LiViTo: a software tool to assess linguistic and visual features of handwritten texts

Aleksej Tikhonov[1] and Klaus Müller[2]

[1] Institut für Slawistik, Humboldt University, Berlin, Germany (HU)
[2] MusterFabrik Berlin, Berlin, Germany (MFB)

Abstract. A mixed method approach for identification of scribes and authors in handwritten documents will be presented by introducing LiViTo, a tool which combines linguistic insights and computer vision techniques in order to assist researchers in the analysis of handwritten historical documents. This report shows that it is feasible to train neural networks for automatic transcription of handwritten documents and to use these transcriptions as input for further analysis. Hypotheses about scribes can be tested effectively by extracting visual handwriting features and clustering them appropriately. The mixed methods system shows the benefits on both sides - linguistics and computer vision. LiViTo was trained with historical Czech texts by 18th century immigrants to Berlin, a total of 564 pages from a corpus of about 5000 handwritten pages without indication of author or scribe. An overview of the development of LiViTo and an introduction into its methodology and its functions will be provided. Findings concerning the corpus of Berlin Czech manuscripts and possible further usage scenarios will be discussed.

Keywords: mixed methods, digital humanities, machine learning, linguistics, authorship attribution, Czech, Slavic, Slavonic

1. Introduction

Manuscripts in small private or parochial archives may contain reports by personal witnesses, new information on everyday culture and language. LiViTo is devoted to exploring handwritten sources of a community of refugees to 18th c. Berlin: the "Czech brethren" (aka *Moravian Church*, *Herrnhuter Brüdergemeine*), who fled from anti-Protestant persecution in the Czech lands to Saxony and Prussia. Research questions that arise in this context are: (i) Are the manuscripts originals or handwritten copies? (ii)

Are the originator/author and the scribe the same person? (iii) How many authors and scribes worked on the manuscripts? (iv) Can these authors and scribes be found in other manuscripts? (v) Are there revisions in the manuscripts and where are they? For the identification of scribes and authors in manuscripts methods from classical linguistic analysis are combined with modern computer vision approaches, such as neural networks to enhance the knowledge discovery process and knowledge representation.

2. Who are the users?

The potential target audience for LiViTo are researchers and students from humanities, social studies, law and medicine. LiViTo is intended to be an assistance system for analysing, comparing and clustering of handwritten (historic) data. Research questions in law and medicine could be the origin or the linguistic and visual interdependence of handwritten legal documents such as birth certificates and clinical records or testaments. In that regard, the questions of the humanities and social sciences are often equal to law and medicine.

Meticulous and close reading, understanding and analysis of handwritten texts is an inalienable part of a research process. Such a qualitative research approach can be combined with the (half-)automatic methods of LiViTo focusing quantitative data, obtained by statistical research methods, resulting in a mixed methods research approach. It should be clear to the user that LiViTo is a data driven assistance system which provides results that can lead the user to both kinds of results - quantitative and qualitative. LiViTo is designed as well for users with only minimal technical knowledge. The intent of this software design is to enable the user to get first insights into the manuscripts and iterate faster through research questions rather than spending time learning a complex tool. This is why it is relevant for archivists, curators, museum employees and genealogists. The users will be provided with a step by step manual which should help in the beginning working with LiViTo.

3. Use cases and interaction

3.1 Preconditions

Since LiViTo processes manuscripts it needs image files. It can process various file formats like tif, png and jpg. It has to be taken into account that for scribe detection the analysis should include at least five pages per potential scribe as well as at least two potential scribes. In order to use the keyword detectors functionality transcripts of the manuscripts for the training of the neural network need to be provided by the user. For keyword detection the minimal amount of pre-transcribed data should be about 150 text line segmentations for each potential scribe. Therefore it does not make sense to use LiViTo if the data set is smaller than the twofold of the minimal amount of data necessary for the keyword detector.

3.2 Software structure

LiViTos functionality shall be explained in three use cases. Fig. 1 gives an overview of LiViTos module dependency and data flow. The main module, which is underlying the main functionalities is the preprocessor module. It needs to be executed before the other modules can be used. The preprocessor takes all input images, which should ideally be densely written pages, and extracts binarized text lines and its corresponding coordinates in the image from the manuscripts, which will be needed for the other modules. The preprocessor also creates a data structure which will be built upon by the other modules.

The scribe detector as well as the keyword detector both need preprocessing steps, as they contain neural networks, which need to be trained on the users data. The revision detector is general enough to be data independent and does not need any training.

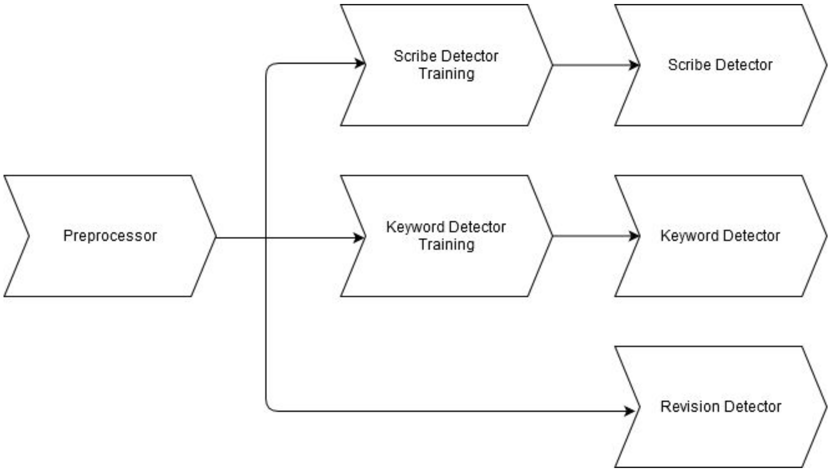


Fig. 1: LiViTos data flow and module dependency.

3.3 Module 1: Scribe Detector:

The user needs to provide a folder of about 100 text line segmentations per scribe to the training system, which were generated in the preprocessing step. Next the neural network will try to differentiate the scribes from each other. The grade to which the hypothetical scribes are distinguishable will be shown in a graph. The scribe detector has two use cases. First it can be used as a tool for hypothesis testing for identifying probable scribes in manuscripts. Fig. 2 shows the users hypothesis/training results for two training processes. On the left side in Fig. 2 the user made a hypothesis, which the network could not verify, as no monocolor clusters can be formed. On the right side in Fig. 2 the user made a hypothesis which can verified to a high degree.

Each data point represents a single text line segmentation, where as the color stands for the respective class attributed by the neural network. 128 features from an intermediate network layer are embedded with t-SNE into a 3 dimensional representation. Therefore the measure on the axes is not as relevant as the clustering property itself (van der Maaten, 2008). If the user is satisfied with the results the trained model can be applied on all documents in order to create a mapping of scribes to manuscripts. This would be the second use case for the scribe detector.

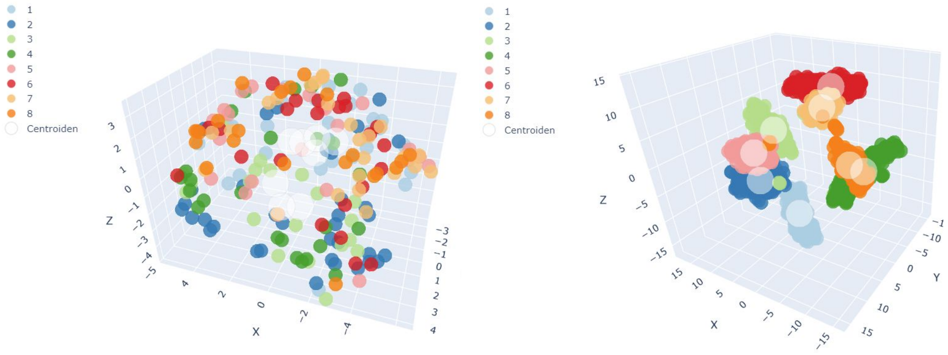


Fig. 2 Scribe Detector output. Left: not differentiable; Right: well differentiable.

The convolutional neural network (CNN) used for the scribe detector is based on DeepWriter. It uses multiple crops from each text line segmentation to augment the dataset and learn character specific features, which makes the scribe identification process text independent. The scribe detector achieves similar accuracy on classifying scribes on the IAM[1] dataset as DeepWriter, which is about 92% (Xing, 2016).

3.4 Module 2: Keyword Detector:

The keyword detector module is a customizable query engine, which needs to be trained on the users data. It needs transcripts of text line segmentation provided by the user to train a neural network. The transcripts need to be in .txt file format and encoded in UTF-8. Once the model is trained the user can query the manuscripts for detecting language features which can be traced back to an individual style of writing in both meanings - author and scribe. In order to clarify how it works some examples are given.

- The use of lexis from colloquial language is concerned with linguistic register variation or dialectology, e.g. Czech pronoun <won>[2] which is marked by the initial prosthetic <w-> as clearly colloquial. That can be traced back to a specific author who used colloquial language in written texts.

But the keyword detector is not only detecting full word forms as might be expected. Word fragments or a single letter can be detected and analysed as well.

- Likely as <won> the adjective ending <-ej> is concerned with linguistic register variation or dialectology in Czech. The query for "*ej*" as an ending of words would show which texts in a particular data sample are written in colloquial Czech or a dialect of Czech.

- The statistics of upper and lower case will show the distribution of absolutely and relatively upper, lower and other[3] characters for the whole data set. The interdependence of the number of upper and lower cases can be attributed to educational background and grammar competence of the scribe or as well the period of the origin of the manuscript, because at different times different rules or norms of Upper-Lower-Case-writing did exist.

[1] http://www.fki.inf.unibe.ch/databases/iam-handwriting-database (Last accessed 2019/12/03)

[2] standard Czech of 18th century & today: *on*; English: *he*

[3] Punctuation characters and numerals.

Among single word forms, word fragments and single letters can be used for analysing a larger language unit as well. The results for a competing query of Czech adjective feminine ending <-ná>[4] and masculine ending <-ný> can be seen in Fig. 3. The top graph shows the distribution of detection over all documents (blue = feminine; orange = masculine). Fig.3 left/right show the top 20 accrued words for each query. In this case there are as well some pronouns among the results (e.g. <ona> (Eng.: she)) and a few number of other word classes, which should be ignored in the analysis. This module also contains a manuscript viewer, which lets the user browse through the query results. Deleting individual results will dynamically adjust the statistical outputs.

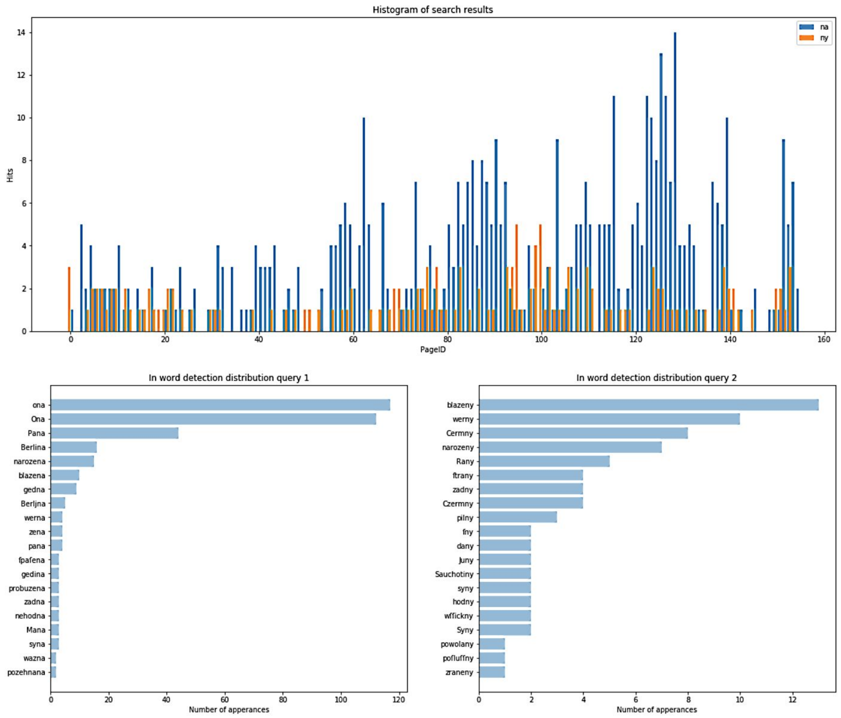
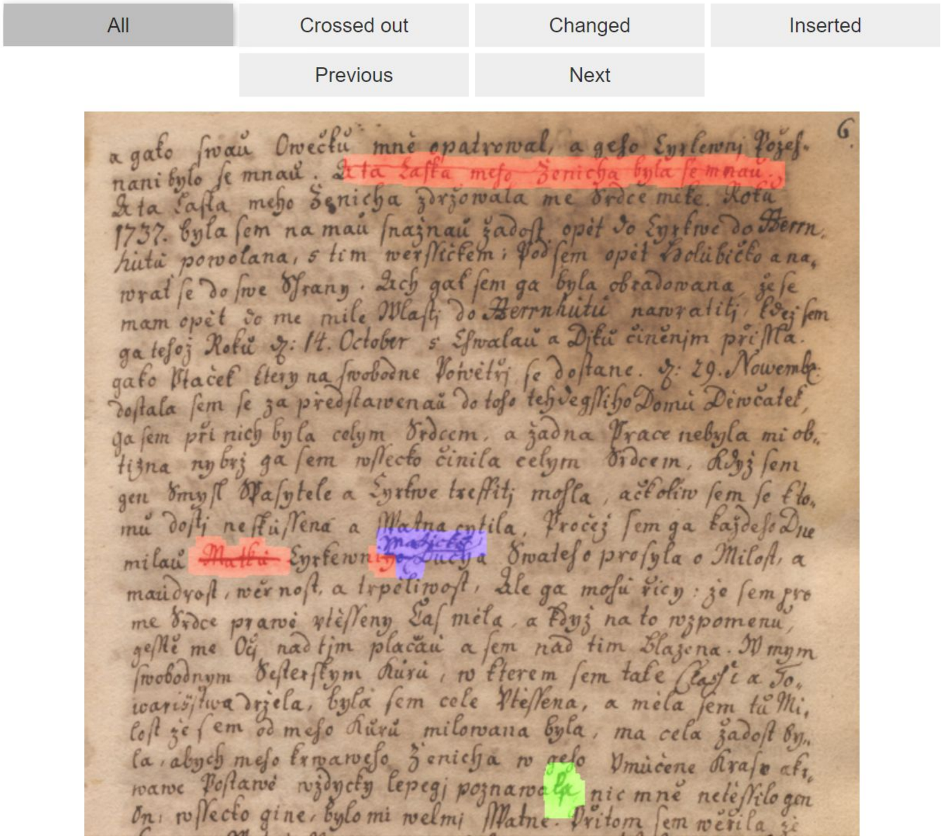


Fig. 3: Keyword detector comparing query results. Top: Histogram of both queries for the whole data set. Bottom left/right: Histogram of the most common words satisfying the query.

[4] There are more feminine / masculine endings in Czech. This is only one example.

The network architecture for the transcription network used for the keyword detector is a CNN-LSTM-CTC. Outputs from the CNN get fed into a special form of recurrent neural network, a long short-term memory (LSTM) network, which is 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 (Graves, 2006). A comparison of LiViTos transcription accuracy in terms of character



6.

a gako ſwau Owečku mně opatrowal, a geho Cyrkewnj Požeh-
nani bylo ſe mnau. ~~A ta čaſka meho Ženicha byla ſe mnau.~~
A ta čaſka meho Ženicha zdržowala me Srdce měke. Roku
1737. byla ſem na mau ſnažnau žadoſt opět do Cyrkwe do Herrn-
hutu powolana, s tim werſſičkem; Pod ſem opět holubičko a na-
wrat ſe do ſwe Strany. Ach gak ſem ga byla obradowana že ſe
mam opět do me mile Wlaſtj do Herrnhutu nawratitj, kdež ſem
ga teho Roku d: 14. October s Chwalau a Djku činěnjm přiſſla.
gako Ptaček který na ſwobodne Powětrj ſe doſtane. d: 29. Nowembr.
doſtala ſem ſe za předſtawenau do toho tehdegſſiho Domu Děwčatek,
ga ſem při nich byla celym Srdcem, a žadna Prace nebyla mi ob-
tižna nybrž ga ſem wſſecko činila celym Srdcem, když ſem
gen vmyſl Spaſytele a Cyrkwe treffitj moſla, ačkoliw ſem ſe kto-
mu doſti neſtuſſena a Matna cytila, Pročež ſem ga každeho Dne
milau ~~Matki~~ Cyrkewnjho matičke Ducha Swateho proſyla o Miloſt, a
maudroſt, wěrnoſt, a trpělivoſt, Ale ga mohu řícy: že ſem pro
me Srdce prawě vtěſſeny Čaſ měla, a když na to wzpomenu,
geſſtě me Očy nad tjm plačau a ſem nad tim Blažena. W mym
ſwobodnym Seſterſkym Kůru, w kterem ſem take Čaſ tj a To-
wariſſtwa držela, byla ſem cele Vtěſſena, a měla ſem tu Mi-
loſt že ſem od meho Kůru milowana byla, ma cela žadoſt by-
la, abych meho krwaweho Ženicha w geho Vmučene Kraſe a kr-
wawe Poſtawě wždycky lepegi poznawala nic mně neteſſilo gen
On: wſſecko gine, bylo mi welmi Matne. Pritom ſem wěřila že

Fig. 4: Revision Detector. Displaying all three kinds of revisions detected on a manuscript.

error rate (CER) with tools like Transkribus[5] on the IAM Dataset resulted in 5% CER for Transkribus and 4% CER for LiViTo (Scheidl, 2018).

3.5 Module 3: Revision Detector:

The model used for revision detection is based on the U-Net architecture (Ronneberger, 2015). It does not need any training as the model is general enough to detect revisions even in other languages and historical handwriting styles. The module is mainly a manuscript viewer, which can be used directly after data preprocessing.

It will highlight three different kinds of revisions in manuscripts, crossed out areas, annotations made above a text line, and probable changes of single letters (e.g. if a scribe changes the letter <a> to <e>), which is shown in Fig. 4. The user can browse through the whole data set looking for all detected revision types or just individual ones. Each revision type is highlighted in a different color, so they are easily distinguishable. A comparison of the revision detector with other tools was not possible as a search for similar technology did not yield any results.

4. Conclusion and future work:

LiViTo is designed as an open source tool which provides assistance for the analysis of historical manuscripts and will be released in March 2020. It allows modification and sharing of changes to the code on GitLab. This tool relieves researchers of much technical work and allows them to focus on the analysis of their data and iterate faster through hypotheses. Moreover, the tool enables researchers with little knowledge of machine learning methods to apply them to their work.

In our special research question about the scribes of the Czech immigrant manuscripts from the 18th century in Berlin LiViTo assisted us in making the following conclusions: (i) Significant revisions were made in the first half of the 19th century. LiViTo helped localizing the revisions without close reading of the manuscripts and categorized their kinds. Especially the revisions of <j> to <í> guided us to the revision moment not earlier than in the 1820s, because of the grammatical regulations of Czech standard language made by Josef Dobrovský in the first 20 years of the 19th century (Dobrovský, 1809, 1819). (ii) We could finally identify 10 scribes who wrote the

[5] www.transkribus.eu (last accessed 2019/12/03)

analysed manuscripts with linguistic and visual features. LiViTo assisted the search for different linguistic features, as the archaic verb form <geſt> versus the modern <ge>. (iii) In the tradition of the Czech brethren in the 18th century these manuscripts should be an autograph of the people the CVs are dealing with. Altogether there are 183 people mentioned, but there are 10 scribes. All manuscripts are probably copies and not the originals. A deeper interpretation of the results will be finished by March 2020.

Future expansion might include stylometric analysis of transcribed text with tools like stylo (Eder & Rybicki, 2016) and general natural language processing applications on the automatically transcribed texts generated by the keyword detector. The combination of LiViTo and stylo would allow to work with shorter and handwritten texts. We have a strong interest in making LiViTo available and intraoperative for GLAM organisations using IIIF in a future version of LiViTo.

Funding

This work was supported by VolkswagenStiftung [A118750, Mixed Methods Program].

References

Dobrovský, J.: Ausführliches Lehrgebäude der böhmischen Sprache, zur gründlichen Erlernung derselben für Deutsche, zur vollkommenern Kenntniss für Böhmen. J. Herrl, (1809).

Dobrovský, J.: Lehrgebäude der Böhmischen Sprache: Zum Theile verkürzt, zum Theile umgearbeitet und vermehrt. Haase, (1819).

Eder, M., Rybicki, J.: Stylometry with R: A Package for Computational Text Analysis (2016). https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf.

Graves, A., Fernandez, S., Gomez F., Schmidhuber J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proc. Int. Conf. on Machine Learning, pages 369–376, (2006).

Ronneberger, O., Fischer P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv preprint arXiv:1505.04597 (2015).

Scheidl, H.: Handwritten Text Recognition in Historical Documents. TU Wien (2018).

van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE, Journal of Machine Learning Research (2008).

Xing, L., Qiao, Y.: DeepWriter A MultiStream Deep CNN for Text-independent Writer Identification, arXiv preprint arXiv:1606.06472 (2016).