Artem Suslov

Dr. Eijun Senaha

Scholar & Scholarship I

2023

An Annotated Bibliography:

Digital Humanities in Literary Studies

Methods Overview (1986 – present)

Table of Contents

Abbreviations 3
Introduction 4
Theory on Digital Literary Studies 11
Textbooks, Tutorials and Companions 30
Distant Reading 36
Stylometric Analysis 42
Other Methods and Approaches 78
Sentiment Analysis 100
Visualization 104
Digital Technologies in Literary Scholarship and Education 107
Appendix 122

Abbreviations

AA	Authorship Attribution
AI	Artificial Intelligence
BoW	Bag of Words
DH	Digital Humanities
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NN	Neural Network
PCA	Principal Component Analysis
SA	Stylometric Analysis
TEI	Text Encoding Initiative

Introduction

This bibliography introduces 234 items, exemplifying digital methods in literary studies, placed in eight sections: “Theory on Digital Literary Studies”, “Textbooks, Tutorials and Companions”, “Distant Reading”, “Stylometric Analysis”, “Other Methods and Approaches”, “Sentiment Analysis”, “Visualization”, “Digital Technologies in Literary Scholarship and Education” which together implement challenging and complicated two-fold task. First, all these sections introduce the development of digital in literary studies, i.e., methods which facilitate quantification and computers dealing with texts. Second, as DH inherited their terminology from a variety of disciplines, including linguistics, mathematical statistics and data science, their apparatus-may be unfamiliar for traditional students of literature and can better be introduced in a clear understandable way without namedropping with terms and concepts.

Before the engagement with these eight parts in details, it takes to apply to the principle of these sections’ composition. They are compiled thematically covering the major problematic fields in the discipline. However, they are typologies, which means that publications are allocated into each section rather conventionally. For example, Franco Moretti’s *Network Theory, Plot Analysis* [198] significantly contributed to literary theory, suggesting analysis of characters’ interactions as a network with nodes and connections, but the bibliography highlight its contribution to visualization practices. Secondly, the sections significantly vary in their size as some directions of DH attracted more researchers’ interest. The literature about SA (which solves the problem of AA analyzing texts quantitative parameters as frequency of the most common words, average word length, etc.) historically was more numerous. However, even if sections devoted to sentiment analysis (evaluates in numbers the emotions of the text and the way they change through narration) or visualizations (practices of picturing some numerical information, intertextual relations in a picturesque form of diagrams, graphs, trees, etc., which makes their understanding clear and easy) are not so voluminous as one on stylometry, it does not mean that these practices should be ignored. Considering this, the structure of the bibliography appears as following.

First, "Theory on Digital Literary Studies" presents the variety of discussions and attitudes of researchers on emergence and development of DH and computer-assisted criticism. The papers from this section outline the development of the discipline, reflecting the invade of SA into literary studies and the discussion of 1993-1994 (see [5, 6, 8, 9, 10]) demonstrated the first problem of statistical computer-assisted methods in literary studies. That crisis lied in the fact that some researchers believed that computers should be used not for precise analysis of a limited number of texts (what stylometry actually did) but for larger generalizations of many texts analysis. One of the answers was Distant Reading which has a separate section in this bibliography.

The other themes which occupied minds of the thinkers were the mutation of the object of literary studies, the text, and how this transformation affects the scholarship [24, 26], eternal problem of position of Humanities in the science [35], reflections on how to use "big data" in literary studies [29] and even the post-colonial critical thoughts on the influence of access to the digital infrastructure (databases, journals, archives, electronic tools) on the success of a digital humanitarian scholar [36]. However, the core theme always remained as the problem of a method in literary studies [1-4, 23, 29, 30, 34].

Second, “Textbook, Tutorials and Companions” introduces the basic books which can be used in (self-)education of how to use digital instruments in literary studies. Although the number of materials is limited and some of them are specified mostly for computer linguists [37], there are two companions [39, 42] which are brief and informative introductions into DH. Also, there are four core textbooks which introduce programming for literary scholars [41, 43-45], three of them are for the R language and one [45] teaches the Python language, however, only [44] is designed for entirely literature students, as previously mentioned ones contain topics for digital humanitarians in general.

Third, “Distant Reading” is devoted to the novel for the beginning of the 21st c. approach in literary studies proposed by Franco Moretti. He argued that close reading was able to cover only insignificant part of world literature which was identified by other researchers as a canon. Instead, Moretti suggested distant reading as an alternative. Before computer-assisted criticism gained its today’s popularity, Moretti proposed that to have a complete picture of literature a literary scholar should deny reading only few canonic novels but face an ocean of less-famous literary works in all countries. Before big data analysis was possible (it was implemented by Jockers and Underwood [52-55] present in the bibliography), Moretti turned from close reading of few canonic texts to reading secondary sources like compound histories of literature, which covered a wider range of non-canonic literary works of many uncommon genres in many countries. This section consists mostly of Moretti’s works, clarifying his approach, one collection of responses to his ideas [50] and other research papers which are inspired by Moretti’s ideas and use “big data” and coding-based instruments [52-55].

Fourth, “Stylometric Analysis” covers the core span of methods in the statistical analysis of literary texts. SA or stylometry is a range of cognate methods analyzing statistical parameters of texts like frequencies of particular text metrics (number of the most frequent words, different ratios (e.g., lexical richness index) and any other calculated parameters, which facilitated AA purposes. This section lists numerous papers implementing SA. However, number of entries in this part do not show the same extreme originality in ideas as number of articles have similar research designs and mostly implement different variations of one stylometric method (in general, Burrow’s Delta [96]) in case of AA between a limited number of texts with known candidates for the authorship). However, this section demonstrates the evolution of stylometry from Burrow’s multivariate analysis of the most frequent words to more complicated Delta method [96], Zeta and Iota methods [105], and the SA based on NN [70, 74].

Fifth, “Other Methods and Approaches” observes papers the methodology of which is problematically distinguishable into other sections. However, all of these books and articles share the usage of statistics and computing for solving any problem of literary research. Some papers [159, 155] apply logical computing for dealing with episodes sequency in problematic texts (with uneven internal chronology) or the initial order of parts of a book or the chronology of manuscripts. A number of ones [142, 153, 158] uses statistics for analysis of prosody in poetry, counting metrical, acoustic features of poetic texts. [163] presents an interesting algorithm of automatic metaphor identification which can benefit numerous literary studies. A range of articles [188, 189] attempts to implement some structuralist and post-structuralist approaches formalizing these ideas in computer programs. Thus, the implementations of computing methods vary and represent many scholars’ vigor.

Sixth, "Sentiment Analysis" introduces different research papers utilizing the approach of the same name. Sentiment analysis separately evaluates every word in text sentences (some its varieties consider complicated sematic inter-sentence relations) and gives the score (usually from +1 to -1) which indicates whether the text tends to be more positive or negative in emotions and the intensity of sentiment in a number. The annotated articles analyze the change of mood in the text (tonality) and imply the results to questions of literary investigation. Although there are just a few articles touching this method and its one of the best implementation was presented in a textbook from other section [49], sentiment analysis still deserves attention.

Seventh, "Visualization" reviews papers demonstrating the representation of numerical or other non-visual data as relations between elements into an illustrative form of schemes, graphs, trees, maps, etc. may contribute to better text understanding. This section is also rather tricky as the majority of papers, presented in the bibliography, have a plenty of maps, graphs, charts, and different whimsical pictures and figures. This section covers articles which use this visualization as an independent instrument of text analysis, which becomes not a supplementary part, but a major instrument not a supplement to literary analysis.

Eighth and final, the publications in "Digital Technologies in Literary Scholarship and Education" contains two groups of articles. The first one shares experience of different researchers and instructors who introduced digital instruments into their classroom. As the majority of these articles were written in the 1990s, all the novel digital instrument like usage of scanned texts (OCR or optical character recognition) and their distribution via email or special services, web-sites of the course with the forum for students' communication, studying materials and more intensive feedback from instructors, experiments with programming courses for humanities students and so on are not so fresh for the current readers of these papers but indicate the evolution of our literature classes. Another part of articles introduces the digital infrastructure of the literary scholarship, digital editions of literary texts, databases, electronic journals, different corpora, annotations of electronic texts and whole corpora via a special mark-up language (TEI project). Thus, this section observes the routine of literary scholarship being transformed via its digitalization.

Chronologically this bibliography starts from the foundation of the *Digital Scholarship in Humanities* Journal – an authoritative journal in this field, currently in Q1 in linguistics and language[1]. Historically the scope can be proved with the history of introducing computers in a wider range of academic tasks: in the late 1980-s academicians started to use rather compact and user-friendly machines not only for calculations in sciences (physics, engineering, etc.) but in literary scholarship, which some items from the bibliography will demonstrate. The papers of the previous period are not extremely numerous, but the approaches demonstrated in the article of the selected period can be effectively applied even in our current research.

The core of this bibliography consists of related articles from the two major journals in the discipline, *Language Resources and Evaluation* (previously *Computers and the Humanities*) and the mentioned *Digital Scholarship in Humanities.* The bibliography, as a rule, does not cover publications in minor journals, conference and workshop proceedings.

As this bibliography intends to be available to a general reader without any experience in specified topics, except literary studies, in “Appendix” there is a special

[1] Quartile 1 which means that the journal stands with top 25% of the most ranked journals in this field. “Digital Scholarship in the Humanities.” *Scimago Journal & Counting Rank*, www.scimagojr.com/journalsearch.php?q=21100465194&tip=sid&clean=0. Accessed 29 May 2023.

“Terms & Concepts” section which contains definitions and explanations of key terms, without which the demonstration of importance and academic contributed mentioned below papers may be merely understood. However, for specialists from these close spheres of computer science, computerized linguistics, statistics and so on these definitions may be not sufficiently precise and too brief, but they represent only the ease of a “complicated” term, the major mechanism of that or so algorithm, but without minor details, unnecessary for a literary scholar.

Theory on Digital Literary Studies

[1] Potter, Rosanne G. "Literary Criticism and Literary Computing: The Difficulties of a Synthesis." *Computers and the Humanities*, vol. 22, no. 2, 1988, pp. 91–97, https://doi.org/10.1007/BF00057648.

Comprehends on two opposite dimensions in using computer-driven approach in literary criticism: utter lack of training in computer methods and deepening into an overcomprehensive jargon of programming and statistics. On the one hand, Potter acknowledges that computer methods can lead critics to discover new perspectives. However, she also states that "the people and the machines are not really ready for each other" and that "literary computing does not replace scholarship" (p. 93). She emphasizes the limitations of computer methods, such as losing the sense of using such methods in jargon and reducing the text to mere statistical parameters. Her main point is that it is essential to "know when to stop using technology" (p. 97), as criticism requires a level of specificity beyond what statistics and NLP can offer.

[2] Corns, Thomas N. "Computers in the Humanities: Methods and Applications in the Study of English Literature." *Literary and Linguistic Computing*, vol. 6, no. 2, 1991.

Critically examines the "disappointing achievements", almost absolute lack of papers conducting computer-assisted methods in academic journals on English Studies, (p. 127) of digital methods in literary studies. The analysis highlights a crucial limitation of sentiment and broader statistic-based approaches, as researchers tend to lose the link between the metrics they study and the actual text of the literary piece.

Corns suggests that digital literary scholars should shift their focus towards developing large literary databases to realize the principle of intertextuality, which was a prominent aspect of the poststructuralist mainstream during that period.

[3] Fortier, P. A. "Theory, Methods and Applications: Some Examples in French Literature." Literary and Linguistic Computing, vol. 6, no. 3, July 1991, pp. 192–96, https://doi.org/10.1093/llc/6.3.192.

Argues that current computer-based criticism is overly fixated on algorithms for data manipulation and lacks the integration of theoretical ideas. Fortier suggests that computer criticism should encompass psycho-criticism, socio-criticism, and formalist-structuralist criticism, paying particularly precise attention to the latter.

[4] Bruce, Donald. "Towards the Implementation of Text and Discourse Theory in Computer-Assisted Textual Analysis." *Computers and the Humanities*, vol. 27, no. 5–6, Sept. 1993, pp. 357–64, https://doi.org/10.1007/BF01829386.

Aligns with Olsen's paper [5], highlighting the issue that humanities computing is excessively focused on close text analysis and fails to introduce new theoretical concepts that necessitate studying not just individual texts but also the relationships between texts and discourse. Furthermore, Bruce proposes the need for the invention of macro-analysis (prior to Jockers and Moretti).

[5] Olsen, Mark. "Critical Theory and Textual Computing: Comments and Suggestions." *Computers and the Humanities*, vol. 27, no. 5–6, Sept. 1993, pp. 395–400, https://doi.org/10.1007/BF01829390.

Provokes a theoretical discussion on the achievements of computer-assisted literary criticism. Olsen argues that “humanities computing has demonstrably failed to have an impact on text-oriented disciplines” (p. 395), which have largely remained unchanged in their methodology. In response to possible objections from other researchers, Olsen contends that current computing practices have not fully implemented the latest theories and are limited to mere counting of strokes and variables. He emphasizes the critical importance of “maintaining that difficult balance between theory, method, and empirical verification” (p. 399). This article shares a similar spirit with Bruce [4].

[6] Goldfield, Joel D. “An Argument for Single-Author and Similar Studies Using Quantitative Methods: Is There Safety in Numbers?” *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 365–74.

Another response to Olsen’s paper [5], concerning the limited success of computer-assisted methods in the analysis of individual pieces of writing, comes from Goldfield. Goldfield argues that even frequency analysis of specific words in texts can make valuable contributions to criticism. The method illustrated involves comparing the z-score of a word in the text with its expected frequency based on a normal mathematical distribution. If the z-score of a word significantly differs, it holds a certain meaning for both the author and the critic. As a result, the article demonstrates that in the future, computer-assisted criticism will play a significant role in micro-analysis and even semantic analysis of texts.

[7] Matsuba, Stephen Naoyuki. “Finding the Range: Linguistic Analysis and Its Role in Computer-Assisted Literary Study.” *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 331–40.

Outlines how syntactic parsing can be employed for the literary analysis of literary texts, using AI as an example to develop specific mechanisms for text comprehension, which serve as the foundation for various literary theories. Although computer programming has not yet been able to create real AI capable of solving Matsuba's problem, the belief is that this direction will pave the way for future computer-based models.

[8] Olsen, Mark. “Critical Theory and Textual Computing: Comments and Suggestions.” *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 395–400.

Responses to the papers criticizing and supporting Olsen's suggestions that “humanities computing has failed to have a significant impact on text-oriented disciplines” (see also [2] which discusses this problem) and that it is necessary to bring new theories into computer-aided research (p. 395).

[9] Olsen, Mark. “Signs, Symbols and Discourses: A New Direction for Computer-Aided Literature Studies.” *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 309–14.

Collects numerous responses in the same volume of the journal. Here, Olsen states that the lack of significant success in current computer-assisted research is not due to the lack of calculating facilities, computer memory, or other internal factors.

Instead, Olsen argues that the existing researchers have been overly focused on individual texts or comparing a limited number of authors. According to Olsen, computer facilities are being used improperly, and computer-reliant literary scholars should shift their focus towards the study of intertextuality and discursivity, which are aspects that were largely overlooked during that period.

[10] Taylor, Dennis. "Literary Texts and the State of the Language: The Role of the Computer." *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 341–47.

Specifies Olsen's idea on using computers for the analysis of larger masses of data. He suggests using computers to track the influence of authors on the development of language. Taylor demonstrates how data from the Oxford English Dictionary can be utilized to identify which writers have contributed more to the introduction of neologisms. He provides evidence that with a greater number of texts digitized, it becomes possible to track the influence of language formation more accurately and in-depth. By leveraging computational methods and larger datasets, researchers can gain valuable insights into the evolution of language and the impact of authors on linguistic development.

[11] Havholm, Peter, and Larry Stewart. "Computer Modeling and Critical Theory." *Computers and the Humanities*, vol. 30, no. 2, 1996, pp. 107–15, https://doi.org/10.1007/BF00419786.

Shares and critically analyzes the author's experience of introducing new computer technologies into the study of their classes on pre-Romantic English literature. They discuss three positive impacts of using these technologies. Firstly, having digital texts accessible on the computer screen made them more readily available to students compared to physical books in a library. Secondly, when studying 17th-century English drama, the authors encouraged students to create hyperlinks to specific fragments of the text that required commentary. This implementation of hyperlinks allowed for the exploration of intertextuality in real-time and motivated students to read more extensively than they might have done with traditional library resources. Thirdly, the authors provided students with prepared literary attributes, which allowed them to apply a formalist approach to the text, specifically conducting Propp's analysis of character functions. Overall, the integration of computer technologies in their teaching enabled a more interactive and engaging learning experience, facilitating access to texts, promoting critical reading, and showcasing the application of literary theories.

[12] Rudman, Joseph. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities*, vol. 31, no. 4, 1998, pp. 351–65.

Critically reviews the SA[2]. "A variety of problems in these studies are listed and discussed: studies governed by expediency; a lack of competent research; flawed statistical techniques; corrupted primary data; lack of expertise in allied fields; a dilettantish approach; inadequate treatment of errors. Various solutions are suggested: construct a correct and complete experimental design; educate the practitioners; study style in its totality; identify and educate the gatekeepers; develop a complete theoretical

[2] Stylometric analysis, a variety of methods used for AA which use textual statistical parameters as frequency of the most common words, average word length, lexical richness indices and statistical instruments.

framework; form an association of practioners" (p. 351). Besides this, contains a good bibliography of SA papers.

[13] Robinson, Peter. "The One Text and the Many Texts." *Literary and Linguistic Computing*, vol. 15, no. 1, Apr. 2000, pp. 5–14. *DOI.org (Crossref)*, https://doi.org/10.1093/llc/15.1.5.

Discusses the role of modern electronic editions, which were intended to convey different versions of texts without prioritizing any of them. However, modern editions, as exemplified by Chaucer, Dante, and the Greek *New Testament*, actually argue the reverse. Robinson argues that editions should provide all the texts, which better explains which lost original could be, assuming that these editions can help readers become better readers.

[14] McCarty, Willard. "Humanities Computing: Essential Problems, Experimental Practice." *Literary and Linguistic Computing*, vol. 17, no. 1, Apr. 2002, pp. 103–25, https://doi.org/10.1093/llc/17.1.103.

Discusses the current status of humanities computing, including its achievements and limitations as a discipline. It raises the question of defining the primitives in humanities computing, which refers to the simplest set of instruments that humanists can use without interference from other computing fields. Additionally, another problem highlighted is the lack of theoretical grounding in the discipline. However, McCarthy suggests that humanities computing should be developed through an experimental approach to knowledge-making.

[15] Merriam, Thomas. “Linguistic Computing in the Shadow of Postmodernism.” *Literary and Linguistic Computing*, vol. 17, no. 2, June 2002, pp. 181–92, https://doi.org/10.1093/llc/17.2.181.

Observers and comprehends the criticism of determination of authorship, hidden under the umbrella term "theory". This paper covers numerous criticisms, especially one of Barthes, Foucault and Derrida because of whom AA is distinguished from authorship ascription.

[16] Pelt, Tamise van. “The Question Concerning Theory: Humanism, Subjectivity, and Computing.” *Computers and the Humanities*, vol. 36, no. 3, 2002, pp. 307–18.

Philosophic research, observing the theoretical shift from Heideggerian and Lacanian anti-humanism to modern posthumanism. The author examines the key concept of this philosophic movements, the relation between human and technology. Referring to philosophers, van Pelt shows that a human is no more controlling the technology and even theories which they produce, but just a “user”, led by these theories and technologies. This observation questions the possibility to use previous approaches to texts transformed from print into e-format. However, the author is rather optimistic, referring to the experience of Havholm’s [11] students who modeled Propp's analysis, implemented non-humanistic approach to theory (which denies author's and researchers’ control under their tool-agency) in a humanistic way, believing they control computer like any instrument. In his words, van Pelt aims to “help readers decide whether today's computing environments can still be approached through [...]

antihumanist theories or whether e-texts demand new, media-specific analysis" (p. 307).

[17] Winder, William. "Industrial Text and French Neo-Structuralism." *Computers and the Humanities*, vol. 36, no. 3, 2002, pp. 295–306.

Approaches the implementation of neo-structuralist ideas in SATOR's topoi dictionary, WinBrill syntactical tagger and Rastier's interpretative semantics. In the article Winder introduces the concept of the industrial text, a text which is produced by a machine and can be read and used by a machine in sake of operation of these machines, like a syntactic markup language. He states that the emergence of these industrial texts (which allows extracting much information from texts) will change the literary criticism.

[18] Gardner, Colin. "Meta-Interpretation and Hypertext Fiction: A Critical Response." *Computers and the Humanities*, vol. 37, no. 1, 2003, pp. 33–56.

Introduces the interest towards a new type of hyperfiction, which differs from the traditional book in multivariativity and unfixness. Gardner states that hyperfiction has no fixed form for every reader, and every reading is individual for a reader who, by choosing the options, constructs the variation of a text. This demands a new methodology of studying this type of fiction. He argues that criticism of such fiction should be observative. He conducts an experiment using a special program to observe and record how real readers read *Afternoon; a Novel* by M. Joyce and analyzes how much time readers spend on episodes and how similar their obtained versions of the text are. Additionally, he discusses crucial limitations of this experiment, such as the "camera effect", "random selection", and "overinterpretation".

[19] Rockwell, Geoffrey. "What Is Text Analysis, Really?" *Literary and Linguistic Computing*, vol. 18, no. 2, June 2003, pp. 209–19, https://doi.org/10.1093/llc/18.2.209.

"[…] revisits the question of what text analysis could be. He traces the tools from their origin in the concordance. He argues that text-analysis tools produce new texts generated from queries through processes implemented on the computer. These new texts come from the decomposition of original texts and recomposition into hybrid new works for interpretation. The author ends by presenting a portal model for how text-analysis tools can be made available to the community" (p. 209).
Quoted from the annotated paper.

[20] Gottschall, Jonathan. *Literature, Science, and a New Humanities*. Palgrave Macmillan, 2008. *DOI.org (Crossref)*, https://doi.org/10.1057/9780230615595.

Literary studies are at a tipping point. There is broad agreement that the discipline is in "crisis" - that it is aimless, that its intellectual energy is spent, that all of the trends are bad, and that fundamental change will be required to set things right. But there is little agreement on what those changes should be, and no one can predict which way things will ultimately tip. Literature, Science, and a New Humanities represents a bold new response to the crisis in academic literary studies. This book presents a total challenge to dominant paradigms of literary analysis and offers a sweeping critique of those paradigms, and sketches outlines of a new paradigm inspired by scientific theories, methods, and attitudes.

Review taken from Literature, Science, and a New Humanities, https://link.springer.com/book/10.1057/9780230615595. Accessed 25 July 2023.

[21] Sculley, D., and Bradley M. Pasanek. “Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities.” *Literary and Linguistic Computing*, vol. 23, no. 4, Sept. 2008, pp. 409–24, https://doi.org/10.1093/llc/fqn019.

Draws theoretical attention to data mining for machine learning algorithms, which are used in digital humanities. It states that, as the discipline is affected by the subjectivity factor, attention to the data and building hypotheses should be more rigorous, even in scientific disciplines. A digital humanist should maintain a clear boundary between the analyzed data and the interpretation of this data.

[22] Berry, Dave M. “The Computational Turn: Thinking About the Digital Humanities.” *Culture Machine*, vol. 12, 2011, pp. 1–22.

Attempts to translate that the current state of digital humanities represents the core of computability in its third wave. If the first wave focused on creating infrastructure of large databases and digital editions for the discipline, and the second wave drew attention to digitally born objects as the subject of digital humanities, the third wave transforms humanities into sciences with the introduction of coding. In general, it emphasizes the ongoing aspiration for the future of humanities in the digitally transforming world.

[23] Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, 2011.

Presents an alternative to Moretti's "distant reading" approach. Ramsay's algorithmic criticism aims to "provide alternative visions" (p. 16) and "shares […] a desire to use the narrowing forces of constraint to enable the liberating visions of potentiality" (p. 32). Thus, he suggests improving close reading practices by introducing algorithms that can extract textual statistical data, aiding researchers in obtaining a new vision of the text. A significant part of the book is devoted to distinguishing humanities from sciences. According to Ramsay, the lack of clear formal methodology in the computing age has resulted in literary researchers struggling to articulate their purposes, not only to external non-expert society but even within their own community. The introduction of such scientific methodologies shall contribute to the dialogue and mutual understanding between literary studies and sciences. Thus, this represents an important milestone in comprehending the inevitable digital transformation.

[24] Eskelinen, Markku. *Cybertext Poetics: The Critical Landscape of New Media Literary Theory*. Continuum, 2012.

Combines ludology and cybertext theory to solve persistent problems and introduce paradigm changes in the fields of literary theory, narratology, game studies, and digital media. The book first integrates theories of print and digital literature within a more comprehensive theory capable of coming to terms with the ever-widening media varieties of literary expression, and then expands narratology far beyond its current confines resulting in multiple new possibilities for both interactive and non-interactive narratives. By focusing on a cultural mode of expression that is formally, cognitively, affectively, socially, aesthetically, ethically and rhetorically different from narratives and stories, Cybertext Poetics constructs a ludological basis for comparative game studies, shows the importance of game studies to the understanding of digital media, and argues for a plurality of transmedial ecologies.

Review taken from *Cybertext Poetics: The Critical Landscape of New Media Literary Theory*, https://books.google.ru/books?id=HsuoAwAAQBAJ&hl=en. Accessed 25 July 2023.

[25] Sanz, Amelia, and Maria Goicoechea. “Literary Reading Rituals and Practices on New Interfaces.” *Literary and Linguistic Computing*, vol. 27, no. 3, Sept. 2012, pp. 331–46, https://doi.org/10.1093/llc/fqs037.

“[…] discusses the reception of electronic literature, focusing on the development of new literary reading rituals in electronic environments in the Hispanic world. It consists of three parts: a revision of the situation of Hispanic reader communities in the electronic environment; the description of the results of a series of literary on-screen reading experiences with university students, and a final exposition of some reading strategies required to bridge the gap between print and online literature” (p. 331).

Quoted from the annotated paper.

[26] Caton, Paul. “On the Term ‘text’ in Digital Humanities.” *Literary and Linguistic Computing*, vol. 28, no. 2, June 2013, pp. 209–20, https://doi.org/10.1093/llc/fqt001.

“In digital humanities, within a core semantic scope, the term ‘text’ occurs ubiquitously, with both mass and count noun senses. This article sets out to define the relationship between the two senses—between some text and a text—and in particular to say what makes a text discrete. Three characteristics of a scholarly edition (considered the normative instance of a countable text) are isolated and discussed in relation to several marginal cases. [...] conclude[s] that two of them—the representation of language and intent to communicate—give us text in the mass sense” (p. 209).

Quoted from the annotated paper.

[27] McGann, Jerome J. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Harvard University Press, 2014.

Manifests the demand of change in literary studies in the digital age. McGann asserts that existing for decades methods of dealing with text should dramatically change and a digital researcher should turn back to philology, which complexly interrogates both with text and the system of thoughts. That defines the direction of his discussion towards the methods of preservation of digital texts and digital projects in general rather than about algorithmic based methods of text analysis.

[28] Earhart, Amy E. *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*. University of Michigan Press, 2015, https://doi.org/10.3998/etlc.13455322.0001.001.

Observers the history of DH from first electronic librarian catalogues and digital editions until the current large projects in DH. Although the book is precise and overviews what has been done in American DH, it overfocuses on ready-made instruments and projects, rather omitting the influence of stylometry and programming on digital literary studies.

[29] Underwood, Ted. "The Literary Uses of High-Dimensional Space." *Big Data & Society*, vol. 2, no. 2, Dec. 2015, p. 205395171560249, https://doi.org/10.1177/2053951715602494.

Theoretically touches the problem of "big data" in humanities. It points that popular statistic methods of the 20th century were mostly appropriate for smaller samplings of structured data. On the contrary big data is an unstructured set of data and machine trained methods are more appropriate for this kind of data.

[30] Erb, Maurice, et al. "Distant Reading and Discourse Analysis." *Le Foucaldien*, vol. 2, no. 1, June 2016, pp. 1-7, https://doi.org/10.16995/lefou.16.

Discusses and criticizes the computerization of two popular approaches, Foucault's historical discourse analysis and Moretti's Distant Reading. Authors criticize Moretti's Distant Reading as "nothing to do with computer-based analysis". His approach does not use any big data and his graphs, trees have no clear methodology of how they processed and need interpretation. Foucault's methods seem to be more formal and suitable for computer analysis but as Foucault did not operationalize key concepts of his model, the current development of computer technologies is unable to commit such non-formalized analysis.

[31] Underwood, Ted, and The NovelTM Research Group. "Genre Theory and Historicism." *Journal of Cultural Analytics*, Oct. 2016, pp. 1–6, https://doi.org/10.22148/16.008.

Introduces the problem of the study of literary genres in current literary studies which often denies to define the genre of a book. The proposed solution lies in historical study of genre development using the computational analysis of large sets of works, which Underwood implements in [54].

[32] Murray, Janet H. *Hamlet on the Holedeck: The Future of Narrative Cyberspace*. The Free Press, 2017.

Shows how the computer is reshaping the stories we live by. Murray discusses the unique properties and pleasures of digital environments and connects them with the traditional satisfactions of narrative. She analyzes the dramatic satisfaction of participatory stories and considers what would be necessary to move interactive fiction from the formats of childish games and confusing labyrinths into a mature and compelling art form. Through a blend of imagination and techno-wizardry, Murray provides both readers and writers with a guide to the storytelling of the future.

Review taken from *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*, https://mitpress.mit.edu/9780262631877/hamlet-on-the-holodeck/. Accessed 25 July 2023.

[33] Gold, Matthew K., and Lauren F. Klein, editors. *Debates in the Digital Humanities 2019*. University of Minnesota Press, 2019.

Comprehensive collective monograph, containing 44 papers by individual contributors. It is divided into 5 sections: first discusses possibilities and constraints of DH and finishes with Ted Underwood's brief paper, that digital humanities no more endanger traditional ones. The second section focuses on theory of digital humanities, like how they interrogate with critical theory, the third part introduces particular methods mostly in neighboring disciplines except for literary studies. The fourth focuses on the institutional structure of DH. The fifth is a discussion on sharp issues as ethics of DH, the heuristic limitations of the discipline, etc.

[34] Kuhn, Jonas. "Computational Text Analysis within the Humanities: How to Combine Working Practices from the Contributing Fields?" *Language Resources and Evaluation*, vol. 53, no. 4, Dec. 2019, pp. 565–602, https://doi.org/10.1007/s10579-019-09459-3.

Another example of the incompatibility between scientific methods and the foundations of humanities. Kuhn states that the reason for the slow progress in introducing scientific methods from Computer Linguistics and Digital Humanities into traditional Humanities is that the latter is based on hermeneutics. He introduces a specific method called "rapid probing" to replace the pre-understanding of text derived from theoretical concepts on which hermeneutic interpretation stands, with pre-understanding based on data. However, he acknowledges that this approach may force researchers to "jump to conclusions" (p. 35).

[35] Papadopoulos, Costas, and Paul Reilly. "The Digital Humanist: Contested Status within Contesting Futures." *Digital Scholarship in the Humanities*, vol. 35, no. 1, Apr. 2020, pp. 127–45, https://doi.org/10.1093/llc/fqy080.

Estimating the current status, achievements, and risks, presents possible strategies for the development of DH. The article concludes that in order to remain relevant and resilient in a world constantly threatened by disruption, DH should adopt more flexible and less hierarchical divisions, open processes, and policies. Embracing flatter organizational structures, incorporating extended and interoperable networks of communities, sources, and technologies, while employing a blended basket of criteria, will prevent identified schisms from becoming "dangerous chasms".

[36] Pawlicka-Deger, Urszula. "Infrastructuring Digital Humanities: On Relational Infrastructure and Global Reconfiguration of the Field." *Digital Scholarship in the Humanities*, vol. 37, no. 2, 2022, pp. 534–50, https://doi.org/10.1093/llc/fqab086.

"[…] intervenes in these questions by discussing social dimensions of global knowledge infrastructure—connection, standardization, and access—to understand the specification and materialization of global DH. As digital practices expand across the world, the DH community struggles to ensure inclusive participation and equal opportunities in developing the field. This article shows that discrepancies in global DH lie at the root of existing infrastructure inequalities. Drawing on science and technology studies, it then argues that in order to overcome these imbalances, the academic community can seek the ‘infrastructuring’ of DH” (p. 534).

Textbooks, Tutorials and Companions

[37] Baayen, R. H. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, 2008.

Introduces a wide range of instruments for statistical text analysis that can be performed using the R language. While these instruments are not specifically oriented towards literary scholarship, they include crucial models such as PCA[3], clustering, and numerous statistical metrics. Additionally, the text can serve as a parallel guide for statistics, as the explanations, although complex, are presented in a manner that can be understood by readers.

[38] Peer, Willie van, and Sonia Zyngier, editors. *Directions in Empirical Literary Studies: In Honor of Willie van Peer*. John Benjamins Pub. Co, 2008.

Devoted to the legacy of Willie van Peer who significantly contributed to the introduction of formal linguistic and quantitative methods into literary research. Part III titled "Computers and the Humanities" contains four chapters by different authors discussing how different tools can be used in literary research. The opening chapter discusses how the ready-made instrument Coh-Metrix can be used in analyzing how van Peer's style of academic papers was being complicated over the years. Another piece analysis how three methods (latent semantic analysis, frequency of shared bigrams, and unigrams yielding) may be used in separating a literary style from non-literary texts. The third text demonstrates how the qualitative coding software ATLAS.ti 5.2 can identify and classify metaphors in literary texts and how this can be applied in

[3] Principle Component Analysis (PCA) is a mathematical method of representation of multidimensional data into a two-dimensional ("plain") form. See more in Appendix.

comparative analysis of numerous literary sources. The fourth paper studies how the lexical analysis of Burrow’s Zeta and Yota [105] (different approaches of combing frequency vocabularies) can characterize the features of modern American poetry. The final text in the section studies the style change of 170 British poets from 1290 to 1949, characterizing how they develop the novelty of their writings to meet the demand for freshness and novelty, expected from high-quality literary works.

[39] Siemens, Ray, and Susan Schreibman, editors. *A Companion to Digital Literary Studies*. Wiley-Blackwell A John Wiley & Sons, Ltd., Publication, 2013.

The text comprises four sections: “Introduction”, “Traditions”, “Textualities”, and “Methodologies”, along with a total of 31 essays. Originally published in 2007, the paper reflects the aspirations of the discipline. In “Traditions”, the focus is primarily on introducing the spread of digital editions and the usage of databases in classical and modern literature. “Textualities” emphasize the transformation of the object of literary studies — the text — into the digital space. It discusses hypertext, visual novels, and digital media, which have become significant areas of interest for modern researchers. In “Methodologies”, Stephen Ramsay's key ideas on “Algorithmic Criticism” are presented. William Winder's essay, “Writing Machines”, showcases a major breakthrough in 2007 — the semantic and syntactic analysis of literary text using NLP (Natural Language Processing) through the NLTK package for Python. David L. Hoover's essay, “Quantitative Analysis and Literary Studies”, introduces stylometric statistical analysis, as pioneered by Burrows (PCA), and its developments up to 2007. The Companion concludes with “Annotated Overview of Selected Electronic Resources” by Tanya Clement and Gretchen Gueguen. Although this section may now be outdated, it still contains collections of digitalized editions of classical literature and ready-made instruments for basic text analysis.

[40] Drucker, Johanna, et al. *Introduction to Digital Humanities: Concepts, Methods, and Tutorial for Students and Instructors*. UCLA, 2014.

Briefly in a form of a textbook with exercises and questions introduces spheres of DH and acquaintances with ongoing research projects in DH. The way of presenting material needs all-time interaction with an instructor, but the book does not demand from a student any vigor for coding, providing them with the list (with brief and illustrative tutorials) of mapping, visualization, network building and text analysis tools.

[41] Arnold, Taylor, and Lauren Tilton. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. Springer International Publishing, 2015, https://doi.org/10.1007/978-3-319-20702-5.

Textbook, instructing several digital instruments which can be implemented with R computing language. The first part is preparatory and introduces indeed all the necessary manuals for R. The second part 5 approaches of digital analysis in humanities: networks, geospatial data, image data, natural language processing and text analysis. For a literary scholar, 3 or 4 are crucial in their use. First, “Networks” demonstrate how to create a graph of characters (characters are vertices of a graph and their interactions with other characters are connections in the form of lines). The instrument is illustrative and allows one to identify hidden core characters. Second, “Natural Language Processing” which except for basic methods of tokenization and lemmatization[4] presents a powerful instrument of automated Name Entity Recognition when in a variety of words algorithm identifies a character's names and combines them in a list. Using a bit more complicated code we can see the distribution of their appearance within the text. Third, “Text Analysis” acquaintances students with basic SA (PCA) which may show how similar or distant texts are (which is valuable in AA). In this part, automated topic identification is also covered (BoW[5], LDA[6]). Forth, “Geospatial Data” introduces an effective way of tagging locations on the maps and visualising them, which can also interest a researcher[7].

[42] Schreibman, Susan, et al., editors. *A New Companion to Digital Humanities*. John Wiley & Sons, Ltd, 2016.

Revisioned and expanded version of the previous companion [39]. If the previous companion mostly focused on literary studies, the content of this edition faces different aspects as digital history, game studies, visualization and other digital spheres, reflecting the turn from literature-centrism of the discipline.

[43] Arnold, Taylor, et al. “Beyond Lexical Frequencies: Using R for Text Analysis in the Digital Humanities.” *Language Resources and Evaluation*, vol. 53, no. 4, Dec. 2019, pp. 707–33. *DOI.org (Crossref)*, https://doi.org/10.1007/s10579-019-09456-6.

[4] Tokenization is an operation of dividing texts into separate words, lemmatization is the operation which converts token into its initial dictionary form. See more in Appendix.

[5] Bag of Words (BoW) is a way of text’s representation into a numerical vector form. See more in Appendix.

[6] Latent Dirichlet Allocation is a mathematical method of automatic identification of major topics in the text (topic modeling). See more in Appendix.

[7] For the history of this approach see Moretti’s [48].

Observes different (mostly NLP lexical) packages for R computing language. In additions, contains the list of basic instruments for text analysis for those who have no coding experience. The book may be used as a guide. Also, introduces the two current digital humanities projects, driven with R, *Photogrammar* and *Federal Writers Project* (FWP).

[44] Jockers, Matthew L., and Rosamond Thalken. *Text Analysis with R: For Students of Literature*. Springer International Publishing, 2020, https://doi.org/10.1007/978-3-030-39643-5.

Textbook, which reveals in detail all methods Jockers used in [52]. The first part of the book deals with so introduction from zero into R and basic metrics of texts, like distribution of tokens (words), analysis of lexical variety, hapax richness, etc, the metrics which can be used in SA. The second part is devoted to collecting text's metadata and it introduces rather effective mechanism of sentiment analysis. The third part instructs how to deal with a large bulk of text (“macronalysis") and mostly step by step explains the methods used in [52]. To some extent, with [45] this textbook nowadays is most suitable for beginners in computer-assisted analysis of literary texts.

[45] Karsdorp, Folgert, et al. *Humanities data analysis case studies with Python*. Princeton University Press, 2021, https://www.humanitiesdataanalysis.org/.

Designed as a textbook in the form of a website, in an easy but not oversimplified way demonstrates how to use Python computing language for dealing with textual data in humanities. First, the textbook almost from the very basics (the material does not demand students to be proficient in programming or mathematical statistics) explains the basic usage of Python, how to extract textual information from different types of files, and how to operate with tabular data. In the part of the textbook devoted to advanced data analysis, the five cases are provided. The major interest for literary students should be put on the “Stylometry and the Voice of Hildegard” chapter, which implements a SA (PCA) in identifying authorship and stylistic feature of a female medieval writer Hildegard of Bingen.

Distant Reading

[46] Moretti, Franco. "Conjectures on World Literature." *New Left Review*, no. 1, Feb. 2000, https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature.

Manifests some ideas of new Moretti's methods of studying world literature. His major assertion is that world literature studied via traditional mechanisms (mostly close reading) becomes segmented (it cannot come through language barrier of different national literary traditions), limited and may operate only with very limited number of canonic texts. The solutions of these limitations are distant reading which trough denial of reading between lines and spreading the number analyzed texts (the great unread) through reading secondary sources of literary history to widen our perception of primary texts. However, the article does not provide any clear model and principles of this distant reading except that "the larger the geographical space one wants to study, the smaller should the unit of analysis be: a concept (in our case), a device, a trope, a limited narrative unit — something like this" (p. 61).

[47] Moretti, Franco. "The Slaughterhouse of Literature." *Modern Language Quarterly*, vol. 61, no. 1, Mar. 2000, pp. 207–28, https://doi.org/10.1215/00267929-61-1-207.

Key manifestation of Moretti's project of distant reading. In this paper Moretti studied why Conan Doyle's detective fiction became a part of literary canon. Moretti demonstrated that in contrast to Doyle's "rivals" he managed to give a reader a clear clue in a detective story. Moretti in details demonstrates this analysis of clues and as always provides rich illustrations of his data.

[48] Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, 2005.

The realization of initial Moretti's vision of distant reading. Not referring to the original texts but explaining the scale of the literary process referring to secondary sources and making some statistical visualizations. Via graphs Moretti attempted to interpret the genre development of the British novel, explaining why the average time of the genre's “life” was 20 years. Via mapping, it shows the benefit of visualizing of novel's space. In Trees Moretti mostly referred to the detective genre, showing its derivation of clues presented in the text. This paper was a fresh vision of dealing with general literary studies, and the history of literature as it introduces rather unfamiliar and alien ways of dealing with literary material for humanitarian. Nevertheless, Moretti is still far from dwelling on computation.

[49] Moretti, Franco. “Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850).” *Critical Inquiry*, vol. 36, no. 1, Jan. 2009, pp. 134–58, https://doi.org/10.1086/606125.

A practical implementation of Moretti manifest on the distant reading [46, 47]. In order to see not only limited canon but whole literature while the digitalization of all printed novel is far from being complete, he uses the bibliography and analyzes titles of the books. Moretti draws attention to some trends in these titles: shift from long annotative titles to short ones, presence of female names ion titles, usage of definite and indefinite articles. All these changes Moretti tries to connect with social history of the British society.

[50] Goodwin, Jonathan, and John Holbo, editors. *Reading Graphs, Maps & Trees: Responses to Franco Moretti*. Parlor Press, 2011.

Contains a number of critical essay-responses to Moretti's [48]. Some of the reviews support Moretti's distant reading, while many others are rather critical and question the basis of Moretti's argumentation, noticing some smaller inaccuracies. Additionally, the edition contains Moretti's responses to these critiques, which makes this book a dialogical work in many ways. Thus, it provides valuable evidence that distant reading was a disputable approach in literary studies.

[51] Moretti, Franco. "Network Theory, Plot Analysis." *New Left Review*, no. 68, 2011, pp. 80–102, https://newleftreview.org/issues/ii68/articles/franco-moretti-network-theory-plot-analysis.

Proposes to use network graphs to illustrate the relationships between characters in literary texts. Moretti suggests to study texts plot by abstracting from the text itself but build a network of characters, linking them with line if they had any dialogic interaction. Thus, these visualizations highlight the core characters and playing the second stage. Also, Moretti introduces his definition of plot symmetry (when characters equally frequently interact with each other), noticing that this symmetric plot model is more frequent in Asian literature.

[52] Jockers, Matthew Lee. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Introduces the quantitative-driven approach to the study of layers of literary texts which he calls “macroanalysis”, comparing it in opposition to macroeconomics to macroeconomics. Using computational analysis to retrieve keywords, phrases, and linguistic features in thousands of digitalized literary texts, demonstrates how to identify trends in the literary development of certain epochs, countries, and ethnic and demographical groups. Although some of his contents appear to be rather formalistic (statistical graph of the number of novels written by males and females in England, the usage of the article “the” in British and American literature, etc.), it attempted to profoundly demonstrate how to cluster into genres different writings and to track influence between numerous writers, comparing distances between literary works, presented as vector space in which these characteristics can be mathematically measured). Also, his papers demonstrate the explanatory power of data visualization.

[53] Moretti, Franco. “‘Operationalizing’: Or, the Function of Measurement in Modern Literary Theory.” *The Journal of English Language and Literature*, vol. 60, no. 1, Mar. 2014, pp. 3–19, https://doi.org/10.15794/JELL.2014.60.1.001.

Considers the problems of operationalizing of texts for reading them distant. Moretti summarizes his ideas present in previous papers and argues to what extent a researcher needs operationalize text, i.e., transfer notions of literary analysis into numerical visualized data and how this “operationalization” contributed literary analysis.

[53] Underwood, Ted. “A Genealogy of Distant Reading.” *Digital Humanities Quarterly*, vol. 11, no. 2, 2017.

“[…] argues that distant reading has a largely distinct genealogy stretching back many decades before the advent of the internet – a genealogy that is not for the most part centrally concerned with computers. It would be better to understand this field as a conversation between literary studies and social science, initiated by scholars like Raymond Williams and Janice Radway, and moving slowly toward an explicitly experimental method. Candor about the social-scientific dimension of distant reading is needed now, in order to refocus a research agenda that can drift into diffuse exploration of digital tools. Clarity on this topic might also reduce miscommunication between distant readers and digital humanists".

Quoted from the annotated paper.

[54] Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. The University of Chicago Press, 2019.

Being to some extent similar to Jockers's [52], demonstrates how using large scopes of texts may contribute to criticism. Underwood on the scale of large statistical amounts undermines many Moretti's opinions that in the “great unread”, Underwood demonstrates that some trends are common both for cannon and less-known texts. He covers several spheres as the development of genre in the English literature, evolution of this genres in diachronic aspect. What was the difference between less known works and bestsellers, applying machine learning methods.

[55] Underwood, Ted, et al. “NovelTM Datasets for English-Language Fiction, 1700-2009.” *Journal of Cultural Analytics*, vol. 5, no. 2, May 2020, pp. 1–30, https://doi.org/10.22148/001c.13147.

"[…] accompanies a collection of 210,266 volumes, predicted to be fiction, that researchers are encouraged to borrow for their own work. We divide the collection into seven subsets with different emphases (for instance, one where books written by men and women are represented equally, and one composed of only the most prominent and widely-held books). Comparing the pictures produced by these different subsets allows us to assess the resilience or fragility of recent quantitative arguments about literary history. Readers can also simply browse the report as a description of English-language fiction in HathiTrust Digital Library" (p. 1).

Quoted from the annotated paper.

Stylometric Analysis

[56] Holmes, David I. “The Analysis of Literary Style--A Review.” *Journal of the Royal Statistical Society. Series A (General)*, vol. 148, no. 4, 1985, pp. 328-341, https://doi.org/10.2307/2981893.

Explains the major statistical parameters, which can be used in the SA (words length, syllables, sentence length, distribution of parts of speech, functional words, vocabulary richness and vocabulary distribution). The list of these metrics for SA is always completely duplicated in [72].

[57] Dierks, Karin. “Automatic Stylistic Analysis of Lyrical Texts.” *Literary and Linguistic Computing*, vol. 1, no. 3, July 1986, pp. 129–35, https://doi.org/10.1093/llc/1.3.129.

An early example of a pivot in SA is the transition from lexical features of the text to the identification of poetic stylistic figures. Dierks presents two programs, STYLE and AKME, capable of identifying such figures as anaphora, epiphora, geminatio (kyklos, symploke, gradatio), and irregular parallelism in poetic texts. Nowadays, the algorithms presented in the article can be implemented faster and more efficiently.

[58] Burrows, John F. “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style.” *Literary and Linguistic Computing*, vol. 2, no. 2, Jan. 1987, pp. 61–70, https://doi.org/10.1093/llc/2.2.61.

Repeats the algorithm of statistical analysis of narrative style based on frequency matrix of the most common words in fragments of 2000 words decomposition into eigenvectors. The article is based on a more limited number of Austen's texts and causes the same questions and problems as Burrows' [64].

[59] Craik, E. M., and D. H. A. Kaferly. "The Computer and Sophocles' Trachiniae." Literary and Linguistic Computing, vol. 2, no. 2, Jan. 1987, pp. 86–97, https://doi.org/10.1093/llc/2.2.86.

SA was employed to accurately date Sophocles' play *Trachinae*. Craik and Kaferly conducted a study where they counted the ratio of vowels to consonants, the position of consonants in words, and consonant group alliteration in trimester sections of lines across 7 plays. They then implemented PCA and identified that *Trachinae* appears to be the earliest work by Sophocles. However, these results are far from being robust and contradict traditional criticism.

[60] Frautschi, Richard L. "Focal Style and the Problem of Authorship Attribution in Eight Prose Tale by Perrault." *Literary and Linguistic Computing*, vol. 2, no. 4, 1987, pp. 213–20.

Proposes a new design for AA and defines a list of several narrational features (connected with narrated and narrating time) in *The Mother Goose Tales*. He analyzes these features statistically, using Markovian chains and chi-square tests of modal length and function words. Although he cannot conclusively determine AA, he identifies the presence of two narrational styles in the *Tales*.

[61] Stratil, M. “A Disputed Authorship Study of Two Plays Attributed to Tirso de Molina.” *Literary and Linguistic Computing*, vol. 2, no. 3, July 1987, pp. 153–60, https://doi.org/10.1093/llc/2.3.153.

Successfully prove the authorship of two 17th-century plays by Tirso de Molina. They conducted six analyses, including sentence length and form, word length and distribution, total word counts, word frequency, word content, and contextual cluster analysis. They pointed out that the traditional stylometric approach of mechanically comparing text categories was not efficient enough for their case. Therefore, they focused on the use of words, including co-occurrence and collocations, as well as an analysis of irregular verb conjugations. While they didn't implement PCA, which was known to researchers at that time, they performed cluster analysis, which demonstrated the close relationship between the three pieces they studied.

[62] Baker, John Charles. “Pace: A Test of Authorship Based on the Rate at Which New Words Enter an Author’s Text.” *Literary and Linguistic Computing*, vol. 3, no. 1, Jan. 1988, pp. 36–39, https://doi.org/10.1093/llc/3.1.36.

Introduces a different stylometric measure called “author's pace”, which represents the speed of appearance of new unique words in the text. Using this metric, the article proves previous criticism regarding the maturity of some Shakespearean works, among others. It states that, compared to Shakespeare, Marlowe introduces new words more rapidly.

[63] Benson, James D., and Barron Brainerd. “Chesterton’s Parodies of Swinburne and Yeats: A Lexical Approach.” *Literary and Linguistic Computing*, vol. 3, no. 4, Oct. 1988, pp. 226–31, https://doi.org/10.1093/llc/3.4.226.

Reinvigorates the use of Bayesian statistics for estimating lexicon frequency in literary texts. Benson and Brainerd employed Bayesian probability to examine the occurrence of lexicon in Chesterton's parodies of Swinburne and Yeats compared to the originals. They concluded that Chesterton's imitations were rather precise in replicating the lexicon of the genuine poets. This research not only showcases the use of computer-based criticism but also exemplifies the precise formation of text samples from Swinburne and Yeats for comparison, combining the spirit of traditional scholarship.

[64] Burrows, John F. “‘An Ocean Where Each Kind. . .’: Statistical Analysis and Some Major Determinants of Literary Style.” *Computers and the Humanities*, vol. 23, no. 4–5, Aug. 1989, pp. 309–21, https://doi.org/10.1007/BF02176636.

Demonstrates identification of the similarity of texts using a method of PCA. Burrows took dialogue lines from Jane Austen and from several other writers (Henry and Sarah Fielding, Cleland, Collins, Dickens, Lennox, Scott, etc.), builds tables of the most frequent words in fragments, found their share in per cent and built correlation matrix for each writer. Then he decomposed this matrix into eigenvectors and eigenvalues. The method is uncomplicated in realization and can demonstrate how close or distant the writing style of writers is. However, it illustrates the danger of such computing formalism. As Burrows himself writes, “such graphs make vivid pictures. Pictures of what?”. This simple statistical analysis of textual data may give a picturesque graph, which demands interpretation which exceeds this analysis.

[65] Frautschi, Richard L. “Focal Patterns in Textual Samples from Balzac, Flaubert, Zola, and Proust.” *Literary and Linguistic Computing*, vol. 4, no. 4, Oct. 1989, pp. 274–80, https://doi.org/10.1093/llc/4.4.274.

Continues Frautschi's experiments by employing the statistical chi-square test in the analysis of literary works. Based on the correspondence of narrating and narrated time, the researcher defined four temporal modes and applied them to four novels by Balzac, Flaubert, Zola, and Proust. By analyzing the parameters of chi-square tests, the study successfully demonstrated narratological differences between older and newer novels. However, it is worth noting that Frautschi's methods appear to be more formalistic, even when compared to Burrow's analysis of Austen's idiolects (see [64]).

[66] Irizarry, Estelle. “Exploring Conscious Imitation of Style with Ready-Made Software.” *Computers and the Humanities*, vol. 23, no. 3, June 1989, pp. 227–33, https://doi.org/10.1007/BF00056145.

Utilizes readily available computer software to analyze the success of imitations of colonial historical chronicles in a novel by Puerto Rican writer Rodriges Julia. Using the software, Irizarry analyzes the average length of sentences in both the authentic chronicles and the pseudo-chronicles present in the novel. Additionally, the study compares vocabularies and examines the context in which interesting collocations are used.

[67] Logan, Harry M., and Grace B. Logan. “The Case of the Canterbury Pilgrims: Sentence Semantics and World View in Frag. I of *The Canterbury Tales*.” *Literary and Linguistic Computing*, vol. 5, no. 3, 1990.

Explores how Chaucer differentiated his characters by comparing the General Prologue to The Canterbury Tales with the Tales of the Miller and the Reeve. The researchers used syntactical markup of action types to analyze the texts. By corresponding the lists of concordances, they concluded, through analyzing the results, that Chaucer indeed developed idiolects for his characters.

[68] Holmes, David I. “Vocabulary Richness and the Prophetic Voice.” *Literary and Linguistic Computing*, vol. 6, no. 4, Oct. 1991, pp. 259–68, https://doi.org/10.1093/llc/6.4.259.

“[…] applies a multivariate approach for measuring the vocabulary richness of a literary text to thirteen textual samples. These samples are taken from both the personal writings and the prophetic voices of Joseph Smith and Joanna Southcott, and from the King James Bible. It is shown that statistical procedures based on vocabulary richness prove sensitive enough to distinguish between samples from the same genre as well as being able to discriminate between samples from different genres The paper also looks for possible biblical influence on the style of the prophetic voices analysed” (p. 259).

Quoted from the annotated paper.

[69] Burrows, John F. “Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information.” *Literary and Linguistic Computing*, vol. 7, no. 2, Apr. 1992, pp. 91–109, https://doi.org/10.1093/llc/7.2.91.

demonstrates the general principles of Burrow's PCA, which involve building a matrix of word counts for the 50 most frequent words in chunks of works. This mechanism is similar to that used in [64]. However, Burrows aims to showcase how his method can be applied to identify authorship (and broader differences between texts) when considering them as anonymous. He successfully demonstrates that the method distinguishes not only between rather dissimilar authors like J. Austen and H. James but also between the three Brontë sisters.

[70] Matthews, Robert A. J., and Thomas V. N. Merriam. “Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher.” *Literary and Linguistic Computing*, vol. 8, no. 4, Oct. 1993, pp. 203–10, https://doi.org/10.1093/llc/8.4.203.

Pioneers the use of NN in AA. It briefly explains the principles of this mechanism and builds a model that takes as input 5 ratios of certain words in texts, along with the author ID, and outputs whether the piece belongs to Shakespeare or Fletcher's authorship. The NN demonstrated sufficient results in accurately attributing authorship.

[71] Binongo, Jose Nilo G. “Joaquin’s Joaquinesquerie, Joaquinesquerie’s Joaquin: A Statistical Expression of a Filipino Writer’s Style.” *Literary and Linguistic Computing*, vol. 9, no. 4, Oct. 1994, pp. 267–80, https://doi.org/10.1093/llc/9.4.267.

Implements PCA analysis of the 36 most frequently used common words to identify whether two different in style texts were written by one person. Nick Joaquin's *Tropical Gothic* is an overcomplicated prose, tending to be written in extremely long sentences with the usage of prolific vocabulary. On the contrary, his *Joaquinesquerie* is composed in simple language accessible for elementary school students. The PCA analysis of the most frequent words successfully proved that both texts were written by one person, making this a powerful tool for AA.

[72] Holmes, David I. “Authorship Attribution.” *Computers and the Humanities*, no. 28, 1994, pp. 87–106.

Provides 18 statistical text metrics which can be used in SA. Also, the article covers how the statistical analysis deals with “change over time” in literary style, reviewing papers on the authorship of Biblical texts.

[73] Ledger, Gerard, and Thomas Merriam. “Shakespeare, Fletcher, and the Two Noble Kinsmen.” *Literary and Linguistic Computing*, vol. 9, no. 3, 1994, pp. 235–48.

“[…] shows how letter frequency measurements provide the means to discriminate between authors. The data is analyzed using Cluster Analysis and other techniques of multivariate analysis. Texts of Shakespeare and Fletcher (and other authors) are found to differ markedly from each other. This information is used to ascertain which parts of Two Noble Kinsmen resemble Fletcher, and which resemble Shakespeare. The twenty-six scenes of the play are then allocated according to the results obtained” (p. 235).

Quoted from the annotated paper.

[74] Merriam, Thomas V. N., and Robert 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, Jan. 1994, pp. 1–6, https://doi.org/10.1093/llc/9.1.1.

“Using principles set out in an earlier paper [70], a NN was constructed to discriminate between the works of Shakespeare and his contemporary Christopher Marlowe. Once trained using works from the core canon of the two dramatists, the network successfully classified works to which it had not been previously exposed. In the light of these favourable results, we used the network to classify a number of anonymous works. Strong support emerged for Tucker Brooke's view that The True Tragedy is the Marlovian original of *Henry VI, Part 3*, the latter being the product of subsequent revisions by Shakespeare” (p. 1).
Quoted from the annotated paper.

[75] Greenwood, H. H. “Common Word Frequencies and Authorship in Luke’s Gospel and Acts.” *Literary and Linguistic Computing*, vol. 10, no. 3, July 1995, pp. 183–87, https://doi.org/10.1093/llc/10.3.183.

Analyzes the frequencies of common words in selected New Testament texts using “multivariate representation by non-linear mapping techniques with cluster analysis” (p. 183). The article found out that texts attributed to St. Luke were composed by several different authors.

[76] Holmes, D. I., and R. S. Forsyth. “The Federalist Revisited: New Directions in Authorship Attribution.” *Literary and Linguistic Computing*, vol. 10, no. 2, Apr. 1995, pp. 111–27, https://doi.org/10.1093/llc/10.2.111.

Demonstrates three AA techniques: a multivariate approach to vocabulary richness using PCA analysis, analysis of frequencies of occurrence of sets of common words (see Burrows [64]), and a machine-learning algorithm that generates rules for classifying between Hamilton or Madison. Additionally, the article observes previous stylometric research on the authorship of *The Federalist Papers*.

[77] Ilsemann, H. “Computerized Drama Analysis.” *Literary and Linguistic Computing*, vol. 10, no. 1, Jan. 1995, pp. 11–21, https://doi.org/10.1093/llc/10.1.11.

Introduces “dramalists.exe”, a program for automated text analysis of plays. Ilsemann demonstrates the application of this program on Lillo's *The London Merchant* and a set of plays for comparison. The study utilizes the statistical metrics collected by dramalists.exe, such as scenes where characters appear, the length of their dialogues, whether they are mono-, dia-, or polilogues, and more, for text analysis. Additionally, the author attempts to build a classification of play types (theatre of ideas, epic play, melodrama, etc.) based on the extracted quantified data.

[78] Laan, Nancy M. “Stylometry and Method. The Case of Euripides.” *Literary and Linguistic Computing*, vol. 10, no. 4, Nov. 1995, pp. 271–78, https://doi.org/10.1093/llc/10.4.271.

"Raises some methodological points concerning stylometry. Thorough description of an author's style should be considered the sine qua non of any stylometric study and, ideally, a study of the differences in style within the works of an author should precede an attribution or chronology study concerning that same author. These points are illustrated by a discussion of the results of some stylometric studies of Euripides (485-406 BC), including a number of preliminary results from my own work on elision in his iambic trimeters" (p. 271).

Quoted from the annotated paper.

[79] Lowe, David, and Robert Matthews. "Shakespeare vs. Fletcher: A Stylometric Analysis by Radial Basis Functions." *Computers and the Humanities*, vol. 29, no. 6, Dec. 1995, pp. 449–61, https://doi.org/10.1007/BF01829876.

Develops the further usage of frequencies of common words (such as *like*, *a*, *the*, *of*, etc.) in AA. Lowe and Matthews extend the method and build NN based on Radial Basis Functions. NN takes the frequency of five common words in a textual fragment (*are*, *in*, *no*, *of*, *the*) as input and outputs whether the text belongs to Shakespeare or Fletcher. The results support the conventional criticism of attribution of disputed plays, while the research opens a new chapter in AI usage in stylometry and literary studies.

[80] Mealand, D. L. "Correspondence Analysis of Luke." *Literary and Linguistic Computing*, vol. 10, no. 3, July 1995, pp. 171–82, https://doi.org/10.1093/llc/10.3.171.

"[…] applies multivariate statistics to the *Gospel of Luke*. The method most used is correspondence analysis. The paper tests source theories current in conventional New Testament scholarship. The text is divided into 500 word samples. A variety of criteria are used: the variables include function words, parts of speech, and some letter variables. The Infancy narratives are distinct from the rest of Luke. Within the main text the Q material is fairly distinct from the Markan material and the L material. Distinguishing the Markan material from L is not so clear. Where good separation was achieved an attempt was made to replicate this with fewer variables to identify a smaller variable set that would still effect separation" (p. 171).

Quoted from the annotated paper.

[81] Baayen, Harald, et al. "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution." *Literary and Linguistic Computing*, vol. 11, no. 3, Sept. 1996, pp. 121–32, https://doi.org/10.1093/llc/11.3.121.

Modifies PCA in AA by proposing to replace real words with other tokens that signify certain syntactic sequences (approximately 2000 types of syntactic phrases). The researchers conclude that if frequencies of the most and less common syntactic replaced words are put into PCA, AA becomes more precise.

[82] Marindale, Colin, and Paul Tuffin. "If Homer Is the Poet of *the Iliad*, Then He May Not Be the Poet of *the Odyssey*." *Literary and Linguistic Computing*, vol. 11, no. 3, Sept. 1996, pp. 109–20, https://doi.org/10.1093/llc/11.3.109.

Conducts discriminant analysis of the most frequent words in *the Iliad* and *the Odyssey*, which are traditionally prescribed to Homer. The analysis revealed significant differences in the frequent words, indicating that both texts have different authors. Marindale and Tuffin compare the results of Homer with different Greek authors and observe that their style variation is not as strong, further supporting the notion that both epics belong to different authors.

[83] Tweedie, F. J., et al. "Neural Network Applications in Stylometry: The Federalist Papers." *Computers and the Humanities*, vol. 30, no. 1, 1996, pp. 1–10, https://doi.org/10.1007/BF00054024.

Utilizes NN for AA of disputed essays from *The Federalist Papers*. The authors use 10 functional words as input for NN. The results of the AA align with previous research, showcasing the benefits of using NN in stylometry.

[84] Barr, George K. "The Use of Cumulative Sum Graphs in Literary Scalometry." *Literary and Linguistic Computing*, vol. 12, no. 2, June 1997, pp. 103–11, https://doi.org/10.1093/llc/12.2.103.

Introduces and utilizes the cumulative sum method (CUSUM) for AA in New Testament texts. CUSUM is used to check the stability of sample parameters, such as sentence length. According to Barr, the curves of CUSUM can be beneficial for comparing texts and AA.

[85] Holmes, David I. “The Evolution of Stylometry in Humanities Scholarship.” *Literary and Linguistic Computing*, vol. 13, no. 3, Sept. 1998, pp. 111–17, https://doi.org/10.1093/llc/13.3.111.

Observes the historical development of SA. Holmes defines stages of its development: the early origins, which were based on analysis of certain textual statistical elements (frequency of a particular term, vocabulary frequency index, etc.), and multivariate approaches, which were first introduced in Burrow's [64]. Additionally, he criticizes the non-working CUSUM method proposed by Barr [84]. The study also introduces new AI-based attributional algorithms, such as [70].

[86] Merriam, Thomas. “Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V.” *Literary and Linguistic Computing*, vol. 13, no. 1, Apr. 1998, pp. 15–28, https://doi.org/10.1093/llc/13.1.15.

Implies a SA based on letter frequencies and the most frequent function words using neural networks for AA of *Titus Andronicus* and *Henry VI* plays. The analysis reveals an influence of Marlowe's (or Peele's) style in these texts, which allows for a hypothesis on different strategies of mutual authorship.

[87] Tweedie, Fiona J., and R. Harald Baayen. “How Variable May a Constant Be? Measures of Lexical Richness in Perspective.” *Computers and the Humanities*, vol. 32, no. 5, 1998, pp. 323–52.

Analyzes different coefficients of lexical vocabulary richness used in SA. While it contains some mathematical apparatus, researchers should focus on this text to use parameters that fit their research design.

[88] Tweedie, Fiona J., et al. "The Provenance of *De Doctrina Christiana*, Attributed to John Milton: A Statistical Investigation." *Literary and Linguistic Computing*, vol. 13, no. 2, June 1998, pp. 77–87, https://doi.org/10.1093/llc/13.2.77.

Studies *De Doctrina Christiana* prescribed to Milton using the SA of the most frequent words. It concluded that only some parts of the book can be definitely attributed with Milton's style, which means that this text cannot reflects Milton's Christology and his religious ideas entirely.

[89] Forsyth, Richard S. "Stylochronometry with Substrings, or: A Poet Young and Old." *Literary and Linguistic Computing*, vol. 14, no. 4, Dec. 1999, pp. 467–78, https://doi.org/10.1093/llc/14.4.467.

Using a novel method in stylometry, Monte-Carlo Feature-Finding (derives from mathematical method of studying random processes) for stylochronometrics. This method is used for dating texts instead of their AA.

[90] Forsyth, Richard S., et al. "Cicero, Sigonio, and Burrows: Investigating the Authenticity of the Consolatio." *Digital Scholarship in the Humanities*, vol. 14, no. 3, Sept. 1999, pp. 375–400, https://doi.org/10.1093/llc/14.3.375.

Applies PCA for the analysis of the attribution of *Consolatio*, rediscovered in the 16th century, to the Latin orator Cicero. The results demonstrate that the discovered text of *Consolatio* is untypical of Cicero's writing style and cannot be attributed to him.

[91] Waugh, Sam, et al. “Computational Stylistics Using Artificial Neural Networks.” *Literary and Linguistic Computing*, vol. 15, no. 2, June 2000, pp. 187–98, https://doi.org/10.1093/llc/15.2.187.

“[…] examines the use of the Cascade-Correlation algorithm for the construction of minimal networks. We find that a number of problems in computational stylistics with a large number of variables but a limited number of training examples may be solved successfully without resorting to large networks. The issue of redundancy in the data is also considered” (p. 187).

[92] Elliot, Ward E. Y., and Robert J. Valenza. “Smoking Guns and Silver Bullets: Could John Ford Have Written *the Funeral Elegy*?” *Literary and Linguistic Computing*, vol. 16, no. 3, Sept. 2001, pp. 205–32, https://doi.org/10.1093/llc/16.3.205.

Applies SA based on a number of parameters to disprove that *the Funeral Elegy* was not written by Shakespeare. Several tests reject Shakespeare's authorship and indicate with higher probability that the elegy was composed by John Ford.

[93] Holmes, David I., et al. “A Widow and Her Soldier: Stylometry and the American Civil War.” *Literary and Linguistic Computing*, vol. 16, no. 4, Nov. 2001, pp. 403–20, https://doi.org/10.1093/llc/16.4.403.

Implies SA of the most frequent words and clustering for AA of *the Pickett Letters*, which are prescribed to General George Pickett. By uniting traditional historical criticism with computer-based analysis, the authors concluded that the letters were indeed published by the general's widow, LaSalle Corbell Pickett. Her motives were explained by traditional scholarship.

[94] Holmes, David I., et al. “Stephen Crane and the ‘New-York Tribune’: A Case Study in Traditional and Non- Traditional Authorship Attribution.” *Computers and the Humanities*, vol. 35, no. 3, 2001, pp. 315–31.

Using PCA confirmed the outcomes of traditional analysis of writing style that the 17 previously unknown newspaper articles by Stephen Crane written by him for the *New York Tribune* between 1889 and 1892.

[95] McKenna, C. W. F., and A. Antonia. “The Statistical Analysis of Style: Reflections on Form, Meaning, and Ideology in the ‘Nausicaa’ Episode of Ulysses.” *Literary and Linguistic Computing*, vol. 16, no. 4, Nov. 2001, pp. 353–73, https://doi.org/10.1093/llc/16.4.353.

Implements statistical SA, based on Student's t-test and Mann-Whitney test, of the 99 most common words to distinguish narration, dialogue, and internal monologue in the text of *Ulysses*. Additionally, using the same most common words, researchers aim to detect ideological influence in the “Nausicaa” Episode.

[96] Burrows, J. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing*, vol. 17, no. 3, Sept. 2002, pp. 267–87, https://doi.org/10.1093/llc/17.3.267.

Suggests a novel method for AA, which nowadays remains a gold standard. Burrows introduces the Delta (originates from the meaning of "delta"(Δ) as difference used in mathematics) coefficient, which indicates how distant texts are from each other. Like the previous Burrows' solution to AA, it uses frequency lists of common words. This method appears to be a breakthrough as it can distinguish not only between preset author sets but also within anonymous candidates for AA.

[97] Hoover, David L. "Frequent Word Sequences and Statistical Stylistics." *Literary and Linguistic Computing*, vol. 17, no. 2, June 2002, pp. 157–80, https://doi.org/10.1093/llc/17.2.157.

Investigates the efficiency of multivariate analysis, especially cluster analysis, for AA. Hoover concluded that instead of using plain frequencies of separate tokens, it is more efficient and accurate to use frequent word sequences (n-grams).

[98] Barr, George K. "Two Styles in the New Testament Epistles." *Literary and Linguistic Computing*, vol. 18, no. 3, Sept. 2003, pp. 235–48. *DOI.org (Crossref)*, https://doi.org/10.1093/llc/18.3.235.

Implements SA to reveal differences between four Pauline and Pastoral Epistles, which, although demonstrating some affinity between these two sets of texts, have shown distinct variations. The article reapproaches the attitude to these differences, which were previously considered as a sign of different authorship, now suggesting that they indicate a style variation within the writings of one person.

[99] Hoover, David L. “Multivariate Analysis and the Study of Style Variation.” *Literary and Linguistic Computing*, vol. 18, no. 4, Nov. 2003, pp. 341–60, https://doi.org/10.1093/llc/18.4.341.

“[…] investigates style variation in George Orwell’s Nineteen Eighty-Four and William Golding’s The Inheritors using multivariate analysis, specifically, cluster analysis of the frequencies of frequent words” (p. 341). Although it still more focuses on the major stylometric task, AA, multivariate analysis of frequent words is used to compare parts within the novel, indicate the sudden change in style, which makes stylometry more compatible with tasks of traditional criticism.

[100] Merriam, Thomas. “Intertextual Distances, Three Authors.” *Literary and Linguistic Computing*, vol. 18, no. 4, Nov. 2003, pp. 379–88, https://doi.org/10.1093/llc/18.4.379.

“Using an adaptation of C. and D. Labbé’s method of intertextual distances, fourteen short texts by three literary scholars are shown to be distinguishable by author. Of equal interest is the proximity of texts which treat similar themes, or employ similar sub-genres. The interaction of the factors which contribute to intertextual distances enriches the approach and goes some way towards bridging the gap between computational stylistics and traditional literary criticism” (p. 379).

Quoted from the annotated paper.

[101] Somers, Harold, and Fiona Tweedie. “Authorship Attribution and Pastiche.” *Computers and the Humanities*, vol. 37, no. 4, 2003, pp. 407–29.

Using several methods including Yule's K and Orlov's Z metrics of lexical richness, PCA of the 40 most frequent words, discriminant analysis, and weighted CUSUMs (cumulative sums), the study investigates the possibility of AA between the original work and its parodies. The algorithm was tested on Carroll's *Alice in Wonderland* and its later parody by G. Adair, *Alice Through the Needle's Eye*, as well as some other texts. Somers and Tweedie concluded that their design of analysis can distinguish between pastiche and original works but is not as efficient in AA.

[102] Can, Fazli, and Jon M. Patton. “Change of Writing Style with Time.” *Computers and the Humanities*, vol. 38, no. 1, Feb. 2004, pp. 61–82, https://doi.org/10.1023/B:CHUM.0000009225.28847.77.

Demonstrates on the material of two modern Turkish writers how the basic instrument of SA with high precision can be used for the characterization of changes in writers' style over the years. Counting the average length of tokens (words), the most frequent ones in earlier and older novels of writers, Can and Patton made two observations. First, with age, the average length of words in writing increases with statistical significance. Second, they mentioned that frequent preferred “favorite” lexicons also change. Although these findings may seem self-evident, the authors managed to develop an algorithm which can identify if writing belongs to an earlier or later period of a writer’s artistic biography.

[103] Hoover, David L. “Testing Burrows’s Delta.” *Literary and Linguistic Computing*, vol. 19, no. 4, Nov. 2004, pp. 453–75, https://doi.org/10.1093/llc/19.4.453.

Tests the method of AA, Delta, proposed by Burrows in [96]. Hoover tested the method on a larger set of heterogenous texts and widened the list of common frequent words, which Delta analyses. As a result, he proved significant efficiency of Delta and confirmed that it continues to become more accurate as a list of frequent words widens.

[104] Mannion, David, and Peter Dixon. “Sentence-Length and Authorship Attribution: The Case of Oliver Goldsmith.” *Literary and Linguistic Computing*, vol. 19, no. 4, Nov. 2004, pp. 497–508, https://doi.org/10.1093/llc/19.4.497.

The chi square test of sentence length was used for AA of 10 essays attributed to Goldsmith compared to his 16 authentic essays. The analysis demonstrated that within these 6 at least 4 can be attributed to Goldsmith with a significant probability.

[105] Burrows, John. “Who Wrote *Shamela*? Verifying the Authorship of a Parodic Text.” *Digital Scholarship in the Humanities*, vol. 20, no. 4, Nov. 2005, pp. 437–50, https://doi.org/10.1093/llc/fqi049.

Implies SA to search authorship of *Shamela* (1741), a parody of Samual Richardson's *Pamela*. In stead of most frequent words, which show mixed results, Burrows suggests analyzing a list of specific words, which further will be developed in his Zeta and Iota methods [108].

[106] Labbe, Cyril, and Doninique Labbe. “A Tool for Literary Studies: Intertextual Distance and Tree Classification.” *Literary and Linguistic Computing*, vol. 21, no. 3, Aug. 2005, pp. 311–26, https://doi.org/10.1093/llc/fqi063.

Suggests and checks the quality of intertextual distance as a method for text contemplation. The researchers propose finding distances between texts (cosine, Manhattan or Euclidean distances) as vectors and demonstrate how this method contributes to the field using the example of Racine's tragedies and an AA experiment.

[107] Rybicki, Jan. “Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz’s Trilogy and Its Two English Translations.” *Digital Scholarship in the Humanities*, vol. 21, no. 1, Apr. 2006, pp. 91–103, https://doi.org/10.1093/llc/fqh051.

Tests if the unique idiolect of every character remains not only in the original text but also in translation. Rybicki, using Burrow's multivariate analysis [64], studies and proves that both in the original novel by Henryk Sienkiewicz and in two English translations, the individual speech of characters remains still distinguishable.

[108] Burrows, John. “All the Way Through: Testing for Authorship in Different Frequency Strata.” *Literary and Linguistic Computing*, vol. 22, no. 1, Apr. 2007, pp. 27–47, https://doi.org/10.1093/llc/fqi067.

Introduces methods that are, in principle, opposite to Burrows's Delta [96]. Instead of analyzing the most frequent common words as he proposed in previous papers, Burrows suggests analyzing the less frequent words and the least frequent ones.

This was implemented in the Zetta and Iota methods, which also demonstrated satisfactory results.

[109] Tearle, Matt, et al. “An Algorithm for Automated Authorship Attribution Using Neural Networks.” *Literary and Linguistic Computing*, vol. 23, no. 4, Sept. 2008, pp. 425–42, https://doi.org/10.1093/llc/fqn022.

Presents an algorithm of AA based on artificial NN. The authors state that, in comparison with traditional stylometry, NN consider non-linear interactions between input variables. The algorithm was tested on the case of Shakespeare and Marlowe, as well as the traditional case of *The Federalist Papers*. In both cases, the algorithm successfully conducted AA.

[110] Hoover, David L., and Shervin Hess. “An Exercise in Non-Ideal Authorship Attribution: The Mysterious Maria Ward.” *Literary and Linguistic Computing*, vol. 24, no. 4, Dec. 2009, pp. 467–89, https://doi.org/10.1093/llc/fqp027.

Investigates the difficulty of AA of 'Female Life Among the Mormons,' prescribed to Maria Warda. Using cluster analysis, Delta analysis, t-test, and PCA, the article attempts to attribute this text to a number of Mormon texts from the first and third person. Although the authorship problem was not resolved, it demonstrated methods for making progress in difficult conditions for AA.

[111] Holmes, David I., and Daniel W. Crofts. “The Diary of a Public Man: A Case Study in Traditional and Non-Traditional Authorship Attribution.” *Literary and*

Linguistic Computing, vol. 25, no. 2, June 2010, pp. 179–97, https://doi.org/10.1093/llc/fqq005.

Describes both traditional and non-traditional AA approaches used to attribute *The Diary of a Public Man*, an anonymous publicist paper written in 1860-61 in the US. The authors hypothesize that *The Diary* was written by William Henry Hurlbert. Using both traditional analysis and stylometry, they present their arguments in favor of Hurlbert's authorship.

[112] Jockers, Matthew L., and Daniela M. Witten. “A Comparative Study of Machine Learning Methods for Authorship Attribution.” *Literary and Linguistic Computing*, vol. 25, no. 2, June 2010, pp. 215–23, https://doi.org/10.1093/llc/fqq001.

Tests several methods of machine learning-based AA. The paper demonstrates that both specific stylometry methods such as Delta, as well as methods from other disciplines, efficiently differentiate Hamilton's and Madison's essays in a traditional case of *The Federalist Papers*.

[113] Forstall, Christopher W., et al. “Evidence of Intertextuality: Investigating Paul the Deacon’s Angustae Vitae.” *Literary and Linguistic Computing*, vol. 26, no. 3, Sept. 2011, pp. 285–96, https://doi.org/10.1093/llc/fqr029.

Uses computational method to trace influence of Latin poet Catullus on a poem by Paul the Deacon. Analyzing n-grams of words, characters and metrical quantities via supporting vector machines (a major algorithm of machine learning) they traced the influence of Latin poet Catullus on Paul the Dean.

[114] Koppel, Moshe, et al. “Authorship Attribution in the Wild.” *Language Resources and Evaluation*, vol. 45, no. 1, 2011, pp. 83–94.

Provides a new method of AA that 1) identifies an author from a set of many ones, whereas previous methods selected the author from a limited list of candidates; 2) considers that the list of candidates may contain no author; and 3) acknowledges that the “known text for each candidate and/or the anonymous text might be very limited” (p. 84). Their method still uses the “naïve” analysis of frequency of 4-grams as previous ones but adds one more mathematical estimation, which modifies AA.

[115] Luyckx, Kim, and Walter Daelemans. “The Effect of Author Set Size and Data Size in Authorship Attribution.” *Literary and Linguistic Computing*, vol. 26, no. 1, Apr. 2011, pp. 35–55, https://doi.org/10.1093/llc/fqq013.

Tests the approach presented in [114] to measure the efficiency of the model on sets of different sizes and consisting of a certain number of authors. It examines how many texts per author suffice for correct AA. “Results show that, as expected, AA accuracy deteriorates as the number of candidate authors increases and the size of training data decreases, although the machine learning approach continues performing significantly above chance” (p. 35).

[116] Pearl, Lisa, and Mark Steyvers. “Detecting Authorship Deception: A Supervised Machine Learning Approach Using Author Writeprints.” *Literary and*

Linguistic Computing, vol. 27, no. 2, June 2012, pp. 183–96, https://doi.org/10.1093/llc/fqs003.

Describes a new machine learning approach for identifying authorship deception, where one writer attempts to imitate the style of another. Their algorithm, analyzing 81 different metrics, was both effective in AA and in the detection of authorship deception.

[117] Sayoud, Halim. “Author Discrimination between the Holy Quran and Prophet’s Statements.” *Literary and Linguistic Computing*, vol. 27, no. 4, Dec. 2012, pp. 427–44, https://doi.org/10.1093/llc/fqs014.

Implements SA to identify if *the Quran* and *the Hadith* (statements by the prophet Muhammad). The results of the analysis demonstrate that the two books have different authors.

[118] Rybicki, Jan, and Magda Heydel. “The Stylistics and Stylometry of Collaborative Translation: Woolf’s Night and Day in Polish.” *Literary and Linguistic Computing*, vol. 28, no. 4, Dec. 2013, pp. 708–17, https://doi.org/10.1093/llc/fqt027.

Implies cluster analysis of the most frequent words in an untraditional stylometry for identifying not an author but a translator, using the example of the Polish translation of Woolf's *Night and Day*. The article demonstrated that using this method, translator attribution can also be successfully achieved.

[119] Feng, Vanessa Wei, and Graeme Hirst. "Patterns of Local Discourse Coherence as a Feature for Authorship Attribution." *Literary and Linguistic Computing*, vol. 29, no. 2, June 2014, pp. 191–98, https://doi.org/10.1093/llc/fqt021.

Defines a model of AA based on Barzilay and Lapata's entity grids. These grids are "based on the assumption that a text naturally makes repeated references to the elements of a set of entities that are central to its topic. It represents local coherence as a sequence of transitions, from one sentence to the next, in the grammatical role of these references" (p. 191). They tested this approach on a corpus of 19th-century novels.

[120] Forsyth, Richard S., and Phoenix W. Y. Lam. "Found in Translation: To What Extent Is Authorial Discriminability Preserved by Translators?" *Literary and Linguistic Computing*, vol. 29, no. 2, June 2014, pp. 199–217, https://doi.org/10.1093/llc/fqt018.

Addresses the problem of AA of sources that have been translated into a different language. It questions to what extent the translation preserves authorial differences and examines the efficiency of AA in such cases. Forsyth and Lam demonstrated that AA is still possible for translated texts, although it may be less efficient and require the use of appropriate techniques.

[121] Rybicki, Jan, et al. "Collaborative Authorship: Conrad, Ford and Rolling Delta." *Literary and Linguistic Computing*, vol. 29, no. 3, Sept. 2014, pp. 422–31, https://doi.org/10.1093/llc/fqu016.

"Based on Burrows's measure of stylometric difference that uses frequencies of most frequent words, Rolling Delta is a method for revealing stylometric signals of two (or more) authors in a collaborative text. It is applied here to study the texts written jointly by Joseph Conrad and Ford Madox Ford, producing results that generally confirm the usual critical consensus on the visibility of the two author's hand. It also confirms that Ford's claims to a sizeable fragment in Nostromo are unfounded" (p. 422).

Quoted from the annotated paper.

[122] Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." *Digital Scholarship in the Humanities*, vol. 30, no. 2, June 2015, pp. 167–82, https://doi.org/10.1093/llc/fqt066.

Raises the question of what the minimum number of words is in a textual fragment for AA. Using different major methods Eder concluded that modern prose requires a minimum of 5000 words, poetic texts need at least 2500 words, and Greek and Latin novels also require a minimum of 2500 words. Furthermore, the most precise results were achieved using BoW methods instead of the classical division of texts into paragraphs or chunks.

[123] Alqurneh, Ahmad, et al. "Stylometric Model for Detecting Oath Expressions: A Case Study for Quranic Texts." *Digital Scholarship in the Humanities*, vol. 31, no. 1, Apr. 2016, pp. 1–20, https://doi.org/10.1093/llc/fqu038.

Utilizes various methods of machine learning-based SA for the identification and stylistic evaluation of specific verses in the Quran, known as oaths. The findings suggest three main points. Firstly, stylometric application-specific features are best utilized as a combination of both structural-based and content-specific elements, rather than treating them as separate entities. Secondly, the application of stylometric features yielded more significant results in Juz' 'Amma, a section of *the Quran* containing 40% of its surahs (chapters) with oath statements. Lastly, the stylometric model was extended for oath styles detection by incorporating three additional stylometric features: syntactic, character, and lexical. The analysis was carried out using a statistical approach.

[124] Gladwin, Alexander A. G., et al. "Stylometry and Collaborative Authorship: Eddy, Lovecraft, and 'The Loved Dead.'" *Digital Scholarship in the Humanities*, vol. 32, no. 1, Apr. 2017, pp. 123–40, https://doi.org/10.1093/llc/fqv026.

The stylometric method was employed for AA of the short story "The Loved Dead", which has been attributed to Cliffoford Nartin Eddy, Jr. and Sunand Tryambak Joshi. The study concludes that this short story is a collaboration between H.P. Lovecraft and Eddy.

[125] Hoover, David L. "The Microanalysis of Style Variation." *Digital Scholarship in the Humanities*, vol. 32, no. supplement 2, Dec. 2017, pp. ii17–30, https://doi.org/10.1093/llc/fqx022.

Implements cluster analysis based on the most frequent common words to research intertextual style variation. Hoover reveals the problem of inconsistency of narrative style within literary works, which may distort the analysis. He suggests mixing the lines of the text and excluding long sentences from the analysis, which will support the consistency of the literary work, a crucial aspect for inter-author comparison.

[126] Eisen, Mark, et al. “Stylometric Analysis of Early Modern Period English Plays.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, Sept. 2018, pp. 500–28, https://doi.org/10.1093/llc/fqx059.

Suggests a new method of SA based on function word adjacency networks (WANs). In WANs, nodes represent function words, and directed edges between them represent the relative frequency of directed co-appearance of the two words. According to the authors, the WAN method outperforms all frequency-based stylometric methods.

[127] Nini, Andrea. “An Authorship Analysis of the Jack the Ripper Letters.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, Sept. 2018, pp. 621–36, https://doi.org/10.1093/llc/fqx065.

Proposes a new approach to AA by examining the corpus of letters written under the name of Jack the Ripper. Since most of the texts were of insufficient size for a traditional SA, Nini suggested focusing on features of 2-grams and their overlap in the corpus. She successfully demonstrated that “the most iconic texts signed as “Jack the Ripper”, the *Dear Boss letter*, and the *Saucy Jacky* postcard, have been written by the same person” (p. 634).

[128] Oakes, Michael P. “Computer Stylometry of C. S. Lewis’s The Dark Tower and Related Texts.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, Sept. 2018, pp. 637–50, https://doi.org/10.1093/llc/fqx043.

Analyzes *The Dark Tower*, attributed to C.S. Lewis, using the *Stylo* package for the R computing language. The SA revealed that the first 7 chapters were consistent in style with Lewis's previous works. However, the final chapters exhibited a change in genre, shifting from narrative to pseudoscientific style, which was attributed via PCA.

[129] Tuzzi, Arjuna, and Michele A. Cortelazzo. “What Is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, Sept. 2018, pp. 685–702, https://doi.org/10.1093/llc/fqx066.

Employs PCA of the most frequent common words and clustering to investigate the phenomenon of the popular Italian writer Elena Ferrante, who writes under a pseudonym. By comparing her novels within a corpus of 150 novels by different modern Italian writers, the authors concluded that Domenico Starnone might be the person behind her pen name. Despite Ferrante claiming to be of Milanian descent, the research demonstrated that she is distant from other writers from Milan.

[130] Rebora, Simone, et al. “Robert Musil, a War Journal, and Stylometry: Tackling the Issue of Short Texts in Authorship Attribution.” *Digital Scholarship in the Humanities*, vol. 34, no. 3, Sept. 2019, pp. 582–605, https://doi.org/10.1093/llc/fqy055.

Addresses a complex case of AA of short articles, with an average length of 500 words, written by the editor of *Tiroler Soldaten-Zeitung*, Robert Musil. Utilizing various machine learning algorithms and the Delta method of AA within the *Stylo* package for R, the authors successfully developed a suitable algorithm for analyzing short texts.

[131] Hoover, David L. *Modes of Composition and the Durability of Style in Literature*. 1st ed., Routledge, 2020, https://doi.org/10.4324/9780429348068.

Follows Burrow's methodology and “employs the tools and methods of computational stylistics to show that style is extremely resistant to changes in how texts are produced. Addressing an array of canonical writers, including William Faulkner, Joseph Conrad, Thomas Hardy, and Henry James, along with popular contemporary writers like Stephen King and Ian McEwan, this volume presents a systematic study of changes in mode of composition and writing technologies. Computational analysis of texts produced in multiple circumstances of composition, such as dictation, handwriting, typewriting, word processing, and translation, reveals the extraordinary durability of authorial style”.
Review taken from *Modes of Composition and the Durability of Style in Literature*, https://www.routledge.com/Modes-of-Composition-and-the-Durability-of-Style-in-Literature/Hoover/p/book/9780367366704. Accessed 25 July 2023.

[132] Barber, Ros. “Big Data or Not Enough? Zeta Test Reliability and the Attribution of *Henry VI*.” *Digital Scholarship in the Humanities*, vol. 36, no. 3, Oct. 2021, pp. 542–64, https://doi.org/10.1093/llc/fqaa041.

Analyzes the usage of the Zeta method, an AA approach based on less-frequent words compared to the Delta method. Barber suggests that this method is quite multifaceted and presents certain challenges, as it requires a relatively large base-set of at least 100,000 words. It would be more beneficial to consider genre influences in the analysis. However, the major obstacle remains that the Zeta test demands real big data.

[133] Labbé, Dominique, and Jacques Savoy. "Stylistic Analysis of the French Presidential Speeches: Is Macron Really Different?" *Digital Scholarship in the Humanities*, vol. 36, no. 1, Apr. 2021, pp. 153–63, https://doi.org/10.1093/llc/fqz090.

Stands in line with other papers on content and SA of politicians' texts [136, 139]. By sharing the methodology used in those articles, this study also identified that Macron's writing style is distinguishable from previous French leaders, as well as from Trump and Obama.

[134] Rotari, Gabriela, et al. "The Grimm Brothers: A Stylometric Network Analysis." *Digital Scholarship in the Humanities*, vol. 36, no. 1, Apr. 2021, pp. 172–86, https://doi.org/10.1093/llc/fqz088.

"Stylometric methods can be used to reveal similarities between texts and, combined with network analysis, to depict the stylistic relations between those texts. The research conducted here focuses on a corpus of letters written by Jacob and Wilhelm Grimm. Using SA, we model the writing styles of the brothers depending on the addressees and chronology. The brothers have individual styles: Wilhelm has a more friendly and personal tone independent on addresses, while Jacob has a more impersonal style, unless he was writing to Wilhelm. Their styles merge at the interactions of their career or personal development" (p. 172).

Quoted from the annotated paper.

[135] Savoy, Jacques. "Stylometric Analysis of Trump's Tweets." *Digital Scholarship in the Humanities*, May 2021, p. 1-20, https://doi.org/10.1093/llc/fqab048.

Analyzes the stylistic and rhetorical aspects of D. Trump's tweets during his period as a businessman, candidate, and President. Trump's tweets are compared with those of other politicians. The study identifies stylistic features, the most frequent words, references to other sources, and the sentiment of the rhetoric in Trump's tweets. Moreover, the article introduces a general approach to content analysis of political tweets.

[136] Beausang, Chris. "Diachronic Delta: A Computational Method for Analysing Periods of Accelerated Change in Literary Datasets." *Digital Scholarship in the Humanities*, vol. 37, no. 3, Aug. 2022, pp. 644–59, https://doi.org/10.1093/llc/fqab041.

Employs the Delta test for estimating the novelty of fictional, poetic, and dramatic texts in British literature from 1700 to 1922. By using the Delta test score (based on word frequencies) as a measure of novelty compared to a value for a one-year milestone, Beausang developed a trained model that, with frequency input, successfully identified whether a book was written before or after the breakpoint year. Additionally, he identified a list of words whose usage correlated with the novelty of a work.

[137] Faltýnek, Dan, and Vladimír Matlach. “Hapax Remains: Regularity of Low-Frequency Words in Authorial Texts.” *Digital Scholarship in the Humanities*, vol. 37, no. 3, Aug. 2022, pp. 693–715, https://doi.org/10.1093/llc/fqab077.

Reinvigorates the method of stylometric AA based on unique rarely used words (hapax legomena). Testing their approach on a larger dataset, the authors highlight its major advantage: the ability to properly cluster authors even in a set of thousands of sources.

[138] Savoy, Jacques, and Marylène Wehren. “Trump’s and Biden’s Styles during the 2020 US Presidential Election.” *Digital Scholarship in the Humanities*, vol. 37, no. 1, Mar. 2022, pp. 229–41, https://doi.org/10.1093/llc/fqab046.

“[…] analyzes the stylistic and rhetorical characteristics of Donald Trump and Joe Biden during the 2020 US presidential election on the basis of their oral speeches, written texts and tweets” (p.229). As the result, the article provides a list of speech characteristics of both nominees and similarity between their texts and the key words for their articulations.

[139] Hernández-Lorenzo, Laura, and Joanna Byszuk. “Challenging Stylometry: The Authorship of the Baroque Play *La Segunda Celestina*.” *Digital Scholarship in the Humanities*, vol. 38, no. 2, May 2023, pp. 544–58, https://doi.org/10.1093/llc/fqac063.

“verif[ies] the possibility of Sor Juana Ine´s de la Cruz authoring the anonymous part of the baroque *play La Segunda Celestina*, commissioned to Agustın de Salazar”

(p. 544). Using ready stylometric tools in *Stylo* package for R language based on common words frequency, the article proves for the authorship of Sor Juana.

[140] Suddaby, Lee, and Gordon J. Ross. “Did Mary Shelley Write *Frankenstein* ? A Stylometric Analysis.” *Digital Scholarship in the Humanities*, vol. 38, no. 2, May 2023, pp. 750–65, https://doi.org/10.1093/llc/fqac061.

Conducts a SA of Mary Shelley's *Frankenstein* using Burrow's Delta method. The findings refute the myth that the novel was completely written by Percy Bysshe Shelley. The study states that it is highly improbable that there are any writings of Percy's in the text.

Other Methods and Approaches

[141] Philippides, Dia M. L. “Literary Detection in the Erotokhtos and The Sacrifice of Abraham.” *Literary and Linguistic Computing*, vol. 3, no. 1, 1988, pp. 1–11.

Statistically evaluates the features of rhymes for AA of two Cretan Renaissance texts, *Erotokritos* and *The Sacrifice of Abraham*. Philippides focuses on studying the types of rhymes, including whether they are composed with the same parts of speech or different ones, and the endings the rhymes have. This research primarily involves the statistical analysis of phonology and morphology of rhymes. It is worth noting that this research is preliminary and does not formulate any conclusions on AA at this stage.

[142] Roman, Gustavo San. “Using OCP: A Study of Characterization in the Two Don Quixotes.” *Literary and Linguistic Computing*, vol. 5, no. 4, Oct. 1990, pp. 314–18, https://doi.org/10.1093/llc/5.4.314.

Conducts an analysis of the original *Don Quixote* by Cervantes and the mimicric text by Avellaneda. Using the Oxford Concordance Programme (OCP), which can create word lists, indices, and concordances of words in texts, the study states that although the characters of Sancho do not differ significantly, Don Quixote in Avellaneda's version does not seem to be a coarser hero. This is an example of a literary analysis with precision on the lexis and speech of characters.

[143] Snelgrove, Teresa. “A Method for the Analysis of the Structure of Narrative Texts.” *Literary and Linguistic Computing*, vol. 5, no. 3, July 1990, pp. 221–25, https://doi.org/10.1093/llc/5.3.221.

Describes a new approach for analyzing the structure of narrative text and illustrating the reader's response to the text. Snelgrove proposes manually marking up the text into narrative categories (performative or evaluative) and into 17 narrative modes (such as description of character, etc.). By analyzing the frequency distribution, she attempts to measure the reader's response to the text. However, her article is too brief and does not cover the complete mechanism of her new analysis.

[144] Irizarry, Estelle. “Some Approaches to Computer Analysis of Dialogue in Theater: Buero Vallejo’s En La Ardiente Oscuridad.” *Computers and the Humanities*, vol. 25, 1991, pp. 15–25.

“ […] describes approaches to the dialogue of a well-known contemporary Spanish play that are neither time-consuming nor labor-intensive. Ready-made software can produce data and graphs for dialogal structuring, repeat rates of semantic fields, distribution and density of theme words and imagery, interaction of characters, and idiolects” (p. 15).

Quoted from the annotated paper.

[145] Sabol, C. Ruth. “Semantic Analysis and Fictive Worlds in Ford and Conrad.” *Literary and Linguistic Computing*, vol. 6, no. 2, 1991, pp. 97–103.

Examines the collocations of the function word *that* with verbs such as “know”, “think”, “guess”, “see”, and “hear” in a fresh view of Ford's *The Good Soldier* and Conrad's *Lord Jim*. Based on the perception verbs she identifies, Sabol distinguishes four fictive worlds: 1) the world of make-believe, 2) the world of thought/belief, 3) the world of performance, and 4) the world of fact or potential fact. This distinction helps to differentiate between the two novels more accurately.

[146] Koch, Christian. “Combining Connectionist and Hypertext Techniques in the Study of Texts: A HyperNet Approach to Literary Scholarship.” *Literary and Linguistic Computing*, vol. 7, no. 4, Oct. 1992, pp. 209–17, https://doi.org/10.1093/llc/7.4.209.

“Connectionism can offer the literary scholar and student a method of indexing hypertext documents in ways which can uncover patterns of similarities among text segments that might not otherwise be noticed. The basic idea of connectionist networks (also called NN) is explained and this idea is then applied to the analysis, or mapping, of texts. Finally, the basic idea of mapping texts with connectionist networks is incorporated into a design for a Macintosh computer application, called ‘HyperNet’, which is explained in detail” (p. 209).

Quoted from the annotated paper.

[147] Landow, George P. *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. Johns Hopkins University Press, 1992.

Key Landow's monograph on different aspect of hypertext. He traces the predesessing theoretical concepts of hypertext from the French school and analyses the shift from the mere printed text to interconnected hypertext which penetrates literature and art. In addition, a separate chapter is devoted to implementation of connected by hyperlinks text network in education.

[148] Fortier, Paul A. “Babies, Bathwater and the Study of Literature.” *Computers and the Humanities*, vol. 27, no. 5/6, 1994, pp. 375–85.

Opposes M. Olsen's point of view that in the computation age, literary studies will become a sub-field of cultural or social history. Instead, Fortier demonstrates that basic elements of statistical analysis, such as the z-score for estimating word frequency, can be effectively used for the analysis of particular texts (not just for macroanalysis). The article also includes a good introduction to the statistical measures used in quantitative text analysis.

[149] Ousaka, Y., et al. “Automatic Analysis of the Canon in Middle Indo-Aryan by Personal Computer.” *Literary and Linguistic Computing*, vol. 9, no. 2, Apr. 1995, pp. 125–36, https://doi.org/10.1093/llc/9.2.125.

Introduces an effective mechanism for meter analysis of the canonic *Uttarajihaya* in Prakrit text using basic programming. While their algorithm successfully identifies meters in Prakrit, in this version, it is not easily applicable to European languages.

[150] Bucher-Gillmayr, Susanne. “A Computer-Aided Quest for Allusions to Biblical Texts within Lyric Poetry.” *Literary and Linguistic Computing*, vol. 11, no. 1, Apr. 1996, pp. 1–8, https://doi.org/10.1093/llc/11.1.1.

Suggests a simple and effective model for the search of Biblical allusions in lyric poetry. The first step involves searching for the overlap of lemmas in a poetic text with a Biblical vocabulary list (excluding function words). The second approach is to divide the entire Bible into chunks of 200 words each and then compare them with each poetic text to identify potential allusions.

[151] Maczewski, Jan-Mirko. “Virginia Woolf’s The Waves in French and German Waters: A Computer Assisted Study in Literary Translation.” *Literary and Linguistic Computing*, vol. 11, no. 4, Dec. 1996, pp. 175–86, https://doi.org/10.1093/llc/11.4.175.

Focuses on computer-assisted literary translation studies and utilizes the PALIMPSET program tools for analyzing Virginia Woolf's first chapter of *The Waves* along with its German and French translations. The program tool allows researchers to follow the original text and observe the strategies chosen by the translators, as well as how they adhere to the syntactic structure of the original text during the translation process.

[152] Roberts, Alan. “Rhythm in Prose and the Serial Correlation of Sentence Lengths: A Joyce Cary Case Study.” *Literary and Linguistic Computing*, vol. 11, no. 1, Apr. 1996, pp. 33–39, https://doi.org/10.1093/llc/11.1.33.

Proposes a new method of analyzing textual pace. It involves analyzing the frequencies of sentence lengths and calculating the coefficient of lags in the sequence of these frequencies. Roberts discovered a pattern for the rhythm in Joyce Cary's work and suggested that this model of analyzing the frequency of particular textual elements could be applied to other ways of analysis as well.

[153] Laffal, Julius. “Union and Separation in Edgar Allan Poe.” *Literary and Linguistic Computing*, vol. 12, no. 1, Apr. 1997, pp. 1–14, https://doi.org/10.1093/llc/12.1.1.

Proposes a computer-based method of concept analysis, aiming to track the development of the idea of death in Edgar Allan Poe's tales and poetry. Laffal created a list of more than one hundred thematic topics (such as annihilation, birth, weakness, work) and correlated them with the vocabulary used in Poe's texts. Then, he statistically analyzed the frequency of these topics in the texts and calculated their z-scores. Through this analysis, he traced the correlation of the concept of death with various ideas and criticized the traditional criticism, which the results of the study disapprove.

[154] Burg, Jennifer, et al. “Using Constraint Logic Programming to Analyze the Chronology in ‘A Rose for Emily.’” *Computers and the Humanities*, vol. 34, no. 4, 2000, pp. 377–92.

Using constraint logic programming, the problem of the chronology of Faulkner's *A Rose for Emily* was solved. This text presents a complicated narrative structure, where episodes do not follow the true chronology. Despite numerous traditional research on “Emily”, a special program that algorithmically evaluates conditions and hints stated in the text was employed to output the correct sequence of events in Faulkner's short story.

[155] DeForest, Mary, and Eric Johnson. “The Density of Latinate Words in the Speeches of Jane Austen’s Characters.” *Literary and Linguistic Computing*, vol. 16, no. 4, Nov. 2001, pp. 389–401, https://doi.org/10.1093/llc/16.4.389.

Analyzes the author's speech and dialogues of characters in Jane Austen's novels. DeFrost and Johnson calculated the percentage of Latinate words (of Latin and Greek origin) and German words in the texts. Latinate words are considered elegant and high-class, while German words express physical features and are considered low-class. Using this approach, they analyzed the characters' speech and identified which ones are closer to the authorial language of Austen.

[156] Spencer, Matthew, and Christopher J. Howe. “Estimating Distances between Manuscripts Based on Copying Errors.” *Literary and Linguistic Computing*, vol. 16, no. 4, Nov. 2001, pp. 467–84, https://doi.org/10.1093/llc/16.4.467.

Uses a mathematical model to estimate the number of changes between numerous manuscripts. This method allows for the reconstruction of the genealogy of manuscripts from one common ancestor, based on the analysis of copying errors.

[157] Mahoney, Anne. “Talking about Meter in SGML.” *Computers and the Humanities*, vol. 37, no. 4, 2003, pp. 469–73.

Now deserving mostly historical attention, describes major principles of coding the poetic meter in SGML (Standard Generalized Markup Language) which was used in TEI (Text Encoding Initiative) project.

[158] Spencer, Matthew, et al. "Analyzing the Order of Items in Manuscripts of 'The Canterbury Tales.'" *Computers and the Humanities*, vol. 37, no. 1, 2003, pp. 97–109.

By analyzing the order of separate stories from *The Canterbury Tales* by Chaucer, the study provides trees (charts) of manuscripts arranged based on the similarity of story orders. The results support the opinion present in some traditional criticism that the original manuscript by Chaucer was highly likely unfinished and disordered.

[159] Stewart, Larry L. "Charles Brockden Brown: Quantitative Analysis and Literary Interpretation." *Literary and Linguistic Computing*, vol. 18, no. 2, June 2003, pp. 129–38, https://doi.org/10.1093/llc/18.2.129.

Argues that quantitative text analysis can provide background and evidence for literary critical discussions. It analyzes chapters from two novels by Charles Brockden Brown using Principal Component Analysis (PCA) and cluster analysis. The analysis reveals that three chapters from *Wieland* are clustered together with *Carwin*, and they share the same narrating character. This finding indicates that the writer developed a separate idiolect for the narrating character, raising interesting questions for literary criticism.

[160] Milkman, Katherine. L., et al. “A Statistical Analysis of Editorial Influence and Author Character Similarities in 1990s New Yorker Fiction.” *Literary and Linguistic Computing*, vol. 22, no. 3, May 2007, pp. 305–28, https://doi.org/10.1093/llc/fqm011.

Implements statistical analysis of 442 fiction pieces in the *New Yorker* magazine during the 1990s to investigate whether the change of editors influenced the fiction and if authors prefer to write about characters with whom they have no common demographic features. The authors manually coded every text into categories such as the theme of the fiction, race, gender, sexuality of the author, and the same of characters. They proved that the change in Fiction Editor had an effect on the fiction in the magazine, and they also found that fiction writers in the *New Yorker* more often write about protagonists with different race, gender, and country of origin compared to the authors.

[161] Ilsemann, Hartmut. “More Statistical Observations on Speech Lengths in Shakespeare’s Plays.” *Literary and Linguistic Computing*, vol. 23, no. 4, Sept. 2008, pp. 397–407, https://doi.org/10.1093/llc/fqn011.

Presents research conducted via the IDAP program, which analyzes frequency distributions of speech length in four of Shakespeare's plays: *The Merry Wives of Windsor*, *King Henry VI, Part 2*, *Much Ado About Nothing*, and *King Henry V*. The results show that Shakespearean style significantly changed after 1599, which corresponds with other traditional criticism.

[162] Markert, Katja, and Malvina Nissim. "Data and Models for Metonymy Resolution." *Language Resources and Evaluation*, vol. 43, no. 2, June 2009, pp. 123–38, https://doi.org/10.1007/s10579-009-9087-y.

Estimates the efficiency of five systems for figurative language resolution in identification of metonymy (a kind of metaphor). The systems they tested are available for researchers and can be used in practical criticism.

[163] Sotove, Alexandre. "Lexical Diversity in a Literary Genre: A Corpus Study of the *Rgveda*." *Literary and Linguistic Computing*, vol. 24, no. 4, Dec. 2009, pp. 435–47, https://doi.org/10.1093/llc/fqn044.

Analyzes the ratio of word frequency in the *Rgveda*, the Indo-Aryan oral poetry, aiming to demonstrate its dependence on characteristic choices of subject matter, usage of refrains, and the attribution of hymns to distinct poetic collectives (p. 435). Sotove concludes that most cultic and popular hymns contain more frequent function words and lack words that appear rarely (hapax legomena). This finding supports the notion that mythological texts remain consistent and regular in word usage.

[164] Hohl Trillini, Regula, and Sixta Quassdorf. "A 'Key to All Quotations'? A Corpus-Based Parameter Model of Intertextuality." *Literary and Linguistic Computing*, vol. 25, no. 3, Sept. 2010, pp. 269–86, https://doi.org/10.1093/llc/fqq003.

Proposes a data-driven approach using the HyperHamlet corpus, which enables dynamic categorization to accommodate diverse phenomena, which allows identification of hidden quotations and references. The advantages of this approach include comprehensive parameter descriptions, recognition of implicit genre definitions, and a better understanding of interactions.

[165] Clement, Tanya, et al. "Distant Listening to Gertrude Stein's 'Melanctha': Using Similarity Analysis in a Discovery Paradigm to Analyze Prosody and Author Influence." *Literary and Linguistic Computing*, vol. 28, no. 4, Dec. 2013, pp. 582–602. *DOI.org (Crossref)*, https://doi.org/10.1093/llc/fqt040.

Innovatively uses prosodic features such as tone, stress, and sequences of sounds for AA. It analyzes certain prosodic sequences that most readers attribute as "noise". The method was tested on the AA of Gertrude Stein's *Three Lives*.

[166] Muralidharan, Aditi, and Marti. A. Hearst. "Supporting Exploratory Text Analysis in Literature Study." *Literary and Linguistic Computing*, vol. 28, no. 2, June 2013, pp. 283–95, https://doi.org/10.1093/llc/fqs044.

Presents WordSeer, an explanatory environment analyzer of literary text. This tool can analyze lexical collocations in the text and provide interpretations. It "now supports grammatical search and contextual similarity determination, visualization of patterns of word context, and examination and organization of the source material for comparison and hypothesis building" (p. 283). Using the example of the Shakespearean corpus, WordSeer identified a distinction between male and female characters and concluded that in romantic plots, words related to female characters become more associated with physical features. Additionally, the program includes a visualization tool.

[167] Van Dalen-Oskam, Karina. “Names in Novels: An Experiment in Computational Stylistics.” *Literary and Linguistic Computing*, vol. 28, no. 2, June 2013, pp. 359–70, https://doi.org/10.1093/llc/fqs007.

Describes how the analysis of proper names in literary texts necessitates a quantitative approach. Dalen-Oskan collected a set of 22 Dutch and 22 English novels, along with 10 translations, and tagged the context of all the names. She presented some initial statistical results and observations on the usage of proper Dutch geographical names.

[168] Weingart, S., and J. Jorgensen. “Computational Analysis of the Body in European Fairy Tales.” *Literary and Linguistic Computing*, vol. 28, no. 3, Sept. 2013, pp. 404–16, https://doi.org/10.1093/llc/fqs015.

Explores how digital methods can be used to analyze the gender representation of the body in the corpus of European fairy tales. To conduct the analysis, the researchers created a hand-coded database of 233 fairy tales, which lists any appearance of the body in the texts. The findings revealed strong indications that the gender and age of fairy-tale protagonists correlate in ways that indicate societal value being placed on certain perspectives. Specifically, youthful and masculine perspectives are validated as universal, whereas feminine and aged bodies are often marked as ‘other’ (p. 404).

[169] Hoover, David L., et al., editors. *Digital Literary Studies: Corpus Approaches to Poetry, Prose and Drama*. Routledge, 2014, https://doi.org/10.4324/9780203698914.

Explores the potential of digital methods in studying literary texts, with a focus on analyzing style, language, characters, and interpretations of individual works and collections of texts. By utilizing large language databases and smaller specialized collections, the study applies statistical techniques to address questions about authorship and style. It covers various literary genres, including poetry, prose, and drama, and employs different techniques and concepts for each. The overarching aim of the book is to inspire the application of these methods to enhance literary studies.

[170] Forstall, Christopher, et al. "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level n-Gram Matching." *Digital Scholarship in the Humanities*, vol. 30, no. 4, Dec. 2015, pp. 503–15, https://doi.org/10.1093/llc/fqu014.

Presents the third version of the Tesserae website, a tool designed for identifying intertextuality in the corpus of Ancient Greek and Latin texts. This tool is capable of finding similar citations (similar n-grams) even if they are modified in two compared texts. The third version also introduces a scoring system that "sorts the found parallels by a formula accounting for word frequency and phrase density" (p. 503).

[171] Herbelot, Aurélie. "The Semantics of Poetry: A Distributional Reading." *Digital Scholarship in the Humanities*, vol. 30, no. 4, Dec. 2015, pp. 516–31, https://doi.org/10.1093/llc/fqu035.

"[…] proposes that at the semantic level, what distinguishes poetry from other uses of language may be its ability to trace conceptual patterns which do not belong to everyday discourse but are latent in our shared language structure. [...] First, the notion of a specific 'semantics of poetry' is discussed, with some help from literary criticism and philosophy. Then, distributionalism is introduced as a theory supporting the notion that the meaning of poetry comes from the meaning of ordinary language. In the second part of the article, experimental results are provided showing that (i) distributional representations can model the link between ordinary and poetic language, (ii) a distributional model can experimentally distinguish between poetic and randomized textual output, regardless of the complexity of the poetry involved, and (iii) there is a stable, but not immediately transparent, layer of meaning in poetry, which can be captured distributionally, across different levels of poetic complexity" (p. 516).

Quoted from the annotated paper.

[172] Bonch-Osmolovskaya, Anastasia, and Daniil Skorinkin. "Text Mining *War and Peace* : Automatic Extraction of Character Traits from Literary Pieces." *Digital Scholarship in the Humanities*, Dec. 2016, pp. i17-i24, https://doi.org/10.1093/llc/fqw052.

Uses automatic syntactic and semantic analysis to identify character roles in Tolstoy's *War and Peace*. By employing ABBYY COMPRENO tool, they identified parameters for agent, object, experiencer, addressee, and possessor semantic roles to characterize their roles in the novel's text, highlighting their functions from different perspectives.

[173] Bruster, Douglas, and Geneviève Smith. “A New Chronology for Shakespeare’s Plays.” *Digital Scholarship in the Humanities*, vol. 31, no. 2, June 2016, pp. 301–20, https://doi.org/10.1093/llc/fqu068.

“It is widely recognized that Shakespeare’s verse lines grew progressively longer as his career unfolded. Scholars have traditionally used this fact, among others, to date the plays. Drawing on the existing and original data relating to their verbal arrangements, this essay constructs a new chronology for 42 dramatic texts, and parts of texts, by Shakespeare. This chronology is based on a constrained correspondence analysis of the plays’ internal pauses, qualified in relation to a principal component analysis of other verbal features and the recorded closings of the London playhouses owing to plague. The result is a more specific ordering of the Shakespeare canon than has previously been available” (p. 301-20).

Quoted from the annotated paper.

[174] Finlayson, Mark A. “*ProppLearner* : Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory.” *Digital Scholarship in the Humanities*, vol. 32, no. 2, 2017, pp. 284–300, https://doi.org/10.1093/llc/fqv067.

Presents a project of annotating the corpus of Russian hero fairytales according to Propp's *Morphology of Fairytale*. Several teams led by Finlayson worked on this project. The annotators created the annotated corpus of texts in XML formats, which combined 18 layers of annotations. Five of these layers were developed specifically to support learning Propp's morphology, including referent attributes, context relationships, event valences, Propp's ‘dramatis personae’, and Propp's functions.

[175] Koppel, Moshe, and Shachar Seidman. “Detecting Pseudepigraphic Texts Using Novel Similarity Measures.” *Digital Scholarship in the Humanities*, vol. 33, no. 1, Apr. 2018, pp. 72–81, https://doi.org/10.1093/llc/fqx011.

Presents an improved method for identifying outlier documents in a set of works by a particular author that do not belong to that writer. The authors base their method on finding distances between texts, represented as vectors of the number of n-grams. Their method, called “sim2”, was tested on Shakespearean and Pauline corpora.

[176] Navarro-Colorado, Borja. “A Metrical Scansion System for Fixed-Metre Spanish Poetry.” *Digital Scholarship in the Humanities*, vol. 33, no. 1, Apr. 2018, pp. 112–27, https://doi.org/10.1093/llc/fqx009.

Presents an automatic hybrid scansion model for fixed-metre Spanish poetry is presented. “The article is mainly focused on the metrical ambiguities produced by synaloephas: verse lines from which it is possible to derive two or more metrical patterns. This metrical ambiguity is resolved through probabilities, assuming a relation between high probabilities and metricality. The system has been evaluated through more than 1,000 lines extracted from a corpus of Golden-Age Spanish sonnets. An accuracy of 95% has been achieved, resulting in not only considerable progress if we compare it to previous proposals, but also in an adequate way of performing the task when compared to human performance” (p. 112).

[177] Elewa, Abdelhamid. “Authorship Verification of Disputed Hadiths in Sahih Al-Bukhari and Muslim.” *Digital Scholarship in the Humanities*, vol. 34, no. 2, June 2019, pp. 261–76, https://doi.org/10.1093/llc/fqy036.

Analyzes three parameters, word length, word type, and lexical richness, to provide suggestions for authorship verification of disputed Hadiths from Prophetic Traditions. However, based on this analysis, Elewa refrains from making judgments on whether the Hadiths are genuine or disputed, leaving the final word to the theologians.

[178] Liu, Mingming, et al. “Literary Intelligence Analysis of Novel Protagonists’ Personality Traits and Development.” *Digital Scholarship in the Humanities*, vol. 34, no. 1, Apr. 2019, pp. 221–29, https://doi.org/10.1093/llc/fqy020.

Proposes a model for predicting the personality traits of characters based on dialogues. The authors extracted dialogue lines of two characters in a Chinese novel, *Ordinary World*, tokenized them, and used a program to distribute them into 88 semantic categories. Based on these categories, they developed a regression model that predicted five factors of personality: neuroticism, extraversion, openness, agreeableness, and conscientiousness.

[179] Reza Mahmoudi, Mohammad, and Ali Abbasalizadeh. “Statistical Analysis about the Order of Quran’s Revelation.” *Digital Scholarship in the Humanities*, vol. 34, no. 1, Apr. 2019, pp. 152–58, https://doi.org/10.1093/llc/fqy030.

Employs statistical regression and hierarchical cluster analysis to examine the order of suras (chapters) in the *Quran*, following 12 orders proposed by various theologians and researchers. The paper demonstrates that different orders show a high correlation of 90% and can be combined into two cluster groups. However, in comparison with other articles, the methodology and outcomes of this article are found to underperform.

[180] Van Cranenburgh, Andreas, et al. “Vector Space Explorations of Literary Language.” *Language Resources and Evaluation*, vol. 53, no. 4, Dec. 2019, pp. 625–50, https://doi.org/10.1007/s10579-018-09442-4.

Provides a model, which analyzing the writing style of the style, evaluates the literariness of the text (based on the reviews of readers and critics). The research was conducted on a corpora of 401 recent Dutch novels. The texts of novel were divided into chunks of 1000 words and were transformed into vectors (see texts as vectors Bag of words, LDA unigram paragraph vectors (Distributed Bag-of-Words Paragraph Vectors). Evaluating several algorithms (bag of words, LDA, DBoW), they used topic models (LDA) and neural documents embeddings (DBoW Paragraph Vectors). Then they built predictive model which identifies specific words and lexical markers related to literariness. It provides a robust research design for a comparison of a large number of literary works and identifying specific entries, typical, for example, a certain literary epoch or genre. Also, the article contains a link to the Github repository, where all the used code for Python is located and available.

[181] Yeung, Chak Yan, and John Lee. “Dialogue Analysis: A Case Study on the New Testament.” *Language Resources and Evaluation*, vol. 53, no. 4, Dec. 2019, pp. 603–23, https://doi.org/10.1007/s10579-019-09461-9.

Conducts an analysis of New Testament dialogues in terms of the length of the dialogue, the initiator, and the finisher of the dialogue. To implement this task, Yeung and Lee developed an algorithm based on machine learning, which automatically extracts direct speech lines from the text and identifies the initiator and character who finished each dialogue. Their algorithm is highly applicable not only for the Bible but also for any literary text.

[182] Murai, Hajime. "Factors of the Detective Story and the Extraction of Plot Patterns Based on Japanese Detective Comics." *Journal of the Japanese Association for Digital Humanities*, vol. 5, no. 1, Nov. 2020, pp. 4–21, https://doi.org/10.17928/jjadh.5.1_4.

Manually identifies 37 types of plot elements, 10 types of tricks, 9 types of criminal motives, and 10 types of relationships between victims and criminals in 134 detective episodes of the *Detective Conan* manga. Using factor analysis, Murai groups and categorizes these features into 11 categories and builds graphs demonstrating typical plots in *Detective Conan*.

[183] Karlińska, Agnieszka. "The Art of Nerves: A Quantitative and Qualitative Analysis of Drama at the Turn of Nineteenth and Twentieth Century." *Digital Scholarship in the Humanities*, vol. 36, no. 1, Apr. 2021, pp. 122–37, https://doi.org/10.1093/llc/fqaa005.

Analyzes 90 Polish modernist plays between 1890 and 1913 to investigate the influence of medical discourse in the texts. Karlinska manually coded the mild, moderate, and severe symptoms of hysteria and used these subcategories to identify them in her set using a ready statistical tool. She correlates the increase of hysteric characters with the scientific publications on this mental disorder.

[184] Shamir, Lior. “UDAT: Compound Quantitative Analysis of Text Using Machine Learning.” *Digital Scholarship in the Humanities*, vol. 36, no. 1, Apr. 2021, pp. 187–208, https://doi.org/10.1093/llc/fqaa007.

Introduces machine learning driven Universal Data Analysis of Text (UDAT) instrument which analyses 24 metrics and can 1) classify text's genre, 2) find the differences between authors and draw distance graphs between texts, 3) conduct sentiment analysis of texts.

[185] Dobson, James E. “Vector Hermeneutics: On the Interpretation of Vector Space Models of Text.” *Digital Scholarship in the Humanities*, vol. 37, no. 1, Mar. 2022, pp. 81–93, https://doi.org/10.1093/llc/fqab079.

“This essay theorizes two major categories of vector space models, the document-term matrix and neural language models, to position these models as not merely descriptions of texts but inscriptive representational objects that perform interpretive work of their own in order to demonstrate the need for a multi-level hermeneutics in computational literary studies” (p. 81).
Quoted from the annotated paper.

[186] Hyytiäinen, Pasi. "A New Method in Establishing Quantitative Relationships between Manuscripts of the New Testament." *Digital Scholarship in the Humanities*, vol. 38, no. 1, Apr. 2023, pp. 151–66, https://doi.org/10.1093/llc/fqac030.

Applies a new method for tracing similarities between New Testament manuscripts. "This is achieved by using a technique called shingling, where the manuscript transcriptions are turned into smaller pieces called tokens or k-grams in a computerized manner. Then, a string metric is used to calculate the similarities between the tokenized strings" (p. 151). The method is efficient and provides similarity values that do not conflict with the traditional approach.

[187] Karlsen, Heidi. "Foucault's Archeological Discourse Analysis with Digital Methodology—Discourse on Women Prior to the First Wave Women's Movement." *Digital Scholarship in the Humanities*, vol. 38, no. 1, Apr. 2023, pp. 195–208, https://doi.org/10.1093/llc/fqac022.

Observes how Foucault's archaeological discourse analysis may be implemented via digital methods. In *L'archeologie du savoir* (1969) Foucault proposes that there is a large set of fiction and non-fiction texts of a certain era, or Archive ("Great Unread"). This Archive contains statements which are typical thoughts which can be expressed at a certain time in the current discursive limitations. The individual expressions of the statement are called enunciations. Studying the massive of 1830s – 1890s Norwegian literature (over 7000 books), the author formulated 6 gender discursive statements, some of which can be rather sudden - e.g., "(4) Woman must be accorded freedom" (p. 204). Karlsen used STM (Sub-Corpus Topic Modeling) and Bag of Words methods to identify on sub-corpus of decided by her texts of 17 authors in order to identify topics (statements), then she applied STM and NMF (Non-Negative Matrix Factorization) to identify women-related topics in the larger set of digitalized texts of Norwegian Library. Finally, she close read the machine-obtained results, she formulated the set of interrelated gender discursive statements. The proposed approach can be implemented in any historical literary research, aiming identification major ideas, and discourses on a large scale of texts analyzed.

[188] Keskustalo, Heikki, et al. “Analyzing Gender Clues in War-Time Letters.” *Digital Scholarship in the Humanities*, vol. 38, no. 1, Apr. 2023, pp. 209–23, https://doi.org/10.1093/llc/fqac035.

Analyzes gender clues qualitatively and presents a typology of them based on a corpus of private letters from the Second World War in Finland. Although authors managed to identify merely applicable to literary studies gender clues in greetings and signatures of letters, their approach can be developed into the identification of women’s literary language, based on specific stylistic gender markers, typical for a genre.

Sentiment Analysis

[189] Miall, David S. “Estimating Changes in Collocations of Key Words across a Large Text: A Case Study of Coleridge’s Notebooks.” *Computers and the Humanities*, vol. 26, no. 1, 1992, pp. 1–12.

Analyzes the change in Coleridge's thoughts on emotions in his criticism through statistical analysis of collocations. Miall collected Coleridge's criticism and defined a vocabulary of ten words meaning emotion (e.g., feel, feels, passion, etc.). He then found collocations and calculated z-scores for these collocations (z-score indicates how often or seldom this collocation appears in comparison with other collocations in the text). By comparing measures for these collocations, he traced the change in Coleridge's view on emotion in a poet's creativity over the years. Although this study tends to be a SA, the procedure of identifying emotional words and estimating the collocations of these words is not a complicated sentiment analysis. It can indeed indicate the emotions expressed in the text, which is the reason to focus on the text in this section.

[190] Whissell, Cynthia M. “A Computer Program for the Objective Analysis of Style and Emotional Connotations of Prose: Hemingway, Galsworthy, and Faulkner Compared.” *Perceptual and Motor Skills*, vol. 79, no. 2, Oct. 1994, pp. 815–24, https://doi.org/10.2466/pms.1994.79.2.815.

Demonstrates a new feasible algorithm for sentiment analysis of prose texts. The TEXT.NLZ program analyzed the texts of Hemingway, Galsworthy, and Faulkner to check the judgments of traditional criticism. The computed metrics validate the characteristics of these authors as described by traditional criticism.

[191] Whissell, Cynthia. “Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon.” *Computers and the Humanities*, vol. 30, no. 3, 1996, pp. 257–65, https://doi.org/10.1007/BF00055109.

Describes a new method of emotional SA. Using the texts of Lennon's and McCartney's songs, Whissell investigates traditional metrics such as word length and frequency, as well as metrics that indicate emotions (e.g., negative words like “no”, “not”, “-n't”, words like “love”, “dead”, “girl”, and nonsense syllables like “nah”, “yah”, “doo”, etc.). She validates the traditional criticism that Lennon's songs were sadder, while McCartney's were cheerful and upbeat. Additionally, the article clarifies the distinction of emotionality in the songs of different periods for these two authors.

[192] Devitt, Ann, and Khurshid Ahmad. “Is There a Language of Sentiment? An Analysis of Lexical Resources for Sentiment Analysis.” *Language Resources and Evaluation*, vol. 47, no. 2, June 2013, pp. 475–511, https://doi.org/10.1007/s10579-013-9223-6.

Observes the basic principles of sentiment analysis using four major sentiment lexicons for estimating text sentiment. These lexicons, in order of complexity, are the General Inquirer Lexicon, Dictionary of Affect in Language, SentiWordNet, and WordNet-Affect. The first two contain fixed lexicons of words, each of which has a numerical sentiment estimation, while the last two can expand their lexicon independently.

[193] Sprugnoli, Rachele, et al. “Towards Sentiment Analysis for Historical Texts.” *Digital Scholarship in the Humanities*, vol. 31, no. 4, 2015, pp. 1–11, https://doi.org/10.1093/llc/fqv027.

“[…] presents the integration of sentiment analysis in ALCIDE, an online platform for historical content analysis. A prior polarity approach has been applied to a corpus of Italian historical texts, and a new lexical resource has been developed with a semi-automatic mapping starting from two English lexica. This article also reports on a first experiment on contextual polarity using both expert annotators and crowdsourced contributors. The long-term goal of our research is to create a system to support historical studies, which is able to analyse the sentiment in historical texts and to discover the opinion about a topic and its change over time” (p. 1).

Quoted from the annotated paper.

[194] Rahimi, Zeinab, et al. “Applying Data Mining and Machine Learning Techniques for Sentiment Shifter Identification.” *Language Resources and Evaluation*, vol. 53, no. 2, June 2019, pp. 279–302, https://doi.org/10.1007/s10579-018-9432-0.

Examines shifters words, which can change the sign (“+” or “-”) of sentiment of expression. Although this paper is in interest of computer linguist, it may raise theoretical interest for literary scholars, implementing sentiment analysis, for their better understanding of the essence of sentiment metrics they use.

[195] Gritsenko, Daria, et al., editors. *The Palgrave Handbook of Digital Russia Studies*. Springer International Publishing, 2021, https://doi.org/10.1007/978-3-030-42855-6.

The section “Automatic Sentiment Analysis of Texts: The Case of Russian” by Natalia Loukachevitch observers the mechanism of SA applicable for the sources in the Russian language, compares their scopes and limitations. The section Contains a list of lemmatizing mechanism. “Social Network Analysis in Russian Literary Studies” by Frank Fischer and Daniil Skorinkin are implementation of Moretti’s character’s network for a corpus of Russians texts. The rest sections of the handbook bare less interest for literary scholars.

Visualization

[196] Jessop, Martyn. “The Visualization of Spatial Data in the Humanities.” *Literary and Linguistic Computing*, vol. 19, no. 3, Sept. 2004, pp. 335–50, https://doi.org/10.1093/llc/19.3.335.

“[…] discusses some of the intellectual, research and practical issues affecting the use of spatial data in Humanities Computing projects. It is illustrated by examples from a number of projects at King’s College. Although it is grounded on specific examples to support the points being made the main aim is to draw out a series of general themes which affect the use of spatial data by researchers throughout the Humanities and increase awareness of what can be achieved with minimal resources and a little creative thought” (p. 335).

Quoted from the annotated paper.

[197] Moretti, Franco. “Network Theory, Plot Analysis.” *New Left Review*, no. 68, 2011, pp. 80–102.

Proposes using network graphs to illustrate the relationships between characters in literary texts. Three categories are composed for this purpose (add that it should be placed in all three). Moretti suggests studying a text's plot by abstracting from the text itself and building a network of characters, linking them with lines if they had any dialogic interaction. This visualization highlights the core characters and identifies those playing secondary roles. Additionally, Moretti introduces his definition of plot symmetry, where characters interact with each other equally frequently. He notes that this symmetric plot model is more frequent in Asian literature.

[198] Borin, Lars, et al. “Geographic Visualization of Place Names in Swedish Literary Texts.” *Literary and Linguistic Computing*, vol. 29, no. 3, Sept. 2014, pp. 400–04, https://doi.org/10.1093/llc/fqu021.

“ […] describes the development of a geographical information system (GIS) at Spra˚kbanken as part of a visualization solution to be used in an archive of historical Swedish literary texts. The research problems we are aiming to address concern orthographic and morphological variation, missing place names, and missing place name coordinates. Some of these problems form a central part in the development of methods and tools for the automatic analysis of historical Swedish literary texts at our research unit. We discuss the advantages and challenges of covering large-scale spelling variation in place names from different sources and in generating maps with focus on different time periods” (p. 400).

Quoted from the annotated paper.

[199] Alves, Daniel, and Ana Isabel Queiroz. “Exploring Literary Landscapes: From Texts to Spatiotemporal Analysis through Collaborative Work and GIS.” *International Journal of Humanities and Arts Computing*, vol. 9, no. 1, Mar. 2015, pp. 57–73, https://doi.org/10.3366/ijhac.2015.0138.

Describes the LITESCAPE.PT project, which is a database of geographical locations mentioned in 350 Portuguese literary works. It locates the inserts into the map and analyzes the frequency of appearance of these locations on the pages of books. The paper specifically focuses on two cases: the difference between formal geographical Lisbon and Lisbon mentioned in literary pieces (which is smaller), and the appearance of wolves in literary works mostly concentrated in the North of the country. The authors state that the design of the database is applicable to any geography and country's literature.

[200] Alex, Beatrice, et al. “Geoparsing Historical and Contemporary Literary Text Set in the City of Edinburgh.” *Language Resources and Evaluation*, vol. 53, no. 4, Dec. 2019, pp. 651–75, https://doi.org/10.1007/s10579-019-09443-x.

Describes an automatic geoparsing of literary locations in Edinburgh. This method involves tagging the locations mentioned in literary texts on the map. The authors presented the automation of extracting this literary geographical data from a bulk of texts and showcased an illustrative literary map of Edinburgh.

[201] Kutty, Sangeetha, et al. “PaperMiner—a Real-Time Spatiotemporal Visualization for Newspaper Articles.” *Digital Scholarship in the Humanities*, Jan. 2019, https://doi.org/10.1093/llc/fqy084.

“[…] reports on the development of PaperMiner,1 a prototype Web-based service enabling the discovery and visual analysis of connections in time and space between people, places, concepts, and many other historic entities within the 60 million articles comprising the National Library of Australia’s (NLA) Australian Newspaper Service. PaperMiner provides these capabilities through using a combination of text mining, entity extraction, geotagging, and crowdsourcing” (p. 2).

Quoted from the annotated paper.

Digital Technologies in Literary Scholarship and Education

[202] Schnelling, H. “Digitizing Shakespeare: Perspectives of Digital Optical Recording of Renaissance Editions in University Libraries.” *Literary and Linguistic Computing*, vol. 2, no. 1, Jan. 1987, pp. 13–18, https://doi.org/10.1093/llc/2.1.13.

Reflects a certain historic stage of introducing digital editions into scholarship. Schnelling points out that due to the limited availability of old editions of Shakespeare, only a few libraries can obtain them. Previously, facsimile editions, critical editions in old and new orthography, as well as microfilming and xerocopy were suggested as solutions, but by 1987, they became outdated. He introduces DOR (digital optical recording) technology, which may make not only Shakespeare but all old Renaissance editions available to every library and scholar.

[203] Tannenbaum, Robert S. “How Should We Teach Computing to Humanists?” *Computers and the Humanities*, vol. 21, no. 4, 1987, pp. 217–25.

Discusses the views of the author and other instructors of computer subjects for humanities students. Despite being published in 1987, it still highlights currently crucial principles: first, to teach general principles of computing rather than particular details, and to focus on usage rather than programming (although the author himself considers that students should focus on programming). The studies are built on three principles: general principles first, more demonstrations in the classroom, and more experimenting and practice outside the classroom.

[204] Bump, Jerome. “Radical Changes in Class Discussion Using Networked Computers.” *Computers and the Humanities*, vol. 24, no. 1/2, 1990, pp. 49–65.

Mainly focuses on behavioral psychology and introduces innovations made at the University of Texas in instructing literature courses. Students were provided with a computer network, similar to modern Learning Management Systems (LMS). Bumpp mentions that students became more involved in discussions in the course forum and tended to collaborate and have more intensive interaction. However, there were also some disadvantages, such as a lack of instructor control and concerns about dehumanization.

[205] Milic, L. T. “The Century of Prose Corpus.” *Literary and Linguistic Computing*, vol. 5, no. 3, July 1990, pp. 203–08, https://doi.org/10.1093/llc/5.3.203.

“The Century of Prose Corpus, a 500,000-word data base of British prose texts, 1680-1780, is described with respect to its origin, history, contents, structure, provenance, organization, usefulness and availability” (p. 203).

Quoted from the annotated paper.

[206] “The Bard in Bits: Electronic Editions of Shakespeare and Programs to Analyze Them.” *Computers and the Humanities*, vol. 24, 1990, pp. 275–87.

Introduces the three existing digital editions of Shakespeare in 1990, mostly available on CD, and discusses text searching programs and other programming tools for very basic text analysis. Although nowadays this paper may not be directly applicable to current research, it demonstrates the tendency towards the usefulness of electronic editions even in conservative fields like Shakespearean studies.

[207] Holland, Simon, and Gordon Burgess. “Beauty and the Beast: New Approaches to Teaching Computing for Humanities Students at the University of Aberdeen.” *Computers and the Humanities*, vol. 26, no. 4, Aug. 1992, pp. 267–74, https://doi.org/10.1007/BF00054272.

Shares the experience of designing computer courses for humanities students. In 1992, they developed what is now considered basic computer literacy courses. These courses covered work in text editors, familiarization with major computer terms, some basic programming, and work with concordances and database programs. The authors further concluded that for humanities students, a design with minimum programming and more usage of ready applications is more appropriate.

[208] Hawthorne, Mark. “The Computer in Literary Analysis: Using TACT with Students.” *Computers and the Humanities*, vol. 28, no. 1, 1994, pp. 19–27.

Introduces the capabilities of the TACT program for text analysis, which can show search words in context, display search words that fit into one screen of text, present search words in a single line of context like a concordance, and show associated search words (collocations). Using classical English texts as examples, Hawthorne demonstrates that TACT “helps literary critics collect material about a text and read the text in different ways” (p. 19).

[209] Katz, Seth R. “Current Uses of Hypertext in Teaching Literature.” *Computers and the Humanities*, vol. 30, no. 2, 1996, pp. 139–48, https://doi.org/10.1007/BF00419790.

Observes the real-life practice of using hypertext (any hypermedia which uses hypertext) in a classroom of literary courses. Katz covers various applications, such as WWW-based online classes (websites created by instructors containing syllabus and study materials), hypertext tutorials (interlinked pages with data), "annotated" hypertext editions of primary sources that expand printed versions, hypertext writing, and studying the hypertext culture and literature.

[210] Mills, Jon, and Balasubramanyam Chandramohan. “Literary Studies: A Computer Assisted Teaching Methodology?” *Computers and the Humanities*, vol. 30, no. 2, 1996, pp. 165–70, https://doi.org/10.1007/BF00419793.

Reflects the usage of TACT [209] in literary classes studying Conrad's *Heart of Darkness*. The ability to search word frequencies and their usage in the text allowed students to look at the text from a different perspective and enrich their analysis. Additionally, reading the digital version of the novel from a screen, rather than in a paper edition, desensitized the text for students and purified their perception of the piece.

[211] Potter, Rosanne G. “What Computers Are Good for in the Literature Classroom.” *Computers and the Humanities*, vol. 30, no. 2, 1996, pp. 181–90, https://doi.org/10.1007/BF00419795.

Critically summarizes the experience of computer usage in a classroom. On the one hand, Potter criticizes the Landow's hypertexts usage in classroom [148], writing that this makes student passive consumers of ready information, often non-critical. On the other hand, she advocates introduction into classroom asynchronous forums, computer-mediated conversations which involve more students into study process, make them teach not only from lecturer but from each other. To some extent, these problems are sharp even nowadays.

[212] Smith, Jonathan. “What’s All This Hype about Hypertext?: Teaching Literature with George P. Landow’s The Dickens Web.” *Computers and the Humanities*, vol. 30, no. 2, 1996, pp. 121–29, https://doi.org/10.1007/BF00419788.

Describes the use of Landow's hypertext in a classroom, using the example of The Dickens Web. It demonstrates how hyperlinks to web sources may widen students' experience and encourage them to delve deeper into the text, enriching their discussions. However, what was a breakthrough in 1996 is now intentionally implemented in numerous classes, as students consult the vast hypertext of the Internet.

[213] Pidd, Michael, et al. “Digital Imaging and the Manuscripts of *The Canterbury Tales*.” *Literary and Linguistic Computing*, vol. 12, no. 3, Sept. 1997, pp. 197–202, https://doi.org/10.1093/llc/12.3.197.

Presents *The Canterbury Tales Project*, which aims to provide access to all 88 14th-century witnesses to the text in a computer-readable format. The project not only intends to distribute the manuscript's high-quality images at a lower cost but also aims to create an entire pictorial history of the manuscript, becoming a powerful tool for researchers.

[214] Prescott, Andrew. "*The Electronic Beowulf* and Digital Restoration." *Literary and Linguistic Computing*, vol. 12, no. 3, Sept. 1997, pp. 185–96. *DOI.org (Crossref)*, https://doi.org/10.1093/llc/12.3.185.

"*The Electronic Beowulf* is a project edited by Professor Kevin S. Kiernan of the University of Kentucky which is creating an archive of digital images of primary evidence for the transmission of the text of Beowulf. This paper describes the nature of this evidence and the way in which digital imaging technology allows aspects of it to be recorded in a way which would not feasible with conventional photographic technology. The use of digital technology facilitates the virtual restoration of the Beowulf manuscript in a fashion which is also applicable to other damaged manuscripts" (p.185).
Quoted from the annotated paper.

[215] Carter, Bryan. "From Imagination to Reality: Using Immersion Technology in an African American Literature Course." *Literary and Linguistic Computing*, vol. 14, no. 1, Apr. 1999, pp. 55–65, https://doi.org/10.1093/llc/14.1.55.

Encourages instructors to more intensively use the Internet and WWW in their classrooms, using the example of a 3D VR (virtual reality) model of Harlem. This model allows users to visualize, explore, and interact with the literary landscape of African American Literature, providing a unique opportunity to immerse themselves in the subject.

[216] Driver, Marta, and Jeanine Meyer. “Beowulf to Lear: Text, Image, and Hypertext.” *Literary and Linguistic Computing*, vol. 14, no. 2, June 1999, pp. 223–36, https://doi.org/10.1093/llc/14.2.223.

Shares the experience of the multidisciplinary course *Beowulf to Lear: Text, Image, and Hypertext* at Pace University. Throughout the course, there was a strong emphasis on involving students in digital technologies, and as a result, course websites were created. These websites encompassed a wide array of materials, including the syllabus, quizzes, and students’ works, making the course highly interactive and engaging.

[217] Duggan, Mary Kay. “Teaching Manuscripts from a Digital Library on the Web.” *Literary and Linguistic Computing*, vol. 14, no. 2, June 1999, pp. 151–60, https://doi.org/10.1093/llc/14.2.151.

Informs about the experience of the inter-university course *Medieval Manuscripts as Primary Sources*, jointly offered by the University of California and Columbia University. The course aimed to explore whether universities could collaborate and teach a mutual course by sharing their librarian resources. It was an interdisciplinary, distance-learning course, led by multiple instructors, made possible through the utilization of digital infrastructure and the digitalization of librarian sources. These digital resources allowed students outside the universities to access the course materials and participate fully in the learning experience.

[218] Talarico, Kathryn Marie. “Cyberspace Without Tears: Fundamental Approaches to the Uses of Technology in the Classroom.” *Literary and Linguistic Computing*, vol. 14, no. 2, June 1999, pp. 199–210, https://doi.org/10.1093/llc/14.2.199.

Discusses the basic rules of Internet usage in a classroom of medieval and Renaissance studies. It suggests three exercises to enhance students' understanding and effective use of the Internet. Firstly, one exercise informs students about the World Wide Web (www) and its significance in academic research. Secondly, another exercise teaches students how to critically analyze Internet sources for their research, ensuring they are reliable and credible. Lastly, a third exercise assigns students to create an annotated list of useful Internet links relevant to their specific area of interest within the field of medieval and Renaissance studies.

[219] Hardwick, Lorna. “Electrifying the Canon: The Impact of Computing on Classical Studies.” *Computers and the Humanities*, vol. 34, no. 3, 2000, pp. 279–95.

Observes the influence of digitalization on Classical Latin and Greek studies. In 2000, Hardwick noted that digital editions of manuscripts, the development of specialized databases, computer-based courses and video conferencing, electronic bibliographies, and special computer programs for language learning significantly facilitated Classical studies. Although she is not overly optimistic, mentioning the higher requirements for students due to the increasing number of tasks and assignments, this article demonstrates the progress made in the two decades before Covid. Many of the instruments we now use were emerging during that period, even in such a conservative field as Classical Studies.

[220] Christmann, R. “Books into Bytes: Jacob and Wilhelm Grimm’s Deutsches Worterbuch on CD-ROM and on the Internet.” *Literary and Linguistic Computing*, vol. 16, no. 2, June 2001, pp. 121–33, https://doi.org/10.1093/llc/16.2.121.

Presents the CD-ROM edition of the *Deutsches Wörterbuch* by Jacob and Wilhelm Grimm, which represents an extensive documentation of the German language. The edition is compiled in accordance with the Standard Generalized Markup Language (SGML) and guidelines of TEI, making it informative and accessible for various research needs.

[221] De Smedt, Koenraad. “Some Reflections on Studies in Humanities Computing.” *Literary and Linguistic Computing*, vol. 17, no. 1, Apr. 2002, pp. 89–101, https://doi.org/10.1093/llc/17.1.89.

Reflects on how the status of humanities as an instructed discipline changes with the spread of computer-assisted courses in university curricula. De Smedt states that this trend promotes interdisciplinarity, introduces the best teaching practices, and transforms both the content and methods of student learning. He emphasizes that this process necessitates both promotion and deep reflection to effectively integrate technology into humanities education.

[222] Siemens, Raymond G. “A New Computer-Assisted Literary Criticism?” *Computers and the Humanities*, vol. 36, no. 3, 2002, pp. 259–67.

Touches on the crucial issue of the coexistence of computer-based literary text analysis with traditional criticism. Supporting the opinion that literary criticism can be divided into Lower, which “is chiefly textual and bibliographical in nature”, and Higher, which “is typified by interpretive studies” (p. 260), R. G. Siemens proves that computers can significantly contribute to Lower Criticism. Additionally, she discusses the distribution of electronic scholarly editions of literary texts, which drastically improves interaction with texts.

[223] Carlquist, Jonas. “Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions.” *Literary and Linguistic Computing*, vol. 19, no. 1, Apr. 2004, pp. 105–18, https://doi.org/10.1093/llc/19.1.105.

Highlights the limitations of printed editions to enclose all the features of hand-written ancient manuscripts as notes on margins, pointing hands and other features. The digital editions should fill this crucial for researchers gap and allow synchronic comparison of different versions of the manuscripts, representing several layers of information.

[224] Losh, Elizabeth. “Reading Room(s): Building a National Archive in Digital Spaces and Physical Places.” *Literary and Linguistic Computing*, vol. 19, no. 3, Sept. 2004, pp. 373–84, https://doi.org/10.1093/llc/19.3.373.

Emphasizes that new digital collections of texts are influenced by the same national discourses, and as a result, they can exhibit varying degrees of regulation. The article problematizes the uneven and complicated access to certain digital collections created by libraries, which contradicts the initial aspiration that electronic databases would make information more democratically accessible. In some cases, the regulations imposed on digital collections can hinder their full availability to the public.

[225] Vanhoutte, E. “An Introduction to the TEI and the TEI Consortium.” *Literary and Linguistic Computing*, vol. 19, no. 1, Apr. 2004, pp. 9–16, https://doi.org/10.1093/llc/19.1.9.

Introduces Text Encoding Initiative (TEI) and the TEI Consortium (TEI-C), which creates standards for TEI mark-up protocols. It in detail describes the development of TEI and the structure of the language and used tags.

[226] Burrows, John. “All the Way Through: Testing for Authorship in Different Frequency Strata.” *Literary and Linguistic Computing*, vol. 22, no. 1, Apr. 2007, pp. 27–47, https://doi.org/10.1093/llc/fqi067.

“[…] describes the operation of two new tests of authorship and offers some results. Both tests rely on controlled contrasts of word-frequency and both exclude the very common words, which have been put to such good use in recent years. One test treats of words used with some consistency by a target-author but more sporadically by others. The second treats of words used sporadically by the target-author but not by most others. (The inclusion of words that some other authors use avoids the strict constraint that has impoverished this form of evidence)” (p. 27).

Quoted from the annotated paper.

[227] Pritchard, David. “Working Papers, Open Access, and Cyber-Infrastructure in Classical Studies.” *Literary and Linguistic Computing*, vol. 23, no. 2, Oct. 2007, pp. 149–62, https://doi.org/10.1093/llc/fqn005.

Discusses the initiative for internet publication of working papers, known as the *Princeton-Stanford Working Papers in Classics* (PSWPC). The article provides information about the initiative and focuses on the benefits of open internet access to academic publications, such as higher quality of texts, improved acquaintance with the content, and other advantages.

[228] Bradley, John. “Thinking about Interpretation: Pliny and Scholarship in the Humanities.” *Literary and Linguistic Computing*, vol. 23, no. 3, Sept. 2008, pp. 263–79, https://doi.org/10.1093/llc/fqn021.

Introduces a special tool for humanists called *Pliny*. Unlike other digital tools that focus on novice methods, *Pliny* aims to promote traditional scholarship. It is primarily a note-taking tool that helps users read and organize text. However, *Pliny* also offers sophisticated instruments, such as correspondence analysis tools, which can arrange notes hierarchically that have been taken.

[229] Jannidis, Fotis. “TEI in a Crystal Ball.” *Literary and Linguistic Computing*, vol. 24, no. 3, Sept. 2009, pp. 253–65, https://doi.org/10.1093/llc/fqp015.

“Text Encoding Initiative (TEI) is an organization, a research community, and a markup language. Looking back into the history of these three TEIs, this article tries to describe what has been achieved and what its future challenges will be. The historical analysis is based on a closer look at the development of the TEI-L and topics covered by the Guidelines. A final section outlines possible roles of the TEI as an infrastructure for digital libraries and disciplinary virtual environments" (p. 253).

Quoted from the annotated paper.

[230] Schlitz, Stephanie A. "The TEI as Luminol: Forensic Philology in a Digital Age." *Literary and Linguistic Computing*, vol. 24, no. 2, June 2009, pp. 173–85, https://doi.org/10.1093/llc/fqp001.

Introduces forensic philology in the context of electronic text editing, using the example of creating a TEI P5 conformant edition of *Hafgeirs* saga *Flateyings*, an alleged Icelandic saga forgery found in an unsigned eighteenth-century paper manuscript. The discussion explores how literary, linguistic, and transmission-level interpretations can be utilized to describe the saga text, its origin, and transmission process. The article also demonstrates how encoding the metadata alongside the near zero-level text can be achieved without compromising the manuscript's role as an artifact or altering the appearance of the text as presented on the page.

[231] Terras, M., et al. "Teaching TEI: The Need for TEI by Example." *Literary and Linguistic Computing*, vol. 24, no. 3, Sept. 2009, pp. 297–306, https://doi.org/10.1093/llc/fqp018.

States the problem that in the TEI community there is a lack of studying materials for learning this coding in TEI mark-up language. It demonstrates the *TEI by Example* project which aims to fill this gap.

[232] Zielinski, Andrea, et al. “TEI Documents in the Grid.” *Literary and Linguistic Computing*, vol. 24, no. 3, Sept. 2009, pp. 267–79, https://doi.org/10.1093/llc/fqp016.

Introduces the TextGrid platform which allows to work with documents in TEI protocols (using special mark-ups, connecting several layers of data about the text).

[233] Goldsyone, Andrew. “Teaching Quantitative Methods: What Makes It Hard (in Literary Studies).” *Debates in the Digital Humanities 2019*, University of Minnesota Press, 2018, https://doi.org/10.5749/j.ctvg251hk.

Deals with the problem of training DH, focusing on three problems. First, only instructing programming is inefficient without a good acknowledgement with the theoretical basis of the discipline. In this case, programming only implements theoretical approaches in a real research. Second, the training faces difficulty as training sets of textual data are extremely limited. Third, overfocus on theory also is not good, as the students should now not only which theories to implement but how do it efficiently from the points of dealing with data.

[234] Bonch-Osmolovskaya, Anastasia, et al. “Tolstoy Semanticized: Constructing a Digital Edition for Knowledge Discovery.” *Journal of Web Semantics*, vol. 59, Dec. 2019, p. 1-9, https://doi.org/10.1016/j.websem.2018.12.001.

Presents the results of the project of creating the digital version of Leo Tolstoy 90-volumes full print critical edition. Authors in details present the process and demonstrate how they compiled with the standards of TEI. They added into text additional mark-up for direct speech, quotations and citations, which aim to facilitate further computer assisted research for easier extraction of the information from the text.

Appendix

Terms & Concepts

Principal Component Analysis (PCA):

a way of decreasing the dimensionality of multidimensional space or simplifying complicated data into a more illustrative form. In SA, for more precise results we analyze numerous metrics (the proportion of consonants and vowels, the average number of words in sentences, etc.). The number of these parameters or variable may exceed several hundred, but not all of them are equally important for the explanation of texts' similarities or differences. Mathematically, all these hundreds of parameters are presented in a form of a matrix, in which the number of columns is the number of texts under analysis, and the number of lines – the number of parameters. We transform this matrix into a multiplication of two smaller (and this means simpler matrixes). After this procedure, we may depict text in two-dimensional space (each dimension is measured with one smaller matrix). In a graph, we will see how close or distant the texts are located, and which clusters they combine, which tells us about the texts' similarity. If we analyze which parameters occurred in that two matrices, we will mention which formal metrics of the texts better represent text features (what is more important for characterizing the author's style, the average number of words in a sentence or a richness and variety of vocabulary, for example).

Latent Dirichlet Allocation (LDA):

in the SA is an algorithm of document topic modelling (automatic extraction of words which refer to a certain topic). Being mathematically rather complicated, LDA assumes that text contains some topics which can be prescribed to this document. The number of topics to identify is set as one parameter of LDA. The algorithm searches through the text and allocates certain words to one of the set topics. The reverse process is always possible: looking at the words in the text algorithm can conclude that with a certain probability, the text contains some topics. As an output LDA provides the list of words. For example, in a text LDA output such two sets of words, related to two topics: “toy”, “play”, “school” and “work”, “salary”, “loan”. It may be concluded that the first associates refer to “kids” topic and the second one corresponds with “adults”.

Python:

is one of the most popular modern programming languages. It is “*simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance”*(italics ours – A.S.)[8]. Along with R language it is widely used in data analysis and has libraries (special modules with ready solutions for typical tasks, which were written by previous programmers) which are used for text analysis (like NLTK – Natural Language Processing Toolkit). Also, there are powerful libraries (as *matplotlib* and *seaborn*) for data visualization. Most of the methods and instruments, implemented in “Stylometric Analysis” have ready solutions in Python packages.

R:

a computing language, “a free software environment for statistical computing and graphics”[9]. It is mostly used for statistical calculations and data analysis, however, it contains instruments for NLP (see Arnold & Tilton [41]) and one of the most popular instruments for SA – *stylo* library.

Natural Language Processing (NLP):

[8] What is Python? Executive Summary, https://www.python.org/doc/essays/blurb/. Accessed 12 June 2023.

[9] The R Project for Statistical Computing, https://www.r-project.org/. Accessed 12 June 2023.

a sphere of computing linguistics and AI which intends to develop universal mechanism of understanding texts and generation responses and other cognizant texts. In broad, it includes all instruments used by computer linguists for implementation for this complicated goal, lemmatization and tokenization tools, instruments for automatic morphological and syntactic analysis.

Tokenization:

a process of text segmentation into separate units, words or tokens. If for a human this task appears primitive, algorithmically it is not so evident, especially for such collocations as ‘ll, ‘d. For languages using hieroglyphic writing, division into separate words always become a non-trivial task.

Lemmatization:

converting a word to its initial form. For instance, told – (to) tell (infinitive verb). However, without a proper syntactic markup, homonyms may cause problems (e.g. reviews – to review (verb) (?) a review (noun)).

Bag of Words (BoW):

a way of representing text in a numerical format in a way of transforming it into a vector (as a fixed sequence of numbers – see **Vector**). This method represents the lexical variety of a text. For example, we have four lines from Dicken's "A Tale of Two Cities"[10]:

It was the best of times,
it was the worst of times,
it was the age of wisdom,

[10] Example are taken from Brownlee, Jason. "A Gentle Introduction to the Bag-of-Words Model." Machine Learning Mastery, 9 October 2017, https://machinelearningmastery.com/gentle-introduction-bag-words-model/. Accessed 12 June 2023.

it was the age of foolishness,

If we want to present these four lines in vector form (BoW) we, first, should find unique tokens, which do not repeat. They are 10 “it”, “was”, “the”, “best”, “of”, “times”, “worst”, “age”, “wisdom”, and “foolishness”. Let us fix the order. Then, the algorithm may vary, but the simplest way is to count the number of entries in every line.

Thus, in the first line “it” :1, “was”: 1, “the”: 1, “best”: 1, “of”: 1, “times”: 1, “worst”: 0, “age”: 0, “wisdom”: 0, “foolishness”: 0. Rewrite it as vector [1, 1, 1, 1, 1, 1, 0, 0, 0, 0].

Analogically, the second line will be [1, 1, 1, 0, 1, 1, 1, 0, 0, 0], the third one – [1, 1, 1, 0, 1, 0, 0, 1, 1, 0], the fourth one – [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]. Now we can conduct all the mathematical operations with these vectors, compare them, and find distances between them.

This method is very simple, but it ignores lots of crucial aspects: ignores word order (that is why it is called Bag of Words), homonyms, misspellings, and synonyms.

n-gram:

any combination of n words. Often in SA, it is more efficient to segment text not into separate words, but count and analyze groups of 2,3,…,n words as they better demonstrate its diversity.

Text vectorization:

a process of representing a text in a form of an ordered numerical sequence, or a vector. The algorithm can be intuitively simple as in **BoW** or rather complicated NN-based as in Word2Vec or BERT, GPT-3, ELMO. The more complicated and perfect tools create more precise vectors which make texts comparison more precise. As resulting vectors may be comprehended geometrically and spatially depicted, it is possible to find distances between these vectors (which can be interpreted as text similarity). Text vectorization is a key element of more complicated SA.

Vector:

formally “a quantity that has both magnitude and direction” or a definition which can arise in memory from school years, a line segment, which has a direction[11]. However, for the listed above methods it is convenient to consider that vector is an ordered series of numbers of any size. For example, sequence [2, 4, 7, 8, 11] is a vector.

Matrix:

“a set of numbers arranged in rows and columns so as to form a rectangular array. The numbers are called the elements, or entries, of the matrix. Matrices have wide applications in engineering, physics, economics, and statistics as well as in various branches of mathematics. Matrices also have important applications in computer graphics, where they have been used to represent rotations and other transformations of images”[12]. Simpler, a matrix is a table containing numerical data. You can perform mathematical operations on matrices and transform them, which simplifies calculations.

Neural Network (NN):

a mathematical model used for implementing different tasks, such as classification of data, prediction of results, etc. It is also defined as supervised machine

[11] Britannica, The Editors of Encyclopaedia. "vector". *Encyclopedia Britannica*, 27 May. 2020, https://www.britannica.com/science/vector-physics. Accessed 12 June 2023.

[12] Britannica, The Editors of Encyclopaedia. "matrix". *Encyclopedia Britannica*, 29 Mar. 2023, https://www.britannica.com/science/matrix-mathematics. Accessed 12 June 2023.

learning algorithm, which means that the researcher “trains” the algorithm providing input with hints (for example, text for AA and the name of its author, i.e., the answer which the algorithm should come to). When the training is finished, NN independently continues AA as in the example. As follows from the instrument’s name, it imitates the structure of the human neural system with neurons (nodes) and connections between them. Matthews and Merriam (1993) suggested using NNs in AA.

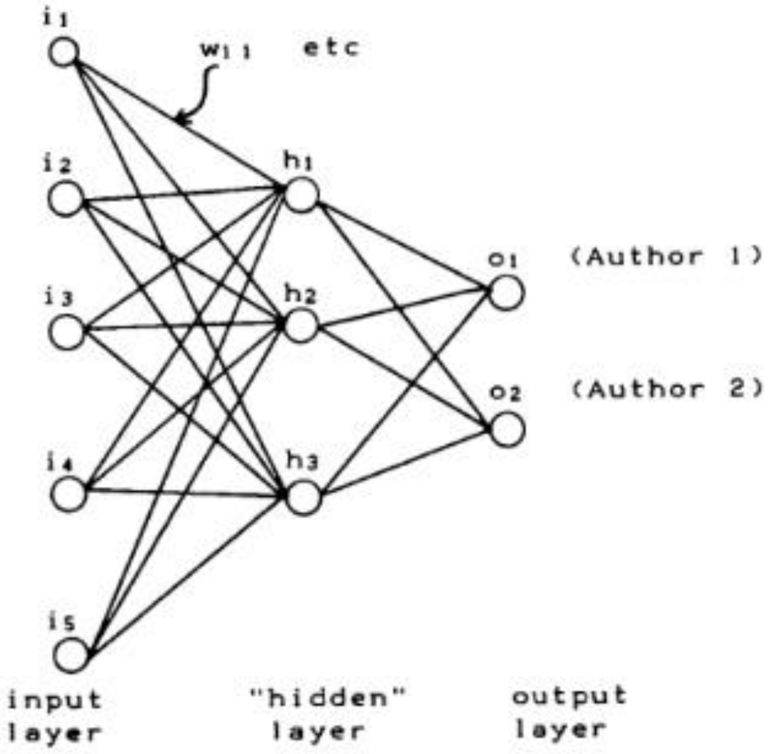
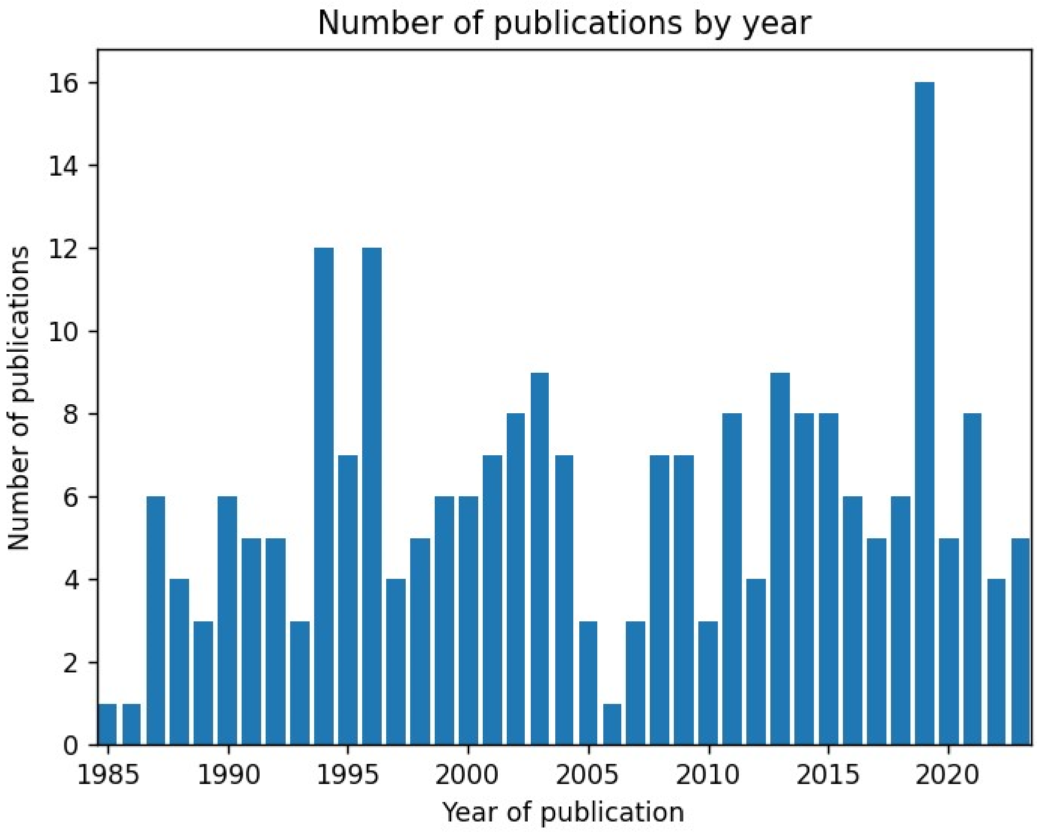


Fig. 1 Topology of a stylometric multi-layer perceptron for classifying works of two authors using five discriminators.

Figure from [70] Matthews, Robert A. J., and Thomas V. N. Merriam. “Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher.” *Literary and Linguistic Computing*, vol. 8, no. 4, Oct. 1993, pp. 203–10, https://doi.org/10.1093/llc/8.4.203.

Their NN had 5 input nodes-neurons which took a certain statistical metric of the text (one of the nodes was input with the familiar frequency of the 10 common words). The output nodes (author A and author B) give a number 0 or 1 which means who is the other of input text (1 for Shakespeare and 0 for Fletcher). In the center of the NN there is hidden layer which also input data from other neurons and output some processed numerical data. The number of these hidden layers may vary. Every node modifies input data, using specific activation function. Excluding mathematical explanations, this function follows whimsical rules which transforms any input into 1 or 0 (that is the reason of such term, as the input signal can activate node and this node produces output 0 or it does not generate any signal and the output will be 0). These activation functions have coefficients which make signal transformation into 1 or 0. During NN training, when a researcher input some metrics of a text, NN is also informed which result (Shakespeare or Fletcher) it was to output. Following mathematical algorithms, NN modifies the coefficients in nodes' activation functions after each training text which was input and in a hidden way formulates its own rule for AA. When NN training stops, the coefficients in activation functions are fixed and NN gets only the statical parameters extracted from the text for genuine AA.

Table 1. Changes in publication activity in DH, based on publications present in Annotated Bibliography



Author Index[13]

Alex, Beatrice [200]

Alqurneh, Ahmad [123]

Antonia, A. [95]

Arnold, Taylor, et al. [43]

Baayen, R. Harald R. [81], [87]

Baker, John Charles. [62]

Barber, Ros. [132]

Barr, George K. [84], [98]

Beausang, Chris [136]

Benson, James D. [63]

Berry, Dave M. [22]

Binongo, Jose Nilo G. [71]

Borin, Lars [198]

Bonch-Osmolovskaya, Anastasia [172], [234]

Bradley, John [228]

Brainerd, Barron [63]

Bruce, Donald [4]

Bruster, Douglas [173]

Bump, Jerome [204]

Burg, Jennifer [154]

Burgess, Gordon [207]

Burrows, John F. [58], [64], [69], [96], [105], [108], [226]

Can, Fazli [102]

Carlquist, Jonas [223]

Carter, Bryan [215]

Caton, Paul [26]

Chandramohan, Balasubramanyam [210]]

Christmann, R. [220]]

Clement, Tanya [165]

Corns, Thomas N. [2]

Cortelazzo, Michele A. [129]

[13] Number if square brackets is a number of entry in the bibliography.

Craik, E. M. [59]

Crofts, Daniel W. [111]

Daelemans, Walter [115]

DeForest, Mary [155]

Devitt, Ann [192]

Dierks, Karin [57]

Dixon, Peter [104]

Dobson, James E. [185]

Driver, Marta,[216]

Duggan, Mary Kay [217]

Eder, Maciej [122]

Eisen, Mark [126]

Elewa, Abdelhamid [177]

Elliot, Ward E. Y . [92]

Erb, Maurice [30]

Feng, Vanessa Wei [119]

Finlayson, Mark A. [174]

Forstall, Christopher W. [113]

Forsyth, Richard S. [76], [89], [90], [120]

Fortier, Paul A. [3], [148]

Frautschi, Richard L. [60], [65]

Gardner, Colin. [18]

Gladwin, Alexander A. G. [124]

Goicoechea, Maria [25]

Goldfield, Joel D. [6]

Greenwood, H. H. [75]

Hardwick, Lorna. [219]

Havholm, Peter [11]

Hawthorne, Mark [208]

Herbelot, Aurélie [171]

Hess, Shervin [110]

Hirst, Graeme [119]

Holland, Simon [207]

Holmes, David I. [56], [68], [72], [76], [85], [93], [94], [111]

Hoover, David L. [97], [99], [103], [110], [125]

Howe, Christopher J. [156]

Ilsemann, Hartmut [77], [161]

Irizarry, Estelle [66], [144]

Jannidis, Fotis [229]

Jessop, Martyn [196]

ockers, Matthew L. [112]

Johnson, Eric [155]

Jorgensen, J. [168]

Kaferly, D. H. A. [59]

Karlsen, Heidi. [187]

Katz, Seth R. [209]

Keskustalo, Heikki [188]

Khurshid Ahmad [192]

Koch, Christian [146]

Koppel, Moshe [114], [175]

Kuhn, Jonas [34]

Kutty, Sangeetha [201]

Labbe, Cyril [106]

Labbe, Doninique [106]

Laffal, Julius [153]

Ledger, Gerard [73]

Lee, John [181]

Liu, Mingming [178]

Logan, Grace B. [67]

Logan, Harry M. [67]

Losh, Elizabeth [224]

Lowe, David [79]

Luyckx, Kim [115]

Maczewski, Jan-Mirko. [151]

Magda Heydel [118]

Mahoney, Anne [157]

Mannion, David [104]

Marindale, Colin [82]

Markert, Katja [162]

Matsuba, Stephen Naoyuki [7]

Matthews, Robert A. J. [70], [74], [79]

McCarty, Willard [14]

McKenna, C. W. F. [95]

Mealand, D. L. [80]

Merriam, Thomas V . N. [15], [70], [73], [74], [86], [100] Meyer, Jeanine [216]

Milic, L. T. [205]

Milkman, Katherine. L. [160]

Mills, Jon [210]

Moretti, Franco [46], [47], [49], [51], [53], [197]

Murai, Hajime [182]

Muralidharan, Aditi [166] Hearst, Marti. A. [166]

Nini, Andrea [127]

Nissim, Malvina [162]

Oakes, Michael P. [128]

Olsen, Mark [5], [8], [9]

Ousaka, Y .[149]

Papadopoulos, Costas [35]

Pasanek, Bradley M. [21]

Patton, John M. [102]

Pearl, Lisa [116]

Pelt, Tamise van [16]

Philippides, Dia M. L. [141]

Pidd, Michael [213]

Potter, Rosanne G. [1], [211]

Prescott, Andrew [214]

Pritchard, David [227]

Rahimi, Zeinab [194]

Rebora, Simone [130]

Reilly, Paul [35]

Roberts, Alan [152]

Robinson, Peter [13]

Rockwell, Geoffrey [19]

Roman, Gustavo San [142]

Ross, Gordon J. [140]

Rotari, Gabriela [134]

Rudman, Joseph [12]

Rybicki, Jan [107], [118], [121]

Sabol, C. Ruth [145]

Sanz, Amelia [25]

Savoy, Jacques [135], [138]

Sayoud, Halim [117]

Schlitz, Stephanie A. [230]

Sculley, D. [21]

Seidman, Shachar [175]

Shamir, Lior [184]

Siemens, Raymond G. [222]

Skorinkin, Daniil [172]

Smith, Geneviève [173]

Smith, Jonathan [212]

Somers, Harold [101]

Spencer, Matthew [156], [158]

Sprugnoli, Rachele [193]

Stewart, Larry L. [11], [159]

Steyvers, Mark [116]

Stratil, M. [61]

Suddaby, Lee [140]

Tannenbaum, Robert S. [203]

Taylor, Dennis [10]

Terras, M. [231]

The NovelTM Research Group [31]

Tuffin, Paul [82]

Tuzzi, Arjuna [129]

Tweedie, Fiona J. [83], [87], [88], [101]

Underwood, Ted [29], [31], [53], [55]

Valenza, Robert J. [92]

Waugh, Sam [91]

Wehren, Marylène [138]

Weingart, S. [168]

Winder, William [17]

Witten, Daniela M. [112]

Yeung, Chak Yan [181]

Zielinski, Andrea [232]