Full text available at: http://dx.doi.org/10.1561/1500000005

Authorship Attribution

Full text available at: http://dx.doi.org/10.1561/1500000005

Authorship Attribution

Patrick Juola

Department of Mathematics and Computer Science
Duquesne University
Pittsburgh, PA 15282
USA
juola@mathcs.duq.edu

now
the essence of knowledge

Boston – Delft

Full text available at: http://dx.doi.org/10.1561/1500000005

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is P. Juola, Authorship Attribution, Foundation and Trends® in Information Retrieval, vol 1, no 3, pp 233–334, 2006

ISBN: 978-1-60198-118-9
© 2008 P. Juola

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1-781-871-0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Full text available at: http://dx.doi.org/10.1561/1500000005

Foundations and Trends® in Information Retrieval

Volume 1 Issue 3, 2006

Editorial Board

Editors-in-Chief:
Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche
fabrizio.sebastiani@isti.cnr.it

Jamie Callan
Carnegie Mellon University
callan@cmu.edu

Editors
Alan Smeaton (Dublin City University)
Andrei Z. Broder (Yahoo! Research)
Bruce Croft (University of Massachusetts, Amherst)
Charles L.A. Clarke (University of Waterloo)
Ellen Voorhees (National Institute of Standards and Technology)
Ian Ruthven (University of Strathclyde, Glasgow)
James Allan (University of Massachusetts, Amherst)
Justin Zobel (RMIT University, Melbourne)
Maarten de Rijke (University of Amsterdam)
Marcello Federico (ITC-irst)
Norbert Fuhr (University of Duisburg-Essen)
Soumen Chakrabarti (Indian Institute of Technology)
Susan Dumais (Microsoft Research)
Wei-Ying Ma (Microsoft Research Asia)
William W. Cohen (CMU)

Full text available at: http://dx.doi.org/10.1561/1500000005

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2006, Volume 1, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/1500000005

Foundations and Trends$^{\circledR}$ in
Information Retrieval
Vol. 1, No. 3 (2006) 233–334
© 2008 P. Juola
DOI: 10.1561/1500000005

the essence of knowledge

Authorship Attribution

Patrick Juola

Department of Mathematics and Computer Science, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282, USA, juola@mathcs.duq.edu

Abstract

Authorship attribution, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and a wide range of application. Recent work in "non-traditional" authorship attribution demonstrates the practicality of automatically analyzing documents based on authorial style, but the state of the art is confusing. Analyses are difficult to apply, little is known about type or rate of errors, and few "best practices" are available. In part because of this confusion, the field has perhaps had less uptake and general acceptance than is its due.

This review surveys the history and present state of the discipline, presenting some comparative results when available. It shows, first, that the discipline is quite successful, even in difficult cases involving small documents in unfamiliar and less studied languages; it further analyzes the types of analysis and features used and tries to determine characteristics of well-performing systems, finally formulating these in a set of recommendations for best practices.

Full text available at: http://dx.doi.org/10.1561/1500000005

Contents

Full text available at: http://dx.doi.org/10.1561/1500000005

Full text available at: http://dx.doi.org/10.1561/1500000005

1

Introduction

1.1 Why “Authorship Attribution”?

In 2004, Potomac Books published *Imperial Hubris: Why the West is Losing the War on Terror*. Drawing on the author’s extensive personal experience, the book described the current situation of the American-led war on terror and argued that much US policy was misguided.

Or did he? The author of the book is technically “Anonymous,” although he claims (on the dust cover) to be “a senior US intelligence official with nearly two decades of experience” as well as the author of the 2003 book *Through Our Enemies*’ *Eyes*. According to the July 2, 2004 edition of the Boston Phoenix, the actual author was Michael Scheuer, a senior CIA officer and head of the CIA’s Osama bin Laden unit in the late 1990s. If true, this would lend substantial credibility to the author’s arguments.

But on the other hand, according to some noted historians such as Hugh Trevor-Roper, the author of the 1983 *Hitler Diaries* was Hitler himself, despite the later discovery that they were written on modern paper and using ink which was unavailable in 1945. Is *Imperial Hubris* another type of sophisticated forgery? Why should we believe historians

Full text available at: http://dx.doi.org/10.1561/1500000005

and journalists, no matter how eminent? What kind of evidence should we demand before we believe?

Determining the author of a particular piece of text has raised methodological questions for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers as in the examples above. Investigative journalism, combined with scientific (e.g., chemical) analysis of documents and simple close reading by experts has traditionally given good results. But recent developments of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option. This field has seen an explosion of scholarship, including several detailed book-length treatments [39, 41, 44, 83, 98, 103, 105, 111, 112, 150]. Papers on authorship attribution routinely appear at conference ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite — or perhaps because of — this interest, the field itself is somewhat in disarray with little overall sense of best practices and techniques.

1.2 Structure of the Review

This review therefore tries to present an overview and survey of the current state of the art. We follow the theoretical model (presented in detail in Section 3.4) of [76] in dividing the task into three major subtasks, each treated independently.

Section 2 presents a more detailed problem statement in conjunction with a historical overview of some approaches and major developments in the science of authorship attribution. Included is a discussion of some of the major issues and obstacles that authorship attribution faces as a problem, without regard to any specific approach, and the characteristics of a hypothetical "good" solution (unfortunately, as will be seen in the rest of the review, we have not yet achieved such a "good" solution).

Section 3 presents some linguistic, mathematical, and algorithmic preliminaries. Section 4 describes some of the major feature sets that

Full text available at: http://dx.doi.org/10.1561/1500000005

have been applied to authorship attribution, while Section 5 describes the methods of analysis applied to these features. Section 6 goes on to present some results in empirical evaluation and comparative testing of authorship attribution methods, focusing mainly on the results from the 2004 *Ad-hoc Authorship Attribution Competition* [75], the largest-scale comparative test to date.

Section 7 presents some other applications of these methods and technology, that, while not (strictly speaking) "authorship" attribution, are closely related. Examples of this include gender attribution or the determination of personality and mental state of the author. Section 8 discusses the specific problems of using authorship attribution in court, in a forensic setting. Finally, for those practical souls who want only to solve problems, Section 9 presents some recommendations about the current state of the art and the best practices available today.

Full text available at: http://dx.doi.org/10.1561/1500000005

References

[1] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 6, pp. 67–75, 2005.

[2] A. Abbasi and H. Chen, *Visualizing Authorship for Identification,* pp. 60–71. Springer, 2006.

[3] American Board of Forensic Document Examiners, "Frequently asked questions," http://www.abfde.org/FAQs.html, accessed January 6, 2007.

[4] Anonymous, "Some anachronisms in the January 4, 1822 Beale letter," http://www.myoutbox.net/bch2.htm, accessed May 31, 2007, 1984.

[5] S. Argamon and S. Levitan, "Measuring the usefulness of function words for authorship attribution," in *Proceedings of ACH/ALLC 2005*, Association for Computing and the Humanities, Victoria, BC, 2005.

[6] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Proceedings of the Classification Society of North America Annual Meeting*, 2005.

[7] A. Argamon-Engleson, M. Koppel, and G. Avneri, "Style-based text categorization: What newspaper am I reading," in *Proceedings of the AAAI Workshop of Learning for Text Categorization*, pp. 1–4, 1998.

[8] R. H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie, "An experiment in authorship attribution," in *Proceedings of JADT 2002*, pp. 29–37, Université de Rennes, St. Malo, 2002.

[9] R. H. Baayen, H. Van Halteren, and F. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, pp. 121–131, 1996.

Full text available at: http://dx.doi.org/10.1561/1500000005

[10] R. E. Bee, "Some methods in the study of the Masoretic text of the Old Testament," *Journal of the Royal Statistical Society*, vol. 134, no. 4, pp. 611–622, 1971.

[11] R. E. Bee, "A statistical study of the Pinai Pericope," *Journal of the Royal Statistical Society*, vol. 135, no. 3, pp. 391–402, 1972.

[12] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, no. 4, p. 048072, 2002.

[13] D. Biber, S. Conrad, and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

[14] J. N. G. Binongo, "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9–17, 2003.

[15] A. F. Bissell, "Weighted cumulative sums for text analysis using word counts," *Journal of the Royal Statistical Society A*, vol. 158, pp. 525–545, 1995.

[16] E. Brill, "A corpus-based approach to language learning," PhD thesis, University of Pennsylvania, 1993.

[17] C. Brown, M. A. Covington, J. Semple, and J. Brown, "Reduced idea density in speech as an indicator of schizophrenia and ketamine intoxication," in *International Congress on Schizophrenia Research*, Savannah, GA, 2005.

[18] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, pp. 79–85, June 1990.

[19] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 955–974, 1998.

[20] J. F. Burrows, "'An ocean where each kind…': Statistical analysis and some major determinants of literary style," *Computers and the Humanities*, vol. 23, no. 4–5, pp. 309–21, 1989.

[21] J. F. Burrows, "Delta: A measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, pp. 267–287, 2002.

[22] J. F. Burrows, "Questions of authorship: Attribution and beyond," *Computers and the Humanities*, vol. 37, no. 1, pp. 5–32, 2003.

[23] F. Can and J. M. Patton, "Change of writing style with time," *Computers and the Humanities*, vol. 28, no. 4, pp. 61–82, 2004.

[24] D. Canter, "An evaluation of 'Cusum' stylistic analysis of confessions," *Expert Evidence*, vol. 1, no. 2, pp. 93–99, 1992.

[25] C. E. Chaski, "Empirical evaluations of language-based author identification," *Forensic Linguistics*, vol. 8, no. 1, pp. 1–65, 2001.

[26] C. E. Chaski, "Who's at the keyboard: Authorship attribution in digital evidence invesigations," *International Journal of Digital Evidence*, vol. 4, no. 1, p. n/a, Electronic-only journal: http://www.ijde.org, accessed May 31, 2007, 2005.

[27] C. E. Chaski, "The keyboard dilemma and forensic authorship attribution," *Advances in Digital Forensics III*, 2007.

[28] D. Coniam, "Concordancing oneself: Constructing individual textual profiles," *International Journal of Corpus Linguistics*, vol. 9, no. 2, pp. 271–298, 2004.

Full text available at: http://dx.doi.org/10.1561/1500000005

[29] M. Corney, O. de Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *Proceedings of Computer Security Applications Conference*, pp. 282–289, 2002.

[30] H. Craig "Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?" *Literary and Linguistic Computing*, vol. 14, no. 1, pp. 103–113, 1999.

[31] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the Third Conference on Applied Natural Lanugage Processing*, Association for Computational Linguistics, Trento, Italy, April 1992. Also available as Xerox PARC technical report SSL-92-01.

[32] A. de Morgan, "Letter to Rev. Heald 18/08/1851," in *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*, (S. Elizabeth and D. Morgan, eds.), London: Longman's Green and Co., 1851/1882.

[33] G. Easson, "The linguistic implications of shibboleths," in *Annual Meeting of the Canadian Linguistics Association*, Toronto, Canada, 2002.

[34] A. Ellegard, *A Statistical Method for Determining Authorship: The Junius Leters 1769–1772*. Gothenburg, Sweden: University of Gothenburg Press, 1962.

[35] W. Elliot and R. J. Valenza, "And then there were none: Winnowing the Shakespeare claimants," *Computers and the Humanities*, vol. 30, pp. 191–245, 1996.

[36] W. Elliot and R. J. Valenza, "The professor doth protest too much, methinks," *Computers and the Humanities*, vol. 32, pp. 425–490, 1998.

[37] W. Elliot and R. J. Valenza, "So many hardballs so few over the plate," *Computers and the Humanities*, vol. 36, pp. 455–460, 2002.

[38] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence," in *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 48–57, San Francisco, California, January 22–24, 1995.

[39] J. M. Farringdon, *Analyzing for Authorship: A Guide to the Cusum Technique*. Cardiff: University of Wales Press, 1996.

[40] R. S. Forsyth, "Towards a text benchmark suite," in *Proceedings of 1997 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 1997)*, Kingston, ON, 1997.

[41] D. Foster, *An Elegy by W.S.: A Study in Attribution*. Newark: University of Delaware Press, 1989.

[42] D. Foster, "Attributing a funeral elegy," *PMLA*, vol. 112, no. 3, pp. 432–434, 1997.

[43] D. Foster, *Author Unknown: Adventures of a Literary Detective*. London: Owl Books, 2000.

[44] D. Foster, *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company, 2001.

Full text available at: http://dx.doi.org/10.1561/1500000005

[45] W. Fucks, "On the mathematical analysis of style," *Biometrika*, vol. 39, pp. 122–129, 1952.

[46] J. Gibbons, *Forensic Linguistics: An Introduction to Language in the Justice System*. Oxford: Blackwell, 2003.

[47] N. Graham, G. Hirst, and B. Marthi, "Segmenting documents by stylistic character," *Natural Language Engineering*, vol. 11, pp. 397–415, 2005.

[48] T. Grant and K. Baker, "Identifying reliable, valid markers of authorship: A reponse to Chaski," *Forensic Linguistics*, vol. 8, no. 1, pp. 66–79, 2001.

[49] T. R. G. Green, "The necessity of syntax markers: Two experiments with artificial languages," *Journal of Verbal Learning and Verbal Behavior*, vol. 18, pp. 481–96, 1979.

[50] J. W. Grieve, "Quantitative authorship attribution: A history and an evaluation of techniques". Master's thesis, Simon Fraser University, 2005. URL: http://hdl.handle.net/1892/2055, accessed May 31, 2007.

[51] J. Hancock, "Digital deception: When, where and how people lie online," in *Oxford Handbook of Internet Psychology*, (K. McKenna, T. Postmes, U. Reips, and A. Joinson, eds.), pp. 287–301, Oxford: Oxford University Press, 2007.

[52] R. A. Hardcastle, "Forensic linguistics: An assessment of the Cusum method for the determination of authorship," *Journal of the Forensic Science Society*, vol. 33, no. 2, pp. 95–106, 1993.

[53] R. A. Hardcastle, "Cusum: A credible method for the determination of authorship?," *Science and Justice*, vol. 37, no. 2, pp. 129–138, 1997.

[54] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison Wesley, 1991.

[55] M. L. Hilton and D. I. Holmes, "An assessment of cumulative sum charts for authorship attribution," *Literary and Linguistic Computing*, vol. 8, pp. 73–80, 1993.

[56] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[57] D. I. Holmes, "The evolution of stylometry in humanities computing," *Literary and Linguistic Computing*, vol. 13, no. 3, pp. 111–117, 1998.

[58] D. I. Holmes and R. S. Forsyth, "The Federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111–127, 1995.

[59] D. I. Holmes, "Stylometry and the civil war: The case of the Pickett letters," *Chance*, vol. 16, no. 2, pp. 18–26, 2003.

[60] D. I. Holmes and F. J. Tweedie, "Forensic stylometry: A review of the CUSUM controversy," in *Revue Informatique et Statistique dans les Science Humaines*, pp. 19–47, University of Liege, Liege, Belgium, 1995.

[61] D. Hoover, "Another perspective on vocabulary richness," *Computers and the Humanities*, vol. 37, no. 2, pp. 151–178, 2003.

[62] D. Hoover, "Stylometry, chronology, and the styles of Henry James," in *Proceedings of Digital Humanities 2006*, pp. 78–80, Paris, 2006.

[63] D. L. Hoover, "Delta prime?," *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 477–495, 2004.

[64] D. L. Hoover, "Testing Burrows's Delta," *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 453–475, 2004.

Full text available at: http://dx.doi.org/10.1561/1500000005

[65] J. Hopcroft and J. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley, 1979.

[66] S. R. Hota, S. Argamon, M. Koppel, and I. Zigdon, "Performing gender: Automatic stylistic analysis of Shakespeare's characters," in *Proceedings of Digital Humanities 2006*, pp. 100–104, Paris, 2006.

[67] IGAS, "IGAS — Our Company," http://www.igas.com/company.asp, accessed May 31, 2007.

[68] M. P. Jackson, "Function words in the 'funeral elegy'," *The Shakespeare Newsletter*, vol. 45, no. 4, p. 74, 1995.

[69] T. Joachims, *Learning to Classify Text Using Support Vector Machines*. Kluwer, 2002.

[70] E. Johnson, *Lexical Change and Variation in the Southeastern United States 1930–1990*. Tuscaloosa, AL: University of Alabama Press, 1996.

[71] P. Juola, "What can we do with small corpora? Document categorization via cross-entropy," in *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK, 1997.

[72] P. Juola, "Cross-entropy and linguistic typology," in *Proceedings of New Methods in Language Processing and Computational Natural Language Learning*, (D. M. W. Powers, ed.), Sydney, Australia: ACL, 1998.

[73] P. Juola, "Measuring linguistic complexity: The morphological tier," *Journal of Quantitative Linguistics*, vol. 5, no. 3, pp. 206–213, 1998.

[74] P. Juola, "The time course of language change," *Computers and the Humanities*, vol. 37, no. 1, pp. 77–96, 2003.

[75] P. Juola, "*Ad-hoc* authorship attribution competition," in *Proceedings of 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.

[76] P. Juola, "On composership attribution," in *Proceedings of 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.

[77] P. Juola, "Compression-based analysis of language complexity," Presented at *Approaches to Complexity in Language*, 2005.

[78] P. Juola, "Authorship attribution for electronic documents," in *Advances in Digital Forensics II*, (M. Olivier and S. Shenoi, eds.), pp. 119–130, Boston: Springer, 2006.

[79] P. Juola, "Becoming Jack London," *Journal of Quantitative Linguistics*, vol. 14, no. 2, pp. 145–147, 2007.

[80] P. Juola and H. Baayen, "A controlled-corpus experiment in authorship attribution by cross-entropy," *Literary and Linguistic Computing*, vol. 20, pp. 59–67, 2005.

[81] P. Juola, J. Sofko, and P. Brennan, "A prototype for authorship attribution studies," *Literary and Linguistic Computing*, vol. 21, no. 2, pp. 169–178, Advance Access published on April 12, 2006; doi: doi:10.1093/llc/fql019, 2006.

Full text available at: http://dx.doi.org/10.1561/1500000005

[82] G. Kacmarcik and M. Gamon, "Obfuscating document stylometry to preserve author anonymity," in *Proceedings of ACL 2006*, 2006.

[83] A. Kenny, *The Computation of Style*. Oxford: Pergamon Press, 1982.

[84] V. Keselj and N. Cercone, "CNG method with weighted voting," in *Ad-hoc Authorship Attribution Contest*, (P. Juola, ed.), ACH/ALLC 2004, 2004.

[85] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING03*, pp. 255–264, Dalhousie University, Halifax, NS, August 2003.

[86] K. Keune, M. Ernestus, R. van Hout, and H. Baayen, "Social, geographical, and register variation in Dutch: From written MOGELIJK to spoken MOK," in *Proceedings of ACH/ALLC 2005*, Victoria, BC, Canada, 2005.

[87] D. V. Khmelev and F. J. Tweedie, "Using markov chains for identification of writers," *Literary and Linguistic Computing*, vol. 16, no. 3, pp. 299–307, 2001.

[88] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1519–1525, 2006.

[89] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, doi:10.1093/llc/17.4.401, 2002.

[90] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003.

[91] M. Koppel and J. Schler, "*Ad-hoc* authorship attribution competition approach outline," in *Ad-hoc Authorship Attribution Contest*, (P. Juola, ed.), ACH/ALLC 2004, 2004.

[92] L. Kruh, "A basic probe of the Beale cipher as a bamboozlement: Part I," *Cryptologia*, vol. 6, no. 4, pp. 378–382, 1982.

[93] L. Kruh, "The Beale cipher as a bamboozlement: Part II," *Cryptologia*, vol. 12, no. 4, pp. 241–246, 1988.

[94] H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English*. Providence: Brown University Press, 1967.

[95] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," *Lecture Notes in Computer Science*, vol. 4243, p. 274283, 2006.

[96] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, "Using literal and grammatical statistics for authorship attribution," *Problemy Peredachi Informatii*, vol. 37, no. 2, pp. 96–198, Translated in "Problems of Information Transmission," pp. 172–184, 2000.

[97] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications. Graduate Texts in Computer Science*, New York: Springer, second ed., 1997.

[98] H. Love, *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press, 2002.

[99] C. Martindale and D. McKenzie, "On the utility of content analysis in authorship attribution: The Federalist Papers," *Computers and the Humanities*, vol. 29, pp. 259–70, 1995.

Full text available at: http://dx.doi.org/10.1561/1500000005

[100] R. A. J. Matthews and T. V. N. Merriam, "Neural computation in stylometry I: An application to the works of Shakespeare and Marlowe," *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 203–209, 1993.

[101] J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1987.

[102] G. McMenamin, "Disputed authorship in US law," *International Journal of Speech, Language and the Law*, vol. 11, no. 1, pp. 73–82, 2004.

[103] G. R. McMenamin, *Forensic Stylistics*. London: Elsevier, 1993.

[104] G. R. McMenamin, "Style markers in authorship studies," *Forensic Linguistics*, vol. 8, no. 2, pp. 93–97, 2001.

[105] G. R. McMenamin, *Forensic Linguistics — Advances in Forensic Stylistics*. Boca Raton, FL: CRC Press, 2002.

[106] T. C. Mendenhall, "The characteristic curves of composition," *Science*, vol. IX, pp. 237–249, 1887.

[107] T. V. N. Merriam and R. A. J. Matthews, "Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe," *Literary and Linguistic Computing*, vol. 9, no. 1, pp. 1–6, 1994.

[108] G. Monsarrat, "A funeral elegy: Ford, W.S., and Shakespeare," *Review of English Studies*, vol. 53, p. 186, 2002.

[109] A. W. Moore, "Support Vector Machines," Online tutorial: http://jmvidal.cse.sc.edu/csce883/svm14.pdf, accessed May 31, 2007, 2001.

[110] J. L. Morgan, *From Simple Input to Complex Grammar*. Cambridge, MA: MIT Press, 1986.

[111] A. Q. Morton, *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribner's, 1978.

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

[113] M. Newman, J. Pennebaker, D. Berry, and J. Richards, "Lying words: Predicting deception from linguistic style," *Personality and Social Psychology Bulletin*, vol. 29, pp. 665–675, 2003.

[114] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large scale personality classifiation of personal weblogs," in *International Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[115] M. Oakes, "Text categorization: Automatic discrimination between US and UK English using the chi-square text and high ratio pairs," *Research in Language*, vol. 1, pp. 143–156, 2003.

[116] J. Oberlander and S. Nowson, "Whose thumb is it anyway? classifying author personality from weblog text," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, pp. 627–634, Sydney, Australia, 2006.

[117] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, "Language independent authorship attribution using character level language models," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 267–274, Budapest: ACL, 2003.

Full text available at: http://dx.doi.org/10.1561/1500000005

[118] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *Journal of Personality and Social Psychology*, vol. 77, pp. 1296–1312, 1999.

[119] J. W. Pennebaker and L. D. Stone, "Words of wisdom: Language use over the life span," *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 291–301, 2003.

[120] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, ourselves," *Annual Review of Psychology*, vol. 54, pp. 547–577, 2003.

[121] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1993.

[122] M. Rockeach, R. Homant, and L. Penner, "A value analysis of the disputed Federalist Papers," *Journal of Personality and Social Psychology*, vol. 16, pp. 245–250, 1970.

[123] S. Rude, E. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition and Emotion*, vol. 18, pp. 1121–1133, 2004.

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

[125] J. Rudman, "Non-traditional authorship attribution studies in eighteenth century literature: Stylistics, statistics and the computer," URL: http://computerphilologie.uni-muenchen.de/jg02/rudman.html, accessed May 31, 2007.

[126] J. Rudman, "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.

[127] J. Rudman, "The non-traditional case for the authorship of the twelve disputed Federalist Papers: A monument built on sand," in *Proceedings of ACH/ALLC 2005*, Association for Computing and the Humanities, Victoria, BC, 2005.

[128] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 318–362, The MIT Press, 1986.

[129] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 379–423, 1948.

[130] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[131] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, p. 688, 1949.

[132] S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*. Anchor, 2000.

[133] M. Smith, "Recent experiences and new developments of methods for the determination of authorship," *Association of Literary and Linguistic Computing Bulletin*, vol. 11, pp. 73–82, 1983.

[134] H. H. Somers, "Statistical methods in literary analysis," in *The Computer and Literary Style*, (J. Leed, ed.), Kent, OH: Kent State University Press, 1972.

Full text available at: http://dx.doi.org/10.1561/1500000005

[135] H. Somers, "An attempt to use weighted cusums to identify sublanguages," in *Proceedings of New Methods in Language Processing 3 and Computational Natural Langauge Learning*, (D. M. W. Powers, ed.), Sydney, Australia: ACL, 1998.

[136] H. Somers and F. Tweedie, "Authorship attribution and pastiche," *Computers and the Humanities*, vol. 37, pp. 407–429, 2003.

[137] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," *Computers and the Humanities*, vol. 35, no. 2, pp. 193–214, 2001.

[138] S. Stein and S. Argamon, "A mathematical explanation of Burrows' Delta," in *Proceedings of Digital Humanities 2006*, Paris, France, July 2006.

[139] D. R. Tallentire, "Towards an archive of lexical norms — a proposal," in *The Computer and Literary Studies*, Cardiff: Unversity of Wales Press, 1976.

[140] S. Thomas, "Attributing *a funeral elegy*," *PMLA*, vol. 112, no. 3, p. 431, 1997.

[141] E. Tufte, *Envisioning Information*. Graphics Press, 1990.

[142] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The Federalist Papers," *Computers and the Humanities*, vol. 30, no. 1, pp. 1–10, 1996.

[143] L. Ule, "Recent progress in computer methods of authorship determination," *Association for Literary and Linguistic Computing Bulletin*, vol. 10, pp. 73–89, 1982.

[144] H. van Halteren, "Author verification by linguistic profiling: An exploration of the parameter space," *ACM Transactions on Speech and Language Processing*, vol. 4, 2007.

[145] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt, "New machine learning methods demonstrate the existence of a human stylome," *Journal of Quantitative Linguistics*, vol. 12, no. 1, pp. 65–77, 2005.

[146] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.

[147] W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 2002.

[148] B. Vickers, *Counterfeiting Shakespeare*. Cambridge: Cambridge University Press, 2002.

[149] F. L. Wellman, *The Art of Cross-Examination*. New York: MacMillan, fourth ed., 1936.

[150] C. B. Williams, *Style and Vocabulary: Numerical Studies*. London: Griffin, 1970.

[151] A. J. Wyner, "Entropy estimation and patterns," in *Proceedings of the 1996 Workshop on Information Theory*, 1996.

[152] B. Yu, Q. Mei, and C. Zhai, "English usage comparison between native and non-native english speakers in academic writing," in *Proceedings of ACH/ALLC 2005*, Victoria, BC, Canada, 2005.

[153] G. U. Yule, "On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship," *Biometrika*, vol. 30, pp. 363–90, 1938.

Full text available at: http://dx.doi.org/10.1561/1500000005

[154] G. U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1944.

[155] P. M. Zatko, "Alternative routes for data acquisition and system compromise," in *3rd Annual IFIP Working Group 11.9 International Conference on Digital Forensics*, Orlando, FL, 2007.

[156] H. Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, (V. Barr and Z. Markov, eds.), Miami Beach, FL: AAAI Press, 2004.

[157] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[158] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. New York: Hafner Publishing Company, 1949. Reprinted 1965.