palgrave pivot

FORENSIC AUTHORSHIP ANALYSIS AND THE WORLD WIDE WEB

Samuel Larner

palgrave▸pivot

Forensic Authorship Analysis and the World Wide Web

Samuel Larner

University of Central Lancashire, UK

palgrave
macmillan

DOI: 10.1057/9781137413758.0001

First published 2014 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN: 978–1–137–41376–5 EPUB
ISBN: 978–1–137–41375–8 PDF
ISBN: 978–1–137–41374–1 Hardback

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

www.palgrave.com/pivot

DOI: 10.1057/9781137413758

Forensic Authorship Analysis and the World Wide Web

DOI: 10.1057/9781137413758.0001

For Simon and Penny

DOI: 10.1057/9781137413758.0001

Other Palgrave Pivot titles

Karen Rich: **Interviewing Rape Victims: Practice and Policy Issues in an International Context**

Vieten M. Ulrike (editor): **Revisiting Iris Marionyoung on Normalisation, Inclusion and Democracy**

Fuchaka Waswa, Christine Ruth Saru Kilalo, and Dominic Mwambi Mwasaru: **Sustainable Community Development: Dilemma of Options in Kenya**

Giovanni Barone Adesi: **Simulating Security Returns: A Filtered Historical Simulation Approach**

Daniel Briggs and Dorina Dobre: **Culture and Immigration in Context: An Ethnography of Romanian Migrant Workers in London**

Toswell, M.J.: **Borges the Unacknowledged Medievalist**

Lack, Anthony: **Martin Heidegger on Technology, Ecology, and the Arts**

Carlos A. Scolari, Paolo Bertetti and Matthew Freeman: **Transmedia Archaeology: Storytelling in the Borderlines of Science Fiction, Comics and Pulp Magazines**

Judy Rohrer: **Queering the Biopolitics of Citizenship in the Age of Obama**

Paul Jackson and Anton Shekhovtsov: **The Post-War Anglo-American Far Right: A Special Relationship of Hate**

Elliot D. Cohen: **Technology of Oppression: Preserving Freedom and Dignity in an Age of Mass, Warrantless Surveillance**

Ilan Alon (editor): **Social Franchising**

Richard Michael O'Meara: **Governing Military Technologies in the 21st Century: Ethics and Operations**

Thomas Birtchnell and William Hoyle: **3D Printing for Development in the Global South: The 3D4D Challenge**

David Fitzgerald and David Ryan: **Obama, US Foreign Policy and the Dilemmas of Intervention**

Lars Elleström: **Media Transformation: The Transfer of Media Characteristics Among Media**

Claudio Povolo: **The Novelist and the Archivist: Fiction and History in Alessandro Manzoni's The Betrothed**

Gerbrand Tholen: **The Changing Nature of the Graduate Labour Market: Media, Policy and Political Discourses in the UK**

Aaron Stoller: **Knowing and Learning as Creative Action: A Reexamination of the Epistemological Foundations of Education**

Carl Packman: **Payday Lending: Global Growth of the High-Cost Credit Market**

Lisa Lau and Om Prakash Dwivedi: **Re-Orientalism and Indian Writing in English**

Chapman Rackaway: **Communicating Politics Online**

G. Douglas Atkins: **T.S. Eliot's Christmas Poems: An Essay in Writing-as-Reading and Other "Impossible Unions"**

Marsha Berry and Mark Schleser: **Mobile Media Making in an Age of Smartphones**

DOI: 10.1057/9781137413758.0001

Acknowledgements

This research has been conducted over a ten-year period, starting in 2004 and concluding in 2014. Over this period, I have talked about my research with several people who offered useful advice. In particular I would like to thank Janet Cotterill for her insightful comments on the first stages of this research, which were carried out in 2004 at Cardiff University, and Mandip Bains for being a superb friend and sounding board for my ideas.

I would also like to thank my colleagues and friends at the University of Central Lancashire for their support and encouragement. I am particularly grateful to my Dean, Isabel Donnelly, for supporting my sabbatical application so that I could complete this work, and my colleagues in the Linguistics team who covered my absence. I am especially grateful to Dawn Archer for her enthusiasm and guidance, and Paul Seager for his advice and for keeping me on track with my writing. I owe my thanks to Beth Richardson who kindly covered some of my teaching which allowed me to push ahead with my writing. Thanks also to Libby Forrest and Rebecca Brennan at Palgrave Macmillan for expertly guiding me through the publication process.

In addition, there are some very important people to thank who kept things ticking over at home whilst I was locked away in my study. Terry and Lynne Larner have always been wonderful parents, but over the past 12

DOI: 10.1057/9781137413758.0002

months in particular have offered a great deal of support and kindness. Thank you very much for helping me and Simon when we needed it the most.

And finally, my biggest thanks goes to Simon Larner for his constant support and reassurance, and Penny for sensing when I needed cuddles.

DOI: 10.1057/9781137413758.0002

Contents

Acknowledgements viii

1 Introduction: The UNABOM Investigation 1
- 1.1 Idiolect 4
- 1.2 Lexis as a marker of authorship 8
- 1.3 Central research question and research aims 9
- 1.4 Overview of the research 10

2 The Web as Corpus and Authorship Attribution 12
- 2.1 The web 14
- 2.2 The nature of search engines 16
- 2.3 The reliability of search engines 22
- 2.4 The web as corpus 28
- 2.5 Searching the web to produce forensic linguistics evidence 31

3 Attributing Documents to Unknown Authors Using Idiolectal Co-selection and the Web 35
- 3.1 Data and method 36
- 3.2 Targeted searches 42
- 3.3 Web searches using commercial search engines 46

4 Attributing Documents to Candidate Authors Using Idiolectal Co-selection and the Web 49
- 4.1 Data and method 50

4.2 Targeted searches 52
4.3 Web searches using commercial search engines 54

5 The Suitability of the Web for Producing Forensic Evidence 58
5.1 UNABOM revisited 60
5.2 Idiolectal co-selection as a marker of authorship 61
5.3 Using the web as a corpus for producing forensic evidence 64
5.4 Conclusion 67

Bibliography 70

Index 76

DOI: 10.1057/9781137413758.0001

1
Introduction: The UNABOM Investigation

Abstract: *Increasingly, forensic linguists are using the web to generate evidence in cases of forensic authorship analysis. A striking example of this occurred during the trial of the Unabomber – a prolific serial bomber – when the web was searched to determine the distinctiveness of a set of idiolectal co-selections. However, to date, questions have not been asked about whether the web can be used reliably in forensic contexts. Therefore, using the Unabomber trial evidence as a case study, this chapter discusses the notion of idiolect and introduces research which explores two issues: (1) whether idiolectal co-selections can be used as a marker of authorship, and (2) whether the web is reliable enough to be used to produce forensic evidence.*

Keywords: authorship attribution; idiolect; idiolectal co-selections; Unabomber; web corpus

Larner, Samuel. *Forensic Authorship Analysis and the World Wide Web*. Basingstoke: Palgrave Macmillan, 2014.
DOI: 10.1057/9781137413758.0003.

Linguists are increasingly utilising the world wide web (henceforth "web"[1]) as a corpus for research purposes (Volk, 2002) and, given that many linguists acknowledge corpus linguistics as a mainstream methodology (Lindquist & Levin, 2000), coupled with the importance of corpus linguistics methods in the field of forensic linguistics (Coulthard, 1994; Hänlein, 1999; Solan & Tiersma, 2004; Woolls & Coulthard, 1998), it stands to reason that forensic linguists increasingly turn to the web for investigative and evidential purposes. This is particularly the case for forensic authorship analysis – determining the author of a document whose authorship is contested, such as in detecting plagiarism and collusion, attributing a criminal text to an author from a list of potential authors, or profiling an unknown author based on linguistic characteristics. In such cases, the linguist may use the web to show the distinctiveness or rarity of particular words and phrases (Coulthard, 2004).

A striking example of this occurred during the trial of Theodore Kaczynski, a prolific American serial bomber. During the period of May 1978 to April 1995, a total of 16 bombing incidents occurred, initially targeted at individuals connected to universities and the airline industry. These specific targets led the FBI to codename the investigation UNABOM. As a result of the bombing campaign, three people were killed and many more were injured. In June 1995, *The New York Times*, *The Washington Post*, *Penthouse*, and *Scientific American* (as well as a Professor of Sociology at the University of California at Berkeley) received a manuscript – a terrorist's manifesto purportedly written by the Unabomber – entitled *An industrial society and its future*. Along with the manuscript was a deal: if the manuscript was published in full, the bombing would stop. *The Washington Post* eventually published the manuscript in September 1995 (Fitzgerald, 2004).

Upon reading an internet version of the published manifesto, Linda Patrik became unnerved. Although she had never met her brother-in-law, there was something about the text that seemed familiar. She asked her husband, David Kaczynski, to read the publication and urged him to compare it with her brother-in-law's writings. David sceptically complied, but started to suspect that his older brother, Theodore, may indeed be the Unabomber. The occurrence of one phrase in particular convinced him: *cool-headed logicians*. David recalled his brother "using that distinctive term on numerous occasions" (Fitzgerald, 2004: 208) and as a result, he contacted the FBI. To assist with the investigation, the Kaczynski family made available many documents known to have

DOI: 10.1057/9781137413758.0003

been written by Theodore for comparative analysis with the manifesto (p. 208).

James Fitzgerald of the FBI Behavioral Analysis Unit led a team of FBI agents and analysts in the comparative analysis of writings known to have been authored by the Unabomber (including so-called ruse letters which incited recipients to open accompanying bombing devices sent through the mail, ideological letters which outlined the Unabomber's rationale for the bombing campaign, brokering letters in which the Unabomber tried to get the manifesto published, and of course the manifesto itself) with writings known to have been authored by Kaczynski (which included, amongst others, Kaczynski's doctoral thesis, personal letters, and short stories) (Fitzgerald, 2004). In April 1996, after reviewing all of the evidence, including the report produced by the comparative analysis team, a federal judge signed a warrant to search Kaczynski's cabin in Montana. The FBI arrested Kaczynski at his home, whereupon they found "a virtual treasure trove of evidentiary materials" (p. 215) including a fully assembled bomb and numerous bomb parts.

In preparation for the trial, Kaczynski's defence team attempted to undermine the basis on which the search warrant had been obtained: "If they could get a ruling from the judge during the pre-trial stage that the search warrant was obtained by the FBI improperly, and that there was not enough probable cause to support the search, the entire case against Kaczynski could conceivably be dismissed" (Fitzgerald, 2004: 216). The comparative analysis team's report came under scrutiny by the defence's expert witness, Robin Lakoff, who outlined seven areas of error, and opined that the claims of common authorship between the known writings of the Unabomber and the known writings of Theodore Kackzynski were "untenable and unreliable at best" (Fitzgerald, 2004: 217). On the other hand, the prosecution's expert witness, Donald Foster, concluded that "Fitzgerald had cautiously understated the case for common authorship" (Foster, 2001: 107). According to Coulthard (2000), Lakoff argued that many of the lexical items which were shared across both groups of texts could "quite easily occur in any argumentative text" (p. 281) and were therefore not indicative of common authorship between the Unabomber's terrorist manifesto and Kaczynski's known writings. In particular, 12 words and phrases were selected for exemplification: *at any rate*, *clearly*, *in practice*, *gotten*, *more or less*, *moreover*, *on the other hand*, *presumably*, *propaganda*, *thereabouts*, and lexemes derived from the lemmas *argu* and *propos* (p. 281).

DOI: 10.1057/9781137413758.0003

To counter the proposal that such words and phrases could occur in any similar text, Coulthard (2000) reports that the web was searched and approximately three million documents were found which included at least one or more of the 12 lexical items. However, when the search was limited to finding only documents which contained all 12 lexical items, only 69 documents were identified and each of these documents was an online copy of the terrorist manifesto (p. 282). Coulthard concludes of this evidence that "a writer's combinations of lexical choices are more unique, diagnostic or idiolectal than people have so far been willing to believe" (p. 282). Furthermore, Coulthard (2004) argues that this evidence is "a powerful example of the idiolectal habit of co-selection and an illustration of the consequent forensic possibilities that idiolectal co-selection affords for authorship attribution" (p. 433).[2]

In order to judge this evidence as "powerful", two assertions must firstly be accepted: (1) that idiolectal co-selection – that is, words which taken in combination appear to characterise an individual author's linguistic style – is a useful marker of authorship; and (2) that the web is a valid and reliable corpus for producing forensic evidence. As will become clear in Section 1.3, the aim of this research is to test empirically both of these assertions, and since testing both assertions rests on the notion of idiolect and the use of lexis as a marker of authorship, it is firstly necessary to discuss both before setting out the scope of this research.

1.1 Idiolect

Although the term *idiolect* was first coined by Bloch (1948), Sapir (1927) laid the groundwork in his discussion of the relationship between speech and personality. Sapir outlined five levels of speech that were indexical of individual personality including voice, dynamics, pronunciation, vocabulary, and style. Of these, *vocabulary* and *style* are the most relevant precursors to the concept of idiolect. Sapir argued of vocabulary that:

> We do not all speak alike. There are certain words which some of us never use. There are other, favorite, words which we are always using... Individual variation exists, but it can properly be appraised only with reference to the social norm. Sometimes we choose words because we like them; sometimes we slight words because they bore or annoy or terrify us. We are not going to be caught by them. All in all, there is room for much subtle analysis in the determination of the social and individual significance of words. (p. 903)

DOI: 10.1057/9781137413758.0003

Here, Sapir clearly draws out the potential for individual variation at the lexical level (further considered in Section 1.2) and the complex relationship between the individual and society, which is further exemplified through his consideration of individual style:

> We all have our individual styles in both conversation and considered address, and they are never the arbitrary and casual things we think them to be. There is always an individual method, however poorly developed, of arranging words into groups and of working these up into larger units. It would be a very complicated problem to disentangle the social and individual determinants of style, but it is a theoretically possible one. (pp. 903–4)

Some linguists have given a more prominent place to writing alongside speech than Sapir (e.g. Coulthard, 2004) whilst for others, the term "style" is instead a recognised term for "idiolect in writing" (e.g. Kredens, 2002). In the context of this research, idiolect should be understood to include written language.

A good early definition for the discussion of idiolect is Hockett's (1958): "the totality of speech habits of a single person at a given time constitutes an idiolect" (p. 321). Hockett's definition raises two issues: potentially, one might need to observe and catalogue every single speech habit before one could fully characterise an individual's idiolect, and secondly that idiolect will change over time. In so far as *totality* means *complete* and *entire*, Hockett appears to suggest that idiolect is the entire repertoire of speech habits available to a single person. However, it is impossible to collect a totality, although for Hockett's purposes this would not have been an issue. In fact, in a later paragraph, Hockett notes that the entire idiolect cannot be observed, only examples of the linguistic output that it generates (1958: 322). In other words, rather than being able to observe the totality of habits, all that the linguist can observe is what a speaker or writer actually does at the particular point of observation.

The second implication of Hockett's definition, that idiolectal features can only be described *at a given time*, implies that idiolect is organic and evolutionary in nature and will differ when observed at different times. This raises the question of by how much and whether the difference is significant. Related to this is the issue of the rate at which such change occurs, a question which so far has received no definitive answer with the exception of Bel et al. (2012) who show that the use of bigrams and trigrams do not vary substantially across a span of between six and ten years for individual authors (ages unknown), indicating that this feature remains sufficiently stable for this limited period of time at least. Corroborative evidence is provided by Barlow (2010) who found that bigrams were used consistently over the shorter period of one year in the spoken language of White House Press Secretaries. The problem with such a definition for forensic purposes is that not only would the idiolect of one individual differ from that of another (as assumed in authorship attribution, although cf. Grant (2010) for an alternative view on the importance of idiolect for forensic purposes), but would also be subject to variation between the same individual when observed at different points. This would make the comparison of documents in the forensic context very difficult because Known Documents (those documents whose authorship is attested, henceforth KD) are rarely authored at the same time as each other or as the Questioned Documents (those documents whose authorship is unknown and under suspicion, henceforth QD).

Sixty years later, Louwerse (2004) claimed that writers "implicitly leave their signature in the document they write" and that idiolects "are person-dependent similarities in language use" (p. 207). He explains that if idiolect exists, texts composed by one author will show more similarities in language than texts composed by different authors (p. 207). However, a potential problem arises in relation to Hockett's definition. Louwerse states that similarities between texts produced by one author will be greater than texts produced by different authors. Hockett proposes that idiolect will change over time. Unless the individual signatures upon which Louwerse's definition relies remain static, the similarities between two pieces of writing by the same individual at different times could be no greater than the similarities between two individuals with similar linguistic backgrounds (a common assumption, e.g. Loakes, 2006). It seems then that the temporal dimension could indeed be a confounding variable in forensic authorship attribution. Through examining a third definition of idiolect, a clearer picture may be gained.

Coulthard (2004) also says that "every native speaker has their own distinct and individual version of the language they speak and write, their own *idiolect*" and that "this *idiolect* will manifest itself through distinctive and idiosyncratic choices in texts" (pp. 431–2, original emphasis). The main difference here is between what Coulthard refers to as *choice* and what Hockett refers to as *habit*. Insofar as *choice* implies conscious decision, *habit* implies an involuntary behaviour pattern. If idiolect is based on habit, it is reasonable to argue that a person's linguistic patterns will remain constant, until such a time when that habit is changed. In this scenario, texts produced during a period when the habit remains the same should be comparable. Choice, however, is more volatile and dependent on many extra-linguistic factors (e.g. mood of the individual, genre of the text, audience of the text, time available to compose the text, and indeed recency) as well as conscious attempts to disguise identity. As such, any features of language that are subject to choice could result in differences between texts produced by the same author, regardless of when they were authored.

These three definitions, somewhat representative of the many that could have been reviewed (e.g. Labov, 1972; Trudgill, 1974, 2003; Wardhaugh, 2006) capture between them the key issue for authorship attribution, namely, the extent to which an individual's idiolect really is a reliable signature irrespective of stylistic choice and change over time. However, it is readily acknowledged that the theory of idiolect, to date, lacks empirical investigation (e.g. Kredens, 2001, 2002; Louwerse, 2004; Kniffka, 2007), and, as discussed above, the totality of linguistic habits for each person can never fully be observed. This point is echoed by Coulthard (2004) who says that "any linguistic sample, even a very large one, provides only very partial information about its creator's idiolect" (p. 432), so it remains a largely theoretical notion. In light of its theoretical basis, it is now necessary to be explicit about how idiolect will be understood and applied in this research. From reviewing the definitions of idiolect provided by Hockett (1958), Louwerse (2004), and Coulthard (2004), the consensus seems to be that either choice, habit, or both are intrinsically linked to idiolect. The definition to be used in this research is therefore as follows:

> Determined and conditioned by a wide and immeasurable range of biological, sociological, cognitive and environmental factors (including *inter alia* age, IQ, occupation, friendship networks, language contact), idiolect is the combination of language choices (planned features) and habits (subconscious features) made by an individual, the sum of which creates a distinctive, albeit oftentimes overlapping, range of choices and habits from another individual. (Tomblin, 2013: 35)

In this conceptualisation of idiolect, the goal of the forensic linguist is to identify those features of idiolect which overlap less with others in order to demonstrate the similarity or difference between a series of authors. It will not be possible to determine in this research what constitutes a choice and what constitutes a habit, but one might surmise that a feature such as using a specific lexical item to mark identity (such as youth vernacular words

DOI: 10.1057/9781137413758.0003