# Stilometrie in der Diplomatik: Ein neues Forschungsfeld?

**Geißel, Pia**
geissel@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Seit rund 20 Jahren steigt die online verfügbare Menge an digitalen Daten in unterschiedlichster Form und Qualität. Dies erleichtert den Geisteswissenschaftler|innen den Zugang zu neuen Materialien und senkt die Hemmschwelle sich mit Big Data zu beschäftigen und der „methodological monoculture of close reading"(Karsdorp et al. 2021, *preface*) zu entkommen. Dadurch müssen sie sich aber auch vermehrt mit den technischen, sprich, computergestützten Anwendungen und Programmiersprachen beschäftigen. Zudem müssen Daten meist noch gesammelt, von überflüssiger Information bereinigt und Ergebnisse visuell aufbereitet werden. Alle diese neuen Schritte sind je nach Datengrundlage aufwändig, da Techniken erst erlernt und korrekt angewendet werden müssen. Als zusätzlicher Faktor werden für Berechnungen großer Datenmengen fachfremde Methoden aus der Mathematik oder Informatik entlehnt. Die Anwendung dieser arithmetischen Methoden werden jedoch oftmals nicht ausreichend hinterfragt. Zusätzlich fehlt zur Überprüfung der Methoden auch die empirische Evidenz vor allem dann, wenn das Untersuchungsmaterial keine faktischen Hinweise liefern kann.

Auch in der Geschichtswissenschaft wenden sich, angeregt durch niedrigschwelligen Zugang zu online verfügbaren Texten, Geisteswissenschaftler:innen neuen Forschungsfragen zu. So eröffnet beispielsweise die Stilometrie neue Möglichkeiten, die Anonymität eines mittelalterlichen Textzeugens aufzuheben und neue Thesen bezüglich Überlieferung und Organisation in Schreibstuben und Kanzleien aufzustellen. Dabei bewegen sich jedoch erstens die Forschungsfragen häufig um eine konkrete Identifizierung eines Stiles, einer Kanzlei oder einer Schreibschule und weniger über Makrosignale wie übergeordnete geographische, kulturelle oder sprachliche Dimensionen. Zweitens vernachlässigt die Arbeit am konkreten Korpus auch die Auseinandersetzung mit der Auswahl der geeigneten mathematischen Verfahren, die hinter den computergestützten Berechnungen stehen. Eine Auseinandersetzung auf der Makroebene über die Effektivität von beispielsweise einem Distanzmaß-Verfahren wie Burrows's Delta auf die konkrete Textgattung der Urkunden findet bis heute noch nicht statt. Zwar gibt es Messungen über den Wirkungsgrad des Verfahrens für Lyrik und Prosa in lateinischer Sprache und kurzer oder auch fehlerhafter Texte[1], dennoch finden darin die spezifischen Eigenschaften von Urkunden nicht ausreichend Berücksichtigung: Individuelle, orthographische Signale, formelhafte Sprache und lückenhafte Überlieferungen sind nur einige der spezifischen Faktoren, welche noch nicht ausreichend für Burrow's Delta untersucht wurden (vgl. Eder 2013, 2015).

Betrachtet man das weitere Forschungsfeld zum Thema Autorschaftsattributionen werden aktuell vermehrt andere statistische Ansätze in Erwägung gezogen, die nicht zwingend auf Textfeatures wie Wortlängen oder inhaltlichen Merkmale wie Worthäufigkeiten basieren. Eine Fokussierung auf syntaktische oder semantische Zusammenhänge beispielsweise könnte bei der Untersuchung der Signale in Urkunden ebenso Distinktionen herausheben. Eine Loslösung vom Vector-Space-Model und Burrows Delta hin zu Topics und Neuronalen Netzen wurde zwar bisher auf dem Typus Urkunden noch nicht im großen Rahmen angewandt, dennoch könnte man dadurch potentiell das Manko der geringen Textlängen und der formelhaften Sprache umgehen. Neuere Ansätze haben sich zudem das Ziel gesetzt, durch eine Kombination mehrerer methodischer und textimmanenter Ansätze Merkmale herauszuarbeiten. Diese umschließen dann folglich nicht mehr nur die klassische Stilometrie, sondern auch die oben genannten syntaktischen und semantischen Features. Ob diese neuen Ansätze bei Urkunden Wirkung zeigen, soll ein Ziel dieser Dissertation werden.

Ein grundsätzliches Problem bezüglich der Urkundentexte ist allerdings, dass diese häufig nicht von ihren originalen Textzeugen, sondern nur aus digitalisierten Editionen entnommen sind, die nicht dem eigentlichen Überlieferungstext entsprechen müssen. Wie sehr vertraut man Texten aus älteren Editionen vor 1945, in denen die Lachmannsche Editionstechnik angewendet wurde, mit der die *emendatio* angeblicher Fremdeingriffe unbedarft angewendet und zudem schlecht oder gänzlich undokumentiert in den Druck gegeben wurde (vgl. Plachta 2006, S. 29)?

Aus diesen Überlegungen heraus ist es naheliegend, die stilometrischen Verfahren einmal aus der Makroperspektive zu untersuchen: Anstatt sich mit einer These zu beschäftigen, die aus dem konkreten Material, vielleicht sogar aus dem close reading-Prozess selbst, entstanden ist, sollten die mathematischen Methoden an einer großen und diversen Menge an Urkundenmaterial untersucht werden. Eine mannigfaltige Auswahl bieten dazu mehrere Urkundenportale, wie monasterium.net, Cartae Europae Medii Aevi, Codice diplomatico della Lombardia medievale oder Telma.[2] Dabei spielen zunächst weder die Urkundentypen, noch die zeitliche Dimension, in der die Urkunden entstanden sind, eine Rolle. Die Annahme ist eher, dass verschiedene Verfahren unterschiedlich starke Distinktionen der Urkunden unterstreichen. Ihre jeweilige Sensibilität für bestimmte Eigenschaften des Textmaterials gilt es herauszuarbeiten und methodisch zu begründen. So lassen sich Stärken und Schwächen der Methoden analysieren und gegenüberstellen, nicht im Sinne eines Rankings (‚bessere vs schlechtere Methode'), sondern im Hinblick auf ihre Eignung für spezifische Korpora und Fragestellungen. Ein solcher Ansatz erlaubt es Forschenden nicht nur, eine fundiertere Entscheidung über die anzuwendende Methode zu treffen, sondern im Idealfall sogar, diese gezielt auf den eigenen Anwendungsfall abzustimmen und zu verfeinern.

# Fußnoten

1. Stichwort fehlerhaftes OCR.
2. https://www.monasterium.net/mom/home, http://telma-chartes.irht.cnrs.fr/, https://cema.lamop.fr/, https://lombardiabeniculturali.it/cdlm/edizioni/.

# Bibliographie


**Argamon, Shlomo**. 2008. „Interpreting Burrows's Delta: Geometric and Probabilistic Foundations." In *Literary and Linguistic Computing,* 23,2. S. 131-147. Oxford: University Press.

**Burrows, John**. 2002. „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." In *Literary and Linguistic Computing,* 17,3. S. 267-287. Oxford: University Press.

**Burrows, John**. 2003. „Questions of authorship: attribution and beyond: a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001, New York." In *Computers and the Humanities,* 37,1. S. 5-32. Dordrecht: Kluwer.

**Eder, Maciej**. 2012. „Computational stylistics and Biblical translation: How reliable can a dendrogram be?" In *The Translator and the Computer,* hgg. von Piotrowski/Grabowski, S. 155-170. Wrocław : Verlag der Hochschule für Philologie in Wrocław .

**Eder, Maciej**. 2013. „Mind your Corpus: Systematic errors in authorship attribution". In *Literary and Linguistic Computing,* 28,4. S. 603-614. Oxford: University Press.

**Eder, Maciej**. 2015. „Does size matter? Authorship attribution, small samples, big problem." In *Digital Scholarship in the Humanities,* 30,2. S. 167-182. Oxford: Oxford Academic.

**Eder, Maciej und Jan Rybicki**. 2015. „Go Set A Watchman while we Kill the Mockingbird In Cold Blood." https://computationalstylistics.github.io/blog/harper-lee/ (zugegriffen: 03. August 2022).

**Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch u nd Thorsten Vitt**. 2017. „Understanding and explaining Delta measures for authorship attribution". In *Digital Scholarship in the Humanities* 32,Suppl. 2. S. ii4- ii16 10.1093/llc/fqx023.

**Jockers, M atthew und Daniela M. Witten**. 2010. „A comparative study of machine learning methods for authorship attribution." In *Literary and Linguistic Computing,* 25,2. S. 215–23. Oxford: University Press.

**Juola, Patrick und Stephen Ramsay**. 2017. „Six Septembers: Mathematics for the Humanist." Zea E-Books Collection, 55. Lincoln: Zea Books.

**Karsdorp, F olgert, Mike Kestemont und Allen Riddel**. 2021. „Humanities Data Analysis: Case Studies with Python." Princeton: Princeton University Press.

**Koppel, Moshe, Jonathan Schler und Shlomo Argamon**. 2009. „Computational Methods in Authorship Attribution." In *Journal of the American Society for Information Science and Technology,* 60,1. S. 9-26. New York: Wiley.

**Koppel, Mosche und Yaron Winter**. 2014. „Determining If Two Documents Are Written by the Same Author." In *Journal of the American Society for Information Science and Technology,* 65,1.
S. 178–187. New York: Wiley.

**Kou, Gang, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, und Fawaz E. Alsaadi**. 2020. „Evaluation of Feature Selection Methods for Text Classification with Small Datasets Using Multiple Criteria Decision-Making Methods". *Applied Soft Computing* 86. 10.1016/j.asoc.2019.105836.

**Modupe, Abiodun, Turgay Celik, Vukosi Marivate, und Oludayo O. Olugbara**. 2022. „Post-Authorship Attribution Using Regularized Deep Neural Network". *Applied Sciences* 12 (15). 10.3390/app12157518.

**Plachta, Bodo**. 2006. Editionswissenschaft eine Einführung in Methode und Praxis der Edition neuerer Texte. 2. Aufl. Stuttgart: Reclam.

**Plakias, Spyridon, und Efstathios Stamatatos**. 2008. „Tensor Space Models for Authorship Identification". In *Artificial Intelligence: Theories, Models and Applications*, herausgegeben von John Darzentas, George A. Vouros, Spyros Vosinakis, und Argyris Arnellos, 239–49. Lecture Notes in Computer Science. Berlin: Springer.

**Potha, Nektaria und Efstathios Stamatatos**. 2017. „An Improved Impostors Method for Authorship Verification." http://www.icsd.aegean.gr/lecturers/stamatatos/papers/CLEF-Potha-2017.pdf (zugegriffen: 03. August 2022).

**Vogeler, Georg**. 2006. „Vom Nutz und Frommen digitaler Urkundeneditionen." In *Archiv für Diplomatik*, 52. S. 449-466. Wien: Böhlau.

**Vogeler, Georg**. o.J. „Die Text Encoding Initiative (TEI) als Werkzeug des Urkundeneditors – Erfahrungen und Desiderate." https://rep.adw-goe.de/bitstream/handle/11858/00-001S-0000-0023-9A13-A/6_Vogeler.pdf?sequence=70. (zugegriffen: 03. August 2022).

**Wu, Haiyan, Zhiqiang Zhang, und Qingfeng Wu**. 2021. „Exploring Syntactic and Semantic Features for Authorship Attribution". *Applied Soft Computing* 111. 10.1016/j.asoc.2021.107815.