry has shown that it is practical to identify authors based on their writing style, but it is equally practical (at present technology) for authors to use a deliberately deceptive style, either obfuscating their own style or mimicking that of another writer, with a strong likelihood of avoiding identification.

In this paper, we investigate the possibility of identifying, not the specific author of a text, but whether or not the author of a text wrote with the (deceptive) intent to disguise their style. Our results strongly suggest that this deceptive intent can itself be identified with greater reliability than the actual author can be.

2 Background

Stylometric authorship attribution — assessing the author of a document by statistical analysis of its contents — has its origins in the 19th century (Mendenhall, 1887; de Morgan, 1851), but has experienced tremendous resurgence since the work of (Mosteller and Wallace, 1964) and the beginnings of the corpus revolution. With the exponential growth of digital-only texts and the increasing need to validate or test the legitimacy of questioned digital documents, this is obviously an area with many potential applications.

The most commonly cited stylometric study is of course that of Mosteller and Wallace (1964), who examined the frequency of appearance of approximately thirty function words within the collection of documents known as *The Federalist Papers*. Using a form of Bayesian analysis, they were able to show significant differences among the various authors in their use of these words and hence infer the probabilities that each document had been written by each author – i.e. infer authorship. Another classic in this field is the study of the *Oz* books by Binongo (2003), where he applied principal component analysis (PCA) to the frequencies of the fifty most frequent words in these books and was able to demonstrate (via the first two principle components) a clear visual separation between the books written by Baum and those written later by Thompson. Recent surveys of this field (Argamon et al., 2009; Koppel et al., 2005; Rudman, 1998; Koppel et al.,

Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection, pages 91–96,
Avignon, France, April 23 - 27 2012.

2009; Juola, 2006; Jockers and Witten, 2010; Stamatatos, 2009) illustrate many techniques of increasing sophistication and accuracy.

What, however, of the person who doesn't want to be identified? Chaski (2005) cites several real-world instances where authorship attribution was applied to the task of detecting miscreants, and in one case a murderer. We assume that these miscreants would have preferred to hide their identities if possible. On a more positive note, activists who fear a tyrannical government would do well to avoid being identified by the political police. Intuitively, it seems plausible that one would be able to write "in a different style," although it also seems intuitively plausible that at least part of one's writing style is fixed and immutable (van Halteren et al., 2005) — you can't pretend, for example, to a bigger vocabulary than you have, as you can't use words that you don't know. On the other hand, the long tradition of pastiche and parody suggests that at least some aspects of style can be copied.

It should be noted that this type of "deception" is different than what most research project study. Traditionally, a "deceptive" statement occurs when a speaker or writer offers an untruth; we instead suggest that another form of "deception" can occur when a speaker or writer offers a statement *that he or she does not want to be identified with.* This statement may be true (a whistleblower identifying a problem, but not wanting to risk being fired) or false (a criminal writing a false confession to incriminate someone else) — the key deception being the identity of the author.

There is little research on the success of "deceptive style" and what little there is should lend hope to activists and whistleblowers. A team of Drexel researchers (Brennan and Greenstadt, 2009; Afroz et al., 2012) developed a small corpus of deceptive writing (described in detail later), but were unable to find any methods to pierce the deception. Larger scale analyses (Juola and Vescovi, 2010; Juola and Vescovi, 2011) similarly failed. '[N]o method [out of more than 1000 tested] was able to perform "significantly" above chance at the standard 0.05 level.... We [...] observe that, yes, there is a confirmed problem here. Although these analyses performed (on average) above chance, they did not do so by robust margins, and there is enough variance in individual performance that we cannot claim even to have "significant" improvement.'

In light of these results, the Drexel team have proposed and developed a tool ["Anonymouth", (Afroz and Brennan, 2011; Perlroth, 2012)] that provides a more formal and systematic method of disguising their writing style. Based in part on the JGAAP tool (Juola et al., 2009; Juola, 2006), this system allows would-be activists to see what aspects of their linguistic fingerprints are more obvious in a document, and guides these same activists to make changes to neutralize their personal style, or even to assume a specific other's style. In some sense, Anonymouth is the "evil twin" countermeasure to JGAAP – while JGAAP detects style, Anonymouth in theory renders style indetectable.

Does it work? The tool is still too new for substantial testing, but we assume based on the earlier work that it will still be difficult to detect the original author under the deception. However, it may be possible to detect the act of deception itself. As will be seen in the following sections, standard stylometric tools themselves can do that.

3 Materials and Methods

One of the most powerful and flexible tools for text analysis and classification is the JGAAP (Java Graphical Authorship Attribution Program) software package. Available for download from `www.evllabs.com`, it is a modular Java-based freeware program that implements a simple pipelined architecture for text classification. We have applied it to the Brennan-Greenstadt (Brennan and Greenstadt, 2009) Adversarial corpus of imitative and obfuscative essays, to determine whether these "imitative and obfuscative" essays have themselves a stable style.

JGAAP itself (Juola, 2006) follows a standard model for statistical authorship attribution. Documents are pre-processed ("canonicized") to eliminate distracting or uninformative variation (for example, the length of a printed line is set by the editor, not the author, so "normalizing whitespace" removes some editorial distraction). Following this, they are broken up into individual events which are subject to some form of statistical analysis and classification.

As an example, Binongo's experiment could be captured within the JGAAP framework as follows: preprocess the *Oz* books to remove extraneous materials (including punctuation), unify case to neutralize syntax-based case distinctions (like

capital letters at the beginning of a sentence), and normalize whitespace as described above. Binongo then broke the books up into individual words and selected only the top fifty most common words as his events of interest. Finally, he applied PCA to get a visual picture of how the documents separated in abstract event space.

The B/G corpus (as published) collects the writings of twelve people, with approximately ten essays (totaling about 5000 words) per person. In addition, it has twelve passages from *The Road* by the professional author Cormac McCarthy. This establishes a baseline of 126 essays of "normal" writing, writing where there is no attempt to disguise style (and hence no deception). Each subject was then asked to write two additional essays: one deliberately written "in such a way that their personal writing style will not be recognized" (a so-called "obfuscation"), and another in the style of Cormac McCarthy as expressed in *The Road* ("imitation"). This creates another twenty-four instances of "deceptive" writing where the writer is attempting to conceal their style.

Of the twelve (thirteen including McCarthy) writers in the sample, we used nine (plus McCarthy) as training data, labeling all instances of the writer using their own style as "HONEST" and all instances of obfuscation/imitation as "DECEPTIVE." The writings of three subjects (Andrew, Chris, and Sara) were used as test subjects, creating six unknown "deceptive" documents and twenty-eight unknown "honest" documents. (This is obviously an unbalanced corpus as the number of honest documents is more than five times the number of deceptive ones; this will be discussed later.)

These documents were analyzed by JGAAP using the following settings, which previous testing (Noecker and Juola, 2009) has shown to work well:

- Canonicizer : Unify Case & Punctuation Separator
- Event Driver : Character 3-grams (Character N-grams with N set to 3)
- Event Culler : <none>
- Analysis Driver : Nearest Neighbor Driver
- Distance : (Normalized) Cosine Distance

"Unify Case" converts all (alphabetic) characters to lower case, thus removing any apparent differences between sentence-initial words and their sentence-internal equivalents. "Punctuation Separator" breaks up groups of consecutive punctuation characters by inserting spaces (e.g. "(!)" would become "(!)"). The events analyzed were strings of three consecutive characters (e.g. the word "there" contains three such 3-grams ("the" "her" and "ere"). These 3-grams were not culled (unlike the Binongo experiment, where the events were culled to include only the top 50) and instead were all used in the analysis. These 3-grams were collected into a histogram for each document and inter-document distances were calculated using the normalized cosine distance (aka dot product distance). Finally, each testing document attributed to (considered to be the same deceptiveness type as) the closest training document.

4 Results

The results are summarized in table 1. Of the six deceptive documents, five (or 83%) were correctly identified, while of the twenty-eight non-deceptive documents, twenty-two (or 79%) were correctly identified. (Of course, due to the imbalance in the test set, only 44% of the documents labeled "deceptive" actually were; we consider this statistic something of an artifact.) This result is of course far above chance: baseline performance would be only two correct on deceptive documents and 19 correct on honest ones. Fisher's exact test on the 2×2 contingency matrix shows a one-tailed probability of $p < 0.00790$ (or a two-tailed probability of double that, of course), confirming the high significance of this result.

Preliminary error analysis is attached as table 2. Most notable is that none of the imitation Cormac McCarthy analyses were misclassified as "normal" writing.

5 Discussion and Future Work

Previous work [(Brennan and Greenstadt, 2009; Juola and Vescovi, 2010; Juola and Vescovi, 2011)] has shown that identifying the author of "deceptively" written materials is extremely difficult. We thus have the highly surprising result that, while identifying the specific author may be difficult, uncovering the mere fact that the author

		Actual Deception	
		Y	N
Detected Deception	Y	5	6
	N	1	22

Table 1: Results from deception-detection experiment

	FP	FN (obfusc)	FN (imit)
Andrew	3	0	0
Chris	1	1	0
Sara	2	0	0

Table 2: Number of incorrect classifications by type

is concerned about being identified is relatively easy. This of course parallels the rather commonplace situation in detective fiction where the fact that the criminal has wiped the fingerprints off the murder weapon is both easy to learn and highly significant, even if the criminal's actual identity must wait five more chapters for the big reveal. Similarly, it appears to be fairly easy to detect the attempt to wipe one's authorial "fingerprints" off of the writing.

This result is all the more surprising in light of the heterogeneity of the corpus; the writing style of ten different people, collectively, created our sample of "normal" writing. The writings of three entirely different people fit that sample relatively well. Astonishingly, the attempts of all twelve people to write "differently" fit into a recognizable and distinctive stylistic pattern; these twelve people seem to have a relatively uniform sense of "the other." This sense of "the other," in turn, persists even when these people model the writings of a professional writer *whose style itself is part of the "normal" sample!*

Put more strongly, when "Chris" (or any of the other test subjects) attempted to write in the style of Cormac McCarthy, the result was actually closer to a third party's attempt to write deceptively than it was to McCarthy's writing himself. In the specific case of "Andrew's" imitative writing, all six of the six closest samples were of deceptive writing, suggesting that "deceptive writing" is itself a recognizable style.

Further investigation is clearly required into the characteristics of the style of deception. For example, there may not be one single style; it may instead be the case that "imitation McCarthy" is a recognizable and distinct style from McCarthy's, but also from "obfuscated style." There may be one or several "obfuscated styles." It is not clear from this study what the characteristics of this style are, and in fact, the inability of JGAAP (and JGAAP's distance-based measures in particular) to produce explanations for what are evidently clear-cut categorizations is one of the major weaknesses of the JGAAP system as currently envisioned. Even simple replication of this experiment would be of value, as while we consider it unlikely that our arbitrary choice of test subjects would have created an unrepresentative result, we can't (yet) confirm that. Indeed, we hope that this finding provides encouragement for the development of larger-scale corpora than the simple twelve-subject Brennan-Greenstadt corpus.

We also hope this finding spurs research into exactly what the stylistic "other" is, and in particular, research from a psychological or psycholinguistic standpoint. For example, Chaski (2005) [see also (Chaski, 2007)] argues that the linguistic concept of "markedness" is a key aspect of author identification. Chaski in particular suggests that the use or non-use of "marked" constructions is a good feature to capture. Following her line of reasoning, if I try to write as "not-myself," does this mean I will deliberately use concepts that I consider to be "marked" and therefore unusual? (If this were true, this would have significant implications for the theory of markedness, as this concept is usually held to be a property of a language as a whole and not of individual idiolects. In particular, if I personally tend to use "marked" constructions, and consider traditionally "unmarked" construtions to be unusual, does this imply that traditional notions of "markedness' are reversed *in my idiolect*, or that my cognitive processing of

this construction is atypical?) Alternatively, if authorship is defined more computationally in terms of probability spaces, can we relate "otherness" to a notion of prototypicality (Rosch and Mervis, 1975) of language?

Even without explanations, our basic results have significant implications for the stylometric arms race. We acknowledge the legitimate need for the good guys to analyze the writings of the bad guys to help find them, while also acknowledging the needs of the good guys (human rights advocates, corporate whistleblowers, etc.) to be free to expose the abuses of the bad guys without fear of retribution. We applaud the development of tools like *Anonymouth* for this reason. On the other hand, if an attempt even to disguise one's style is detectable, it may equally be suspicious — especially in the mind of one who believes that the innocent have no reason to disguise themselves. In this regard tools like *Anonymouth* may be similar to encryption programs like PGP. Encrypted email may be suspected due to its very rarity. Zimmermann (nd) has suggested that "it would be nice if everyone routinely used encryption for all their E-mail, innocent or not, so that no one drew suspicion by asserting their [right to] E-mail privacy with encryption."

This result may also have significant implications for (linguistic) forensic practices. The question of reliability is key for any evidence. Any defense lawyer will ask whether or not it's possible that someone could have imitated the style of his client when writing the incriminating document. The results of repeated analysis of the Brennan-Greenstadt corpus suggest that it is, in fact, possible to fool stylometric analysis. The results presented here, however, show that such deception is detectable — the analyst can respond "yes, it may be possible, but such imitation would leave traces that were not found in the document." By showing a lack of deceptive intent, one can enhance the *de facto* reliability of a report.

A key technical question that remains is whether tools like *Anonymouth* will produce "strongly" stylistic masking – and whether the use of such tools is as detectable as more freestyle approaches to stylistic matching, where the author is simply told "write like so-and-so." In theory *Anonymouth* could guide a writer to specific types of stylistic difference ("you use words that are too short; use longer words") – in practice (Greenstadt, personal communication) this has so far been shown to be very cumbersome. (Of course, *Anonymouth* itself is barely out of prototype stage and can probably be improved.) A worst-case scenario would be where the use of *Anonymouth* itself left the equivalent of stylistic "toolmarks," allowing people to identify that the message had been altered by this specific software package (and possibly even a specific version). This could, in turn, provide investigators with information and evidence that actually makes it easier to identify the origin of a given text (e.g., how many people have *Anonymouth* on their systems?).

6 Conclusions

The results of this study, despite being preliminary, show that attempts to disguise one's writing style can be detected with relatively high accuracy. While these results technically only apply to freestyle deception as opposed to tool-based deception, we expect that similar findings would apply to the use of anti-stylometric tools. Similarly, we have only shown one particular method is capable of performing this detection, but we expect that there are others as well and invite large-scale testing to find the most accurate way to detect deceptive writing, which may or may not be the best way to identify the author of non-deceptive writing (or the author of deceptive writing, for that matter).

From the standpoint of security technologies, this creates another level in the countermeasures/counter-countermeasures/etc. loop. If the use of a tool provides security at one level, it is likely to create a weakness at another; disguising one's writing style may at the same time make it obvious to an appropriate observer that you are trying to conceal something. With interest in stylometry and stylometric security growing, we acknowledge the need for stylistic masking, but argue here that using such tools may actually put the masked writer at risk.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Numbers OCI-0721667 and OCI-1032683. Any opinions, findings, and conclusions or recommendations expressed in this material are those

of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Sadia Afroz and Michael Brennan. 2011. Deceiving authorship detection. In *28th Annual Meeting of the Chaos Computer Club (28C3)*, Berlin.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 33rd conference on* IEEE Symposium on Security and Privacy*, pages=To appear*. IEEE.

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

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.

Michael Brennan and Rachel Greenstadt. 2009. Practical attacks against authorship recognition techniques. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI)*, Pasadena, CA.

Carole E. Chaski. 2005. Who's at the keyboard: Authorship attribution in digital evidence invesigations. *International Journal of Digital Evidence*, 4(1):n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Carole E. Chaski. 2007. The keyboard dilemma and forensic authorship attribution. Advances in Digital Forensics III.

Augustus de Morgan. 1851. Letter to Rev. Heald 18/08/1851. In Sophia Elizabeth. De Morgan (Ed.) *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters.*

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

Patrick Juola and Darren Vescovi. 2010. Empirical evaluation of authorship obfuscation using JGAAP. In *Proceedings of the Third Workshop on Artificial Intelligence and Security*, Chicago, IL USA, October.

Patrick Juola and Darren Vescovi. 2011. Authorship attribution for electronic documents. In Gilbert Petersen and Sujeet Shenoi, editors, *Advances in Digital Forensics VII*, International Federal for Information Processing, chapter 9, pages 115–129. Springer, Boston.

Patrick Juola, John Noecker, Jr., Mike Ryan, and Sandy Speer. 2009. Jgaap 4.0 — a revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Moshe Koppel, Johnathan Schler, and K. Zigdon. 2005. Determining an author's native language by mining a text for errors (short paper). In *Proceedings of KDD*, Chicago,IL, August.

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

T. C. Mendenhall. 1887. The characteristic curves of composition. *Science*, IX:237–49.

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

John Noecker, Jr. and Patrick Juola. 2009. Cosine distance nearest-neighbor classification for authorship attribution. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Nicole Perlroth. 2012. Software helps identify anonymous writers or helps them stay that way, January. New York Times article of 3 January, 2012.

Eleanor Rosch and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

J. Rudman. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–56.

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

Phil Zimmermann. n.d. Why do you need PGP? `http://www.pgpi.org/doc/whypgp/en`. Retrieved 18 January, 2012.