

A computational approach to authorship verification of Johann Wolfgang Goethe's contributions to the *Frankfurter gelehrte Anzeigen* 1772/73

Mike Kestemont, Gunther Martens, Thorsten Ries

Introduction

The present article proposes a computational stylometric authorship verification approach to contribute to the solution of the longstanding philological problem concerning Johann Wolfgang Goethe's (1749-1832) anonymous contributions to the journal *Frankfurter gelehrte Anzeigen* of 1772/73 (*FgA*). A plethora of philological research since the late 19th century has dealt with the question which of the *FgA*'s *Rezensionen* has been penned by which author. Currently, the authorship of many of the anonymously published *Rezensionen* of the *FgA* is either unclear or attribution has to be regarded as tentative, as often based on rather vague philological and stylistic indicators. We present the research design for a more comprehensive project to identify all of Goethe's contributions to the *FgA* of 1772/73 using a well-tested and consistent stylometric method. Our preliminary goal is to test previous attribution attempts and to complement further stylometric arguments for cases where philological arguments did not provide sufficient evidence. Hence, the present article mainly documents and discusses the methods, considerations and results of the first project phase. First, we present and evaluate an authorship verification method, inspired by the so-called *impostor approach* against a corpus subset in order to determine whether the results indicate the approach will solve the research problem. The results, which we discuss in this article, did not only show that this is indeed the case, but also give the opportunity to observe how the proposed method performs on a rather difficult case and corpus. This case study will furthermore discuss modern computational stylometry in the context of linguistic and statistical approaches previously proposed to solve the *FgA* problem since 1903.

Status quaestionis, the historical problem

After Johann Conrad Deinet had bought the *Frankfurter Gelehrtenzeitung* (founded 1736 by Samuel Tobias Hocker) in 1771, the intellectual and literary journal was renamed *Frankfurter gelehrte Anzeigen* (*FgA*) and became the flagship journal of the "Sturm und Drang" movement in 1772, with Johann Wolfgang Goethe as the new editor-in-chief. The *FgA* authors vigorously engaged with the political debates of the emerging literary public of its time, in which literary journals played a key role¹ (Martus 2007; Wolf 2013). Next to Goethe, the journal had a number of renowned main co-editors and contributors such as Johann Heinrich Merck, Johann Georg Schlosser, Johann Gottfried Herder and later Karl Friedrich Bahrdt. Goethe most likely wrote for the *FgA* until 1773, but the exact number and time

¹ Steffen Martus, *Werkpolitik: Zur Literaturgeschichte Kritischer Kommunikation Vom 17. bis Ins 20. Jahrhundert ; Mit Studien Zu Klopstock, Tieck, Goethe Und George*, Historia Hermeneutica, 3 (Berlin ; New York: W. de Gruyter, 2007); Norbert Christian Wolf, 'Heinrich Christian Boies Göttinger Musenalmanach und Johann Heinrich Mercks Frankfurter gelehrte Anzeigen. Medienkämpfe im literarischen Feld des Sturm und Drang', in *Sturm und Drang: Epoche, Autoren, Werke*, ed. by Matthias Buschmeier and Kai Kauffmann (Darmstadt: WBG, 2013), pp. 10–28.

period of his contributions are yet to be determined. The so-called “Rezensionen” (i.e. review articles) of the *FgA* were published anonymously and were often redacted by the editors; some were even written collaboratively by multiple authors (“Protokoll-Rezensionen”², protocol reviews, Wolf 2013). The anonymity of the contributions was on the one hand a conceptual decision in the spirit of new forms of collaboration that enable truthfulness (“Wahrheit”), deliberately avoiding the bias of authorial self-positioning and group formation (“Autorfesseln und Waffenträgerverbindungen”, J.H. Merck³). On the other hand, it served as a protection for the authors, especially when they touched upon religious and political issues. The owner of the journal, J. C. Deinet, was implicated in a number of religious legal complaints and lawsuits against particular contributions in the journal, which grew to a public legal struggle about the freedom of the press. Ultimately, Deinet was fined but spared from further damage through amnesty by the Frankfurt administration. After this conflict, the most prolific authors - among them Goethe - left *FgA* in 1773.⁴

As a consequence, Goethe philology is still today confronted with a longstanding problem: the *FgA* of 1772/73 comprise nearly 900 pages of anonymous journal text, of which some is Goethe’s. Around 40 authors wrote, co-authored or redacted the 396 *Rezensionen*, leaving the authorship question open in many cases (Haenelt 1984). Previous authorship attribution attempts used philological arguments such as attribution in letters by the authors or editors wherever possible, but in the majority of cases had to rely on rather vague interpretations of ambiguous indicators. Attribution is further complicated by the fact that Goethe as editor-in-chief redacted numerous *Rezensionen* by other authors, and, furthermore, reported in *Dichtung und Wahrheit* that he served as the keeper of the minutes during the discussion sessions that were used as the basis of the collaborative “protocol reviews” (Haenelt 1984). Goethe self-attributed some of the *FgA-Rezensionen* by including them in his edition of his own works (*Goethe's Werke. Vollständige Ausgabe letzter Hand*, Goethe 1827-30), but the evidential value of his late self-attribution remained controversial. The authorship of the majority of the *FgA's Rezensionen* has never been verified with systematically controlled or tested methods, despite considerable philological efforts to develop methods to attribute the texts based on philological, stylistic and linguistic features to the main authors of the 1772/1773 volumes.

In his 1865 study, von Biedermann claimed to have identified a review of *Götzens erbauliche Betrachtungen* as Goethe’s that the author had not attributed to himself.⁵ Ever since, any scholarly edition of Goethe’s works has had to define which *Rezensionen* they would attribute to the author. A daunting task, as Georg Witkowski later notes: “With respect to the authorship of these *Rezensionen* we

² Hermann Bräuning-Oktavio, *Herausgeber Und Mitarbeiter Der Frankfurter Gelehrten Anzeigen 1772* (Tübingen: Niemeyer, 1966), pp. 259–66.

³ Cited in Stefan Knödler, ‘Frankfurter gelehrte Anzeigen’, in *Handbuch Sturm und Drang*, ed. by Matthias Luserke-Jaqui (Berlin ; New York: de Gruyter, 2017), pp. 422–28 (p. 424).

⁴ Knödler *ibid.*; Hermann Dechent, ‘Die Streitigkeiten Der Frankfurter Geistlichkeit Mit Den Frankfurter Gelehrten Anzeigen Im Jahre 1772’, *Goethe-Jahrbuch*, 10 (1889), 169–95; Dahnke, Hans-Dietrich: ‘Intentionen und Resultate des Jahrgangs 1772 der Frankfurter Gelehrten Anzeigen’, in *Sturm und Drang: Geistiger Aufbruch 1770–1790 im Spiegel der Literatur*, ed. by Bodo Plachta and Winfried Woesler (Berlin: de Gruyter, 2015), pp. 87–99.

⁵ Von Biedermann, Woldemar, *Goethe und Leipzig*, vol. 2, Leipzig 1865, p. 20.

probably have to say: Ignoramus and also: Ignorabimus!”.⁶ An early attempt at a linguistic, quantifying approach to this problem was made by Carl Ritter with his article *Anwendung der Sprachstatistik auf die Recensionen in den Frankfurter Gelehrten-Anzeigen von 1772*.⁷ Ritter described a method to identify authors by statistical analysis of linguistic features of their style such as orthography, adverbs and conjunctions. This effort, however, was later rebutted for its allegedly questionable choice of base texts for these features (Oktavio-Bräuning 1911: 24-25). The psychologist Karl Marbe ventured into early experiments with quantitative linguistics and phonetics in order to - without much scholarly success - discern Goethe’s specific prose rhythm (Karl Marbe, 1904, 1912).⁸ Following Ludwig Hirzel’s discovery of letters confirming Goethe’s authorship of two *Rezensionen* on Lavater and *Geßners Idyllen*,⁹ Bernhard Seuffert, Wilhelm Scherer and Ludwig Geiger continued to secure new attributions of *FgA Rezensionen* to Goethe using additional sources (Oktavio-Bräuning 1911). The seminal scholarly re-edition of the 1772/1773 volumes of the *FgA* by Bernhard Seuffert, with an introduction by Wilhelm Scherer, made important contributions to the matter of Goethe’s authorship, defining the state of the art for years to come (Seuffert 1882/83, Scherer 1883).¹⁰ The article by Otto Trieloff (1908) made scholars aware of the fact that some of the reviews were actually translated re-publications from articles in English journals, first and foremost *Gentlemen’s Magazine* and *Monthly Review*.¹¹

Max Morris (1909) and Hermann Bräuning-Oktavio (1911) dedicated large parts of their academic lives to this authorship attribution question, gathering large amounts of philological evidence in a number of monographs and articles. Where material, direct evidence for authorship attribution was missing, scholars had to rely on the often shaky grounds of attributing by notions of style and thematic preference, essentially hermeneutic arguments that based the attribution argument on the recurrence of meaningful, distinctive opinions, topics, distinctive phrasings (“kennzeichnende[] Lieblingswendungen”¹²) and (allegedly) individual spelling characteristics (“warrlich”, “Schäckespear”, etc). Bräuning-Oktavio repeatedly criticized especially Morris for attributing *FgA* texts based on the weak evidential basis of distinctive phrasings. In many cases, thus, attribution of *FgA* texts largely relied on either direct external evidence or low-frequency, striking lexical or thematic characteristics, the latter posing a methodological problem in itself (Kestemont et al. 2016: 87).

In his later years, as late as 1966, Bräuning-Octavio aimed to delineate a set of ‘typical features’ of Goethe’s style; language rhythm and melody, favourite expressions, rhetorical features such as specifics

⁶ Georg Witkowski, ‘Einleitung’. Goethes Werke, ed by H. Düntzer, G. Witkowski, K.J. Schröer, A.G.Meyer, vol. 26, (Stuttgart, Tübingen: Cotta 1892), 47-48.

⁷ Ritter, Carl, ‘Anwendung der Sprachstatistik auf die Recensionen in den Frankfurter Gelehrten-Anzeigen von 1772’. *Goethe-Jahrbuch* 24 (1903), pp 185-203.

⁸ Karl Marbe, *Ueber Den Rhythmus Der Prosa: Vortrag, Gehalten Auf Dem 1. Deutschen Kongress Für Experimentelle Psychologie Zu Giessen* (J. Ricker, 1904).

⁹ Hirzel, Ludwig, ‘Goetheana’. *Im neuen Reich*, 8, 1878, II, 597/611.

¹⁰ Bernhard Seuffert (Ed.), *Frankfurter gelehrte Anzeigen*. Nachdruckausgabe. Mit einer Einleitung von Wilhelm Scherer. 2 vols. (Heilbronn: Henninger 1882/83). Wilhelm Scherer, introduction, *ibid*, vol. 2, III-XC. Reprinted 1970 with a concordance and a new introduction by Bräuning-Oktavio.

¹¹ Trieloff, Otto, *Die Entstehung der Rezensionen in den Frankfurter Gelehrten-Anzeigen vom Jahre 1772* (Münster: Schöningh 1908).

¹² Max Morris, *Goethes und Herders Anteil an dem Jahrgang 1772 der Frankfurter Gelehrten Anzeigen* (Stuttgart, Berlin: Cotta, 1909), p. 347.

of exclamation, questions, address, double negation, accumulation and enumeration, anaphora, parenthesis, typical review beginnings, Goethe's grammar during the "Werther Periode", sentences omitting verb, parallelisms, inversion, emphatic sentence endings, Latin quotes etc. The results - beyond the direct external proof found - remained vague. But Bräuning-Octavio already worked on a prototype of stylometry, as his private archive collection in the Archive of the *Technische Universität Darmstadt* shows. For instance, Bräuning-Oktavio's archive contains a typescript on the "Statistik der Füllwörter in den FGA" (statistics of expletives in the *FgA*).¹³ Just in the same year, Joachim Thiele published a brief description of an approach to linguistic statistical aesthetics ("Verfahren der statistischen Ästhetik") with his *Untersuchung der Goethe zugeschriebenen Rezensionen in den Frankfurter Gelehrten Anzeigen mit Hilfe einfacher Textcharakteristiken* (1966).¹⁴ Four years later, Herbert Sparmann tried to distinguish Goethe from Merck by the frequency of the definite article in their writings. He detected that Merck used the definite article 40% more frequently than Goethe.¹⁵ This conclusion was reached on the basis of a very small corpus, taken from *FgA* itself only. The first scholar to propose a computational approach to the *FgA* authorship problem was Karin Haenelt with her article *Die Verfasser Der Frankfurter Gelehrten Anzeigen von 1772. Ermittlung von Kriterien Zu Ihrer Unterscheidung Durch Maschinelle Stilanalyse* in *Euphorion* (1984).¹⁶ Haenelt establishes stylistic profiles by categorizing the frequency of word functions such as nouns, adjectives, by taking into account lexicon variation and by analyzing words in 1st, 2nd, last position in the sentence. The formulation and weight of the latter feature is based on a hermeneutic assumption that these are the most significant for Goethe's individual style. The study used a software tool called LDVLIB developed by R. Drewek, a scarcely documented, early text statistical processor.¹⁷ Here as well, the base corpus of the study was very small and selective, with no control group measures in place, as again the texts were solely taken from the *FgA*.

The *FgA* volumes of 1772/1773 are a corpus of 396 anonymous *Rezensionen*, between 1 and 7 pages in length. The articles have been penned and redacted by up to 40 authors, while there is a small number of known main contributors, amongst which the editor-in-chief Johann Wolfgang Goethe. Only for a few of the *Rezensionen* hard philological evidence is available to corroborate the authorship attribution. The idea that linguistic and statistical methods and style analysis might solve this problem has been around since 1903, but the initiatives were either short-lived or remained in a proof-of-concept state, while the early attempts operated on a small corpus basis from within *FgA*, did not have otherwise tested

¹³ See record at the archive of the Darmstadt University Library, p. 9, URL: https://www.ulb.tu-darmstadt.de/media/ulb/spezialabteilungen/handschriften_1/nachlaesse_1/Braeuning-Oktavio.PDF

¹⁴ Joachim Thiele, *Untersuchung der Goethe zugeschriebenen Rezensionen in den Frankfurter Gelehrten Anzeigen mit Hilfe einfacher Textcharakteristiken*, in: *Studia Linguistica* 20 (1966), 83–85.

¹⁵ Herbert Sparmann, 'Häufigkeitsuntersuchungen, Ein Hilfsmittel Für Den Vergleich von Texten Und Für Die Feststellung Der Verfasserschaft', *STUF - Language Typology and Universals*, 23.1–6 (1970), 227–231 <<https://doi.org/10.1524/stuf.1970.23.16.227>>.

¹⁶ Karin Haenelt, 'Die Verfasser Der Frankfurter Gelehrten Anzeigen von 1772. Ermittlung von Kriterien Zu Ihrer Unterscheidung Durch Maschinelle Stilanalyse', *Euphorion*, 78 (1984), 368–382.

¹⁷ Raimund Drewek, LDVLIK-Textanalyse mit System, in Walter Lehmacher, Allmut Hörmann, *Statistik-Software. 3. Konferenz über die wissenschaftliche Anwendung von Statistik-Software 1985*, Stuttgart, New York: Fischer 1986, pp 283-295; Raimund Drewek and M. Erni, 'LDVLIB (LEM): A System for Interactive Lemmatizing and Its Application', in *Coling 1982 ABSTRACTS: Proceedings of the Ninth International Conference on Computational Linguistics Abstracts*, 1982.

methods at their disposal and the question of a linguistic definition of style remained unanswered. Karin Haenelt's small-scale study was the first to attempt to introduce computational methods, leading to interesting results despite using only a small corpus and not being able to independently test the assumptions of the features and method used. We are convinced that with the recent advances in computational stylometric authorship verification, which provided research with well-tested methods, it is now possible to close in on the *FgA* authorship problem.

Present-day computational stylometric methods of stylometric authorship attribution and verification are benchmarked independently from the particular corpus, their linguistic definition of style and linguistic features abstracts from the individual author and their training data includes much larger corpora than ever before. Well-tested methods focus on inconspicuous, high-frequency and less variant features of an individual's style (Kestemont et al. 2016: 87). Especially Goethe's writings and 18th, 19th century literature and journal corpora offer large amounts of training data needed to achieve reasonably precise results. The particular task of identifying Goethe's contributions to the *FgA* is an interesting challenge, as the writing samples are relatively short in length (1-7pp) and Goethe, as editor-in-chief, has redacted many articles and was involved in collaborative writing, e.g. as the minutes-keeper of collaborative "protocol reviews". The fact that Goethe was such a prolific writer is an advantage for the application of stylometric methods, and at the same time a challenge. The stylometric fingerprinting might be affected by the fact that the other writers to compare Goethe's style to have been much less productive, resulting in smaller corpora than Goethe's. His stylistic features may have changed over time, as the productive literary period of his life was exceptionally long.

Methodology

Setup

In order to verify the suitability of the particular authorship verification method for the purpose of securely attributing *FgA Rezensionen* to Goethe or reject the attribution, we ran a variant of a well-tested verification method¹⁸ against a number of controversial and challenging cases discussed by Karin Haenelt as well as a few texts that have been attributed to Goethe and Herder on solid evidential basis in the past in a blind test setting. Blind test setting meant here that the person conducting the process was not aware of previous research and attributions of the unattributed texts to primarily test. The controversial and challenging test cases were selected to represent different lengths from 2 to 7 *FgA*

¹⁸ Kestemont, Mike, Justin Stover, Moshe Koppel, Folger Karsdorp, Walter Daelemans, 'Authenticating the writings of Julius Caesar', *Expert Systems With Applications* 63 (2016), pp. 86-96.

Kestemont, Mike, Els Stronks, Martine de Bruin, Tim de Winkel, 'Did a Poet with Donkey Ears Write the Oldest Anthem in the World? Ideological Implications of the Computational Attribution of the Dutch National Anthem to Petrus Dathenus'. *Digital Humanities 2017, DH 2017, Conference Abstracts*, McGill University & Université de Montréal, Montréal, Canada, August 8-11, 2017. URL: <https://dh2017.adho.org/abstracts/079/079.pdf>.

Koppel, Moshe, and Yaron Winter, 'Determining If Two Documents Are Written by the Same Author', *Journal of the Association for Information Science and Technology*, 65 (2014), 178–187;

Greta Franzini and others, 'Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm', *Frontiers in Digital Humanities*, 5 (2018), 4.

pp, different previous self-attribution by Goethe and different degrees of attribution confidence by Haenelt. Our goal was to benchmark the method against previous research, especially regarding short text samples and cases known to be ambiguous in previous stylometric attribution attempts.

The focus of our analysis is a selection of four texts that Haenelt marked as previously unattributed, suggesting an attribution by her own linguistic features:

1. Cymbelline, ein Trauerspiel nach einem von Schäckespear erfundnen Stoffe [von J.G. Sulzer], *FgA* 1772, 591-592 - self-attributed by Goethe, 2 pp (*FgA*) length.
2. [Johann Gottfried Schummel], Empfindsame Reisen durch Deutschland von S. 2. Teil. *FgA* 1772, 141-144 - self-attributed by Goethe, 4 pp (*FgA*) length.
3. [John Aikin], Essays on Song-Writing; with a collection of such English Songs, as are most eminent for poetical merit. To which are added some Original pieces, *FgA* 1772, 745-749 - not self-attributed by Goethe, 5 pp (*FgA*) length.
4. Die schönen Künste in ihrem Ursprung, ihrer wahren Natur und besten Anwendung, betrachtet von J.G. Sulzer 1772. *FgA* 1772, 801-807 - self-attributed by Goethe, 7 pp (*FgA*) length.

Based on her own score matrix of five Goethe-specific features - average sentence length (words/syllables), vocabulary composition and distribution, sentence transition position, sentence last position, sentence second position - Haenelt's results for her closed-set test based on a *FgA*-only corpus were as follows:

	Feature match author profile	Results (Haenelt)
1 Cymbelline, ein Trauerspiel nach einem [...] - 2 pp	2 x Merck 2 x Herder 1 x Goethe	Ambiguous case, she decided Herder is most probable.
2 [Johann Gottfried Schummel], Empfindsame Reisen durch [...] - 4 pp	4 x Goethe 1 x Herder	Highest probability: Goethe.
3 [John Aikin], Essays on Song-Writing [...] - 5 pp	5 x Herder	Positive: Herder.
4 Die schönen Künste in ihrem Ursprung [...] - 7 pp	5 x Goethe	Positive: Goethe.

We furthermore tested a number of *FgA* texts that have been attributed to Goethe and Herder on solid external evidential basis in order to determine how strong the statistical signal for both authors are. These are:

Johann Wolfgang Goethe:

1. Leben und Charakter Herrn Christian Adolph Klotzens, entworfen von Carl Renatus Hausen. 1772. 8. 93 S. *FgA* 1772, 342-343.
2. Moralische Erzählungen und Idyllen von Diderot und S. Geßner. 1772. 8. 273 S. *FgA* 1772, 537-534
3. Gedichte von einem Polnischen Juden, 8. 1772. 96 S. *FgA* 1772, 555-558.
4. Nachrede statt der versprochenen Vorrede. *FgA* 1772, 830-832.

Johann Gottfried Herder:

1. J. D. Michaelis Mosaifches Recht. 1ster Theil, 1770 Th. 2. 1771. 8. bey Garbe. *FgA* 1772, 265-269.
2. Staatsveränderungen von Italien, in 24 Büchern entworfen. Von Karl Denina. Erster Band. Aus dein Italienischen von D. J. J. Volkmann. Bey Schwickert 1771. gr. 8-1 Alph. 11 Bogen. *FgA* 1772, 425-430.
3. Schlötzers Vorstellung seiner Universalhistorie. Bey Dietrich 1772. 8. 16 Bogen. *FgA* 1772, 474-478.
4. D.J. Sal. Semleri, Paraphrasis Evangelii Johannis, cum notis, & cantabrigiensis codicis latino textu. Bey Hemmerde, 1771. 1772. 8. 2 Theile, *FgA* 1772, 482-486.
5. J. D. Michaelis Versuch über die siebenzig Wochen Daniels. Ein Auszug dessen, was er in seinen deutschen Collegia über das neunte Kapitel Daniels neues bemerkt hat. Bey Dietrich 1771 8. 17 Bögen, *FgA* 1772, 505-509.
6. Bemerkungen über den Unterschied der Stände in der bürgerlichen Gesellschaft von Joh. Millar Esq. Aus dem Englischen. Bey Schwickert 1772. 8. 15 Bogen. *FgA* 1772, 609-614.
7. James Beattie. Versuch über die Natur und Unveränderlichkeit der Wahrheit; im Gegensatze der Klügeley und der Zweifelsucht. Aus dem Englischen. *FgA* 1772, 673-677.

Digitising the *FgA* Rezensionen proved to be a challenging task, as *FgA* has been printed in *Fraktur* (German Gothic type). The precision of output of OCR-engines for *Fraktur* type is up till today not up to par with Antiqua type. As a result, we corrected the abovementioned controversial or unattributed *FgA* samples by hand, but the overall corpus had to be redarded as “dirty” OCR. The model of Goethe’s style was based on a realtively “dirty” OCR corpus of Goethe’s works. The style fingerprint of Herder and others were also trained based on texts external to *FgA*.

Attribution vs verification

Traditionally, stylometric authorship studies have been dominated by a problem setup that is today known as “closed-set attribution”.¹⁹ In this approach, an authorship problem is cast as a conventional classification task in text categorization. First, a standard algorithm from the field of text categorization is being trained on a set of reference documents or “training” material, for which the authorship is uncontested. The authorship of these documents are considered class labels, or a series of mutually exclusive categories to which each document belongs. Next, for evaluation purposes, the trained

¹⁹ An excellent survey of the field of computational authorship studies can be found in: Efstathios Stamatas, ‘A Survey of Modern Authorship Attribution Methods’, *Journal of the American Society for Information Science and Technology*, 60.3 (2009), 538–556..

algorithm is applied to a set of previously unseen test documents (or “held out items”) that have to be attributed to one of the candidate categories. The attribution results can then be compared against the ground truth for the test documents, which allows us to assess the performance of such a classifier.

This kind of simulation in closed-set attribution tasks is meant to approximate the real-life situation where an anonymous document has to be attributed to one of a series of previously known candidate authors. Naturally, the caveat associated with this type of simulation - which is often compared to a line-up situation - is that this does not correspond to many real world scenarios: often it cannot be guaranteed that the actual author of an anonymous documents is among the candidate authors that the classification has analyzed during training. Recently, the field of stylometry has therefore turned its attention towards more demanding, but also more realistic experimental setups.²⁰

Authorship verification, also known as open-set attribution, is the experimental setup where attribution algorithms can no longer assume that the author of a test document is necessarily among the available set of candidate authors. Regarding classification, this setup is essentially identical to the attribution problem, but it introduces an additional classification option: “none of the above” (i.e. the label which is applicable in the case that an anonymous document can and should not be assigned to any of the known candidate authors). Our experiments, reported below, are strongly indebted to some of the methodological innovations proposed in recent verification research. All code and data necessary to reproduce our experiments are available without restrictions.²¹

Preprocessing

The material available for calibrating our system comes in the form of OCR’ed texts of which the digitization strongly diverges and which abound in OCR artefacts. Although recent work has demonstrated empirically that automatically OCR’ed texts can certainly serve as a useful proxy for manually digitized data,²² we have tried to aggressively reduce the presence of digitization artifacts in the data.²³ The shortest text in the anonymous corpus counted 2090 characters (after applying the preprocessing steps outlined above). We have therefore divided all longer texts in the reference corpus into equal-sized, consecutive slices of 2090 characters each. We converted each slice into a numeric

²⁰ See, in particular, the recent work by Moshe Koppel and collaborators: Moshe Koppel and Yaron Winter, ‘Determining If Two Documents Are Written by the Same Author’, *Journal of the Association for Information Science and Technology*, 65.1 (2014), 178–187; Justin Anthony Stover and others, ‘Computational Authorship Verification Method Attributes a New Work to a Major 2nd Century African Author’, *Journal of the Association for Information Science and Technology*, 67.1 (2016), 239–42 <<https://doi.org/10.1002/asi.23460>>..

²¹ URL: <https://github.com/mikekestemont/goethe/>

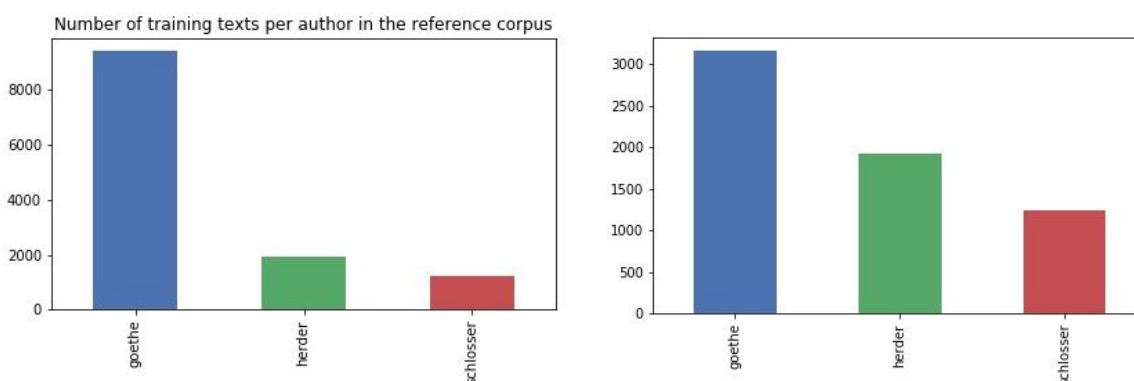
²² Greta Franzini and others, ‘Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm’, *Frontiers in Digital Humanities*, 5 (2018), 4.

²³ The following preprocessing steps were taken: 1) Hyphenated word breaks at the end of lines have been corrected; 2) All texts were lowercased and all whitespace was converted to single spaces; 3) Instances of the long-s were converted into a standard s; 4) The unidecode package was applied to ensure a consistent encoding throughout the corpus (<https://pypi.org/project/Unidecode/>); 5) Only alphabetical characters were retained (to avoid artefacts relating to differences in e.g. quotation rendering).

vector using a traditional bag-of-words approach.²⁴ We extracted a vocabulary to represent our data, coinciding with the 10,000 most common character tetragrams: these are overlapping sequences of 4 characters long that are extracted via windowing over the text (e.g. for this sentence, the first tetragrams would be ‘we e’, ‘e ex’, ‘ext’, ‘extr’, and so). In our model, we represent each slice through counting how often this vocabulary items appear in the slice.²⁵

Calibration

The reference data, after splitting, amounted to 12,543 instances. We divided the available training texts into the 9387 slices by Goethe and the 3156 texts (certainly) not written by Goethe. Goethe slices significantly outnumbered the non-Goethe slices. To correct this imbalance, we resampled the Goethe data through restricting the Goethe samples to the most prototypical ones for his writing style.²⁶ Both the original and resampled per-author distribution of these slices are plotted below. After this resampling, we revectorized all slices using the same approach as above and created a stratified train-test split of the material, with 20% of the available slices as “anonymous cases” in test split.



We experimented with three attribution approaches on the resampled corpus:

1. **Plain:** a simple nearest neighbour classifier, that attributes a test text to one of our three authors as follows: for each test text, we retrieve the closest neighbour's in the training set, i.e. we determine which slice in the training text has the most similar tetragram frequencies (according to the cosine distance). We then extrapolate the authorship of this nearest neighbour to the test item.

²⁴ F. Pedregosa and others, ‘Scikit-Learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12 (2011), 2825–2830.

²⁵ As in previous work, we made use of a conventional TF-IDF weighting scheme to reinforce the weight of less frequently occurring, significant tetragrams. Finally, we also applied L1-normalization: consult the code for further details. All weights needed for this transformation were fitted on the basis of the reference texts only.

²⁶ We calculated the mean feature vector (or geometric centroid) for the Goethe slices and only allowed the n samples which were closest to the mean (using the cosine distance as metric), with n being equal to the number of non-Goethe slices in the material.

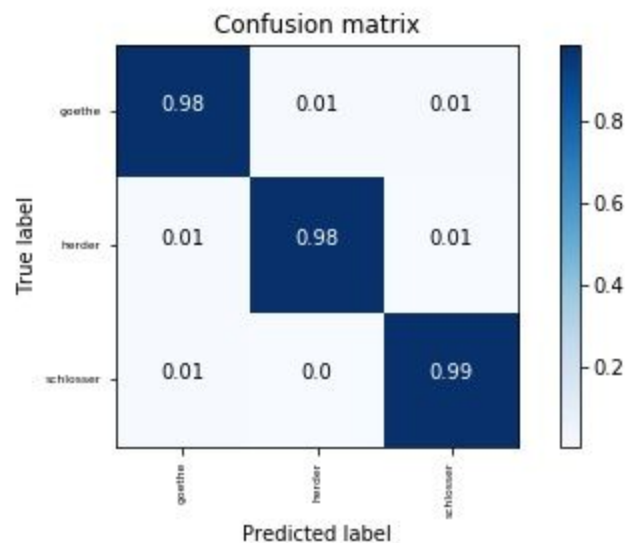
2. **Feature dropout:** We applied the plain classification scheme as above, but now 250: in each iteration, however, the distance calculation for determining the nearest neighbor was restricted to a random sample of 50% of the available character tetragrams (meaning that the other 50% of the tetragrams are “dropped out”, hence the name). The test item is classified according to the most frequent classification in the series of iterations.
3. **Slice dropout:** We use the dropout scheme as in the previous approach, but, additionally, in each iteration, we limit the training material (from which the algorithm is allowed to select a nearest neighbor) to a random selection of 100 slices from the training material (not even guaranteeing that the test item’s correct author would be in that set). Again, the test item is classified according to the most frequent classification in the series of iterations.

This methodology is indebted to the so-called “impostors” approach to authorship verification (although no impostors are included in our approach). The sampling of features in the dropout approach captures the intuition that evidence of common authorship should be stable enough to be visible across different random samples of features. The sampling of training material in each iteration (added in the *slice dropout*) ensures that an attribution to an author is based on a stylistic similarity between test and training material that goes beyond the superficial similarity between specific pairs of texts in the data.

We evaluated our three approaches using the accuracy score (i.e. the simple proportion of correct attributions for the test text) and F1 score (a more technical variant of accuracy, that takes into account the fact that not every author is equally well represented in the material):

	Accuracy score	F1 score
Plain	0.9152810768012668	0.9152975527774266
Feature dropout	0.9572784810126582	0.9574360857325933
Slice dropout	0.9841646872525732	0.9832402618073663

The table above indicates the performance of the three attribution schemes discussed above in terms of accuracy and the F1 score (for a randomly selected test of 20% of the slices in the training material). These scores clearly show that applying the dropout-versions of the system leads to a significant performance improvement (with accuracies and F1-scores generally being very close to each other). The feature dropout adds quite a bit of performance to the naive

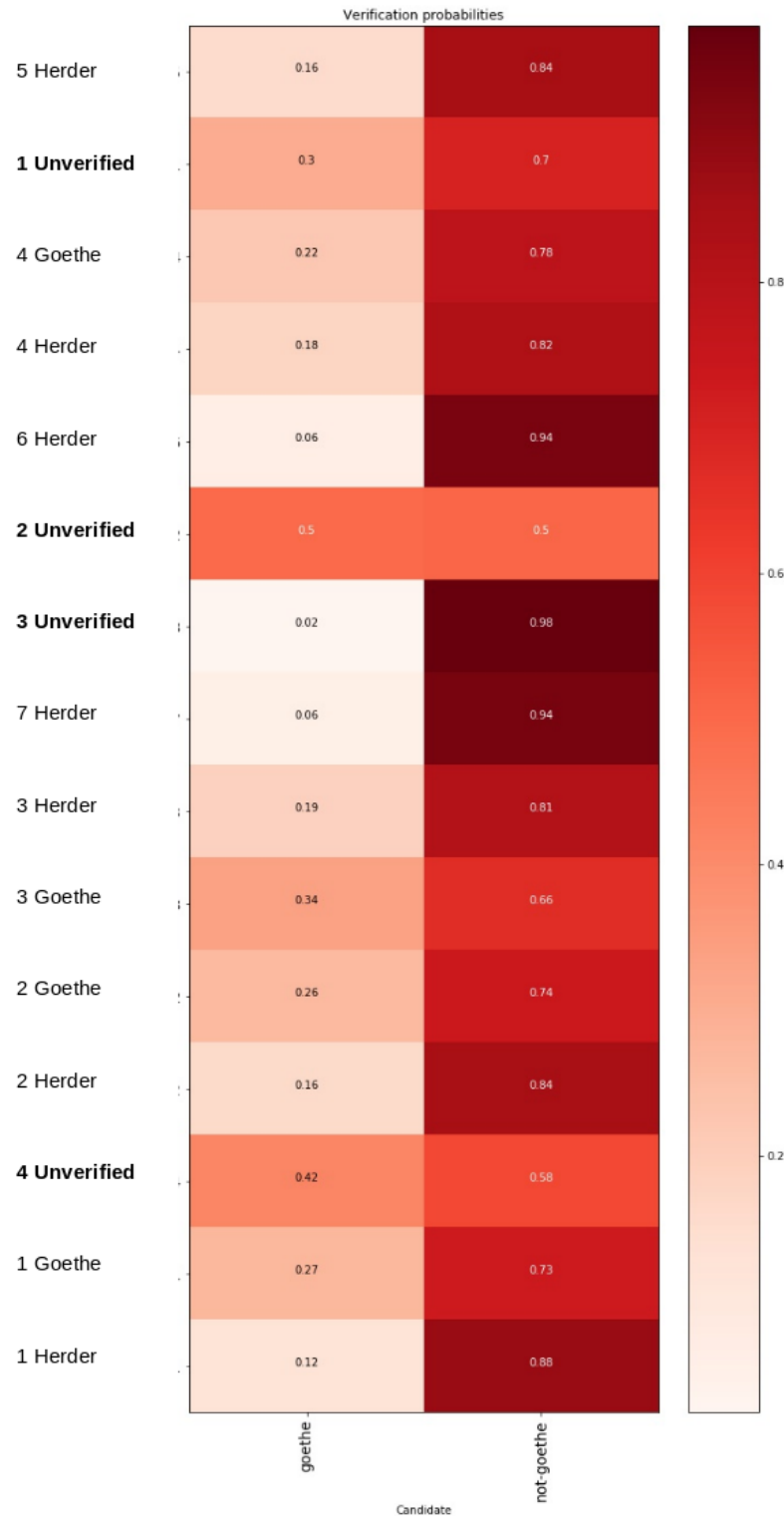


nearest neighbour learner (+4% in accuracy) and adding also the slice dropout adds an additional improvement (+3% in accuracy). Below we plot the confusion matrix for the best experiment (slice dropout), visually supporting the intuition that the confusion between our three authors is minimal.

One interesting methodological caveat concerns the following: entire texts have been divided into slices before the train test division. This means that some samples in the test and training material will in fact have been extracted from the same text, meaning that our models might also capture this common textual origin (instead of *only* their authorship). In spite of this caveat, it is interesting to note that our results show that the sampling approach (both with the feature and slice dropout) still yield clearly the best results, indicating the robustness of the bootstrapped method for text classification. Note that this artefact, of course, cannot have an effect on the actual results for the anonymous test texts reported below.

Application to the anonymous texts

After calibrating the system, we applied the best performing setup (slice dropout) to a test set of anonymous texts, for which the authorship cannot be verified using traditional philological means. We focused on the texts marked as doubt cases by Haenelt. Three of them has been self-attributed by Goethe to himself, although the ascription has remained controversial. All four texts vary in length in the range of 2 to 7 pages (*FgA*). In the heatmap below, we plot the verification scores which can be obtained from applying the slice dropout setup: the columns show the proportion of Goethe and non-Goethe attributions during 250 iterations. This number can be interpreted as a probability score, indicating the robustness of an attribution to the category represented in the respective columns.



Note that these scores might generally seem low for Goethe, because the two other authors are collapsed into a single category. The above heatmap shows that the computational method applied consistently attributed known Goethe contributions to *FgA* (1-4 Goethe) with an attribution value larger than or equal to 0.22. The known Herder samples (1-7 Herder) all scored lower than this value (≤ 0.19).

On these samples, the method performed consistent and - despite dirty OCR data - resulted in higher values for the known Goethe *Rezensionen*.

The following table shows our results for the unattributed or controversial cases (1-4 Unverified), to schematize and compare the confidence of our results with Haenelt's.

		Haenelt	Kestemont, Martens, Ries	Result
1	Cymbelline [...] Length: 2pp Self-attributed by Goethe: yes	Ambiguous case, she decided Herder is most probable. 2 x Merck 2 x Herder 1 x Goethe	Borderline case, reaches the attribution threshold for Goethe with current data. Verification probability: 0.31	Similar result, which might lean more towards Goethe.
2	Empfindsame Reisen durch Deutschland [...] Length: 4pp Self-attributed by Goethe: yes	Highest probability: Goethe. 4 x Goethe 1 x Herder	Positive: Goethe. Verification probability: 0.5	Almost same result.
3	Essays on song-writing: [...] Length: 5pp Self-attributed by Goethe: no	Positive: Herder. 5 x Herder	Positive: not Goethe. Verification probability: 0.02	Same result.
4	Die schönen Künste in ihrem Ursprung, [...] Length: 7pp Self-attributed by Goethe: yes	Positive: Goethe. 5 x Goethe	Reasonable chance Goethe authored. Verification probability: 0.42	Almost same result.

Haenelt came to her results based on her score matrix of five, somewhat more intuitive features - average sentence length, vocabulary composition and distribution, sentence transition position, sentence last positions, sentence second position. From the above table and the heatmap, it can be deduced that we come to similar results compared to Haenelt's findings, which she achieved with this set of features. Like Haenelt, our method clearly excludes Goethe as author of *Essays on song-writing: [...]* (3 Unverified). Likewise, it strongly suggests him as the author of *Die schönen Künste in ihrem Ursprung, [...]* (4 Unverified) and even more clearly than Haenelt's approach as the author of *Empfindsame Reisen durch Deutschland [...]* (2 Unverified).

The result for *Cymbelline [...]* (1 Unverified) is also similar to Haenelt's, it is a borderline case - but suggests a different conclusion: While Haenelt's criteria point to either Merck or Herder with equal scores, there is also one scorepoint for Goethe, although she then without a clear argument decides to state that Herder is the most probable candidate. With 0.31 as a score for Goethe in our approach, the threshold for attribution to Goethe would be reached, as it is in the margin of other known Goethe

Rezensionen - in fact, 3 of 4 known Goethe *Rezensionen* in our test set scored lower (0.22, 0.26, 0.27, 0.34) -, but it is still much lower than the very clear attribution values of 2 and 4 Unverified (0.5, 0.42). With the caveat that *Cymbelline* [...] (1 Unverified) is only 2 pages long and that we worked with a dirty OCR corpus, we can tentatively suggest that Goethe's self-attribution might have been correct and that this text should be cross-verified for Herder and Merck as potential authors with a cleaned corpus to corroborate or challenge this hypothesis. If Herder happens to score equally high, this might be an indicator for collaboration between the two authors in this case. This could seem plausible given the spelling of "Schäckespear", usually regarded as typical for Goethe, and the contextual fact that Herder had just, early 1772, finished the second version of his famous Shakespeare-article, published in 1773.²⁷ This possibility is especially relevant in the light of the almost erratic attribution history of *Cymbelline* [...] (1 Unverified). Goethe attributed the text to himself; Scherer and Biedermann confirmed this attribution, but Trieloff saw both Goethe's and Merck's style in the text, while Morris decided in favour of Herder's authorship. Bräuning-Octavio concluded 1966 that it must be Goethe, mainly based on content-arguments and the spelling "Schäckespear".²⁸ Haenelt, finally, overrules this intuition again in favor of Herder, relying on an early version of stylometry.

Conclusions

Considering the status of Goethe as one of the most important authors of German literature and world literature, it is striking that previous research did not reach a consensus on the authorship attribution question concerning his *Rezensionen* in the *Frankfurter gelehrte Anzeigen* of 1772/73 (*FgA*). Our article traced the research tradition of the authorship attribution problem concerning Goethe's contributions to the *FgA* and his editorship, indicating that despite the fact that linguistic, style identification and even early stylometric methods came into the view of philologists since 1903, none of the discussed approaches has been applied to the whole corpus at large scale. These early proof-of-concept studies often used intuitive definitions of style, relied on closed-set approaches and operated on a limited corpus set defined by the *FgA* itself and few, isolated style examples taken from elsewhere in Goethe's works. 1966 and 1970, Bräuning-Oktavio, Thiele and Sparmann proposed "simple text characteristics" such as word frequency of definite article and expletives as stylistic markers to solve the authorship attribution problem, but did not reach a generalised, consistent definition of linguistic style markers. Sparmann was the first to apply such a linguistic method to a very small, *FgA*-only corpus, finding that texts by Merck might be distinguishable from Goethe's by his more frequent use of the definite article. Karin Haenelt's study, published 1984, was the first to propose computational methods and put them to work on a small number of test cases, based on a *FgA*-only corpus, yet her linguistic definition of Goethe's stylistic features was not tested and lacked methodological foundation. With the recent innovations in computational stylometric authorship verification research, a new, open-set road to solve the problem can be taken which utilises a generalised and well-tested definition of stylistic features that is being trained on large corpora.

²⁷ Franz Zinkernagel, *Herders Shakespeare-Aufsatz. Mit Anmerkungen herausgegeben von F.Z.* (Bonn: Marcus und Weber Verlag 1912), 2, passim.

²⁸ Bräuning-Oktavio 1966, p. 526, 528.

In our blind test trial run, the stylometric authorship verification method proved to be effective to confirm the attribution of known Goethe *Rezensionen* in the *FgA*, clearly distinguish Goethe's style fingerprint from Herder's, as well as largely confirm results for controversial and unattributed cases achieved by Haenelt's closed-set study. Methodologically, this is a large step forward, as the stylometric authorship verification approach as an open-set approach takes into account that the tested texts might have been penned by one of the almost 40 other authors, is based on a well-tested linguistic style model and is trained on a large corpus that is not *FgA*. The test run proved that the method works with reasonable accuracy despite the fact that for pragmatic reasons we had to work with a dirty OCR corpus in this case.

In three of the disputed or previously unattributed cases (2-4 Unverified), our method comes to the same conclusions as Haenelt's. In the case of 2 Unverified, the resulting score even suggests a stronger Goethe style signal than Haenelt's results. The controversial case of *Cymbelline [...]* yielded a result similar to Haenelt's: it is a borderline case. But in our test setup, the Goethe style signal just reached the attribution threshold and was actually stronger than the signal in some texts that have been attributed to Goethe based on external evidence. Therefore, while Haenelt had equal attribution scores for Merck and Herder, and decided to name Herder as most probable candidate, our results suggest to re-run our test with clean OCR training data and cross-verify for Merck and Herder in order to determine either Goethe's authorship or a possible collaboration between Goethe and Herder.

We conclude this essay with an assessment of what our trial run means for future research on Goethe's contributions to *FgA* and what can be expected of the future application of this method. Our results suggest that it would be promising to subject more texts from the corpus to the same test and to test the whole *FgA* 1772/73, which is beyond the scope of the present article. Much work remains to be done before we can settle the question of Goethe's "hand" in the corpus. An obvious extension of the present research would be to draw on a wider range of "impostors", i.e. texts authored by contemporary authors writing on similar content matters and within similar genre characteristics, but excluding texts by Goethe himself (e.g. from the post-1773 editions of *FgA*). For the sake of our limited test-bed experiment, it is worthwhile to stick to the following conclusions: in view of periodical studies, thanks to the advent of huge corpora and digitisation, it has become more feasible to resituate even canonical authors like Goethe in the "thick of things". Even if the method does not provide unanimous certitude, it can help us to establish very precious estimates of the error margins and the degree of incertitude besetting the matter. A significant circumstance is the imperfect nature of the OCR digitisation of the sources. Dealing with texts even in their imperfect shape might seem heretic from the point of philology, but earlier experiments have shown that additional post-processing does not lead to statistically significant changes in the results of authorship attribution.²⁹ An additional caveat applies to the challenge of dealing with short texts and with data cleaning. Since the presentation of our preliminary results, new research has seen the light of day on "Short Samples in Authorship Attribution".

²⁹ Greta Franzini and others, 'Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm', *Frontiers in Digital Humanities*, 5 (2018), 4.

Acknowledgements

We want to extend special thanks to the Heidelberg University Library, which kindly provided the project with high quality scans of the *Frankfurter gelehrte Anzeigen* of 1772/73 (FgA). We would also like to thank Dr. Mirko Nottscheid, Johannes Gutenberg University Mainz, for important hints that led to the initiation of this project, and Sven Vandendriessche, who helped us out with the OCR-postprocessing. The scans have been made available on the University Library's website, Heidelberger historische Bestände - digital. 1772: <https://doi.org/10.11588/diglit.27329> 1773: <https://doi.org/10.11588/diglit.27330>.

Bibliography

- Biedermann, Woldemar von, *Goethe und Leipzig*, 2 vols, Leipzig: Brockhaus 1865.
- Bräuning-Oktavio, Hermann, *Studien zu den Frankfurter Gelehrten Anzeigen vom Jahre 1882* (Gießen et al. 1911).
- Bräuning-Oktavio, Hermann, *Herausgeber Und Mitarbeiter Der Frankfurter Gelehrten Anzeigen 1772* (Tübingen: Niemeyer, 1966).
- Bräuning-Oktavio, Hermann, Wetterleuchten der literarischen Revolution. Johann Heinrich Merck und seine Mitarbeiter an den Frankfurter Gelehrten Anzeigen 1772 in Bild und Wort (Darmstadt 1972).
- Dahnke, Hans-Dietrich, 'Intentionen und Resultate des Jahrgangs 1772 der Frankfurter Gelehrten Anzeigen', in *Sturm und Drang: Geistiger Aufbruch 1770–1790 im Spiegel der Literatur*, ed. by Bodo Plachta and Winfried Woesler (Berlin: de Gruyter, 2015), pp. 87–99.
- Dechent, Hermann, 'Die Streitigkeiten Der Frankfurter Geistlichkeit Mit Den Frankfurter Gelehrten Anzeigen Im Jahre 1772', *Goethe-Jahrbuch*, 10 (1889), 169–95.
- Drewek, Raimund and M. Erni, 'LDVLIB (LEM): A System for Interactive Lemmatizing and Its Application', in Coling 1982 *ABSTRACTS: Proceedings of the Ninth International Conference on Computational Linguistics Abstracts*, 1982.
- Drewek, Raimund, LDVLK-Textanalyse mit System, in Walter Lehmacher, Allmut Hörmann, Statistik-Software. 3. Konferenz über die wissenschaftliche Anwendung von Statistik-Software 1985, Stuttgart, New York: Fischer 1986, pp 283-295.
- Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, and others, 'Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm', *Frontiers in Digital Humanities*, 5 (2018), 4, doi: 10.3389/fdigh.2018.00004.
- Haenelt, Karin, 'Die Verfasser Der Frankfurter Gelehrten Anzeigen von 1772. Ermittlung von Kriterien Zu Ihrer Unterscheidung Durch Maschinelle Stilanalyse', *Euphorion*, 78 (1984), 368–382.
- Hirzel, Ludwig, 'Goetheana'. *Im neuen Reich*, 8, 1878, II, 597/611.
- Knödler, Stefan, 'Frankfurter gelehrte Anzeigen', in *Handbuch Sturm und Drang*, ed. by Matthias Luserke-Jaqui (Berlin ; New York: de Gruyter, 2017), pp. 422–28.
- Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, Walter Daelemans, 'Authenticating the writings of Julius Caesar', *Expert Systems With Applications* 63 (2016), pp. 86-96.
- Kestemont, Mike, Els Stronks, Martine de Bruin, Tim de Winkel, 'Did a Poet with Donkey Ears Write the Oldest Anthem in the World? Ideological Implications of the Computational Attribution of the Dutch National Anthem to Petrus Dathenus'. *Digital Humanities 2017, DH 2017, Conference Abstracts*, McGill University & Université de Montréal, Montréal, Canada, August 8-11, 2017. URL: <https://dh2017.adho.org/abstracts/079/079.pdf>.

- Koppel, Moshe, and Yaron Winter, 'Determining If Two Documents Are Written by the Same Author', *Journal of the Association for Information Science and Technology*, 65 (2014), 178–187.
- Marbe, Karl, Ueber Den Rhythmus Der Prosa. Vortrag, Gehalten Auf Dem 1. Deutschen Kongress Für Experimentelle Psychologie Zu Giessen (J. Ricker, 1904).
- Martus, Steffen, *Werkpolitik: Zur Literaturgeschichte Kritischer Kommunikation Vom 17. bis Ins 20. Jahrhundert ; Mit Studien Zu Klopstock, Tieck, Goethe Und George*, Historia Hermeneutica, 3 (Berlin ; New York: W. de Gruyter, 2007).
- Morris, Max, *Goethes und Herders Anteil an dem Jahrgang 1772 der Frankfurter Gelehrten Anzeigen* (Stuttgart, Berlin: Cotta, 1909).
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and others, 'Scikit-Learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12 (2011), 2825–2830.
- Ritter, Carl, 'Anwendung der Sprachstatistik auf die Recensionen in den Frankfurter Gelehrten-Anzeigen von 1772'. *Goethe-Jahrbuch* 24 (1903), pp 185-203.
- Scherer, Wilhelm, 'Einleitung', Bernhard Seuffert (Ed.), *Frankfurter gelehrte Anzeigen. Nachdruckausgabe*. 1. (Heilbronn: Henninger 1882/83), III-XC.
- Seuffert, Bernhard (Ed.), *Frankfurter gelehrte Anzeigen. Nachdruckausgabe*. Mit einer Einleitung von Wilhelm Scherer. 2 vols. Deutsche Litteraturdenkmale des 18. Jahrhunderts in Neudrucken herausgegeben von Bernhard Seuffert (Heilbronn: Henninger 1882/83).
- Sparmann, Herbert, 'Häufigkeitsuntersuchungen, Ein Hilfsmittel Für Den Vergleich von Texten Und Für Die Feststellung Der Verfasserschaft', *STUF - Language Typology and Universals*, 23.1–6 (1970), 227–231 <<https://doi.org/10.1524/stuf.1970.23.16.227>>.
- Stamatatos, Efstathios, 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology*, 60 (2009), 538–556.
- Stover, Justin Anthony, Yaron Winter, Moshe Koppel, and Mike Kestemont, 'Computational Authorship Verification Method Attributes a New Work to a Major 2nd Century African Author', *Journal of the Association for Information Science and Technology*, 67 (2016), 239–42 <<https://doi.org/10.1002/asi.23460>>.
- Thiele, Joachim, Untersuchung der Goethe zugeschriebenen Rezensionen in den Frankfurter Gelehrten Anzeigen mit Hilfe einfacher Textcharakteristiken, in: *Studia Linguistica* 20 (1966), 83–85.
- Trieloff, Otto, *Die Entstehung der Rezensionen in den Frankfurter Gelehrten-Anzeigen vom Jahre 1772* (Münster: Schöningh 1908).
- Wolf, Norbert Christian, 'Heinrich Christian Boies Göttinger Musenalmanach und Johann Heinrich Mercks Frankfurter gelehrte Anzeigen. Medienkämpfe im literarischen Feld des Sturm und Drang', in *Sturm und Drang: Epoche, Autoren, Werke*, ed. by Matthias Buschmeier and Kai Kauffmann (Darmstadt: WBG, 2013), pp. 10–28.
- Witkowski, Georg, 'Einleitung'. *Goethes Werke*, ed by H. Düntzer, G. Witkowski, K.J. Schröer, A.G.Meyer, vol. 26, (Stuttgart, Tübingen: Cotta 1892).
- Zinkernagel, Franz, *Herders Shakespeare-Aufsatz*. Mit Anmerkungen herausgegeben von F.Z. (Bonn: Marcus und Weber Verlag 1912).