Title: Tracing patterns of contact and change: Philological vs. computational approaches to the handwritings of a 18th century migrant community in Berlin
Team: Humboldt-Universität zu Berlin: Prof. Dr. Roland Meyer (Primary investigator); Aleksej Tikhonov, M.A.; Ewa Kolbik, B.A.; Dr. Robert Hammel (Consultant) – Fraunhofer IPK Berlin: Dr.-Ing. Bertram Nickolay (Primary investigator); Dr. Jan Schneider (Consultant) – MusterFabrik Berlin: Klaus Müller, M.Sc.; Dipl.-Ing. Maxim Schaubert; Luisa Esguerra, M.Sc.; Dr. Marc von der Linden (Consultant) – Archive in the "Bohemian village", Berlin: Stefan Butt (Consultant)
Corpus: Digitized handwritten personal vitae and sermon manuscripts, 18th-early 19th c., from the Archive in the "Bohemian village", Berlin – ca. 5000 pages
Field of Study: (Historical) philology, palaeography, Czech linguistics, language contact; digital restoration of documents; image recognition, authorship attribution, handwriting recognition
Institution: Humboldt-Universität zu Berlin, Fraunhofer IPK Berlin, MusterFabrik Berlin
Methods: High-quality document scanning (own developments of Fraunhofer IPK and MusterFabrik); Image recognition and analysis, Optical character recognition, Machine-learning, Neural networks; Linguistic analysis, historical morphology, Annotation of document images
Tools: Assistance system LiViTo (own development)
Technology: Supervised machine-learning, neural networks, Python programming, tagging, corpus linguistics

Detecting authorship, hands, and corrections in historical manuscripts. A mixed-methods approach towards the unpublished writings of an 18th century Czech emigré community in Berlin

Roland Meyer, Aleksej Tikhonov, Robert Hammel

When one starts working philologically with historical manuscripts, one faces important first questions involving authorship, writers' hands and the history of document transmission. These issues are especially thorny with documents remaining outside the established canon, such as private manuscripts, about which we have very restricted text-external information. In this area – so we argue – it is especially fruitful to employ a mixed-methods approach, combining tailored automatic methods from image recognition/analysis with philological and linguistic knowledge. While image analysis captures writers' hands, linguistic/philological research mainly addresses textual authorship; the two cross-fertilize and obtain a coherent interpretation which may then be evaluated against the available text-external historical evidence. Departing from our 'lab case', which is a corpus of unedited Czech manuscripts from the archive of a small 18th century migrant community, the Herrnhuter Brüdergemeine (Brethren parish) in Berlin-Neukölln, our project has developed an assistance system which

aids philologists in working with digitized (scanned) hand-written historical sources. We present its application and discuss its general potential and methodological implications.

Project description

For all humanities, historical manuscripts are an essential source of knowledge. Interest in textual history has gone so far as to make philologists fear that textual content may become backgrounded in comparison to issues of textual genesis, versions and manuscripts. This fear is however unfounded, as long as researchers are aware of versions of texts and do not rely only on apparent originals or *urtexts*[1]. Traditional philology and other text-centred humanities have developed a received methodology of accessing old manuscripts which involves research on the text-external context, close reading, transcription, critical edition and time-consuming textological 'detective work'. On the other hand, modern image and pattern recognition techniques promise to be able to distinguish personal handwritings and isolate pre-defined graphic templates in them in an automatic (supervised) way. The present project aimed at confronting these two methodologies, reflecting systematically upon the question as to which of the two approaches was more adequate and successful, and combining their benefits. While large collections of handwritten texts produced by different unknown scribes pose a great challenge, they are also an ideal testing ground for applying new research methods which unify computational and linguistic approaches. It is instrumental here to distinguish between the 'scribe' as the material producer of the manuscript at hand and the 'author' as its intellectual originator. The author either delivered the original version of the text on which the manuscript is based, or (s)he dictated its contents to the scribe. Consequently, optically recognizable handwriting features of the manuscript can be attributed to the scribe whereas linguistic features should rather be ascribed to its author. The manuscripts of 18th c. Czech Protestant immigrants in Prussia are an example of such collections of texts. The manuscripts, which include autobiographies of parishioners of the Moravian Church (*Herrnhuter Brüdergemeine*) as well as a large number of so-called Choir speeches (a type of sermons typical for that Church), are written in Czech using *Kurrent* script then common not only among Germans but also among Czech speakers.

[1] Livia Kleinwächter, „The Literary Manuscript: A Challenge for Philological Knowledge Production", in: Pál Kelemen and Nicolas Pethes (Ed.): *Philology in the Making Vol. 1*, (Bielefeld: transcript Verlag, 2019), 109–128.

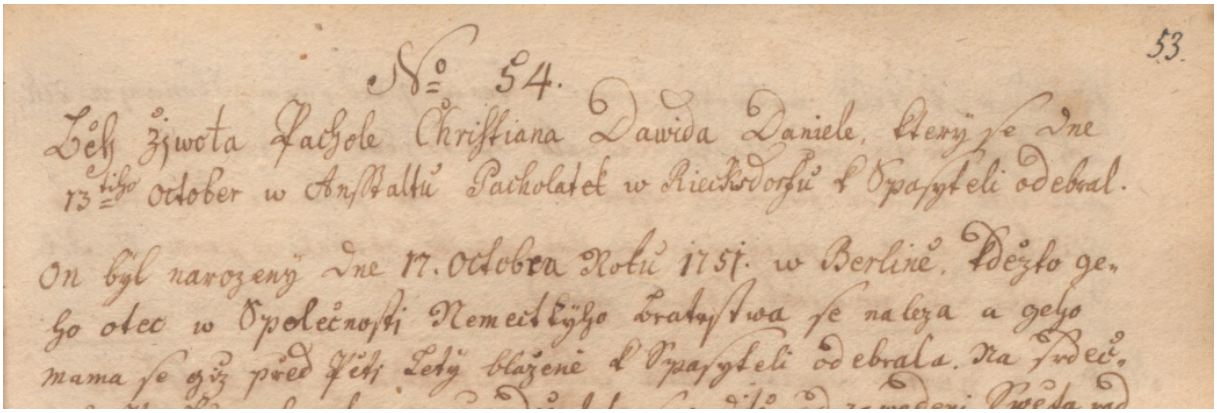

Fig. 1: Excerpt from a biography manuscript from the Rixdorf archive. The headline script differs slightly from the body text, as it probably had a decorative function.

The manuscripts are kept in the archive of the *Herrnhuter Brüdergemeine* in Berlin-Neukölln. This material object of study consists of about 5,000 hand-written pages in (historical) Czech, dating from about 1740 to 1830, from a small community of religiously persecuted migrants (called *exulants*) to Berlin, the ancestors and founders of the present-day Moravian Church Parish.

The Czech immigrants, originally adherents to the Protestant Unity of Brethren, escaped from Catholic counter-reformation in Bohemia and Moravia and, after a many-year odyssey, were eventually accepted in Prussia by King Frederick William I. In 1737, some of them settled in the small village of Rixdorf, then on the outskirts of Berlin but now part of it. In 1756, most of the immigrants joined Count Nikolaus Ludwig Zinzendorf's Moravian Church[2]. The Czech language was commonly used by these immigrants at least until the early 19th c. before it was gradually replaced by German. Surrounded by a German-speaking environment, Czech speakers in Rixdorf thus formed a particularly interesting language enclave during a period when the language in the Czech mainland was suffering a considerable decline.

Our project focused on the autobiographies of Rixdorf parishioners as these documents were of special historical and linguistic interest. Up to the present day, writing an autobiography is part of the religious duties of every member of the Moravian Church. The autobiography is supposed to cover important stages of the parishioner's life with particular emphasis on spiritual aspects[3]. We initially supposed that processes similar to those described by Mettele also applied to Czech immigrants' autobiographies; parishioners with little practice in using the Czech written language, such as peasants and craftsmen, probably needed the assistance of

[2] for an account of the history of Czech settlement in Rixdorf see Manfred Motel, *Das böhmische Dorf in Berlin: die Geschichte eines Phänomens,* (Berlin: Darge Verlag, 1983).

[3] on German brethren autobiographies see Gisela Mettele, *Weltbürgertum oder Gottesreich: die Herrnhuter Brüdergemeine als globale Gemeinschaft 1727 – 1857,* (Göttingen: Vandenhoeck & Ruprecht, 2009).

educated authors who would turn their oral accounts into written texts. The texts were later copied and corrected, the latter being evident especially from numerous subsequent amendments[4]. The text corpus comprises 183 autobiographies covering a period between 1760 and 1819 with a total number of 660 handwritten pages[5]. A selection of these were published in abridged form by E. Štěříková in modern Czech orthography[6] and later also translated into German[7]. Unfortunately, the autobiographies contain no explicit indication of their authors or scribes. For the community, it is important to uncover the content of these texts, the people who were able or authorized to write them, and the history of their transmission. Given the influence of the Herrnhuter Brüdergemeine on the Lutheran Church and that of the *exulant* communities on the Berlin city history, the record of linguistic and cultural adaptation implicit in the texts earns a broader general interest.

Given the various participants involved in the completion of an autobiography, a major goal of the project was to determine the number of different authors and scribes engaged in it, and thus to reconstruct the history of the manuscript. Crucial clues to the reconstruction are provided by linguistic features of the autobiographies, on the one hand, and by visual features of the handwritings on the other. The twofold analysis of both types of features revealed that the 183 autobiographies had been produced by a total of 26 different authors and 12 different scribes. The results of the research project are summarized in Aleksej Tikhonov's PhD thesis[8] and in a number of recently published papers[9].

Moreover, an open-source software tool called LiViTo[10] was developed to provide an assistance system for the analysis of historical manuscripts. The tool comprises modules for scribe and keyword detection as well as modules for revision detection and linguistic feature

---

[4] for more information on the manuscripts held at the Rixdorf archive see Aleksej Tikhonov, *Sprachen der Exilgemeinde in Rixdorf (Berlin): Autorenidentifikation und linguistische Merkmale anhand von tschechischen Manuskripten aus dem 18./19. Jahrhundert,* (Heidelberg: Winter Verlag, 2022), 83–97.

[5] Tikhonov, *Sprachen*, 58.

[6] Edita Štěříková, *Běh života českých emigrantů v Berlíně v 18. Století*, (Praha: Kalich, 1999).

[7] Edita Sterik, *Die böhmischen Exulanten in Berlin*, (Herrnhut: Herrnhuter Verlag, 2016).

[8] Aleksej Tikhonov, *Autorenidentifikation und linguistische Merkmale der Rixdorfer Handschriften: Eine Untersuchung anhand von Manuskripten aus dem 18./19. Jahrhundert (Dissertation)*, (Berlin: Humboldt-Universität zu Berlin, 2020). Tikhonov, *Sprachen*, (2022).

[9] Aleksej Tikhonov and Klaus Müller, „LiViTo: A software tool to assess linguistic and visual features of handwritten texts", in Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, Lydia Pintscher (Ed.): *Qurator - Conference on Digital Curation Technologies 2020*, (Berlin: Online-Open-Access-Publication, 2020), https://ceur-ws.org/Vol-2535/paper_8.pdf.

Klaus Müller, Aleksej Tikhonov, Roland Meyer, „LiViTo: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts", in: *Proceedings of the Twelfth Language Resources and Evaluation* Conference, (Marseille: European Language Resources Association, 2020), 885–890.

Aleksej Tikhonov and Klaus Müller, „Scribe versus authorship attribution and clustering in historic Czech manuscripts: a case study with visual and linguistic features", in: *Digital Scholarship in the Humanities*, (Oxford: Oxford University Press, 2022), 254–263.

[10] Tikhonov and Müller, *LiViTo*, (2020).

analysis[11]. Researchers from both teams – linguists and engineers – jointly developed the tool. It is language-independent and was published as adaptable open-source software in order to make it useable beyond the 'lab case' addressed in the project. An unexpected achievement was the rapid improvement made in the optical character recognition (OCR) of historical individualized handwriting, by using machine-learning techniques with neural networks. Thus, we can now actually search textually in the digitalized document images and identify repeated occurrences of keywords. Finally, based also on neural network technology, LiViTo is able to find various types of corrections and amendments by detecting layers of handwriting on the basis of image processing. This helps linguists to group texts by potential later correctors and form hypotheses about their identity; conversely, linguists' classifications provide training data for the refinement of the image processing component.

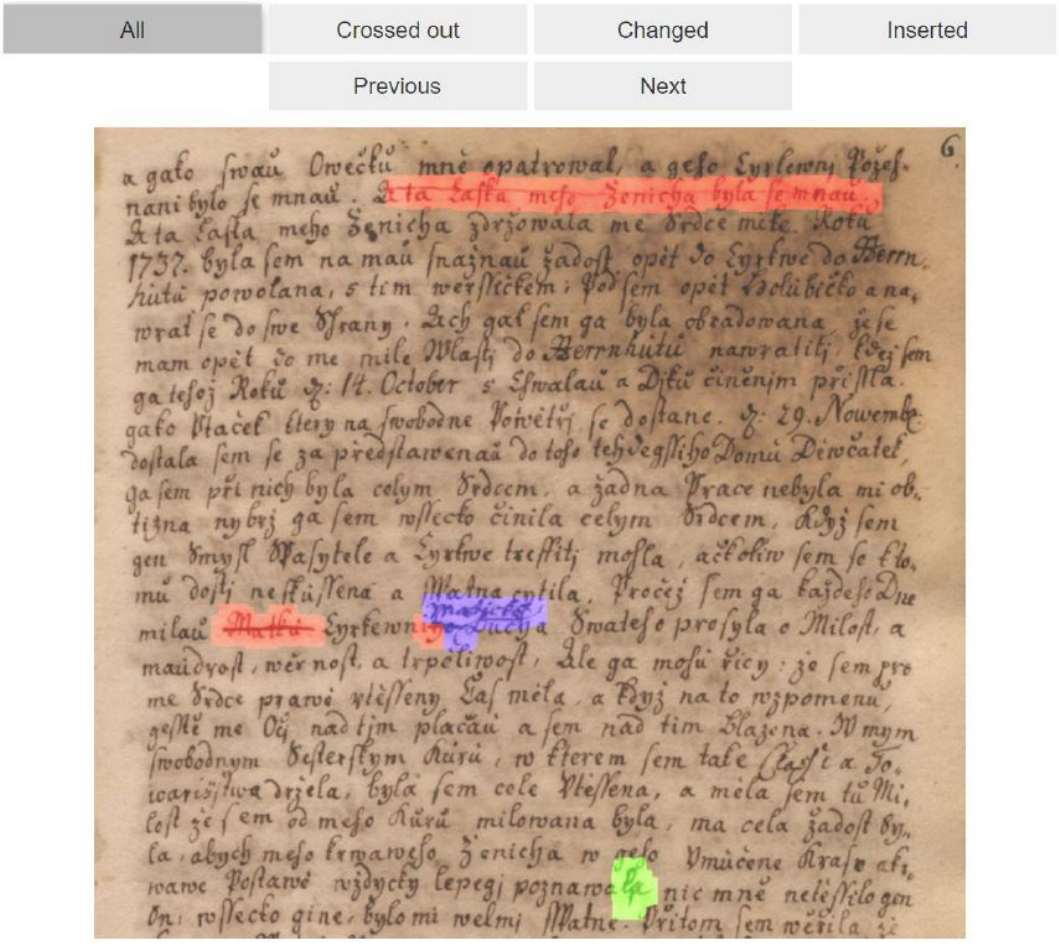

Fig. 2 An example of LiViTo's revision detection function.

[11] Tikhonov and Müller, *LiViTo*, 885–890.

The present paper provides an outline of the methods applied to the analysis of the manuscripts and a discussion of the results.

Both methodologically and content-wise, the project has turned out extremely fruitful. After high-quality digitalization of the documents, both teams used their respective methods to add information that could help to identify scribes and/or original authors: annotation of specific linguistic properties for team (i) and graphics-based machine-learning techniques for team (ii). Both approaches were systematically examined during regular weekly common work sessions, and led to a mutual refinement of the methodology (e.g., as to which parts of the script were distinctive) and to a deeper understanding of the respective results. An interesting and unexpected outcome was that only for part of the autobiographies was there strong agreement between linguistics-based and graphics-based classifications. Another set of texts, however, was considered diverse by the linguists, but homogeneous by the image processing group; the obvious explanation was that these texts had been written up by one scribe or copied from the original sources later. Clearly, none of the two approaches could have achieved this result without the other — both necessarily complement each other in detecting document history. However, it also proved important in the final phase of the project to confront both findings with historical background knowledge from the archives in order to achieve a sound explanation.

State of related research

Since the present project combines various scientific methods and disciplines, current research must be taken into account in at least three[12]: (i) linguistic and visual author and scribe attribution, (ii) stylometric research and (ii) computer aided keyword analysis/search in digital documents. Burrows designed a method of analysing word frequencies to visualize the distance between two or more texts in terms of authorship[13]. Another comparable measurement is Kullback-Leibler divergence (KL divergence). KL divergence is of greater importance because it is not based solely on the relationship between individual word frequencies, but on the stochastic Markov chain and the probability distance[14].The central role

[12] for a detailed overview: Tikhonov and Müller, *LiViTo*, (2020); Müller et al., *LiViTo*, 885–890; Tikhonov and Müller, *Scribe*, 254–263.

[13] John F. Burrows, "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style", in: *Literary and Linguistic Computing 2* (Oxford: Oxford University Press, 1987), 61–70.
John F. Burrows, ""An Ocean Where Each Kind...": Statistical Analysis and Some Major Determinants of Literary Style", in: *Computers and the Humanities 23, (New York/Heidelberg/AA Dordrecht: Springer*, 1989), 309–321.

[14] Moshe Koppel, Jonathan Schler, Shlomo Argamon, "Computational Methods in Authorship Attribution", in: Steven Sawyer (Ed.): JASIST 60 (Hoboken: Wiley-Blackwell, 2009), 9–26.

of function words in multivariate analysis is implemented in machine learning approaches in which text categorization is based on neural networks. This method of author and scribe assignment has been widely used in various disciplines since 1993[15]. Hope[16] studied the authorship of Shakespeare's plays, exploring the connections between John Fletcher, Thomas Middleton, and Shakespeare[17]. The R package 'stylo', developed by Eder, Rybicki & Kestemont[18] is a tool for statistical analysis of the style of one or more texts. In recent years, stylometric techniques in combination with 'stylo' have become popular among scholars in humanities who are concerned with the question of authorship of texts and with language statistics[19]. The use of tools for authorship analysis needs digital input data, but as most historical documents are not digitized and the manual transcription process itself is very time consuming, there has been considerable research on automatic optical character recognition (OCR). One of the first systems capable of transcribing more than just single well separated characters was the omni-font software developed by Kurzweil Computer Products in 1974[20]. A prominent free open-source tool for OCR, which can transcribe various languages and styles is Tesseract[21]. Recent development in machine learning led to first research results on algorithmic handwritten text recognition (HTR), which are on human level accuracy[22]. Inspired by these technological improvements Transkribus, a service platform for computer-aided transcription, was developed in 2017[23].

---

[15] Koppel et al., *Computational*, 11.
[16] Jonathan Hope, *The authorship of Shakespeare's plays. A socio-linguistic study*, (Cambridge: Cambridge University Press, 1994).
[17] for a recent statistical account see also Petr Plecháč, „Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns", in: *Digital Scholarship in the Humanities*, (Oxford: Oxford University Press, 2020).
[18] Maciej Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Digital Scholarship in the Humanities 30*, (Oxford: Oxford University Press, 2010), 167–182.
Maciej Eder, Jan Rybicki, Mike Kestemont, „Stylometry with R: a package for computational text analysis", in: *R Journal 8 (1)*, (Online-Open-Access-Publication, 2016), 107–121.
[19] see the stylometric analysis of direct speech in the television series *The Big Bang Theory*: Maryka van Zyl and Yolande Botha, "Stylometry and Characterisation in The Big Bang Theory", in: *Literator 37/ 2* (Cape Town: Aosis Publishing, 2016), 11.
[20] J. Scott Hauger. *Reading Machines for the Blind: A Study of Federally Supported Technology Development and Innovation (Dissertation)*, (Blacksburg: Virginia Polytechnic Institute and State University, 1995).
[21] Anthony Kay, "Tesseract: An Open-Source Optical Character Recognition Engine", in: *Linux Journal*, (Online-Open-Access-Publication, 2007).
[22] Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks", in: *Proceedings of the 23rd International Conference on Machine Learning*, (Pittsburgh: Carnegie Mellon University, 2006), 369–376.
[23] Philip Kahle, Sebastian Colutto, Günter Hackl, Günter Mühlberger. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", in: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, (Kyoto: IEEE, 2017), 19–24.

Method reflection

Participants of the project, prerequisites

The project was jointly headed by Roland Meyer, chair of West Slavic languages at Humboldt University (HU) in Berlin, and Bertram Nickolay of Fraunhofer Institute for Production Systems and Design Technology (Fraunhofer IPK) in Berlin. A team under Bertram Nickolay is known to have designed an efficient technology and software assistance system for piecing together torn and shredded paper documents of former East German State Security. Musterfabrik Ltd, a company affiliated with Fraunhofer IPK, continues the digital reconstruction of (two-dimensional) cultural assets, including, for example, the written remains of the recently collapsed Cologne city archive, or the fragmented hand-written notes of G.W. Leibniz[24]. Klaus Müller of Musterfabrik Ltd mainly carried out the research for the part of the present project involving optical pattern recognition. He was accompanied by Maxim Schaubert and head of Musterfabrik Marc von der Linden (consultant). As a prerequisite for the project, all of the Czech manuscripts kept at the archive in Berlin-Neukölln were scanned at Musterfabrik by Luisa Esguerra Rodriguez with an overhead scanner using a resolution of 400dpi and a bit depth of 24-bit colour. The quality of the scans proved sufficient for a computational analysis of handwriting features.

Linguistic research for the project was conducted by the team at HU, which has a strong background in Czech (historical) linguistics and in corpus linguistics. The research was undertaken primarily by Aleksej Tikhonov with the assistance of Ewa Kolbik. Slavic and computational linguists Roland Meyer and Robert Hammel regularly contributed their expertise and acted as supervisors. There was a close exchange both during the preparation and training of models of visual variation, and during statistical analysis across the teams.

An absolutely essential ingredient of the research was cooperation with the archive of the *Herrnhuter Brüdergemeine* in Berlin-Neukölln and with *Archiv im Böhmischen Dorf*, headed by Stefan Butt. Butt generously provided advice and orientation in Brethren traditions; and the *Brüdergemeine* parish kindly made available their manuscripts for digitization, handwriting recognition, and analysis of authorship. The project remunerated them with professional digital preservation of their manuscripts, archival contract research, joint outreach activities and, last but not least, unlocking of the contents of the documents which is very important for the community's historical record.

[24] "Analyse der "Rixdorfer-Predigten,"" MusterFabrik Berlin, accessed February 2, 2023, https://musterfabrik-berlin.de/landingpage/index.php/rixdorfer-predigten/

A. Quantification

Our initial goal was to match and align large quantities of linguistic and visual data in order to identify authors and scribes in our corpus of 18th c. Czech Brethren autobiographies. As already mentioned, we departed from the assumption that the production of the autobiographies involved at least two more parts than that of the oral autobiographical account itself, namely, the part of the author as producer of a coherent text and the part of the scribe as producer of the manuscript, the latter above all imposing characteristic orthographic features on the text.

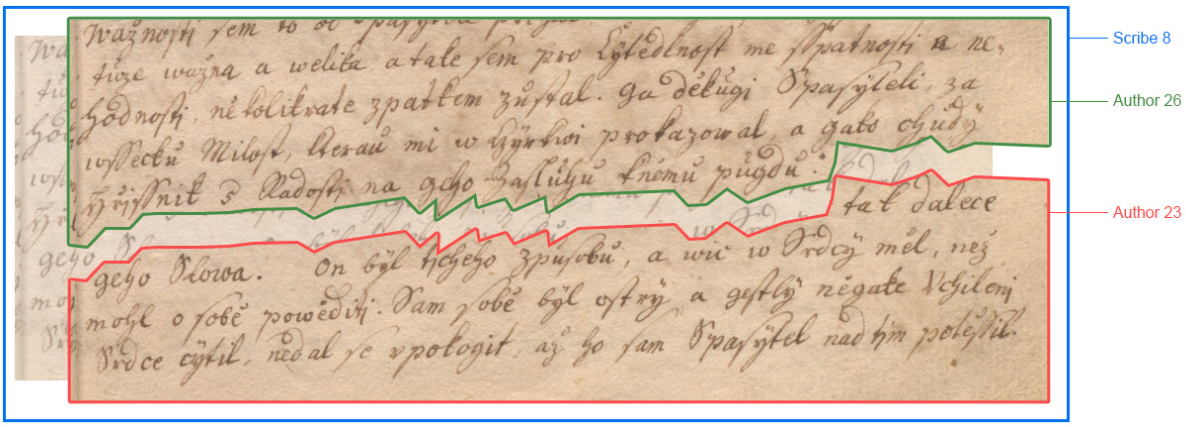

Fig. 3 Example of the distinction between author and scribe.

Mixed methods in the case of the Rixdorf autobiographies thus require a well-defined set of linguistic features believed to sufficiently characterize authors' languages, a set of distinctive orthographic features of scribes[25], and a set of visual handwriting features which can distinguish scribes[26]. Both approaches involve mathematical methods of determining the similarity between different manuscripts by way of data clustering. Data visualization is used to present the distances between data clusters of similar manuscripts (see G below). In the case of linguistic authorship and scribe identification, linguistic stylometry provides well-established quantitative methods and even appropriate software packages. The present project mostly relied on the software package Stylo[27]. In the case of optical pattern recognition, a computer-aided analysis is the only possible method of coping with a large quantity of data since retrieving similar handwriting features from a large corpus of texts is well beyond the limits of manual analysis.

[25] Tikhonov, *Sprachen,* 137–155.
[26] Tikhonov & Müller, *LiViTo*, (2020).
[27] Eder et al., *Stylometry*, (2016).

B. Qualitative data from a linguistic and from a visual perspective

In the present project, the identification of authors and scribes is based on a simultaneous analysis of two qualitatively different data sets, that is linguistic (including orthographic) and visual data. The linguistic features entering the analysis include morphological and syntactic parameters such as particular noun desinences, different infinitive forms, average sentence length, the omission of subject pronouns ('pro-drop'), colloquial vs. literary lexical elements etc. Orthographic features taken into consideration comprise the use of particular orthographic systems, different spellings of geographical names and also the presence of various types of revisions such as visibly marked deletions or additions to the manuscript.

The linguistic and orthographic similarity of different manuscripts was calculated based on Euclidean distance between the feature vectors. Both types of qualitative data allow a classification of the manuscripts at hand according to how many authors and scribes were involved in their production.

There is, however, a heuristic difference between the two types of data sets. While the linguistic and orthographic features used in the stylometric analysis of the manuscripts were deliberately chosen by the researcher on the basis of his knowledge of Czech language history, the optical pattern recognition rests on an analysis of no less than 128 different visual handwriting features automatically chosen by the computer program[28]. In comparison, in her handbook of forensic handwriting analysis Seibt[29] discusses only 60 different characteristic features of individual handwriting that should be noticed by examiners when they compare documents. These include, for example, pen pressure, beginning and end strokes, spacing between words etc. The present research on the Rixdorf manuscripts took into account more than twice as many visual features. A final, truly qualitative source of data for the project were historical records about the Brethren in research literature, which allowed Tikhonov (2020) to finally ascribe most of the identified anonymous authors and scribes plausibly to actual historical persons.

C. Uncertainty

Not only do stylometric linguistic feature analysis and optical pattern recognition require different models to interpret the data, but both models have also to be eventually merged in order to develop a unified picture of distance and similarity between the different manuscripts. It turned out that stylometric linguistic analysis and semi-automatic optical

[28] Müller et al., *LiViTo*, 887.

[29] Angelika Seibt, *Unterschriften und Testamente – Praxis der forensischen Schriftuntersuchung*, (München: Beck, 2008), 97–142.

pattern recognition of handwriting did not always produce identical results, so researchers had to clarify the fuzziness between the results of both analyses. This was accomplished in the following manner:

In the initial phase of the project both linguists and computer scientists defined their own sets of potentially relevant features. While visual handwriting features were obtained automatically by Musterfabrik Ltd software, characteristic linguistic features, including orthographic ones, were devised by the researchers and subsequently tested on a small sample of texts. Preliminary results of both approaches were then compared. While the linguistic analysis yielded 12 subclusters of similar manuscripts equalling 12 different potential authors/scribes, optical pattern recognition of handwriting features resulted in only 10 different subclusters (scribes).

A close comparison of both results revealed that 10 out of the 12 'linguistic' subclusters essentially matched the subclusters identified by visual pattern recognition. However, a number of texts, which were assigned to different scribes by the two approaches, were in fact at the statistical boundary between two separate subclusters and thus could belong to either of two scribes. Finally, 9 out of 12 scribes could be plausibly identified with historical persons, whereas three scribes remain either controversial or completely unknown. The corresponding subclusters may be classified as hybrid since they do not allow unequivocal identification of author or scribe.

D. Interpretable models

Computational (machine-learning) methods and linguistic/stylometric methods generally focus on different aspects of our research question: scribe detection based on visual features, on the one hand, and authorship attribution based on features of language and style on the other. However, there is also an overlap, especially when (computational) word or grapheme detection or revision detection assist linguistic analysis[30]. Both visual pattern recognition and linguistic/stylometric analysis start out with sets of features which function as vectors or dimensions along which texts vary. In the case of visual patterns these features are machine-learned, but in the linguistic/stylometric case they are deliberately chosen and annotated.

A dimension reduction technique (T-distributed Stochastic Neighbour Embedding, t-SNE) is applied in order to visualize the clustering of texts in a 3-dimensional space. The clusters are then interpreted as texts belonging to the same scribe; this concludes the modelling of visual patterns.

---

[30] Müller et al., *LiViTo*, (2020).

Linguistic/stylometric modelling also starts by clustering, but then continues by qualitative analysis of many aspects of the manuscripts, from inspection of single features to historical background knowledge about the persons involved. The linguistic characteristics which form the basis of the clustering are often immediately interpretable. For example, certain endings of words point to a colloquial rather than formal register. Certain word orders (e.g., verb-final in embedded clauses) or a relatively low frequency of null subjects and a high amount of third person subject pronouns would point to German influence. In other cases, dimensions of variation 'just work' in distinguishing individual styles, but a comprehensive interpretation is hard to devise; this would apply, for example, to certain spellings of names or to measures such as average sentence length. In any event, the interpretation of the *model* of authorship and document transmission essentially involves sets of triples of text, author and probability of authorship; but in many cases it also involves individual histories of rewriting and copying.

E. The status of machine learning

While the analysis of linguistic and orthographic features is done more or less manually, the optical pattern recognition technique mainly relies on machine-learning algorithms. Machine-based detection of similar visual handwriting features requires preliminary training based on limited samples of at least five pages from two distinct scribes, that is, about 10 pages of handwritten text.

It is not yet clearly understood which handwriting features are selected by the computer program in the course of training for detecting similarities between different handwritings. Machine learning thus effectively replaces a process of forensic handwriting analysis which relies on a smaller set of features, careful attention, and knowledge by experience, but reaches its limits when it comes to large collections of unknown sources. At the same time, it is clear that due to the complexity of text production — potentially involving distinct autobiographic reporters, authors, scribes, later copyists and correctors —, machine learning of handwriting differences can only contribute partially to the actual research issue of document histories. It must be integrated with independent stylometric/linguistic, textological and historical knowledge, calling for a mixed-methods approach.

F. The 'human in the loop'

Human intervention is necessary at many points in the workflow: The user initially uploads manually transcribed texts in order to form a ground truth. LiViTo uses transliterations to train itself for the particular case. The user then has to form hypotheses about the number of

possible scribes by uploading automatically created line segmentations that probably belong together into the system folders (exact instructions are available on Gitlab).

The statistical methods of stylometry leave many parameters to be determined by the researcher (including, e.g., the choice of distance measure, clustering method, or set of most frequent words); the selection of linguistic features is based on philological wisdom; and historical aspects are investigated by classical rather than digital methods. The classic methods would be for example, research into personalities who were able to write in the community, reconstruction of the history of the handwriting, formulation of possible educational paths in the community. Similarly, the process of machine-based optical pattern recognition involves several steps of manual control during which the recognition process is halted and intermediate results are checked and possibly corrected along the way. Tikhonov's[31] method exemplifies this procedure: Initially, the user estimates a number of potential writers and manually transcribes a sample of the manuscripts, consisting of at least five pages or 100 lines per potential writer. LiViTo trains a neural network and transcribes further texts from the collected examined. The network architecture for the transcription network is a CNN-LSTM-CTC. Outputs from the convolutional neural network (CNN) are fed into a special form of a recurrent neural network, a long short-term memory (LSTM) network designed to handle temporal data structures. The connectionist temporal classification (CTC) function then interprets the sequence of the LSTM outputs as a probability distribution over all possible transcriptions for a given input sequence and trains the network by maximizing the log probabilities of the correct transcriptions on the training set[32]. The scribe identification network achieved an identification accuracy of 85% on our dataset. To make the results human-readable and interpretable, we took the network's output and embedded the 128 automatically chosen visual features to get a three-dimensional vector.

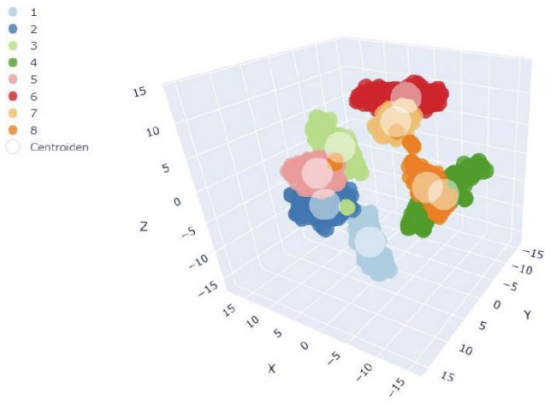

Fig. 4: Visually based writer clustering in LiViTo.

[31] Tikhonov, *Sprachen*, (2022).
[32] Graves et al., *Connectionist*, (2006).

This vector results from several loops of exchange between the user and the machine. During the development phase, these were repeatedly tested by the number of writers and their associated texts. After about five attempts, the figure of ten scribes came as a plausible result, in which the methods of machine and philological classification were compared.

G. Status of data visualization

The stylometric analysis of the manuscripts uses a rather limited set of linguistic features, deliberately chosen by the researcher, to calculate distance matrices between the documents. They can be presented either in tables or in various types of graphs; and graphs constitute a virtually indispensable tool for the identification of clusters in the data. Machine-based optical pattern recognition, moreover, is based on an analysis of no less than 216 different visual handwriting features. Here, the only appropriate way of depicting the results is to visualize the distances between various clusters of similar manuscripts.

We ensure visualization in LiViTo as a research assistance tool by using open-source software Jupyter Notebook. The web-based interactive environment as part of the Jupyter Project makes LiViTo available as web browser-based application[33]. All three functions of LiViTo — localization of revisions, keyword spotting, and clustering according to visual characteristics — can be started and used as three separate applications in the Jupyter Notebook. The users have to be generally open to programming languages, but they do not have to be able to code. A detailed ReadMe document contains commands and preparatory installation steps that must be conducted by the user. Once these pre-settings are done, the functions of the program run with very little effort. All three LiViTo functions have a maximum of 10 short steps that lead to a result or partial result and are described in the ReadMe document. The user can extract and download her/his results from the Jupyter Notebook with just a few clicks (cf. https://gitlab.com/musterworker/livito).

[33] Adam Rule, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, et al., "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks", in: *PLOS Computational Biology 15/ 7,* (San Francisco: PLOS, 2019).

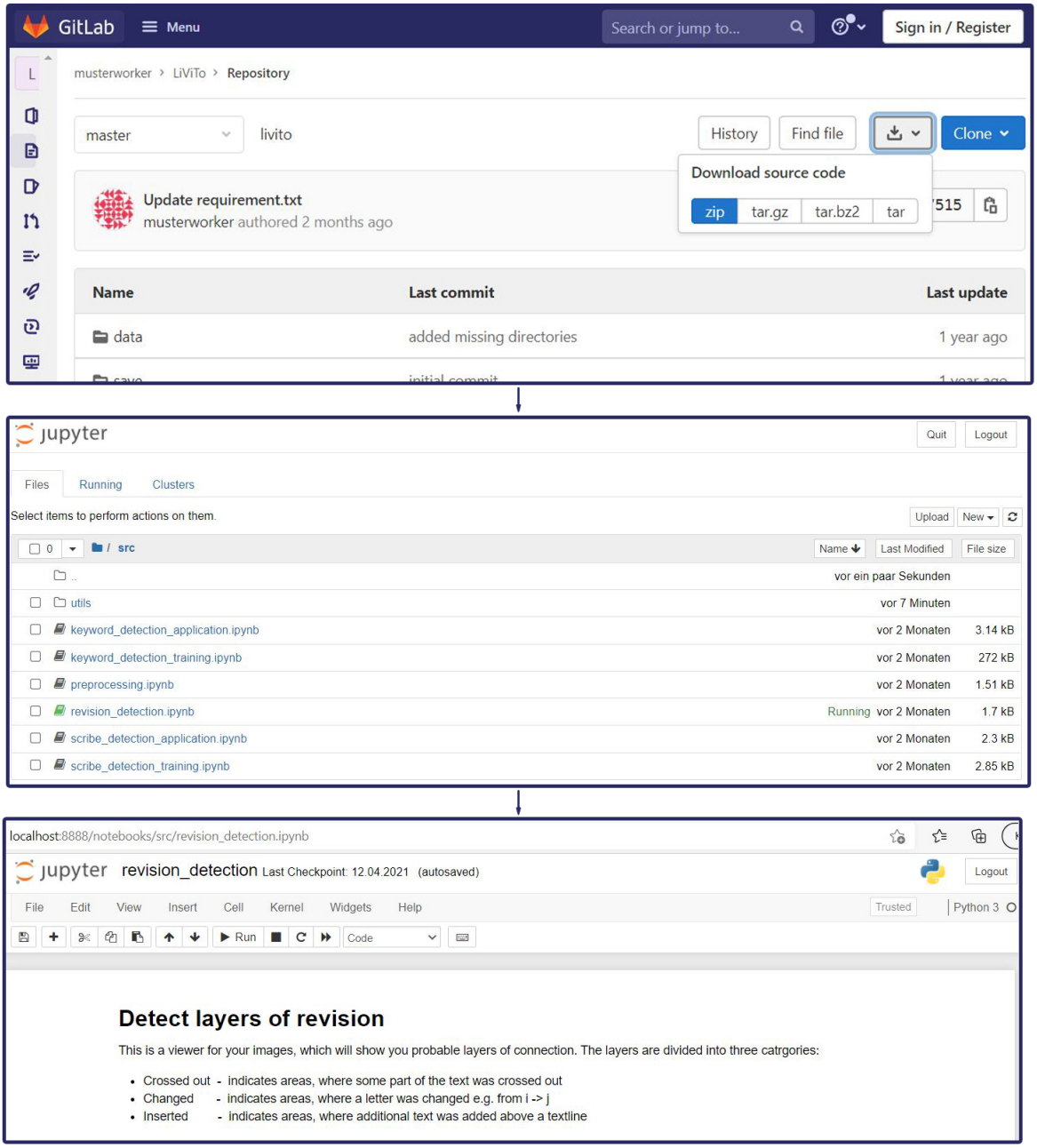

Fig. 5: LiViTo in the jupyter notebook‘s interface in a browser on Windows 10; top down: downloading LiViTo repository for the intallation, LiViTo‘s functions.

Discussion

A. Digitalization:[34] Increase in efficiency and change of research perspective

Normally, working with old manuscripts implies going to an archive or library, putting on white gloves, and leafing carefully through pages, taking notes or transliterating the content.

[34] For a terminological differentiation between digitization and digitalization see Mary Anne M. Gobble, “Digitalization, Digitization, and Innovation”, in: *Research-Technology Management 61/4*, (Virginia: Industrial Research Institute, 2018), 56–59.

This method applies to philologists, historians, theologians and many other scholars who work with such documents. Some institutions offer scanning services, or researchers are allowed to take photos of the manuscripts for free or for a fee so that they can work with scans or photos without time restriction. All the traditional methods are rather impractical in cases of research on hundreds or thousands of pages.

Digitization of manuscripts brings benefits to both sides — to computer science and to the humanities. Digitization forms the ground truth for machine learning and for the definition and collection of quantitative linguistic features, and it allows a very detailed examination of the documents which is essential for developing the revision detector function of LiViTo, so that it works for every examination case without uploading transliterated documents. The function has been successfully tested with German, Czech, French, Hebrew and Latin. In addition, digitization was helpful for the linguistic part of the project, because it made legible marginalia that were no longer discernible to the naked eye. This revealed new facts about the manuscripts. Several handwritten copies or versions of these manuscripts were localized in the Czech Republic, and the complete history of the manuscripts could be traced. With only classical Close Reading methods of the material or photographed manuscripts, none of these results could have been achieved. Working with digitized documents also enabled simultaneous and efficient co-working on the same pages or parts of the manuscripts.

Both sides of the project made different visual segmentations of the documents. Line course detection and line comparison became possible and further development of the keyword spotting function could be witnessed, in which not only whole words but also letters in the beginning, end, or in the middle of the word could be searched. This enables queries for roots, stems, prefixes or derivative affixes in terms of morphology. It also allows the identification of certain registers that are characterized by specific endings. In addition, a layout analysis was carried out at some points in the handwritten books. Subsequently inserted lines or entire passages were recognized. The texts could also be separated according to different principles (e.g. grammatical person — first or third)[35], whereby the hybrid authorship or collective vs. individual genesis of the manuscripts was proved.

B. Methodological controversies

Regarding our philological and historical scientific community, there were no problems presenting the project at colloquia and conferences. The absolute majority of colleagues reacted with great interest and eagerly awaited the results. Only one specific

[35] Tikhonov, *Sprachen*, 109.

misunderstanding, which concerns the combination of quantitative and qualitative methods, came up several times and had to be clarified. Some colleagues conceived of the project goal as a complete switch to quantitative methods and optimization of research tools and procedures. On the contrary, the quantitative approach without the qualitative one would only yield partial results (and vice versa) — the combination of both was absolutely instrumental. The interaction of computational and linguistic methods was decisive for the success of the project.

To demonstrate this with a concrete example, the 3D graphical representation of the clustering in scribe identification is the result of at least three large comparison tests over approximately 12 months. In the beginning, linguistic features were combined with the visual features automatically recognized by AI methods. Next, the results of independent computational and linguistic analyses were compared. After each comparison, the analytical criteria were improved in accordance with the partner method. The 3D clustering then became a manageable result for the analyses.

However, a profound interpretation of this clustering is not possible without a deeper philological analysis. Quantitative visual and linguistic features were used on both sides in order to achieve a common quantitative result. This result then has to be translated into qualitative findings on both sides. In the literal sense of the word, we must zoom in on each individual point of the cluster diagram in the application and take into account the non-visual and quantitative-linguistic features in order to ultimately state concretely how many people wrote the documents and who these people were. So classical qualitative methods are by no means irrelevant — they just need to be combined with quantitative approaches.

C. Details versus abstraction

LiViTo provides both options: details and abstractions. The search results can be presented as a general overview or in detail. Depending on the research question, there are different relevant types of results — small but meaningful details or general overviews of large amounts of analyzed research data. As for the question of quantity and quality, we do not argue for an ‘either/or’ principle, but rather for a balanced combination. Both approaches benefit from each other. The task of the researchers is to use the right method at the right point of investigation.

Often it cannot be defined from the beginning that the research question will only be answered qualitatively or quantitatively, but there can be different scenarios. In the case at hand, qualitative preliminary examinations were carried out both in the computational and

linguistics parts of the project. We first went into detail, that is, recognized prominent linguistic features and the regularities or irregularities in their occurrence; at the same time, we selected representative manuscript pages for first visual tests. In a stepwise process, we enlarged the amount of research data to be handled until we were able to take into account all the necessary features and all of the document pages. As soon as we achieved a first result for the full range of data, we checked whether it was realistic or it contained obvious errors both at a macro- and micro-level (overview vs. detail). When details led to corrections, they had to be scaled up again in order to check for improvements at the more abstract level.

D. Towards a prototype DH laboratory

We are certainly no laboratory in the sense of a permanent institution. To us, a laboratory involves a larger set of researchers from the institutions to which the partners belong (HU, Fraunhofer IPK and MusterFabrik Ltd.), who contribute expertise from a wide range of fields. But we are certainly a team of scientists from different disciplines (including computer science and linguistics), who jointly and regularly conduct research on a common question, by using a mix of methods from their respective fields, in order to produce a joint result.

The most important phase in this common endeavour is the integration of research methods on the way to the concrete answer to a research question. Both sides complement each other with their competence in theoretical and practical areas; the result, however, is a common analysis rather than a confrontation of (computational vs. humanities') standpoints. In our experience, the integration phase has been the most time-consuming and rewarding part of our work, more intense than the actual formulation of results. It seems that such a level of intensity of exchange distinguishes a laboratory from a more loosely defined research group. In this sense, the project can be seen as a prototype DH laboratory. Based on this and several similar smaller-scale projects in the humanities and social sciences, HU Berlin has recently launched a long-term centre for "Digitality and digital methods at Central Campus", headed by Roland Meyer and Torsten Hiltmann.

Major outcomes and prospects for future DH research

We consider the major outcomes of our project to be

(i) a better understanding of the respective contributions of machine-learning and linguistic/stylometric approaches to the task of detecting scribes and authors of historical manuscripts;

(ii) an open-source software package which may assist researchers in detecting authors and scribes on larger sets of unknown historical documents;
(iii) the concrete analysis of document origin and transmission for the 18th c. Czech autobiographies from the *Archiv im Böhmischen Dorf*, Berlin; and
(iv) implications of this analysis for the history of Czech-German language and cultural contact in Berlin, and for the history of the Brethren.

If we focus on the more general DH-related aspects (i)–(ii) here, the obvious future prospect is the application of the mixed-methods approach of this project and its software prototype to other cases of author/scribe detection in other languages and historical periods. Already, within the small field of Slavic philology, many instances of unclear or disputed document origins come to mind, for example the older Church Slavonic witnesses that exist only in numerous partially overlapping later versions[36], or texts of doubtful authenticity such as the Czech *Rukopisy královédvorský a zelenohorský*. Since the LiViTo tool is basically language-independent and requires only a very limited amount of training data, these possibilities will certainly be explored.
In the case of the Rixdorf Czech manuscripts, we have started to apply these techniques to the large set of sermons with promising first results. While most of them are obviously translated from German, their origin and transmission is interesting for the history of the Brethren mission; and there are many issues in historical linguistics which can be fruitfully approached on the basis of such a translation corpus. At present, we are exploring the translations of the Brethren sermons into other early modern languages (even rather exotic targets of missionaries) and their potential for creating a historical parallel corpus.
Sustainable access to digitized sources and to research data in general is becoming increasingly important. A significant branch of DH focuses on document preservation and digital archiving. In order to provide sustained availability of the digitized sources developer in the project, we intend to explore integration into the Laudatio repository[37] after consultation with the Brethren Archive.
During the last few years, character recognition technology (OCR) for manuscripts has witnessed most impressive developments that have opened up possibilities unheard of at the

[36] for a fundamental non-DH treatment of the *Slovo o zakone i blagodati* see Giorgio Ziffer, "Jazyk i stil' slova "O zakone i blagodati"", in: *Učënye zapiski Kazanskogo universiteta 155 (5),* (Kazan': Kazanskij (Privolzhskij) federal'nyj universitet, 2013), 7–16.
[37] „LAUDATIO - Long-term Access and Usage of Deeply Annotated Information", Humboldt University Berlin, accessed February 2, 2023, https://www.laudatio-repository.org/.

outset. In LiViTo, this has already led to effective string comparison even for untranscribed texts with a truly manageable training effort, enabling searches in such documents. Projects such as Transkribus[38] or eScriptorium[39] continue to foster progress in domains like line detection and OCR in historical manuscripts. It will certainly be rewarding to integrate components of these projects into the workflow of LiViTo in order to further improve our scribe detection.

Independently, the linguistic side of the coin has witnessed considerable progress in the application of stylometry, which we intend to reflect in further research in the historical domain. Altogether, it seems that the main idea of the project — to combine pattern recognition for scribe detection and linguistic/stylometric analysis for authorship in order to uncover document origin and transmission for historical manuscripts — is as interesting and topical as ever. We hope that the integration of mixed methods achieved in the project together with the LiViTo tool will make a useful contribution to this area of research.

Bibliography

Burrows, John F. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style". In: *Literary and Linguistic Computing 2*, 61–70. Oxford: Oxford University Press, 1987.

_____ ""An Ocean Where Each Kind...": Statistical Analysis and Some Major Determinants of Literary Style". In: *Computers and the Humanities 23,* 309–321. New York/Heidelberg/AA Dordrecht: Springer, 1989.

Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem". In: *Digital Scholarship in the Humanities 30*, 167–182. Oxford: Oxford University Press, 2010.

Eder, Maciej, Jan Rybicki, Mike Kestemont. "Stylometry with R: a package for computational text analysis". In: *R Journal 8 (1)*, 107–121. Online-Open-Access-Publication, 2016. https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf.

Graves, Alex, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. Pittsburgh: Carnegie Mellon University, 2006.

Hauger, J. Scott. *Reading Machines for the Blind: A Study of Federally Supported Technology Development and Innovation (Dissertation)*. Blacksburg: Virginia Polytechnic Institute and State University, 1995.

Hope, Jonathan. *The authorship of Shakespeare's plays. A socio-linguistic study*. Cambridge: Cambridge University Press, 1994.

Kahle, Philip, Sebastian Colutto, Günter Hackl, Günter Mühlberger. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In:

[38] "Transkribus – Where AI meets historical documents", READ-COOP, accessed February 2, 2023, https://readcoop.eu/transkribus/.

[39] "eScriptorium", GitLab, accessed February 2, 2023, https://gitlab.inria.fr/scripta/escriptorium.

14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 19–24. Kyoto: IEEE, 2017.
Kay, Anthony. "Tesseract: An Open-Source Optical Character Recognition Engine". In: *Linux Journal*. Online-Open-Access-Publication, 2007. https://www.linuxjournal.com/article/9676.
Kleinwächter, Livia. "The Literary Manuscript: A Challenge for Philological Knowledge Production". In: *Philology in the Making Vol. 1*, edited by Pál Kelemen and Nicolas Pethes, 109–128. Bielefeld: transcript Verlag, 2019.
Koppel, Moshe, Jonathan Schler, Shlomo Argamon. "Computational Methods in Authorship Attribution". In: JASIST 60, edited by Steven Sawyer, 9–26. Hoboken: Wiley-Blackwell, 2009.
Mettele, Gisela. *Weltbürgertum oder Gottesreich: die Herrnhuter Brüdergemeine als globale Gemeinschaft 1727 – 1857.* Göttingen: Vandenhoeck & Ruprecht, 2009.
Motel, Manfred. *Das böhmische Dorf in Berlin: die Geschichte eines Phänomens.* Berlin: Darge Verlag, 1983.
Müller, Klaus, Aleksej Tikhonov, Roland Meyer. „LiViTo: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts". In: *Proceedings of the Twelfth Language Resources and Evaluation* Conference, 885–890. Marseille: European Language Resources Association, 2020.
Plecháč, Petr. "Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns". In: *Digital Scholarship in the Humanities*. Oxford: Oxford University Press, 2020.
Rule, Adam, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, et al. "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks". In: *PLOS Computational Biology 15/ 7.* San Francisco: PLOS, 2019. https://doi.org/10.1371/journal.pcbi.1007007.
Seibt, Angelika. *Unterschriften und Testamente – Praxis der forensischen Schriftuntersuchung*. München: Beck, 2008.
Sterik, Edita. *Die böhmischen Exulanten in Berlin*. Herrnhut: Herrnhuter Verlag, 2016.
Štěříková, Edita. *Běh života českých emigrantů v Berlíně v 18. Století*. Praha: Kalich, 1999.
Tikhonov, Aleksej. *Autorenidentifikation und linguistische Merkmale der Rixdorfer Handschriften: Eine Untersuchung anhand von Manuskripten aus dem 18./19. Jahrhundert (Dissertation)*, (Berlin: Humboldt-Universität zu Berlin, 2020.
Tikhonov, Aleksej. *Sprachen der Exilgemeinde in Rixdorf (Berlin): Autorenidentifikation und linguistische Merkmale anhand von tschechischen Manuskripten aus dem 18./19. Jahrhundert.* Heidelberg: Winter Verlag, 2022.
Tikhonov, Aleksej and Klaus Müller. „LiViTo: A software tool to assess linguistic and visual features of handwritten texts". In: *Qurator - Conference on Digital Curation Technologies 2020*, edited by Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, Lydia Pintscher. Berlin: Online-Open-Access-Publication, 2020. https://ceur-ws.org/Vol-2535/paper_8.pdf.
Tikhonov, Aleksej and Klaus Müller. „Scribe versus authorship attribution and clustering in historic Czech manuscripts: a case study with visual and linguistic features". In: *Digital Scholarship in the Humanities*, 254–263. Oxford: Oxford University Press, 2022.
Ziffer, Giorgio. "Jazyk i stil' slova "O zakone i blagodati"". In: *Učënye zapiski Kazanskogo universiteta 155 (5),* 7–16. Kazan': Kazanskij (Privolzhskij) federal'nyj universitet, 2013.
Zyl, Maryka van and Yolande Botha. "Stylometry and Characterisation in The Big Bang Theory". In: *Literator 37/ 2*, 1-11. Cape Town: Aosis Publishing, 2016.

Betreff: Mixing Methods. Practical Insights from the Humanities in the Digital Age
Von: Tino Mager <tino.mager@rug.nl>
Datum: 20.04.23, 12:03
An: roland.meyer@hu-berlin.de

Dear Roland,

As editors, we hereby confirm that the article

Detecting authorship, hands, and corrections in historical manuscripts. A mixed-methods approach towards the unpublished writings of an 18th century Czech emigré community in Berlin

by Roland Meyer, Aleksej Tikhonov, Robert Hammel

has been accepted for publication in the collective volume

Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein (eds.): *Mixing Methods. Practical Insights from the Humanities in the Digital Age*. Bielefeld: Transcript, 2023.

Best wishes,
Tino

Tino Mager

Asst. Prof. Dr.
History and Theory of Architecture and Urbanism

University of Groningen | **Faculty of Arts**
Department of History of Art, Architecture and Landscape
Oude Boteringestraat 34
9712 GK Groningen
The Netherlands

President, ICOMOS Germany
Secretary General, ICOMOS International Scientific Committee on Water and Heritage