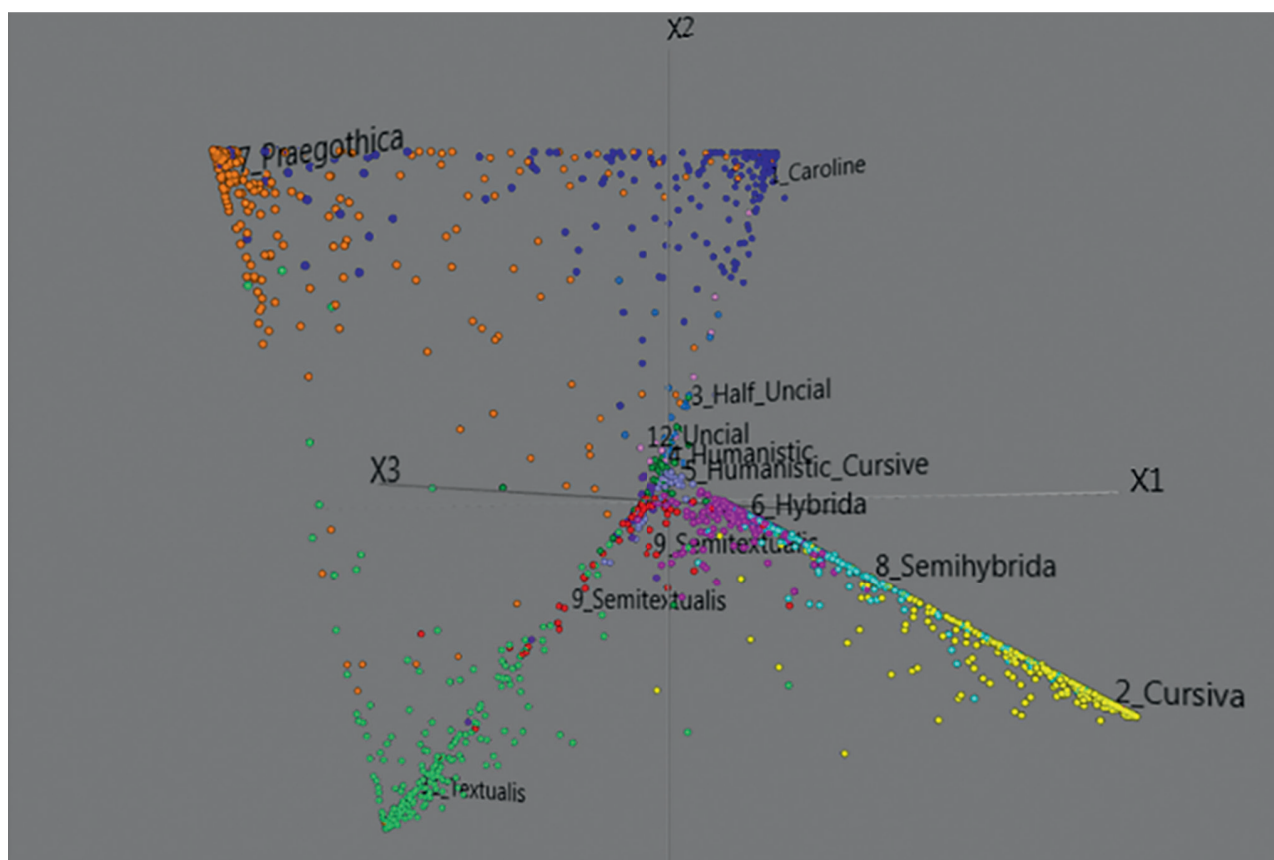


manuscript cultures

Hamburg | Centre for the Study of Manuscript Cultures

ISSN 1867-9617



Publishing Information

Natural Sciences, Technology and Informatics in Manuscript Analysis

Edited by Oliver Hahn, Volker Märgner, Ira Rabin, and H. Siegfried Stiehl

Proceedings of the third *International Conference on Natural Sciences and Technology in Manuscript Analysis* and the workshop *OpenX for Interdisciplinary Computational Manuscript Research* that took place at the University of Hamburg, Centre for the Study of Manuscript Cultures, on 12–14 June 2018.

Editors

Prof. Dr Michael Friedrich
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg

Tel. No.: +49 (0)40 42838 7127
Fax No.: +49 (0)40 42838 4899
michael.friedrich@uni-hamburg.de

Prof Dr Jörg Quenzer
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg
Tell. No.: +49 40 42838 - 7203
Fax No.: +49 40 42838 - 6200
joerg.quenzer@uni-hamburg.de

Editorial Office

Dr Irina Wandrey
Universität Hamburg
Centre for the Study of Manuscript Cultures
Warburgstraße 26
D-20354 Hamburg
Tel. No.: +49 (0)40 42838 9420
Fax No.: +49 (0)40 42838 4899
irina.wandrey@uni-hamburg.de

Layout

Miriam Gerdes

Cover

Image of ICDAR2017 Tensmeyer's distance matrix (axes 2 and (1 and 3)), see article by Dominique Stutzmann, Christopher Tensmeyer and Vincent Christlein in this volume.

Translation and Copy-editing

Amper Translation Service, Carl Carter, Fürstenfeldbrück

Print

AZ Druck und Datentechnik GmbH, Kempten
Printed in Germany

www.csmc.uni-hamburg.de

ISSN 1867–9617

© 2020

Centre for the Study of Manuscript Cultures
Universität Hamburg
Warburgstraße 26
D-20354 Hamburg

CONTENTS

3 | Editorial

by Oliver Hahn, Volker Märgner, Ira Rabin, and H. Siegfried Stiehl

ARTICLES

5 | On Avoiding Segmentation in Handwritten Keyword Spotting: Overview and Perspectives

Marçal Rusiñol*

11 | Writer Identification and Script Classification: Two Tasks for a Common Understanding of Cultural Heritage

Dominique Stutzmann, Christopher Tensmeyer, and Vincent Christlein*

25 | Z-Profile: Holistic Preprocessing Applied to Hebrew Manuscripts for HTR with Ocropy and Kraken

Daniel Stökl Ben Ezra and Hayim Lapin*

37 | On Digital and Computational Approaches to Palaeography: Where Have we Been, Where Are we Going?

Peter A. Stokes*

47 | Creating Workflows with a Human in the Loop for Document Image Analysis

Marcel Gygli (Würsch), Mathias Seuret, Lukas Imstepf, Andreas Fischer, Rolf Ingold*

53 | Building an Evaluation Framework Researchers Will (Want to) Use

Joseph Chazalon*

61 | Turning Black into White through Visual Programming: Peeking into the Black Box of Computational Manuscript Analysis

Vinodh Rajan Sampath and H. Siegfried Stiehl*

73 | Legally Open: Copyright, Licensing, and Data Privacy Issues

Vanessa Hanneschläger*

77 | A Comparison of Arabic Handwriting-Style Analysis Using Conventional and Computational Methods

Hussein Mohammed, Volker Märgner, and Tilman Seidensticker

87 | Illuminating Techniques from the Sinai Desert

Damianos Kasotakis, Michael Phelps, and Ken Boydston

91 | Image Quality in Cultural Heritage

Tyler R. Peery, Roger L. Easton Jr., Rolando Raqueno, Michael Gartley, and David Messinger

105 | When Erased Iron Gall Characters Misbehave

Keith T. Knox

115 | 'Dürer's Young Hare' in Weimar – A Pilot Study

Oliver Hahn, Uwe Golle, Carsten Wintermann, and Ira Rabin

123 | Material-Technical Details on Papyrus as Writing Support

Myriam Krutzsch

133 | The Techniques and Materials Used in Making Lao and Tai Paper Manuscripts

Agnieszka Helman-Ważny, Volker Grabowsky, Direk Injan and Khamvone Boulyaphonh

163 | Inks Used to Write the Divine Name in a Thirteenth-Century Ashkenazic Torah Scroll: Erfurt 7 (Staatsbibliothek zu Berlin, Or. fol. 1216)

Nehemia Gordon, Olivier Bonnerot, and Ira Rabin

185 | Contributors

Article

Z-Profile: Holistic Preprocessing Applied to Hebrew Manuscripts for HTR with Ocropy and Kraken

Daniel Stökl Ben Ezra und Hayim Lapin | Paris

While considerable high quality textual and lexical data is openly available for many languages, such as Greek or Latin (although there is room for improvement here as well), researchers of classical Hebrew and Aramaic together with many other important European languages such as Armenian or Georgian are groping in the dark.¹ Most of the texts available online are vulgate editions, not scholarly reliable texts.²

Among the most important classical Hebrew texts are those redacted during the tannaitic Rabbinic period, around the third century CE: the Mishnah and the Tosefta, two juridical works, and the ‘Halakhic’ or ‘Tannaitic’ Midrashim, commentaries to the Bible (Exodus–Deuteronomy).³ The Tosefta is a text closely related to the Mishnah that follows the same overall structure and clearly ‘knows’ the Mishnah, but it incorporates legal traditions with a complex intertextual relationship to those included in the Mishnah.

All of these sources have a strong interest in legal matters (the first two are in fact juridical texts) that illuminate Jewish life in Palestine, an Eastern province of the Roman Empire. Better known to the world outside of Jewish studies is the Babylonian Talmud, a later text, which is in fact a

commentary to the Mishnah.⁴ (An over-simplified analogy might be the relationship between two synoptic Gospels: the works share a recognizable outline and common materials but also are also distinct from one another.) The length of these texts is substantial, e.g. about two hundred thousand words for the Mishnah, about three hundred thousand words for the Tosefta, and they probably represent the most extensive sources from the pre-Christian Roman Empire that are still extant and not written in Greek or Latin. Their importance for our understanding of the development not only of classic rabbinic Judaism, but also of Roman provincial law and social history cannot be overstated.

Despite the importance of these texts there are neither full critical editions in either digital or print format. While there are several low quality online open source texts, there are no high-quality transcriptions that are openly available. An excellent linguistically annotated transcription of one manuscript of the Hebrew part of almost all texts can be accessed via the website of the Israel Academy of the Hebrew Language.⁵ While access to its resources is free of charge, there are significant restrictions put on the use of the transcriptions by the Academy. Other projects put their transcriptions behind a substantial pay wall.⁶

Three years ago, Hayim Lapin from the University of Maryland and Daniel Stökl Ben Ezra from the École pratique des hautes études / Paris Sciences & Lettres in Paris have joined their respective projects on these texts in the eRabbinica project to start closing this gap. They secured funding from different sources for different subprojects. A pilot edition of three treatises of the

¹ E.g. Perseus Digital Library at <<https://github.com/PerseusDL>>. The ERC Project Lila at <<https://lila-erc.eu/>>. McGillivray 2013. The Classical Language Toolkit (cltk.org). Bizzoni et al. 2014. The HIMANIS (HISTorical MANuscript Indexing for user-controlled Search) and ORIFLAMMS (Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts) projects have produced some Latin script data sets, see at <<https://github.com/oriflamms/>>. Dominique Stutzmann’s most recent project, HORAE (The HORAE project: A textual exploration of Books of Hours, <https://horae.digital/>) will produce an even vaster data set.

² See e.g. the texts in Sefaria: A Living Library of Jewish Texts <<https://sefaria.org>> or on <<https://en.wikisource.org>>. This is certainly not due to disdain for such texts. Yet, OCR of out of copyright print editions are easier to acquire.

³ Good introductions are Stemberger 2011, Ben-Eliyahu, Cohn and Millar 2013.

⁴ ‘The Talmud’ usually refers to the Babylonian Talmud. In fact, there is a separate somewhat earlier Talmud from Byzantine Galilee called the Palestinian or Jerusalem Talmud.

⁵ *Ma’agarim* (The Academy of the Hebrew Language, The Historical Dictionary Project), <<http://maagarim.hebrew-academy.org.il/>>.

⁶ *Cooperative Development Initiative* <www.lieberman-institute.com>.

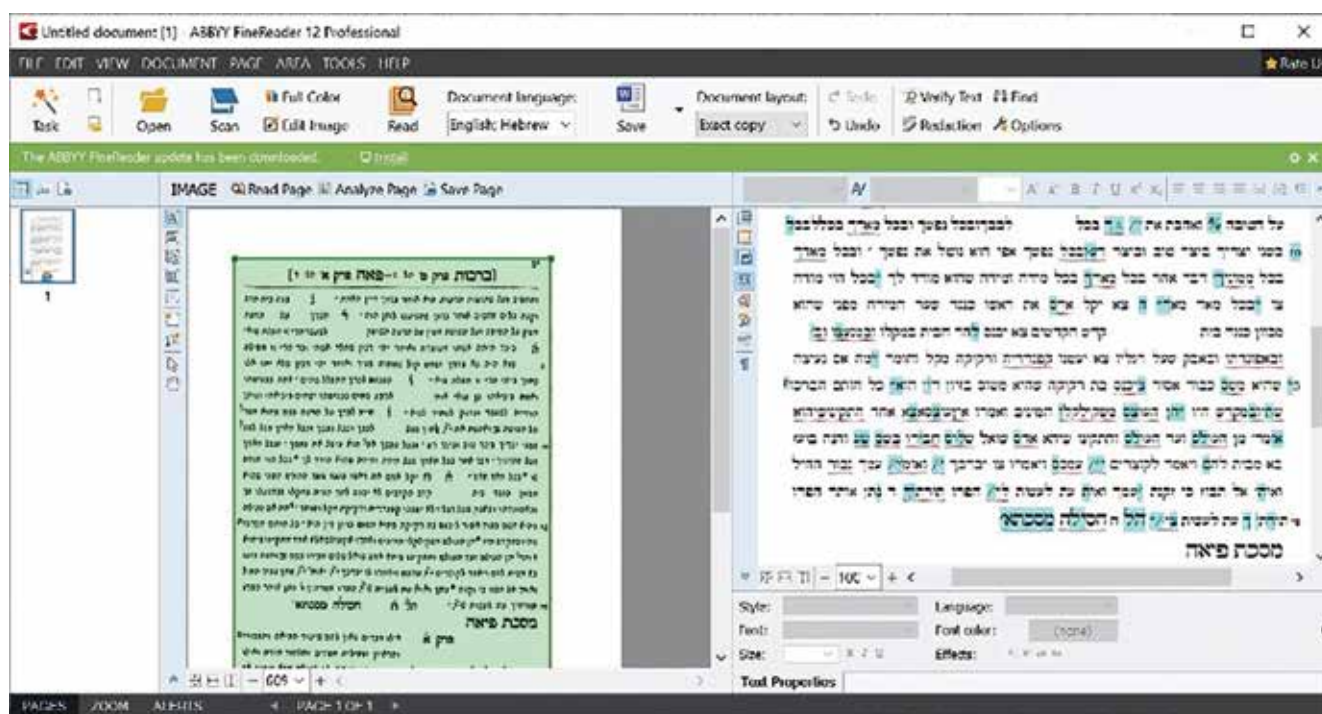


Fig. 1: Screen capture of a random page of Lowe OCR'd with ABBYY.

Mishnah with transcription, automatic textual criticism, French and English translation and linguistic annotation, based on TEI/XML and the open source edition software TEI-Publisher by eXist⁷ has recently been published.⁸ This paper briefly introduces a selection of the computer vision and machine learning algorithms applied in our project following a chronological sequence to perhaps encourage others with similar projects, especially those coming from the humanities. If a given infrastructure constructed for a general purpose achieves bad results, there are means to arrive at good local solutions with open source code.

1. The original motivation: a tailor-made OCR

One of the most important manuscripts of the Mishnah is the Cambridge Ms. Add 470.1 from fifteenth century Byzantium.⁹ In 1883, William H. Lowe published an extremely precise transcription that represents faithfully not only the text of the manuscript but also changes in writing

style using various fonts and special placement of characters above the line for interlinear additions and at the end of lines or of paragraphs in the margins for marginal additions. Dots above letters indicate abbreviations and corrections.

At first, we tried to train a commercial OCR of this nineteenth century transcription, yet the multiplicity of fonts and special characters and ligatures and the use of the less common font (commonly, but imprecisely called *Rashi*) did not give very good results (Fig. 1).

Furthermore, it would have been difficult to preserve all the precious semantic information conveyed in the letter positions. Therefore, in 2016 Stökl Ben Ezra developed a tailor-made simple but very effective OCR engine. In a first step, horizontal projections were used to locate headers and footers, both of which were subsequently excluded from further analysis. The next step was the creation of a huge database of all connected components on the main part of all pages. The central architecture consists of a k-means clustering of 335 classes based on HOG-features (Histogram of Gradients) of connected components (letters or, in the case of ligatures, letter groups).¹⁰ The vector for the Euclidean distance k-means clustering consisted of a concatenation

⁷ TEI-Publisher <<https://teipublisher.com/index.html>>, accessed 26 May 2019.

⁸ For the life interface, see *Digital Mishnah* <editions.erabbinica.org>. For the web interface code, see: <<https://gitlab.existsolutions.com/mishnah/mishnah/>>. For the data see: <<https://gitlab.existsolutions.com/mishnah/mishnah-data/>>.

⁹ Accessible at <<https://cudl.lib.cam.ac.uk/view/MS-ADD-00470-00001>>.

¹⁰ More precisely, letters consisting of several unconnected strokes, such as *he* or *qof* or all letters with a dot or stroke above or inside where combined into a single connected component via an intermediate vertical morphological transformation.



Fig. 2: Sample of Letter Clusters.

of HOG-features of 3 resized representations of each single connected component: 64×64 square and a flat 32×128 and a tall 128×32 rectangle with a cell size of 4×4 and 8×8 , plus the height, width and the height-width proportion of each connected component. The 335 clusters were identified with characters manually. Clusters representing the same glyph with only visual but not semantic differences (e.g., the result of broken type) were grouped into one cluster, while clusters representing glyphs with semantically important differences (i.e. letters of different typefaces and sizes or with or without diacritical dots) were kept separate (Fig. 2).

Paragraph segmentation and identification of marginal additions were done with vertical profiles. Row segmentation was based on horizontal profiles. All connected components could then be assigned to rows. A combination of the clustering result and the centroid position plus the top and bottom boundaries vis-à-vis the row base-line served to evaluate whether a letter was superscripted and where it was to be placed on the horizontal axis. All this semantic layout position and font information was inserted into the transcription of the letters via tags. These tags were translated into Microsoft Word styles for visualization. Subsequently, the automatic transcription was corrected manually. We estimate that the precision of the automatic transcription was higher than 99.5% (Fig. 3).

2. Holistic manuscript layout analysis for writing block detection

Originally, we had anticipated applying the above-mentioned system to manuscripts for automatic alignment and transcription. After some manual clustering, the system indeed attained a transcription precision of about 85%. This was not high enough to replace hand transcription or even to provide a searchable ‘background’ transcription for the publication of images. The main challenge consisted in the letter segmentation of the connected script. With the help of morphological transformations we made some progress in dissecting connected letters, but the process was completely manuscript and scribe dependent and too labor intensive. Transcription-glyph alignment based on synthetic ‘manuscriptization’ of the transcription was more successful, but still not precise enough for production.¹¹

In a lecture in the e-philologie lecture series at PSL Université Paris, Marcus Liwicki mentioned Ocropy.¹² Despite the statement of Thomas Breuel that Ocropy was not suited for transcription of handwritten documents, the biLSTM of Ocropy is in fact quite powerful at least for certain medieval manuscripts.¹³ Its central problem

¹¹ Some of these steps were presented as a poster in the CSMC conference in February/March 2016 by Stökl Ben Ezra.

¹² Breuel 2014.

¹³ We made our first attempts in autumn 2016. Jean-Baptiste Camps reported useful results in 2017, too.

והחטיב ועל שמעונו הרעות הוא אומר ברוך דין האמת ג' כנה בית חדש
 וקנה כלים חדשים אומר ברוך שהגיענו לזמן הזה ד' מברך על הרעה
 מעין על הטובה ועל הטובה מעין על הרעה הצועק לשעבר הררי זו תפלת שוא
 ה' כיצד היתה אשתו מעוברת ואומר יהי רצון שתלד אשתי זכר הררי זו תפילת
 שוא היה בא בדרך ושמע קול צווחות בעיר ואומר יהי רצון שלא יהו אלו
 בתוך ביתי הררי זו תפלת שוא ר' הנכנס לכרך מתפלל שתיים אחת בכניסתו
 ואחת ביציאתו בן עזאי אומר ארבע שתיים בכניסתו ושתיים ביציאתו ונותן
 הודיה לשעבר וצועק לעתיד לבוא ז' חייב לברך על הרעה כשם שהוא מברך
 על הטובה ומהר את אלהיך בכל לברך ובכל נפשך ובכל מאדך בכל לברך
 בשני יצריך ביצר טוב וביצר רע ובכל נפשך אפי' הוא נוטל את נפשך ובכל מאדך
 בכל מחונך דבר אחר בכל מאדך בכל מידה ומידה שהוא מודד לך בכל הוי מודה
 לו בכל מאד מאד ה' לא יקל אדם את ראשו כנגד שער המזרח מפני שהוא
 מחונך כנגד בית קדש הקדשים לא יכנס להר הבית במקלו ובחנעלו ור'
 וראפודתו וראק שטל דגליו לא יעשו קפודיה ורקיקה מקל וחומר מה אח ועילה
 שהיא משם כבוד אסור ליכנס בה רקיקה שהיא משום בזיון דין הוא כל חותם הברכו
 שהיו במקדש היו חן העולם חשק לקלו המינים ואחריו אין עולם אלא אחד המקינו שיהוא
 אומי חן העולם ועו העולם והתקינו שיהא אדם שואל שלום חבירו בשם שני והנה בועז
 בא מבית לחם ויאמר לקוצרים עמכם ויאמרו לו יברכך ואומי עמד גבור החיל
 ואומי אל תבוז כי זקנה עמד ואומי עת לעשות לפרו תורתך ר' נתן אומר הפרו
 תורתך עת לעשות ל' ה' ח' הסילה מסכתא

</div2>
 <div2.xml:id="S01520.1.2">
 מסכת פיאה
 פרק א' אילו דברים שאין להם שיעורי פיאה והבכורים
 והראיון וגמילות חסדים ותלמוד תורה ואילו

Fig. 3: Same page from Lowe in our OCR in Microsoft WORD output.

is the layout analysis, which suits the needs of printed documents but not the small and larger irregularities of manuscripts as well as binarization. The solution was to develop the binarization and column/writing block and line segmentation ourselves and to subsequently feed the preprocessing results into Ocropy. We should note, that we have since then moved from Ocropy to the more advanced OCR-engine Kraken by Benjamin Kiessling, because it is natively RTL (right-to-left), bidirectional and unicode enabled (but allows also for non-unicode codecs) and has a superior recognizer.¹⁴ In the frame of the Scripta-PSL project, we are in the final stages of creating an open-source web-based infrastructure that integrates Kraken and will in the future also permit deep annotation of philological (additions, deletions etc.), historical (e.g. named entities), linguistic and palaeographical nature.¹⁵ It is to this system that we will turn in the near future.

Our binarization is based on a sequence of well-known morphological transformations (Fig. 4): 1. closing with a

structuring element in the form of a disk of a size depending on resolution and script size to calculate background, usually 30 or 50 gave excellent results; 2. deducing background from image to create foreground; 3. adjusting image intensity values of the foreground; 4. Otsu binarization of the resulting image. Despite its simplicity, the results were good enough on our material.¹⁶

With regard to layout analysis, Stökl Ben Ezra's approach was to better exploit the regularity of literary manuscripts. Strangely, documents are frequently considered as two-dimensional objects (even though pages are warped) in automatic document analysis and as a collection of individual pages. However, in particular for our corpus of Hebrew manuscripts, without illuminations, it seemed absurd to deal with pages of a manuscript one by one as if each one was completely new and unrelated to the others. Even if lines can be slightly oblique or curved, or paragraphs can be oblique, or there are frequent marginal additions, the page layout of these manuscripts of literary texts tends to be highly regular in plan. Columns, too, have a relatively constant position,

¹⁴ Kiessling 2019.

¹⁵ Stokes et al. forthcoming.

¹⁶ Stökl Ben Ezra 2018.



Fig. 4: Binarization process: a) input; b) closing for background calculation; c) deducing b; d) image intensity values adjustment; e) Otsu binarization and inversion of d; f) direct Otsu binarization of a for comparison.

width and height. They can be interrupted by intermediary titles or empty space, but in principle their position on the page is quite regular.

Our ‘holistic approach’ takes into serious consideration the overlooked third dimension of manuscripts, the z-axis in addition to y and x.¹⁷ Instead of a horizontal or a vertical profile of single pages, the method consists of calculating a z-profile from the vertical superposition of the two-dimensional images, as if an X-ray was looking through the manuscript and then applying horizontal and vertical profiles.¹⁸

In a first step, a cube is calculated from all images of the manuscript.¹⁹ They are transformed to grayscale and padded in order to fit the width and height of the largest picture. The z-profile is simply the sum of all images divided by their number. If the manuscript has been photographed as single

pages, one can calculate the z-profile for even and odd pages apart. Average columns are calculated by setting all values below the median value of the average image to 0 and all those equal to or above the median value to 255. In a final step all connected components beyond the expected number of columns are deleted.

The resulting image is applied as a mask to each binarized individual page of the cube. It defines the area suspected to contain the centroid of the effective column(s) of each individual page. The median height of the remaining connected components is used for a vertical dilatation. Only connected components inside the masked area(s) are kept. All connected components inside an area become linked to each other via their centroids and holes are filled. The coordinates of the resulting connected component(s) are expected to agree to the coordinate of the major column(s) (Fig. 5).

The z-profile provides excellent information about the regularity of writing block disposition in a manuscript or printed book. Areas that are more frequently part of a writing block have higher z-profile values than those that are not. While this idea may appear extremely simple, it has proven very efficient in practice (Fig. 6).

This approach also makes it possible to calculate the variability of writing block width and height and the distance to marginal additions. Manuscripts with a very regular

¹⁷ This absence is probably not by chance. There is no English adjective that would express the z-axis of an object in the analogue way as the terms horizontal and vertical. In fact, our current usage of the terminology in document analysis presumes a sheet held vertically in the hand, not lying on a table, because vertical actually refers to the axis going up and down. In the Cartesian system the technical terms are abscissa axis and ordinate axis for x and y and applicate axis for z.

¹⁸ Code available on GitHub at <https://github.com/ephenum/z-profile_column_segmentation>.

¹⁹ They may have to be reduced in size in order to fit the memory of the CPU.

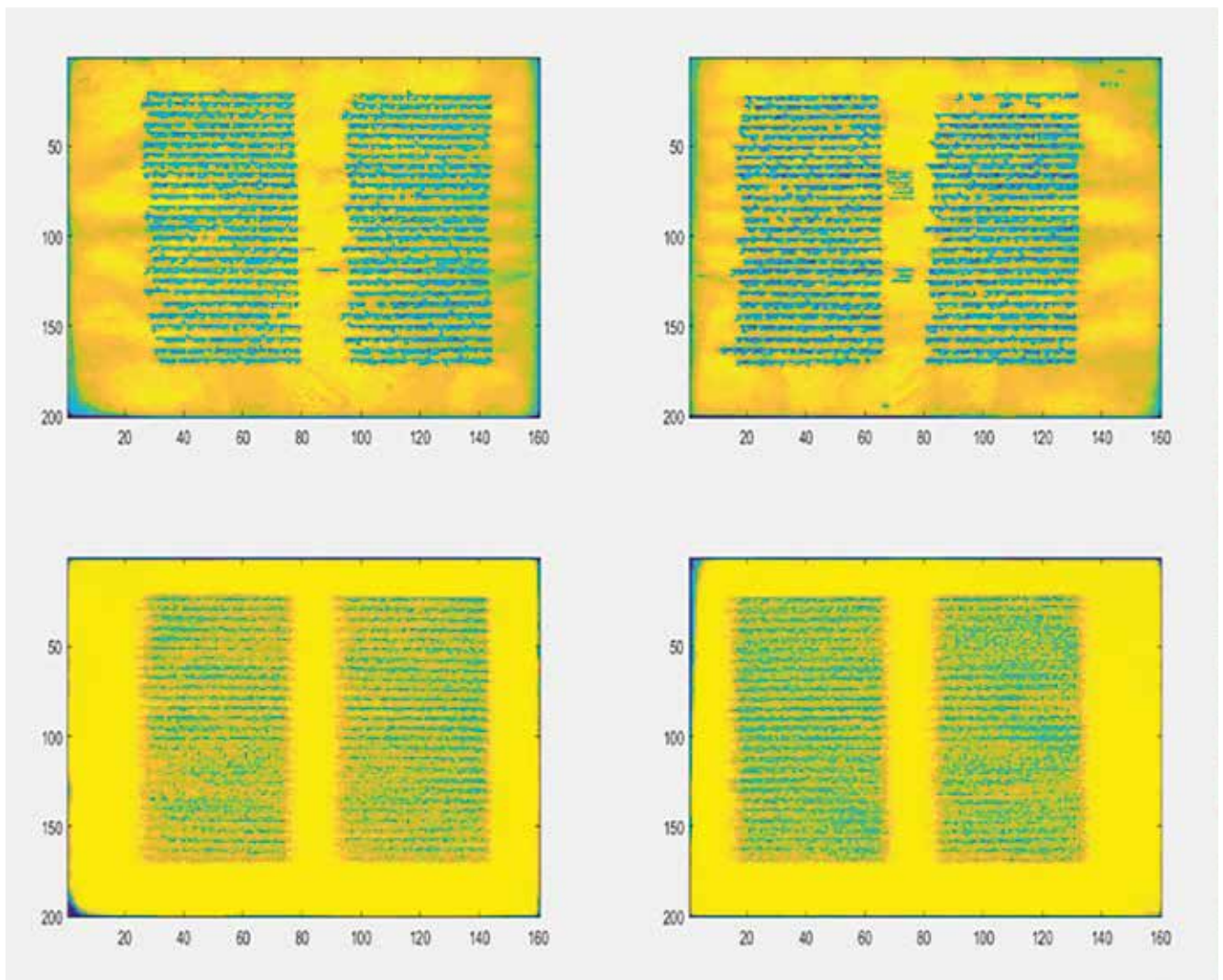


Fig. 5: The upper row shows two manuscript images of a manuscript written in two columns, where one can well discern the marginal additions particular to each page. The lower row shows the z-profile of the even and odd pages of the manuscript where only the main columns remain visible (Ms. Kaufmann A50 from the Library of the Hungarian Academy of the Sciences, Budapest).

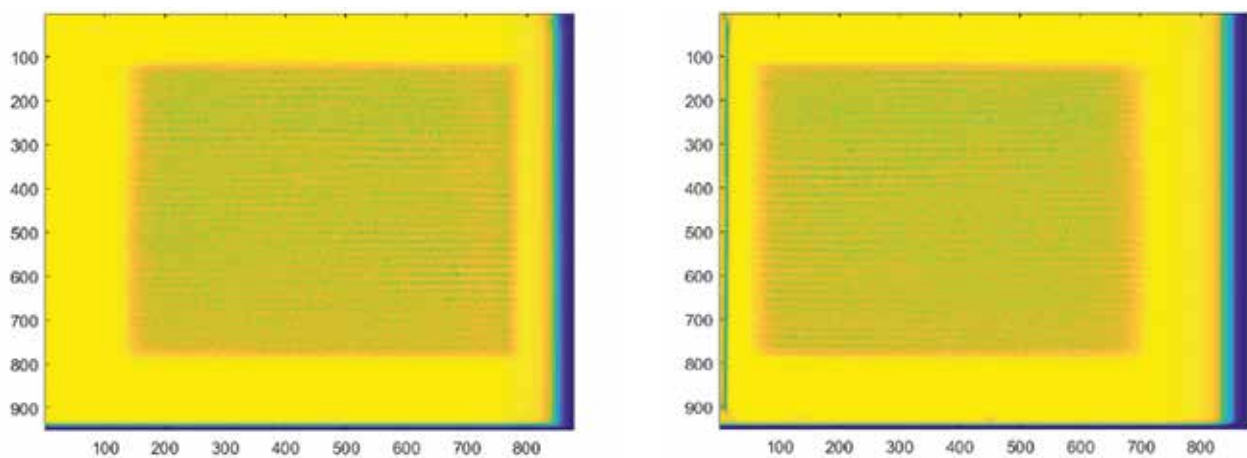


Fig. 6: Erfurt Tosefta z-profile. Even (left picture) and odd (right picture) pages.

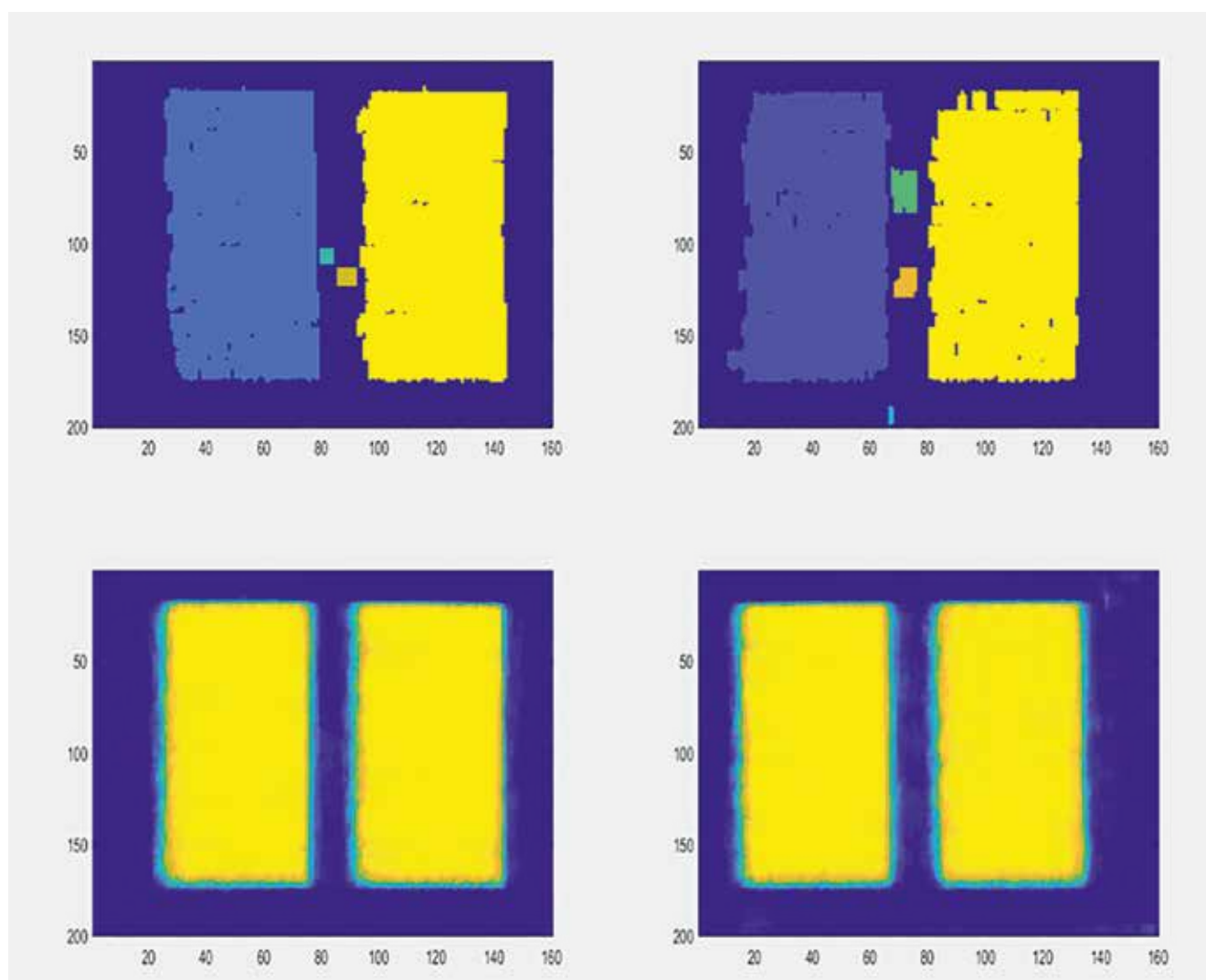


Fig. 7: In the top row an even and an odd page of the two column Kaufmann A50 manuscript after a morphological transformation. Main columns and marginal additions have been discerned with connected components into separate entities. In the bottom the z-profile after a morphological transformation.

layout have a very sharp z-profile, while manuscripts with a less regular layout have a more blurred z-profile. Marginal additions that are difficult to detect in a one-page-at-time approach, become discernible with the z-profile that distinguishes the normative basis from the addition. We should note that the system depends on a good binarization in order to distinguish between marginal additions and other dark areas, e.g. through deterioration of the manuscript or shadows. For this reason, we have started training an off-the-shelf Convolutional Neural Networks (CNNs) to distinguish between marginal areas with and without ink. However, as the main aim is the main text, this distinction is the cherry on the cake (Fig. 7).

Distinguishing between the z-profile of even and odd pages further sharpens the z-profile since many if not

most manuscripts have a mirrored layout. Calculating the distance from the z-profile for each page can subsequently help to establish different z-profiles for different parts of the manuscript, e.g. for the material in the beginning and the end of the book, or for pages that commence a new chapter, pages with illustrations or tables etc. Exploiting this information might also improve existing algorithms, probably even in the age of convolutional neural network layout analysis.

3. Line segmentation with the heartbeat-seamcarve algorithm

Most recent line-segmentation approaches strive to find a general solution for the ultimate problem of finding any line in any orientation and on any position of any document



Fig. 8: Ms. Kaufmann A50 164^a with automatic line segmentation of seemingly easy material. On the left side Transkribus. On the right side heartbeat seamcarve.

image. CNNs have achieved excellent results on such cases.²⁰ However, on the main texts of regular manuscripts standard approaches may be just as good or even superior because current CNNs have a problem with large blank spaces (*vacats*) inside a semantic line, especially, e.g. if one considers poetic manuscripts. Nevertheless, the standard approaches can be unsatisfactory even on a completely regular, quite simple manuscript as the following screenshots show (Fig. 8). The left image was produced using the state-of-the-art infrastructure Transkribus, while the image on the right uses the heartbeat seamcarve method. Both show the same random page of the test manuscript (Ms. Kaufmann A50).

The Transkribus looks clean from afar, yet, it has eight errors (lines 1, 2 (2×), 9 (2×), 10 (2×), 13), while the heartbeat seamcarve has no mistake in line recognition.²¹ Tests with the CNN algorithm resulted in significant errors that did not arise with the heartbeat seamcarve (Fig. 9).

In the Middle Ages, preparing parchment and paper for writing was an expert task. In literary manuscripts, pages very frequently have been carefully ruled before the inscription, indicating lines, columns and/or writing blocks. In literary manuscripts, the distance between these lines is mostly very regular. While the lines can be empty, end early or start in the middle of the column or be interrupted by a large *vacat*, the vertical distance is mostly as constant as the width of the columns (but not the length of each individual line). Very often, the z-profile picture reveals not only the columns but even the number of lines (see Fig. 5, above).

One of the line segmentation algorithms is seam-carving.²² The success of the seam-carving algorithm depends to a large extent on single column segmentation and on the correctness of the detection of median lines. If, however, a line ends early and the subsequent line starts late, the simple median line approach will consider the second a continuation of the first which will result in the two lines being seam-carved as a single line.

²⁰ As shown, e.g. in Diem et al. 2017 and Kiessling et al. (submitted).

²¹ It cuts the head of two *lameds* in line 9 and 13 but transkribus seems to cut all ascenders and descenders (at least in its visualization).

²² Saabni and El-Sana 2011.

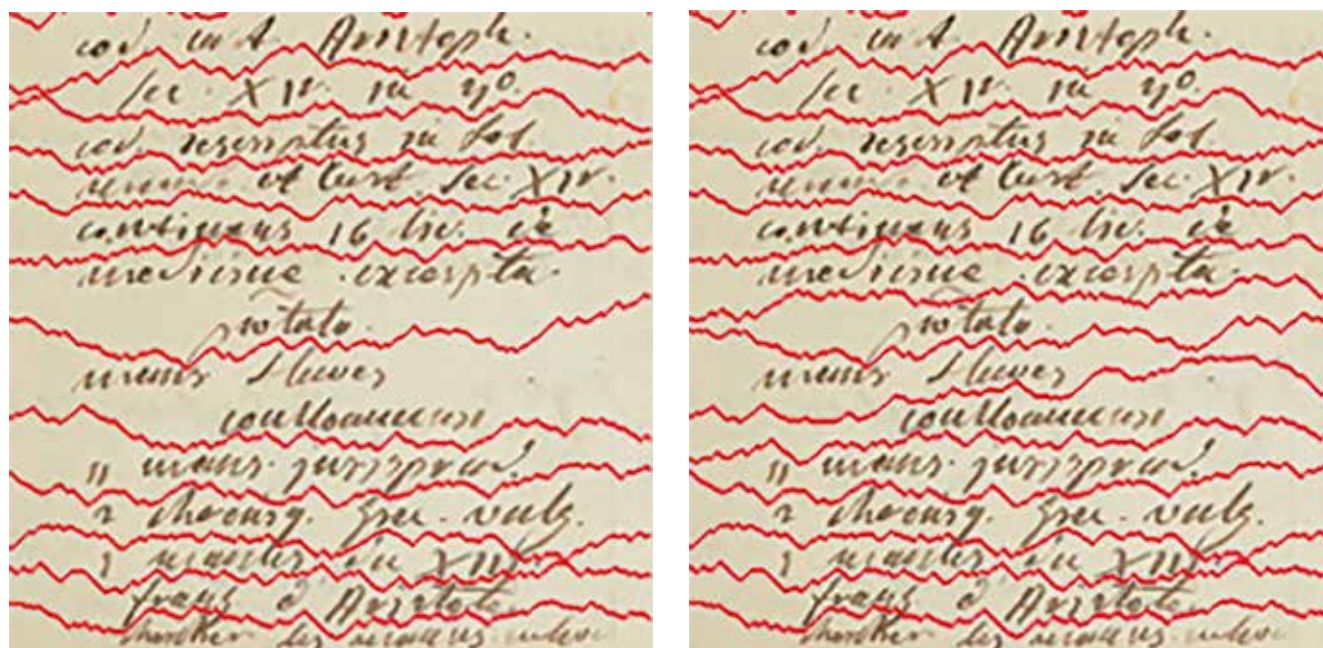


Fig. 9: Seamcarve on a sample from the Washington²⁵ dataset. Left without, right with heartbeat.

Mathias Seuret, Marcus Liwicki and Stökl Ben Ezra started to improve the seam-carving algorithm by the assumption of regularity in the analysis of the manuscript.²³ Based on a Fourier transformation of the horizontal projection of each of n slices of a writing block, the procedure calculates the median line length. Wherever the line is too short or empty and therefore the horizontal profile misses a peak, the algorithm adds one or several artificial peaks according to the regular line distance with regard to the lines above and below. The algorithm is now implemented in the DIVAServices.²⁴

4. Manuscript transcription and transcription-glyph alignment

Once a pipeline for the production of relatively clean manuscript-line-image and transcription was established, we were able to train models with Ocropy showing useful results. A preliminary step was data augmentation. We used the well-known methods of salt and pepper as well as shearing of the manuscript line image in different dosages, angles and combinations to multiply input pairs by a factor of nine.

A challenging stage was the production of transcription text lines that correspond to the visible signs in the main text block. All marginal or interlinear additions had to be deleted. On the other hand, all deletions of the main text by simple

strikethrough had to be kept. Numbers and paratext such as *eschatocols* of chapters or treatises had to be kept. Letters functioning as simple line fillers without importance for the linguistic text, a frequent practice in Hebrew manuscripts, had to be kept, too. Abbreviations had to remain unresolved. Ligatures had to be represented by special marks. Our transcription markup distinguished between the various forms of addition and deletion and it was mainly a question of the order of transformation steps.

For the preparation of the most complex manuscript, we used Microsoft Word with numerous styles to emulate XML tagging because XML editors like Oxygen are still difficult to manage with RTL scripts whose writing direction counters that of the tags. Hayim Lapin wrote a series of conversion scripts to convert the Word documents (docx) to XML/TEI that used Visual Basic to prepare the transcription for transformation to raw XML, and subsequent conversion to a TEI schema-conformant document using XSLT. The forthcoming eScriptorium platform will permit us to combine automatic layout analysis and HTR with deep annotation.

So far, we have applied our pipeline to the following Hebrew manuscripts:

- K: Mishnah: Ms. Kaufmann A50 in the Library of the Hungarian Academy of the Sciences, Budapest. Written in Italian script from the eleventh or twelfth century. 256 folios.

²³ Seuret, Stökl Ben Ezra, and Liwicki 2017.

²⁴ Würsch, Ingold, and Liwicki 2017.

²⁵ Fischer et al. 2012.

- C: Mishnah: Ms. Add. 470.1, Cambridge University Library. Written in Byzantine script from the fifteenth century. 250 folios.
- M: Mishnah: Cod. Ebr. 95 of the Bayerische Staatsbibliothek, Munich (on the part of the Mishnah only because the resolution is very low for the tiny script of the Talmud itself). The manuscript was written in 1342 probably in France. 576 folios. Semi-manual manuscript layout segmentation of the complex Talmudic layout was very kindly provided with by the Larex team around Christian Reul.
- V: Tosefta: Cod. Hebr. 20, Austrian National Library, Vienna. Written around the fourteenth century in square Sephardic script. 327 folios.
- L: Tosefta: Add. 27296, British Library, London. Written in fifteenth century Sephardic script. 73 folios.

This makes altogether ca. 1,400 pages. For most manuscripts, we could achieve a character error rate (CER) <5%, sometimes <3%. We tried different sizes for the hidden layer, different learning rates, different sizes of training data. We also mixed training data from different manuscripts with encouraging results. 5 columns (171 lines, 200 neurons) achieved a CER <10% for the Vienna manuscript, while 19 columns (645 lines) sufficed for 2.1% CER (43k iterations). Due to limited manpower and calculation power (Ocropy runs on CPUs only and demands the transformation of all RTL scripts into LTR), we did not apply all tests on all materials. The main aim was not to improve the LSTM but to arrive at exploitable results.²⁶ Of course, in a manuscript of 1,000,000 characters, 3% CER still means 30,000 errors to spot and to correct. However, where a vulgate text is available or one manuscript of a text is already transcribed, we can automatically align both versions with CollateX or with Shmidman-Koppel-Porat algorithm.²⁷

5. Outlook to the Future

Since June respectively September 2018, two new projects have started around eRabbinica, both with the National Library of Israel and their manuscript portal Ktiv.²⁸ In Tikkoun Sofrim ('scribal error correction') with Haifa University

we have worked on correction of automatic transcriptions of post-classical Midrashim of the Tanhuma-Yelamdenu genre via crowdsourcing.²⁹ We have so far transcribed four manuscripts with CER's of 2.8%, 2.9%, 6.9% and 8.9%.³⁰ The manuscript with 8.9% CER has been the first to be submitted to the crowdsourcing process and we have been able to reconstruct a complete text with the help of CollateX and a majority vote on the word level.³¹ The plan is to link coordinates for words, and where possible glyphs via IIIF to the manuscript images and to integrate them with the help of the Mirador viewer at Ktiv. The National Library will serve as repository for long-term preservation.

In the second project, Sofer Mahir (*tachygraph*, or 'rapid [i.e., skilled] scribe') with the University of Maryland and Dicta³² we collaborate on the creation of a pipeline to produce open source manuscript transcriptions of all major manuscripts of the principal tannaitic compositions: approximately twenty substantial manuscripts with about 6,000 pages. In collaboration with Dicta, the texts will be automatically analyzed linguistically. We hope to be able to integrate the linguistic analysis directly into the transcription pipeline to further reduce the error ratio. In the LAKME project, we have already annotated 25,000 words lexically and morphologically and created the corresponding lexicon in French, English and German in order to apply a neural network architecture developed by Dicta on all of our transcriptions.³³ In a related project, Lapin is creating a database of shared text (identified by string matching) among the corpora that will be part of the infrastructure of future editions.

The resulting text will be presented according to the developing Distributed Text Service (DTS) API (Application Programming Interface).³⁴ An extension of the Canonical Text Service (CTS) specification, first developed for the

²⁶ Results will be published on the eRabbinica.org website and on GitHub after full transition from Ocropy to Kraken.

²⁷ Shmidman, Koppel and Porat 2018.

²⁸ <<http://web.nli.org.il/sites/nlis/en/manuscript>>.

²⁹ Wecker et al. 2019. Tikkoun Sofrim website: <<https://tikkoun-sofrim.firebaseio.com/>>. Tikkoun Sofrim GitHub: <<https://github.com/drone/tikkoun>>.

³⁰ Kuflik et al. 2019.

³¹ Dekker et al. 2015, Decker and Midell 2011.

³² <<http://dicta.org.il/>>.

³³ Stökl Ben Ezra et al. 2018.

³⁴ <<https://distributed-text-services.github.io/specifications/>> (accessed 26 May 2019).

Homer Multitext Project,³⁵ the DTS specification provides a common, predictable protocol for sharing texts at various levels of granularity, as well as information about texts and the collections they appear in. A number of significant projects have begun to use the CTS/DTS specification, among them the Perseus, Kitab (Knowledge, Information Technology, and the Arabic Book), and ARTFL Encyclopédie.³⁶ This approach has internal advantages for the project (for instance, it allows us to build applications around texts without constructing a purpose-built querying system). As the adoption by multiple projects working in diverse languages and types of texts suggests, the CTS/DTS model provides a standardized way of sharing information between projects, so that (as we have learned all too well in our own work) each project does not have to reproduce the work of every other.

The amount of work put into the pipeline, the preparation of the training data and the correction is very substantial. Our preliminary measurements agree with results of previous teams that it is more time-consuming to correct a text with a CER > 10% than to transcribe it from scratch. For the London Tosefta, e.g. the correction of a page created with the first model, trained with only 15 pages (376 lines), took about 30 minutes per page, about the same time as a transcription from scratch. However, after the correction of 20 pages, we trained a new model whose correction only takes ca. 15 minutes per page on the average and we hope to further reduce this amount of time with the next model. The amount of human labor involved in extracting the data from the current interface and launching a retraining is about 2 hours. In the very near future, however, this procedure will just demand a series of button clicks inside the eScriptorium platform. While for short manuscripts the effort to automatize the transcription and then correct it manually may be greater than the time needed to prepare a completely manual transcription, the immediate humanist gain is the direct link between image and transcription. Doing this by hand for manuscripts would usually involve much more human labor. Secondly, the automatization of the transcription will improve with the quantity of ground truth acquired and with future improvement of algorithms for layout segmentation and transcription as well as language modelling. Ergonomics for ground-truth-creation and for post-correction tools will be

further improved. We will attain a stage soon, where we can transcribe most of the historical manuscripts automatically with a grade of precision sufficiently high for human reading and machine exploitation. For Medieval Hebrew, we are on the brink of doing it.

REFERENCES

- Ben-Eliyahu, Eyal, Yehudah Cohn, and Fergus Millar (2013), *Handbook of Jewish Literature from Late Antiquity: 135–700 CE* (Oxford: Oxford University Press).
- Bizzoni, Yuri, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane (2014), ‘The Making of Ancient Greek WordNet’, in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (European Language Resources Association), 1140–1147.
- Breuel, Thomas (2014), *Ocropy: Python-based tools for document analysis and OCR*. <<https://github.com/tmbdev/ocropy>> (accessed 6 November 2019).
- Camps, Jean-Baptiste (2017), ‘Homemade Manuscript OCR (1): Ocropy’. <<https://graal.hypotheses.org/786>> (accessed 6 November 2019).
- Dekker, Ronald H., and Gregor Middell (2011), ‘Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements’, *Supporting Digital Humanities*, University of Copenhagen, Denmark, 17–18 November 2011. <<https://doc.anet.be/docman/docman.phtml?file=.irua.4ec3d9.935894d4.pdf>> (accessed 6 November 2019).
- , Dirk van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert (2015), ‘Computer-supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project’, *Literary and Linguistic Computing*, 30: 452–470.
- Diem, Markus, Florian Kleber, Stefan Fiel, Tobias Grüning, and Basilios Gatos (2017), ‘Cbad: ICDAR 2017 Competition on Baseline Detection’, *International Conference of Document Analysis and Recognition*, 1355–1360. doi: 10.1109/ICDAR.2017.222.
- Fischer, Andreas, Andreas Keller, Volkmar Frinken, and Horst Bunke (2012), ‘Lexicon-Free Handwritten Word Spotting Using Character HMMs’, *Pattern Recognition Letters*, 33.7: 934–942.

³⁵ <<https://www.homermultitext.org/>> (accessed 26 May 2019).

³⁶ Perseus: <<http://www.perseus.tufts.edu/hopper>>; Kitab: <<http://kitab-project.org>>; Encyclopédie: <http://encyclopedie.uchicago.edu>.

- Kiessling, Benjamin, Daniel Stökl Ben-Ezra, and Matthew T. Miller (2019), 'BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts', *Historical Document Imaging and Processing, International Conference of Document Analysis and Recognition*. *arXiv.org*: <arXiv:1907.04041> (accessed 6 November 2019).
- (2019), 'Kraken – an Universal Text Recognizer for the Humanities', *DH2019*, Utrecht University, 8–12 July 2019. <<https://dev.clariah.nl/files/dh2019/boa/0673.html>> (accessed 6 November 2019).
- Kuflik, Tsvi, Moshe Lavee, Daniel Stökl Ben Ezra, Avigail Ohali, Vered Raziel-Kretzmer, Uri Schor, Alan Wecker, Elena Lolli, and Signoret, Pauline (2019), 'Tikkoun Sofrim – Combining HTR and Crowdsourcing for Automated Transcription of Hebrew Medieval Manuscripts', *DH2019*, Utrecht University, 8–12 July 2019. <<https://dev.clariah.nl/files/dh2019/boa/0568.html>> (accessed 6 November 2019).
- Lowe, William H. (1883), *The Mishnah on Which the Palestinian Talmud Rests*, (Cambridge: Cambridge University Press).
- McGillivray, Barbara (2013), *Methods in Latin Computational Linguistics* (Leiden: Brill).
- Saabni, Raid, and Jihad El-Sana (2011), 'Language-Independent Text Lines Extraction Using Seam Carving', *International Conference on Document Analysis and Recognition*, 563–568. doi: 10.1109/ICDAR.2011.119.
- Seuret, Mathias, Daniel Stökl Ben Ezra, and Marcus Liwicki (2017), 'Robust Heartbeat-based Line Segmentation Methods for Regular Texts and Paratextual Elements', *Historical Document Imaging and Processing, International Conference of Document Analysis and Recognition*, 71–76. doi: 10.1145/3151509.3151521 (accessed 6 November 2019).
- , Daniel Stökl Ben Ezra, and Marcus Liwicki (2018), 'Heartbeat Seamcarve Line Segmentation'. <https://github.com/ephenum/heartbeat_seamcarve_line_segmentation> (accessed 6 November 2019).
- Shmidman, Avi, Moshe Koppel, and Ely Porat (2018), 'Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus', *Journal of Data Mining & Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities*. *arXiv.org*: <arXiv:1602.08715v2>.
- Stemberger, Günter (2011), *Einleitung in Talmud und Midrasch* (Munich: C.H. Beck).
- Stokes, Peter Anthony, Daniel Stökl Ben Ezra, Benjamin Kiessling, and Robin Tissot (2019, forthcoming), 'EScripta: A New Digital Platform for the Study of Historical Texts and Writing', *DH2019*, Utrecht University, 8–12 July 2019.
- Stökl Ben Ezra, Daniel (2016), *Why Should Philologists Learn Computer Vision*, CSMC Conference Poster, February/March 2016. <https://www.academia.edu/22887575/Why_should_philologists_learn_computer_vision> (accessed 6 November 2019).
- (2018), 'Simple Binarization for Manuscript Images'. <<https://github.com/ephenum/binarization>> (accessed 6 November 2019).
- , Avigail Ohali, Emmanuelle Main, Hayim Lapin, Avi Shmidman, Shaltiel Shmidman, Meni Adler, Yael Netzer, and Thierry Poibeau (2018), 'The Creation of a Lemmatized and Morphologically Annotated Open Source Corpus for Rabbinic Hebrew', in Jean-Baptiste Camps, Thierry Poibeau, and Daniel Stökl Ben Ezra, *Creating Richly Annotated Linguistic Corpora for Languages with Few Linguistic Resources*, Multiple Paper Panels, Conference of the European Association for Digital Humanities (EADH), Oral presentation, Galway, 2018. <<https://eadh2018.exordo.com/programme/session/41>> (accessed 6 November 2019).
- (2019), 'z-Profile Column Segmentation'. <https://github.com/ephenum/z-profile_column_segmentation> (accessed 6 November 2019).
- Wecker, Alan, Daniel Stökl Ben Ezra, Vered Raziel-Kretzmer, Uri Schor, Tsvi Kuflik, Avigail Ohali, Dror Elovits, Moshe Lavee, and Pauline Stevenson (2019), 'Tikkoun Sofrim: A WebApp for Personalization and Adaptation of Crowdsourcing Transcriptions', *27th Conference on User Modeling, Adaptation and Personalization UMAP'19*, Adjunct Publication (Larnaca. New York: ACM Press). doi: 10.1145/3314183.3324972.
- Würsch, Marcel, Rolf Ingold, and Marcus Liwicki (2017), 'DIVAServices—A RESTful Web Service for Document Image Analysis Methods', *Digital Scholarship in the Humanities*, 32.suppl.1: i150–i156. doi: 10.1093/lle/fqw051.

Written Artefacts as Cultural Heritage

Ed. by Michael Friedrich and Doreen Schröter

Written Artefacts as Cultural Heritage was established in 2020. The series is dedicated to the double role of written artefacts as representations and generators of humankind's cultural heritage. Its thematic scope embraces aspects of preservation, the identity-defining role of artefacts as well as ethical questions.


The mix of practical guides, colloquium papers and project reports is specifically intended for staff at libraries and archives, curators at museums and art galleries, and

scholars working in the fields of manuscript cultures and heritage studies.


Every volume of *Written Artefacts as Cultural Heritage* has been peer-reviewed and is openly accessible. There is an online and a printed version..

If you wish to receive a copy or to present your research, please contact the editorial office:

<https://www.csmc.uni-hamburg.de/publications/cultural-heritage.html>




Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



CENTRE FOR THE
STUDY OF
MANUSCRIPT
CULTURES

**WRITTEN ARTEFACTS
AS CULTURAL HERITAGE**

CENTRE FOR THE STUDY
OF MANUSCRIPT CULTURES
UNIVERSITÄT HAMBURG



**STEP BY STEP GUIDE TO MANUSCRIPT SURFACE
CLEANING AND MAKING E-FLUTE PHASE BOXES
FOR MANUSCRIPTS**

BIDUR BHATTARAI | MICHAELLE BIDDLE

№ 1



manuscript cultures (mc)

Editors: Michael Friedrich and Jörg B. Quenzer
Editorial office: Irina Wandrey

CSMC's academic journal was established as newsletter of the research unit 'Manuscript Cultures in Asia and Africa' in 2008 and transformed into a scholarly journal with the appearance of volume 4 in 2011. *manuscript cultures* publishes exhibition catalogues and articles contributing to the study of written artefacts. This field of study embraces disciplines such as art history, codicology, epigraphy, history, material analysis, palaeography and philology, informatics and multispectral imaging.

manuscript cultures encourages comparative approaches, without regional, linguistic, temporal or other limitations

on the objects studied; it contributes to a larger historical and systematic survey of the role of written artefacts in ancient and modern cultures, and in so doing provides a new foundation for ongoing discussions in cultural studies.

Every volume of *manuscript cultures* has been peer-reviewed and is openly accessible:

<https://www.csmc.uni-hamburg.de/publications/mc.html>

If you wish to receive a copy or to present your research in our journal, please contact the editorial office:
irina.wandrey@uni-hamburg.de

manuscript cultures 14



manuscript cultures 13



manuscript cultures 11



manuscript cultures 10



manuscript cultures 9



manuscript cultures 8



manuscript cultures 7



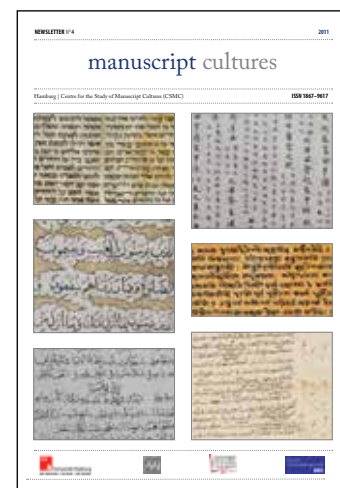
manuscript cultures 6



manuscript cultures 5



manuscript cultures 4



Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

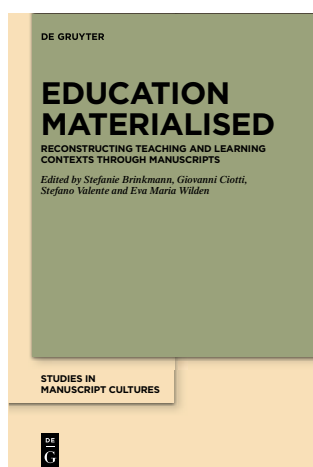
From volume 4 onwards all volumes are available as open access books on the De Gruyter website:

<https://www.degruyter.com/view/serial/43546>

<https://www.csmc.uni-hamburg.de/>



Forthcoming



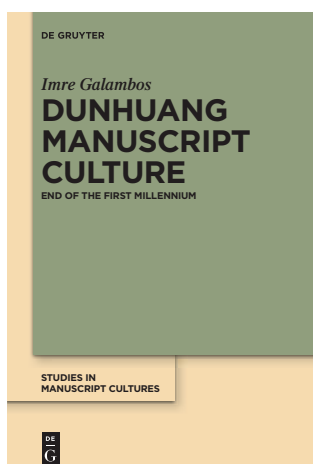
23 – Education Materialised: Reconstructing Teaching and Learning Contexts through Manuscripts, edited by Stefanie Brinkmann, Giovanni Ciotti, Stefano Valente and Eva Maria Wilden

Manuscripts have played a crucial role in the educational practices of virtually all cultures that have a history of using them. As learning and teaching tools, manuscripts become primary witnesses for reconstructing and studying didactic and research activities and methodologies from elementary levels to the most advanced.

The present volume investigates the relation between manuscripts and educational practices focusing on four particular research topics: educational settings: teachers, students and their manuscripts; organising knowledge: syllabi; exegetical practices: annotations; modifying tradition: adaptations.

The volume offers a number of case studies stretching across geophysical boundaries from Western Europe to South-East Asia, with a time span ranging from the second millennium BCE to the twentieth century CE.

New release



22 – Dunhuang Manuscript Culture: End of the First Millennium, by Imre Galambos

Dunhuang Manuscript Culture explores the world of Chinese manuscripts from ninth–tenth century Dunhuang, an oasis city along the network of pre-modern routes known today collectively as the Silk Roads. The manuscripts have been discovered in 1900 in a sealed-off side-chamber of a Buddhist cave temple, where they had lain undisturbed for almost nine hundred years. The discovery comprised tens of thousands of texts, written in over twenty different languages and scripts, including Chinese, Tibetan, Old Uighur, Khotanese, Sogdian and Sanskrit. This study centres around four groups of manuscripts from the mid-ninth to the late tenth centuries, a period when the region was an independent kingdom ruled by local families. The central argument is that the manuscripts attest to the unique cultural diversity of the region during this period, exhibiting – alongside obvious Chinese elements – the heavy influence of Central Asian cultures. As a result, it was much less ‘Chinese’ than commonly portrayed in modern scholarship. The book makes a contribution to the study of cultural and linguistic interaction along the Silk Roads.

Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

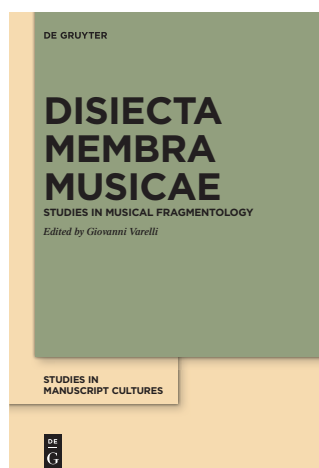
From volume 4 onwards all volumes are available as open access books on the De Gruyter website:

<https://www.degruyter.com/view/serial/43546>

<https://www.csmc.uni-hamburg.de/>



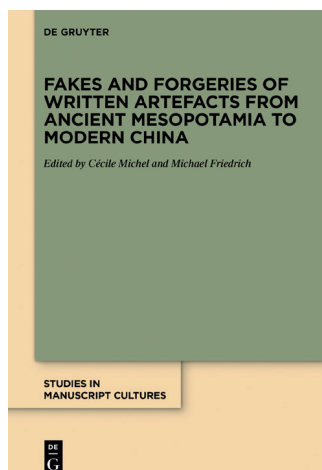
New release



21 – *Disiecta Membra Musicae: Studies in Musical Fragmentology*, edited by Giovanni Varelli

Although fragments from music manuscripts have occupied a place of considerable importance since the very early days of modern musicology, a collective, up-to-date, and comprehensive discussion of the various techniques and approaches for their study was lacking. On-line resources have also become increasingly crucial for the identification, study, and textual/musical reconstruction of fragmentary sources. *Disiecta Membra Musicae. Studies in Musical Fragmentology* aims at reviewing the state of the art in the study of medieval music fragments in Europe, the variety of methodologies for studying the repertory and its transmission, musical palaeography, codicology, liturgy, historical and cultural contexts, etc. This collection of essays provides an opportunity to reflect also on broader issues, such as the role of fragments in last century's musicology, how fragmentary material shaped our conception of the written transmission of early European music, and how new fragments are being discovered in the digital age. Known fragments and new technology, new discoveries and traditional methodology alternate in this collection of essays, whose topics range from plainchant to *ars nova* and fifteenth- to sixteenth-century polyphony.

New release



20 – *Fakes and Forgeries of Written Artefacts from Ancient*

Mesopotamia to Modern China, edited by Cécile Michel and Michael Friedrich

Fakes and forgeries are objects of fascination. This volume contains a series of thirteen articles devoted to fakes and forgeries of written artefacts from the beginnings of writing in Mesopotamia to modern China. The studies emphasise the subtle distinctions conveyed by an established vocabulary relating to the reproduction of ancient artefacts and production of artefacts claiming to be ancient: from copies, replicas and imitations to fakes and forgeries. Fakes are often a response to a demand from the public or scholarly milieu, or even both. The motives behind their production may be economic, political, religious or personal – aspiring to fame or simply playing a joke. Fakes may be revealed by combining the study of their contents, codicological, epigraphic and palaeographic analyses, and scientific investigations. However, certain famous unsolved cases still continue to defy technology today, no matter how advanced it is. Nowadays, one can find fakes in museums and private collections alike; they abound on the antique market, mixed with real artefacts that have often been looted. The scientific community's attitude to such objects calls for ethical reflection.

ISSN 1867–9617

© 2020

Centre for the Study of Manuscript Cultures

Universität Hamburg

Warburgstraße 26

D-20354 Hamburg

www.csmc.uni-hamburg.de