

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von *Transkribus* bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken

von Ina Serif

Zusammenfassung: *In diesem Beitrag wird ‚Transkribus‘ vorgestellt, ein Programm zur Transkription handschriftlicher und gedruckter Dokumente. Die einzelnen Schritte vom Digitalisat zum manuell oder (semi-)automatisch layout- und texterkannten Dokument, die die Durchsuchbarkeit größerer Quellenbestände ermöglichen, werden erklärt. Vorteile des Programms und mögliche Schwierigkeiten, weiterführende Hinweise und Perspektiven für die eigene Forschungsarbeit und Anwendungsmöglichkeiten werden aufgezeigt.*

Abstract: *The paper provides an introduction to ‘Transkribus’, a software for the transcription of handwritten and printed documents. It explains the single steps from a digitized image to a document with manually or (semi)automated recognized layout and text, enabling the searchability of large bodies of sources. Advantages of the program and possible difficulties are discussed, recommendations on further uses and perspectives for the own research and possible applications are shown.*

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Einleitung

Transkribus, eine Plattform und ein Programm zur Erkennung von Handschriften, Bildern und Strukturen, „hat sich zum Ziel gesetzt, all jene zu unterstützen, die mit der Transkription historischer, gedruckter oder handschriftlicher Dokumente befasst sind.“¹ Die großen Zielgruppen, die auf der Startseite angesprochen werden, sind Wissenschaftler*innen, Freiwillige und Gedächtnisinstitutionen, die mit Textquellen aus verschiedensten Epochen und Sprachräumen arbeiten, wobei die Bedürfnisse nicht nur zwischen den unterschiedlichen Gruppen, sondern auch innerhalb der einzelnen Communities zum Teil stark auseinanderklaffen. Dass *Transkribus* sowohl im Kleinen als auch im Großen verschiedenste Anwendungsmöglichkeiten bei der Arbeit mit Schriftquellen hat, die für unterschiedliche Forschungsfragen mit zum Teil recht beachtlichem (Zeit-)Gewinn genutzt werden können, soll in diesem Beitrag an vormodernen Quellen aufgezeigt werden.² Die häufig anzutreffende Skepsis gegenüber *den* Digital Humanities, die nach großen Versprechungen nun endlich Ergebnisse liefern müssten,³ die mit zu einer pauschal ablehnenden Haltung gegenüber digitalen Werkzeugen mit beiträgt, ist ein weiterer Grund für die Abfassung dieses Textes. Und nicht zuletzt hat der große Zuwachs an einfach

¹ *Transkribus*-Wiki, <https://transkribus.eu/wikiDe/index.php?title=Hauptseite&oldid=446>. Hier gibt sich auch einen kurzen Überblick zum Programm. *Transkribus* ist hervorgegangen aus dem *READ-Projekt* (siehe <https://read.transkribus.eu/>), EU-gefördert im Rahmen von *Horizon 2020* (siehe <https://ec.europa.eu/programmes/horizon2020/en>). Eine knappe Übersicht über Struktur und Technik findet sich in Philipp Kale [u.a.], *Transkribus – A service platform for transcription, recognition and retrieval of historical documents*, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), DOI: [10.1109/ICDAR.2017.307](https://doi.org/10.1109/ICDAR.2017.307) und in Günter Mühlberger und Tamara Terbul, *Handschriftenerkennung für historische Schriften. Die Transkribus Plattform*, in: b.i.t. online 21 (2018), S. 218–222, <https://www.b-i-t-online.de/heft/2018-03/fachbeitrag-muehlberger.pdf>. Ich nutze *Transkribus* seit mehreren Jahren in verschiedenen Projekten; die vorgestellten Funktionen entsprechen zum größten Teil meiner eigenen Praxis, die Vernachlässigung von Funktionen resultiert vornehmlich aus dem Nichtgebrauch bzw. schlichter Unkenntnis – es gibt noch viel zu entdecken. Für alle Fragen, die ich während des Schreibens hatte, stand Tobias Hodel (Universität Zürich/Staatsarchiv Zürich, *READ* Projektmitglied) stets mit einer Auskunft bereit – herzlichen Dank! Dank geht auch an die Leser*innen früherer Versionen dieses Tutorials: Eric Decker, Pia Eckhart, Eva Ferro, Anna Reimann und Michael Schonhardt.

² Die größten Kinderkrankheiten des Programms sind mittlerweile überwunden, regelmäßige Updates und die implementierte Möglichkeit der direkten Rückmeldung von Bugs verbessern die Anwendung kontinuierlich. Dennoch treten immer wieder kleinere Probleme bzw. Serviceunterbrechungen auf, die meist durch ein wenig Warten und erneutes Probieren gelöst werden können. Sollte ein anhaltendes Problem bestehen, kann der Support unter email@transkribus.eu kontaktiert werden.

³ Auf dem Historikertag 2018 widmete sich das Panel „Digital Humanities in der Analyse gespaltener Gesellschaften. Beispiele aus der Praxis“ dieser Problematik. Vgl. dazu den Tagungsbericht von Martin Schmitt, *Tagungsbericht: HT 2018: Digital Humanities in der Analyse gespaltener Gesellschaften. Beispiele aus der Praxis*, 25.09.2018 – 28.09.2018 Münster, in: *H-Soz-Kult*, 07.12.2018, www.hsozkult.de/conferencereport/id/tagungsberichte-8009.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



verfügbaren Techniken und Programmen für Geisteswissenschaftler*innen in den letzten Jahren auch zu einem Übersichtsproblem geführt: Bei so viel Auswahl ist aller Anfang schwer.

Die hier vorgestellten Anwendungen sollen einen ersten Zugang zu dem Programm bieten und aufzeigen, wie *Transkribus* als Arbeits- und als Analysewerkzeug für die eigene Forschung, von Einrichtungen wie Bibliotheken und Archiven ebenso wie in der Arbeit mit Studierenden oder in Citizen-Science-Projekten gewinnbringend und erkenntnisfördernd eingesetzt werden kann. *Transkribus* bietet neben einer Transkriptionsoberfläche unter anderem auch die Möglichkeit zur Anwendung von Texterkennung auf historische Dokumente, die in einem ersten Schritt zur Durchsuchbarkeit auch größerer Quellenbestände genutzt werden kann. Auch kollaboratives Arbeiten – ortsunabhängig und plattformübergreifend – an gemeinsam genutztem Material und der Export der transkribierten/annotierten Dokumente in verschiedene Formate zur Weiternutzung ist möglich.

Obwohl bereits diverse Anleitungen und Handbücher zur Nutzung von *Transkribus* von Projektseite aus existieren, ist die Bedienung mäßig intuitiv und schreckt, nach der ersten Verzückung über das Wolpertingerwappentier, viele Nutzer*innen ab.⁴ Im Folgenden sollen daher Bilder und Videos die ersten Schritte erleichtern und Berührungsängste überwinden. (Sollte zukünftig in die Gestaltung der Oberfläche investiert werden, dürfte der Großteil dieses Tutorials obsolet sein – verlieren Sie also keine Zeit!)

⁴ Z.B. Jan Horstmann, *Transkribus*, in: *forText. Literatur digital erforschen*, 15. Oktober 2018, <https://fortext.net/tools/tools/transkribus>; Martin Prell, *Erfahrungs- und Ergebnisbericht zum Transkribusprojekt. Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung*, Mai 2018, online verfügbar unter http://www.histinst.uni-jena.de/histinstmedia/Bereiche/Lehrstuhl+Geschlechtergeschichte/Dokumente/Downloads/Transkribusbericht_2018_06_02.pdf; Stefan Karcher: *Transkribus – The Best Idea to Procrastinate I’ve Ever Had*, in: *INFODITEX*, 20. Juli 2018, <https://infoditex.hypotheses.org/192>. Zum Einsatz von *Transkribus* in der Lehre siehe Tobias Hodel, *IMC Leeds Paper: Sending 15th century missives through algorithms. Testing and evaluating HTR with 2200 documents*, in: *Schrift im Kloster. Ein Blog zum Kloster Königfelden, klösterlicher Verwaltung im Mittelalter und digitaler Geschichtswissenschaft*, 11. Juli 2017, <https://so-lascriptum.wordpress.com/2017/07/11/imc-leeds-paper-sending-15th-century-missives-through-algorithms-testing-and-evaluating-htr-with-2200-documents/>. Ein Webinar ist hier zugänglich: https://idig-bio.adobeconnect.com/_a1130716096/p65yarolyd7.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



1. Registrierung und Installation

Um *Transkribus* nutzen zu können, ist zuerst eine Registrierung unter transkribus.eu notwendig.⁵ Daraufhin kann das kostenlose (nicht *open source*) Programm⁶ für Windows, IOS oder Linux heruntergeladen werden. Eine ausführliche Installationsanleitung wird vom Entwicklerteam auf Deutsch⁷ und Englisch⁸ bereitgestellt, ein Video zeigt die wichtigsten Schritte von der Registrierung über den Download bis zum ersten Öffnen des Programm, bei dem ein Login mit den Benutzerdaten erfolgen muss:

<https://www.youtube.com/embed/8Ei0a7WIITl?start=29&end=60>.

Probleme beim Programmstart sind meist auf die installierte Java-Version bzw. das Environment zurückzuführen; sollte hier ein Update⁹ keine Besserung bringen, kann man sich an email@transkribus.eu wenden.

Neben der Desktopversion gibt es mittlerweile auch einen Webclient. Nach dem Login können dort zum einen die eigenen Dokumente ohne Installation des Programms direkt im Browser bearbeitet werden (allerdings nicht mit allen Funktionen wie Layout- und Texterkennung). Zum anderen kann auch an Projekten mitgearbeitet werden, die für Kollaborationen geöffnet sind; auf die Kollaborationsmöglichkeiten im Allgemeinen wird an anderer Stelle noch eingegangen.

⁵ Zu den Nutzungsbedingungen siehe folgendes Dokument: https://transkribus.eu/Transkribus/docs/Transkribus-TermsOfUse_v04-2016.pdf. Zur Nutzung aller Funktionen von *Transkribus* müssen Dokumente auf die projekt-eigenen Server hochgeladen werden; die Rechenleistung erfolgt dort. Dabei hat niemand Zugriff auf die hochgeladenen Dokumente (Aber, natürlich gilt: „Other Internet users may under certain circumstances have the technical know how to interfere with network security and gain control of the data transmitted. The User is entirely and solely responsible for the safety and backup of the data he transmitted to the Internet and stored on servers.“ Ebd.) – außer, man erteilt Zugriffsrechte, siehe weiter unten. Der Upload von Dokumenten mit sensiblen Inhalten sollte daher abgewogen werden. Dokumente, die mit einem Copyright belegt sind, können aber unbedenklich genutzt werden.

⁶ Nach Ablauf der EU-Förderung soll das Projekt in eine Kooperative überführt werden; die Nutzung für den*die Einzelnen wird bis zu einer Seitenzahl von 1.000 Seiten voraussichtlich kostenlos bleiben. Endgültige Aussagen hierzu sind aber erst nach Gründung der Kooperative möglich. Informationen zum aktuellen Stand sind unter <https://read.transkribus.eu/coop/> verfügbar.

⁷ Online unter: <https://transkribus.eu/wiki/De/index.php/Hauptseite#Installation>.

⁸ Online unter: https://transkribus.eu/wiki/index.php/Users_guide.

⁹ Java SE Runtime Environment 8: <https://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html>. Bei einer anhaltenden Fehlermeldung über ein fehlendes Java-Entwicklerpaket findet sich evtl. hier Hilfe: https://transkribus.eu/wiki/index.php/Questions_and_Answers#Download_and_Installation.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



2. Start und erste Schritte

Die Oberfläche von *Transkribus* besteht aus mehreren Bereichen. Es gibt eine zentrale Menüleiste, eine Tableiste für die Verwaltung von Dokumenten, einen Bildbereich mit separater Menüleiste und einen Text-/Transkriptionsbereich, ebenfalls mit separater Menüleiste.

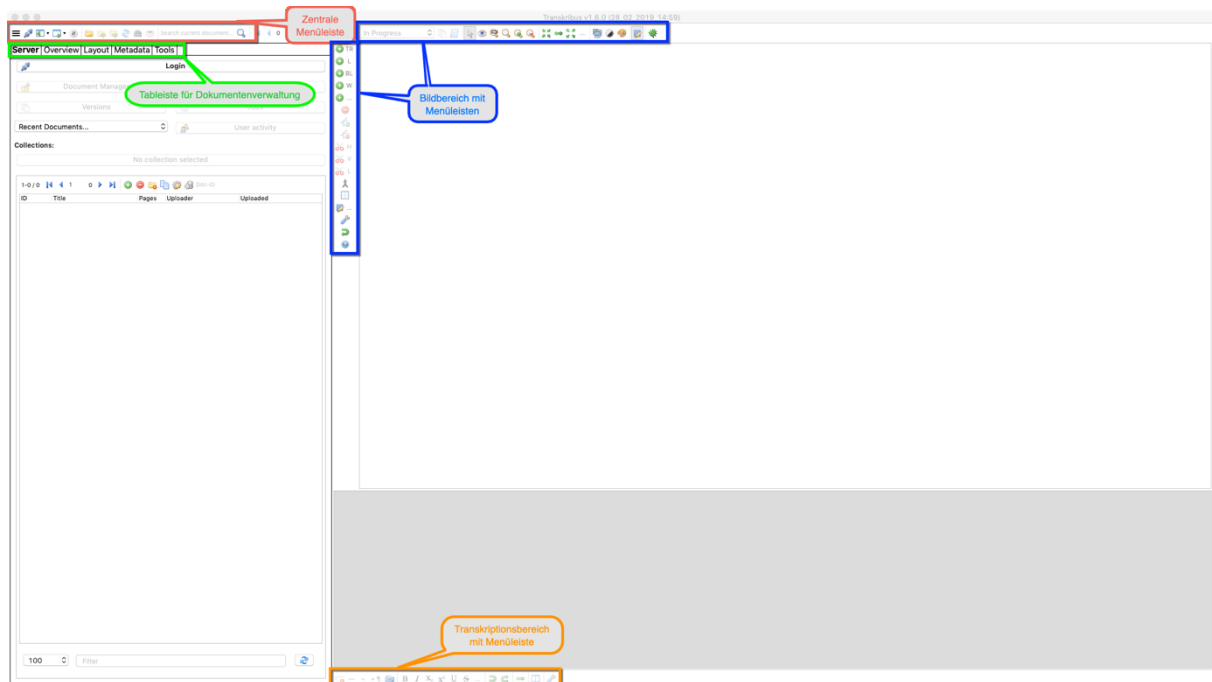


Abb. 1: Arbeitsbereiche und Menüleisten in *Transkribus* vor dem Login (eigener Screenshot, bearbeitet).

Eine detaillierte Übersicht über die Funktionen der einzelnen Leisten ist in der *Transkribus*-Benutzeranleitung¹⁰ aufgelistet.

Nach dem Start und Login – der Button versteckt sich unter der Tableiste für die Dokumentenverwaltung – besteht der erste Schritt im Anlegen von Kollektionen, in die beliebig viele Dokumente, bestehend aus einzelnen Seiten oder auch aus kompletten Handschriften/Drucken, eingefügt werden können.

Jede*r Nutzer*in verfügt von Beginn an über eine Beispielkollektion mit vier verschiedenen Dokumenten:

¹⁰ *Transkribus*. Benutzerhandbuch für die graphische Oberfläche von *Transkribus*. 2016. Onlineresource: <https://transkribus.eu/wikiDe/index.php/Benutzeranleitung>.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

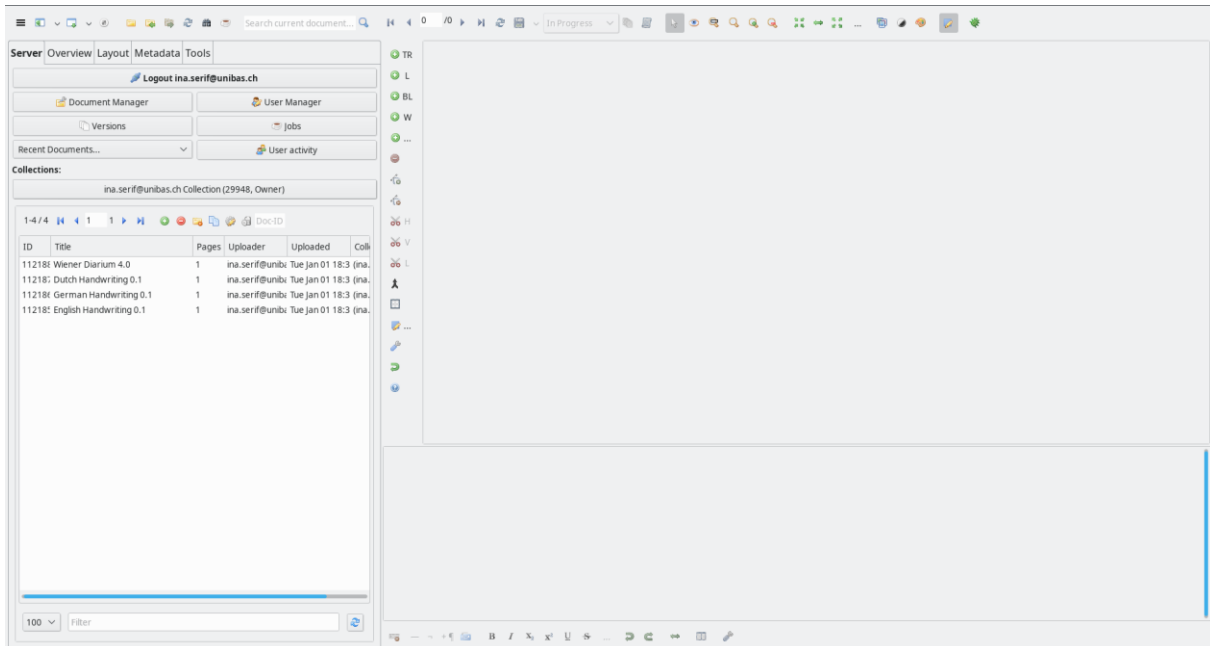


Abb. 2: Ansicht nach erfolgreichem Login mit Übersicht über die integrierte Beispielkollektion (eigener Screenshot, bearbeitet).

Diese Beispielkollektion kann auch entfernt werden; allerdings stehen über diese bereits vier verschiedene Texterkennungsmodelle zur Verfügung, die für eigene Dokumente genutzt und in neue Kollektionen importiert werden können.¹¹

Unter „Collections“¹² können neue Sammlungen erstellt werden, und, falls gewünscht, über „Manage users“ andere *Transkribus*-Nutzer*innen hinzugefügt werden.¹³ Die geteilte Sammlung erscheint dann auch in der „Collections“-Übersicht der Mitnutzer*innen.

Da die Dokumente auf den Projektservern gespeichert und verarbeitet werden, wird der Festplatten- und Arbeitsspeicher des eigenen Geräts nicht belastet, wodurch sich *Transkribus* auch

¹¹ Zu den Modellen mehr unter Punkt 3. Vorab nur: Modelle aus der eigenen Kollektion können kollektionsübergreifend exportiert und genutzt werden. In mit anderen geteilten Kollektionen können eigene Modelle importiert werden, die so dann auch den Mitnutzer*innen zur Verfügung stehen. Eine freundliche Anfrage mit Quellenbeispiel an email@transkribus.eu bezüglich bereits vorhandener Modelle für eine spezifische Schrift(-art) wird womöglich positiv beantwortet, erspart also einige Transkriptionsmühen.

¹² Fachbegriffe aus dem *Transkribus*-Vokabular sind in diesem Artikel in Anführungszeichen gesetzt. Alle weiteren, nicht aus dem Programm zitierten Fachbegriffe, Softwarebezeichnungen und Projektnamen sind kursiv gesetzt.

¹³ Die Möglichkeit, nach Nutzer*innen über Mailadresse oder Name zu suchen, ist freilich bequem; allerdings wissen im Zweifel alle Kolleg*innen, dass man *Transkribus* benutzt.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



zur Verwaltung von Dokumentensammlungen anbietet, die unter „Metadata“ auch mit entsprechenden Informationen angereichert und verwaltet werden können.

Als letzter Schritt, bevor die eigentliche Arbeit der Spaß beginnt, werden eine oder mehrere Dateien in eine „Collection“ eingefügt. Über „Import document(s)“ stehen dabei mehrere Optionen zur Auswahl:

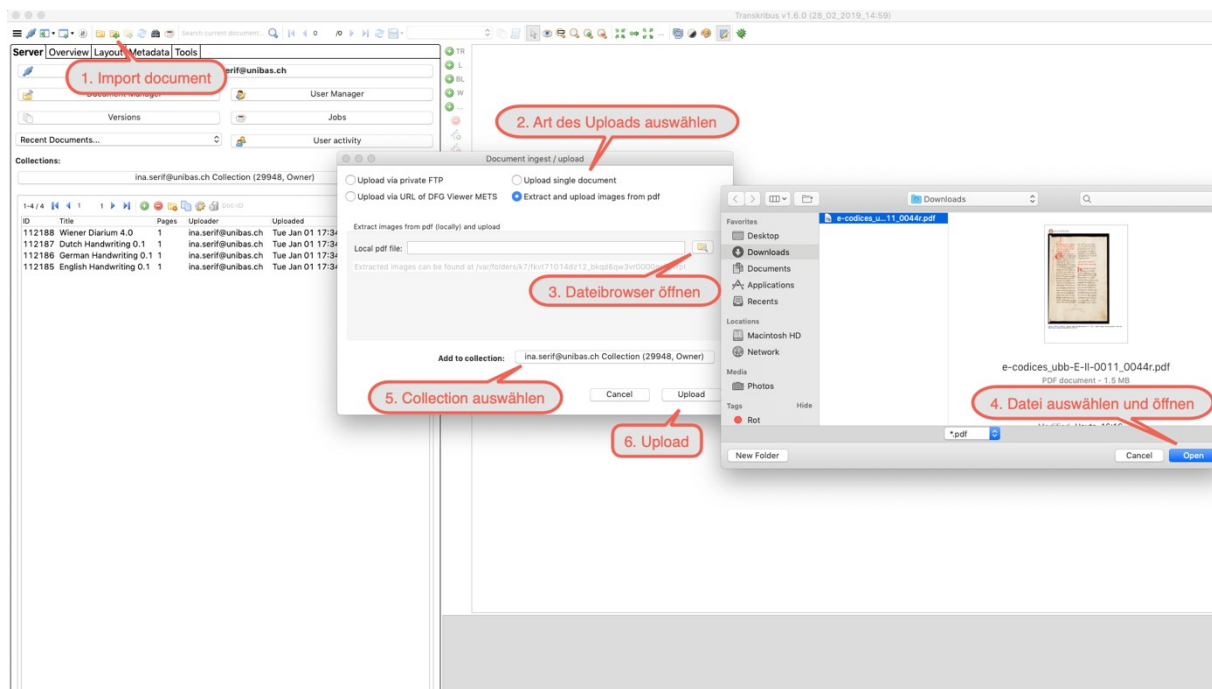


Abb. 3: Upload eines Dokuments in eine Kollektion (eigener Screenshot, bearbeitet).

Standardmäßig ist „Upload single document“ eingestellt, das für Bilddateien in den gängigen Formaten JPG, PNG, TIFF und JP2 geeignet ist (GIF- und RAW-Dateien werden nicht unterstützt). Zu beachten ist bei dieser Option, dass nur der Ordner, nicht eine einzelne Datei ausgewählt werden kann, der Pfad muss also zu einem Ordner führen. Hat man ein mehrseitiges PDF-Dokument, kann dieses direkt angewählt und mit der Auswahl „Extract and upload images from pdf“ hochgeladen werden. Online verfügbare Digitalisate können mit einem DFG-Viewer- oder METS-Link ohne vorheriges Speichern auf dem eigenen Computer hochgeladen werden, wobei hier, je nach Serverauslastung, der Upload nicht immer auf Anhieb gelingt; bei einer

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



entsprechenden Fehlermeldung einfach nochmals versuchen. Ein Upload über ein *IIIF-Manifest*¹⁴ ist (noch) nicht implementiert, das betrifft z.B. große Digitalisat-Sammlungen wie *e-codices*¹⁵; hier muss vorerst noch der Umweg über den Download auf den eigenen Rechner und anschließenden Upload gegangen werden. Für große bzw. viele Bilder eignet sich der Upload mittels eines FTP-Clients (es gibt eine große Auswahl an *free-* und *open-source*-Clients zum Download,¹⁶ die mit Anleitungen versehen sind). Auch selbst erstellte Digitalisate können hochgeladen werden, die Auflösung für eine korrekte Layout- und Texterkennung muss nicht sonderlich hoch sein. Da schummrige Archivlicht aber nicht immer akzeptable Aufnahmen garantiert, wurde als *Spin-Off* innerhalb des *READ*-Projekts das *ScanTent*¹⁷ entwickelt:

¹⁴ *IIIF – International Image Interoperability Framework*, siehe <https://iiif.io/>.

¹⁵ *e-codices – Virtuelle Handschriftenbibliothek der Schweiz*, Fribourg: Universität Freiburg Schweiz, 2008–, <https://www.e-codices.ch/de>.

¹⁶ Eine Übersicht findet sich in Art. „Comparison of FTP client software“, zuletzt bearb. von „Grimm665“, in: Wikipedia. The Free Encyclopedia, https://en.wikipedia.org/wiki/Comparison_of_FTP_client_software#Free_and_open-source_software.

¹⁷ Siehe <https://scantent.cvl.tuwien.ac.at/en/>.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Abb. 4: Einsatz des ScanTent beim Vienna Scanathon am 7.11.2018 (Foto: Sandra Lehecka; Lizenz: [CC-BY 4.0](https://creativecommons.org/licenses/by-sa/4.0/)).

Eine Beleuchtung im Innern und eine fixierte Auflagefläche für Smartphones oder Tablets liefern einheitliche Bilder.¹⁸ Eine App¹⁹ wurde ebenfalls entwickelt, mittels derer die Bilder direkt auf die *Transkribus*-Server hochgeladen werden können. Einziger Nachteil: Richtige Kameras mit Objektiv können nicht in der oberen Öffnung eingespannt werden. Die Qualität von Smartphone-Kameras reicht allerdings bei Weitem aus, um akzeptable Digitalisate zu erhalten.

Nach dem im Programm oder über die App gestarteten Upload wird die Datei auf den Projektservern verarbeitet, was je nach Größe eine Zeit dauern kann. Der aktuelle Status ist über das Kaffeetassensymbol in der zentralen Menüleiste ersichtlich; durch die Aktualisierung der

¹⁸ Hier geht es zu einem kurzen Video: <https://youtu.be/dfQvV6XXeVg>. Ein längerer Bericht wurde verfasst von Florian Kleber [u.a.], *Mass Digitization of Archival Documents using Mobile Phones*, in: *HIP2017 Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, S. 65–70, <https://dl.acm.org/citation.cfm?doi=3151509.3151526>.

¹⁹ Verfügbar unter: <https://play.google.com/store/apps/details?id=at.ac.tuwien.caa.docscan&hl=de>.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Kollektion bzw. der „document list“ kann der Abschluss der Verarbeitung nachvollzogen werden:

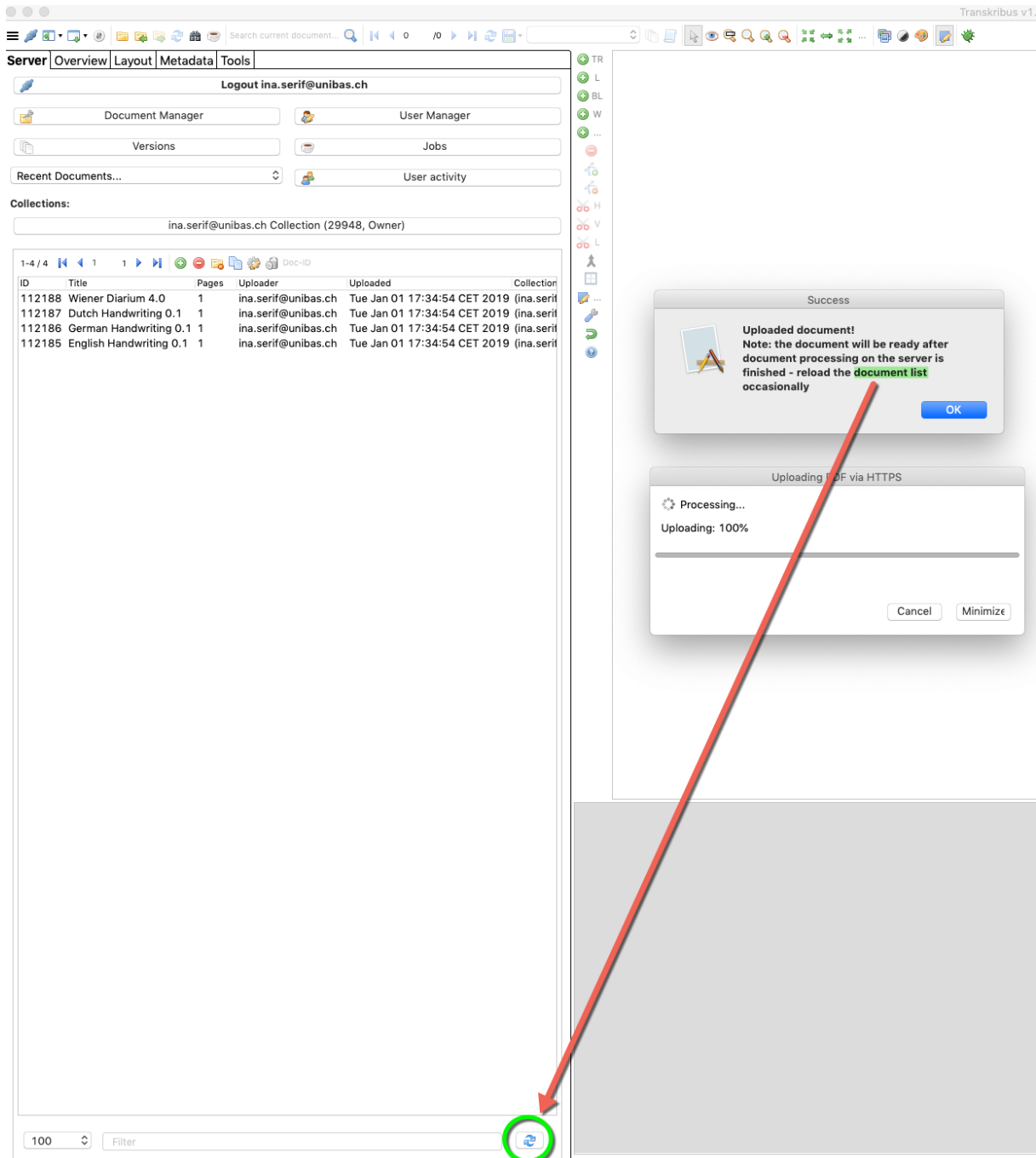


Abb. 5: Ansicht nach gestartetem Upload eines Dokuments (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



3. Layout Analysis und Text Recognition (HTR)

Als Voraussetzung für die automatisierte Texterkennung eines Dokuments muss zuerst definiert werden, in welchen Seitenbereichen sich überhaupt Text befindet. Vor einer Texterkennung steht also eine Layouterkennung. Nach dem Auswählen eines Dokuments aus einer Kollektion kann unter dem Reiter „Tools“ die „Layout Analysis“ angewählt werden; an der Standardeinstellung „CITlab Advanced“ muss nichts geändert werden. Dann folgt die Auswahl, welche Seiten des Dokuments erkannt und in Textregionen segmentiert werden soll.

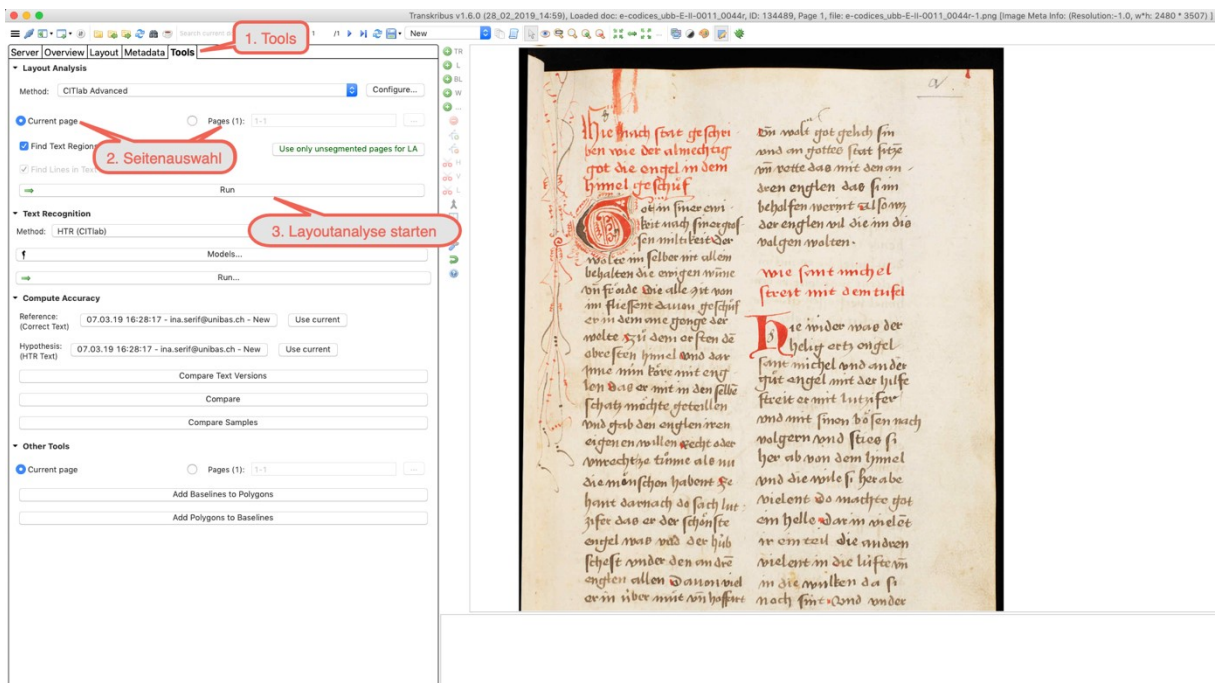


Abb. 6: Auswahl der Seiten für die „Layout Analysis“ und Start der Analyse (eigener Screenshot, bearbeitet); Handschrift im Beispiel: ms. Basel, Universitätsbibliothek, E II 11, fol. 44r. Digitalisat via [DOI: 10.5076/e-codices-ubb-E-II-0011](https://doi.org/10.5076/e-codices-ubb-E-II-0011). Lizenz: [CC0 1.0](https://creativecommons.org/licenses/by-sa/4.0/) (gültig für sämtliche Abbildungen der Seite in diesem Tutorial).

Ist dies getan, erledigen die Server in Innsbruck die Arbeit und es ist, je nach Seitenanzahl, Zeit für eine erste Pause.²⁰ Mit Klick auf die Kaffeetasse kann man jederzeit den Fortschritt der aktuellen Aufgabe einsehen:

²⁰ Der eigene Rechner kann auch eine Pause einlegen: Nach dem Start von Prozessen wie Layout- oder Texterkennung, die extern ablaufen, kann der Computer ggf. auch ausgeschaltet werden.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 2* (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

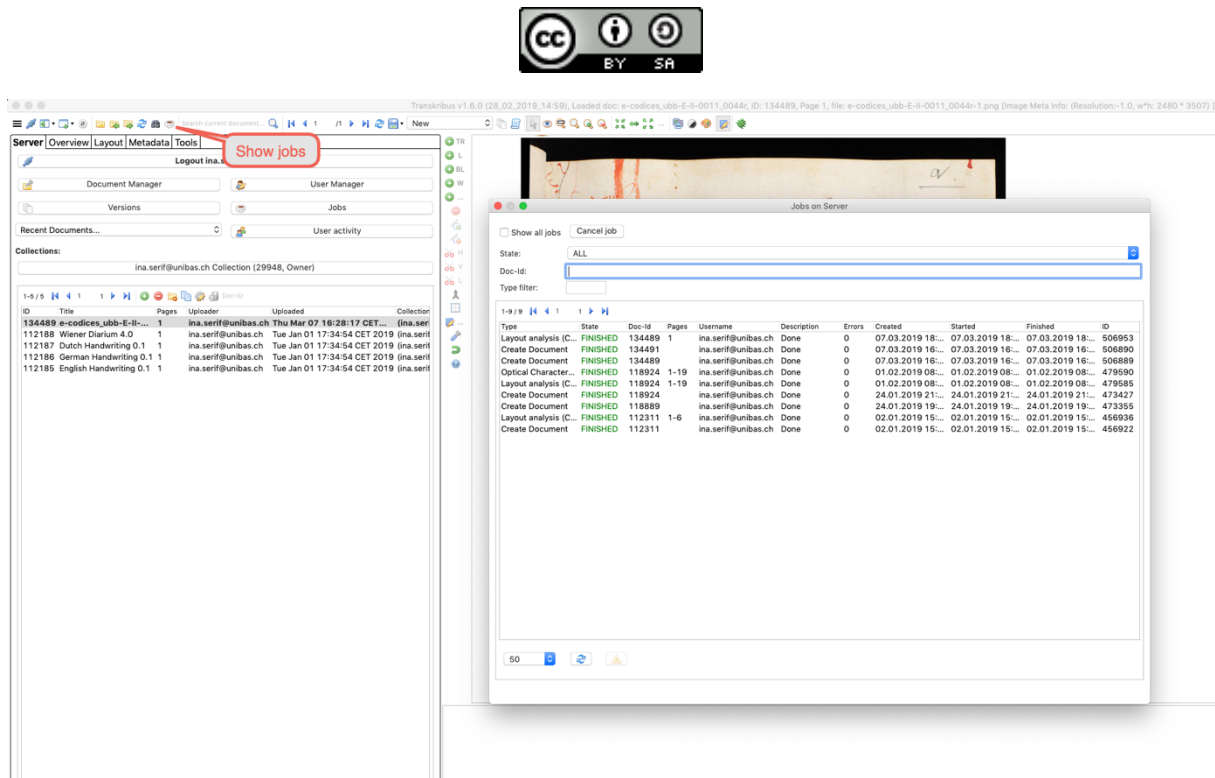


Abb. 7: Anzeige der aktuellen und abgeschlossenen Vorgänge auf dem Server (eigener Screenshot, bearbeitet).

Nach der Layoutanalyse wird die Segmentierung des Dokuments in einzelne Layoutbereiche angezeigt. Bei großen Initialen oder Tabellen treten hier mitunter Schwierigkeiten auf. Gewöhnliches Layout, auch mehrspaltig, wird mittlerweile aber meist problemlos korrekt erkannt:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

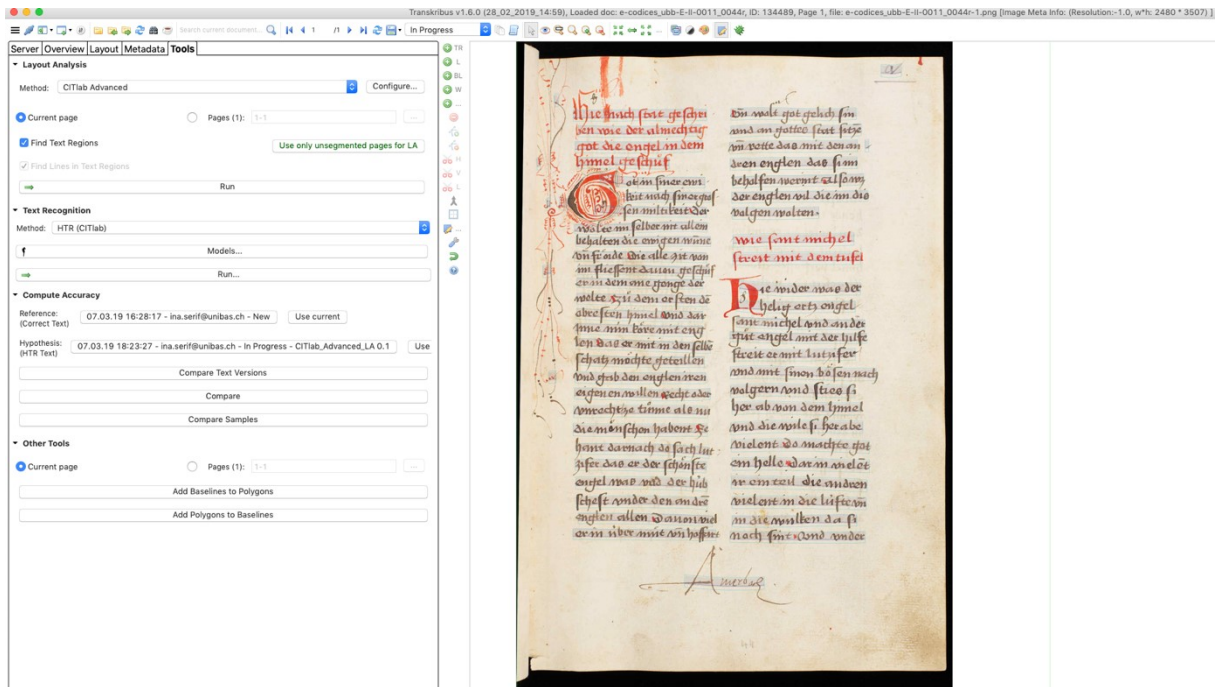


Abb. 8: Ansicht nach erfolgter Layouterkennung: Die blaue Markierung kennzeichnet die erkannten Zeilen (eigener Screenshot, bearbeitet).

Die Reihenfolge der erkannten Zeilen („Lines“) oder Regionen („Regions“) kann über das Augensymbol angezeigt werden:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

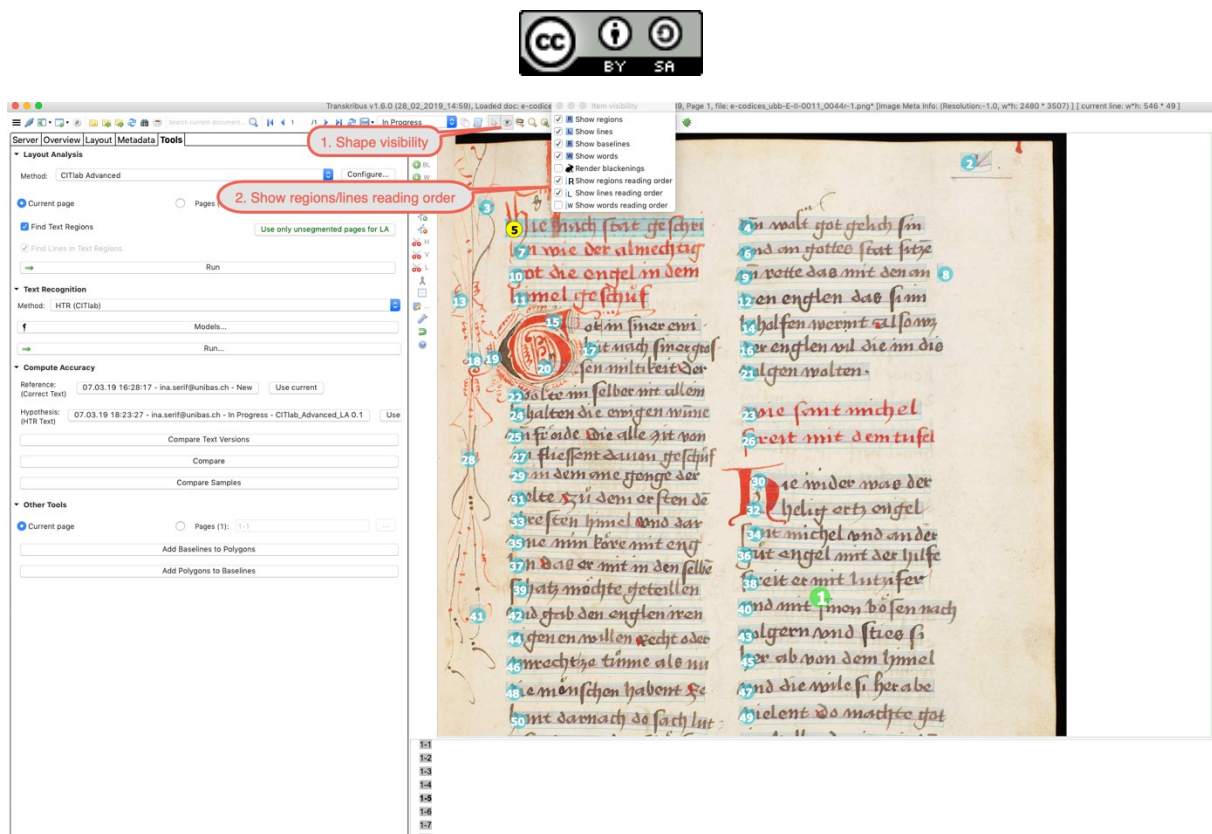


Abb. 9: Anzeige der erkannten Textregionen (grüne Kreise) und Zeilen (blaue Kreise) (eigener Screenshot, bearbeitet).

Für den Export von transkribiertem Text in ein anderes Format – dazu später mehr – ist die Reihenfolge der Zeilen bzw. Textregionen wichtig. Will man nur durchsuchbare Dokumente erhalten, ist sie vernachlässigbar. Beim obigen Beispiel hat das Erkennen der Regionen- und Lesereihenfolge nicht funktioniert; bei den meisten Dokumenten mit ähnlichem Layout ist die Erkennung aber mittlerweile relativ zuverlässig:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

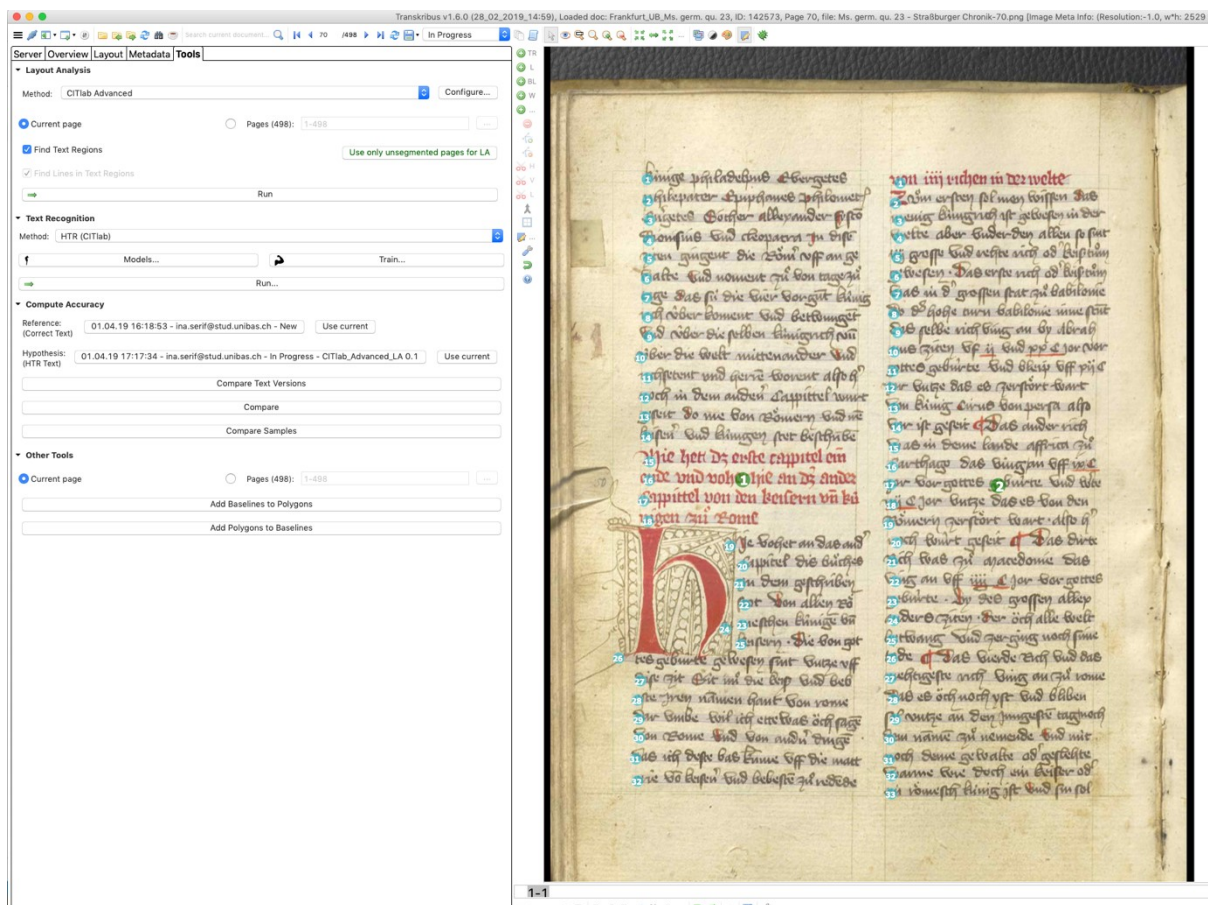


Abb. 10: Ergebnis der Layoutanalyse mit korrekt erkannter Abfolge der Textregionen und Zeilen (eigener Screenshot, bearbeitet; Handschrift im Beispiel: Frankfurt, Universitätsbibliothek, Ms. germ. qu. 23, fol. 30v. Digitalisiert via URN: [urn:nbn:de:hebis:30:2-14660](https://nbn-resolving.org/urn:nbn:de:hebis:30:2-14660). Lizenz: keine (Public Domain).

Die linke Spalte wurde als Region 1 identifiziert, die rechte als Region 2, die einzelnen Zeilen wurden ebenfalls in der richtigen Reihenfolge erkannt.

Bei nicht erkannten Regionen bzw. falscher Zeilenreihenfolge können Regionen und Zeilen hinzugefügt und die Lesereihenfolge verändert werden:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 2* (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

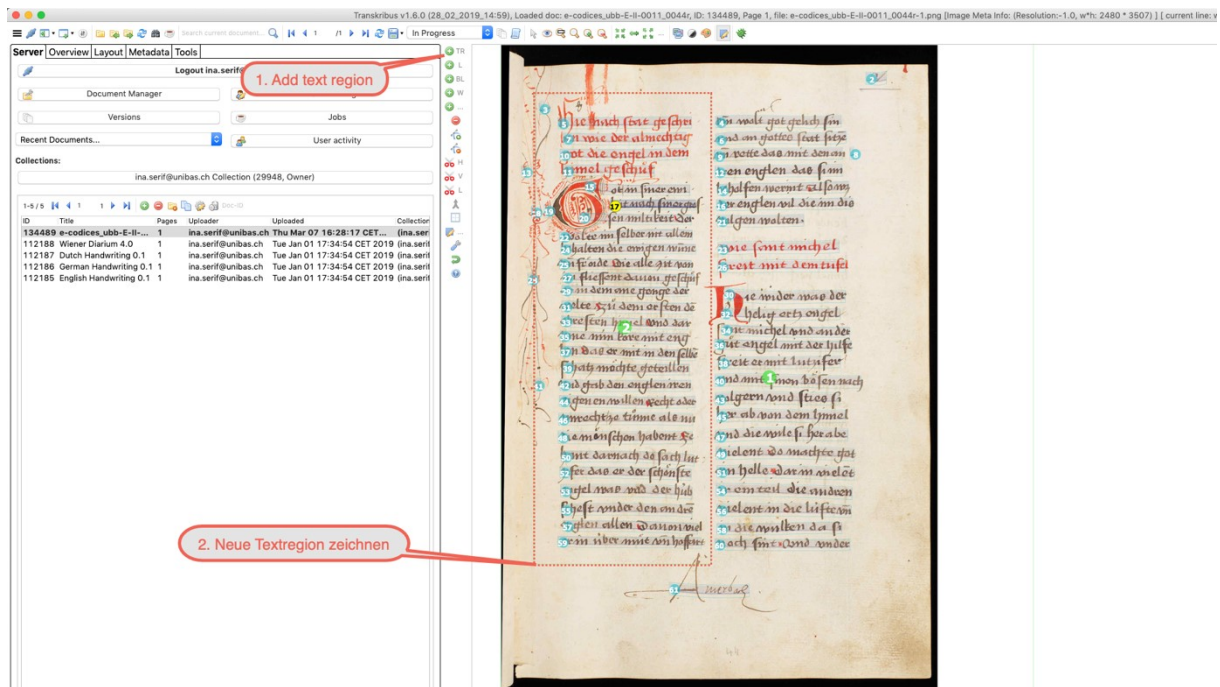


Abb. 11: Hinzufügen neuer Textregionen (eigener Screenshot, bearbeitet).

Nachdem die neue Region gezeichnet wurde, muss die Reihenfolge korrigiert werden:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

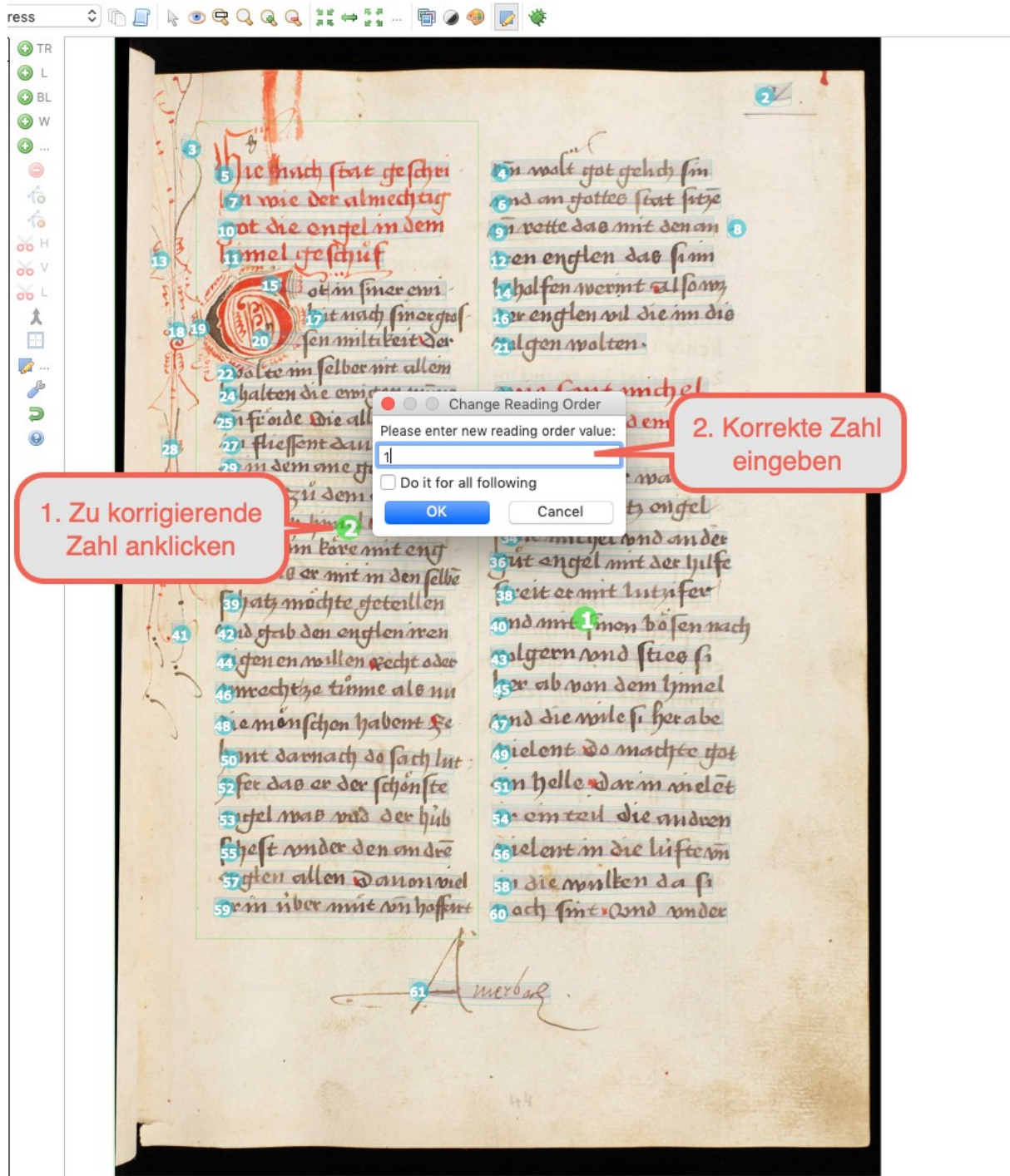


Abb. 12: Korrektur der Reihenfolge der Textregionen (eigener Screenshot, bearbeitet).

Dasselbe gilt für falsch erkannte Zeilenabfolgen:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

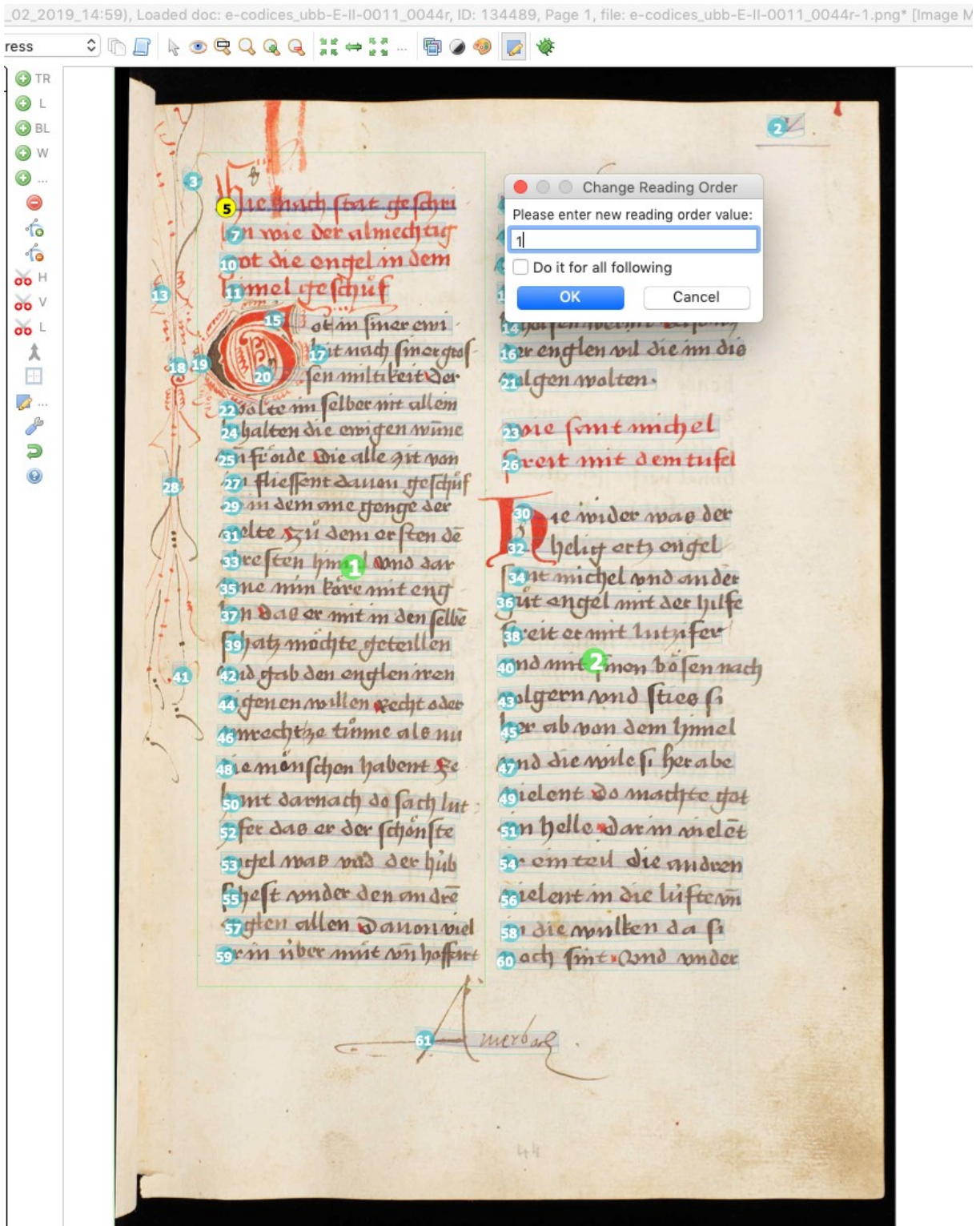


Abb. 13: Korrektur der Zeilenabfolge (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Überzählig erkannte Zeilen können, nachdem sie angewählt (= gelb eingefärbt) wurden, per Rechtsklick gelöscht werden:

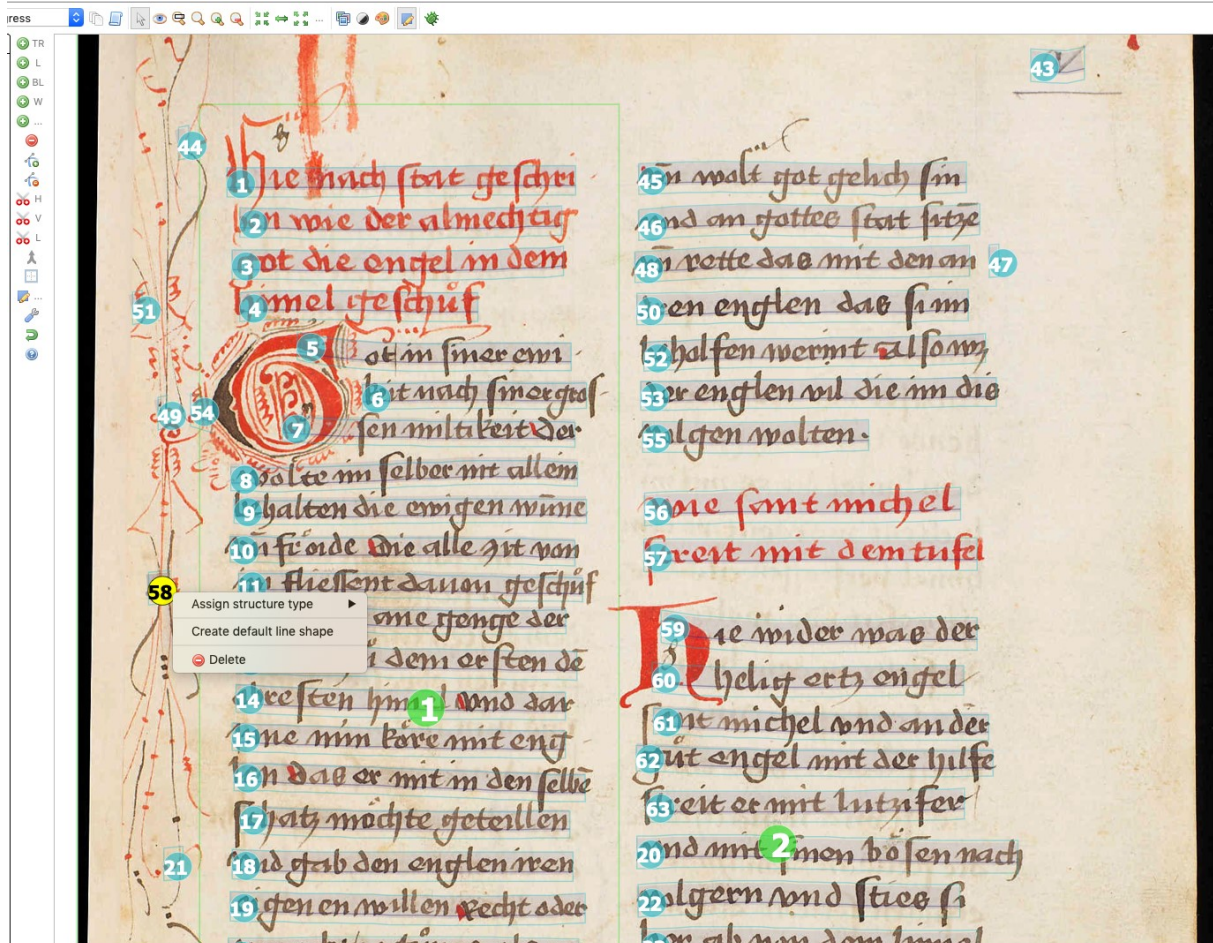


Abb. 14: Löschen falsch erkanntter Zeilen (eigener Screenshot, bearbeitet).

Nach der Korrektur sieht die Beispielseite so aus:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

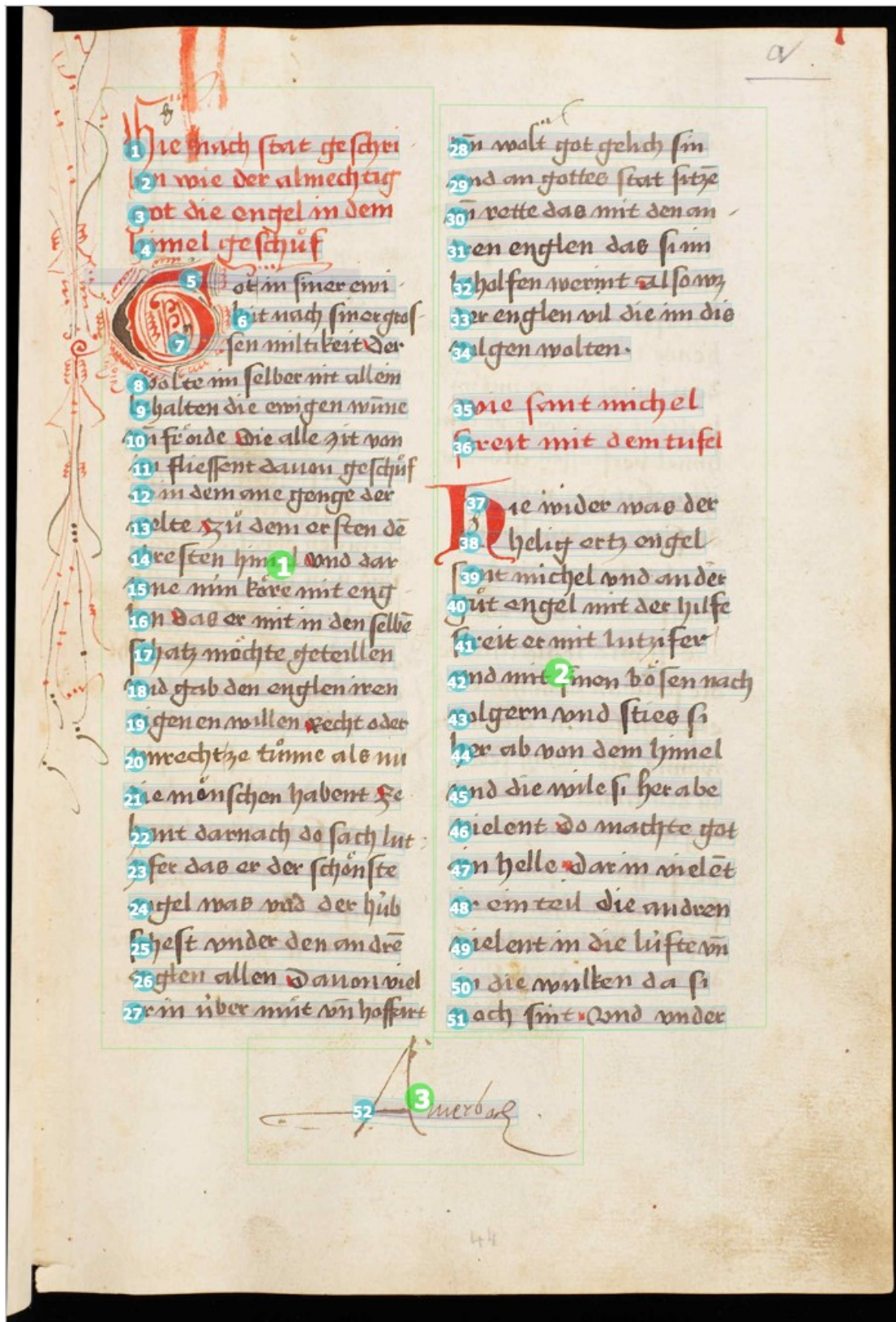


Abb. 15: Ansicht einer Seite mit korrekter Regionen- und Zeilenreihenfolge (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Je nach Ergebnis der Layouterkennung ist es ggf. schneller, bereits Erkanntes ganz zu löschen und die Seite selbst manuell mit Layouteinheiten zu versehen: Über die Menüleiste des Bildbereichs können verschiedene Strukturelemente hinzugefügt werden. Wichtig für die Texterkennung/Transkription sind hier

- Textregion hinzufügen („Add a text region“)
- Grundlinie hinzufügen („Add a baseline“): Auf der Basis von Grundlinien werden Zeilen automatisch erstellt (die Option „Add a line“, also Zeile hinzufügen, ist daher eigentlich obsolet); für die Texterkennung sind sie der zentrale Bezugspunkt.

Die zwei Bildbereich-Menüleisten haben weitere Funktionen, die je nach Dokument nützlich sind. Die vermutlich wichtigsten sind:

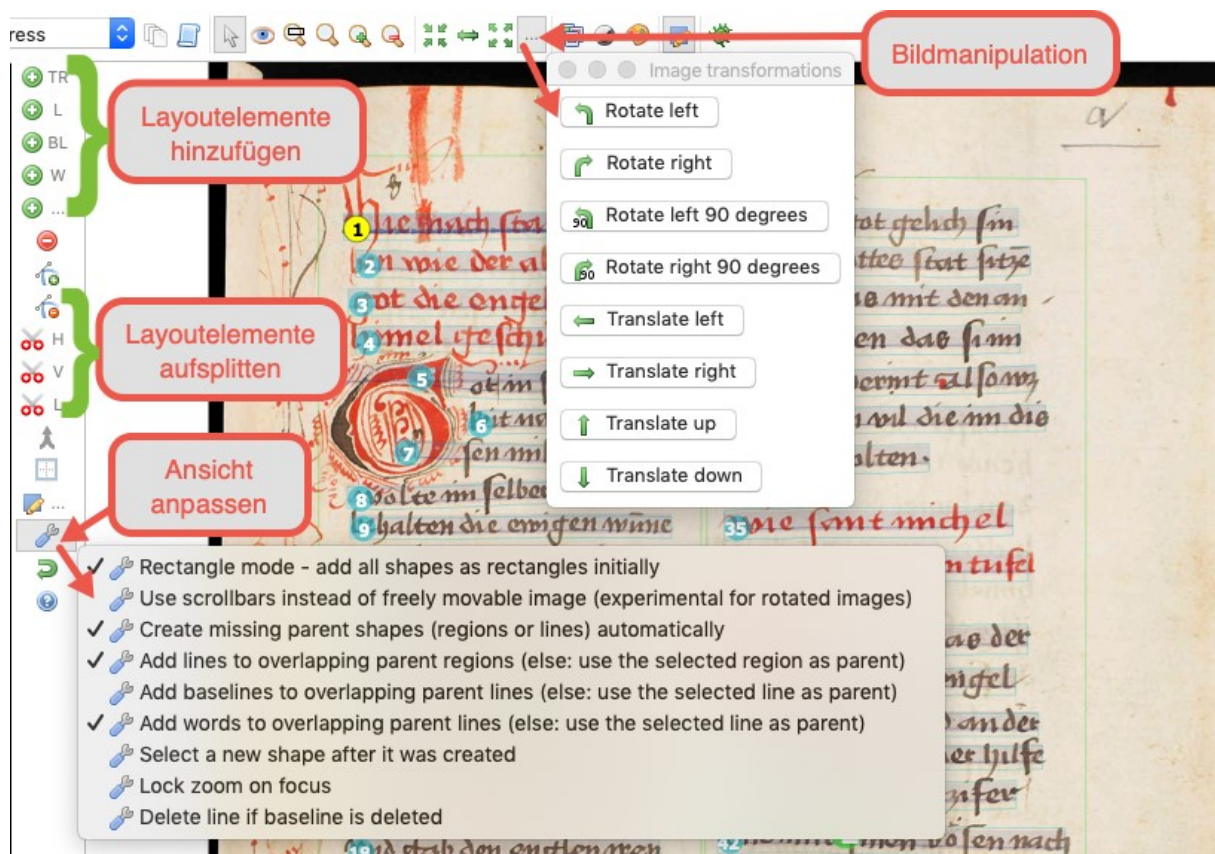


Abb. 16: Funktionen der Menüleiste bei der Layoutanalyse (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Die nächsten Schritte hängen nun stark vom Quellenmaterial und den konkreten Absichten bzw. Forschungsfragen ab: Geht es darum, *Transkribus* vor allem als Transkriptions- bzw. Editions-
werkzeug zu nutzen, kann nun Zeile für Zeile transkribiert werden. Der Transkriptionsmodus, der in der zentralen Menüleiste oben links unter „Profiles“ ausgewählt werden kann, bietet dabei auf einen Blick die zu transkribierende Zeile, optisch nochmals hervorgehoben, und den transkribierten Text („Transcription“ bzw. „Transcription invisible tabs“):

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

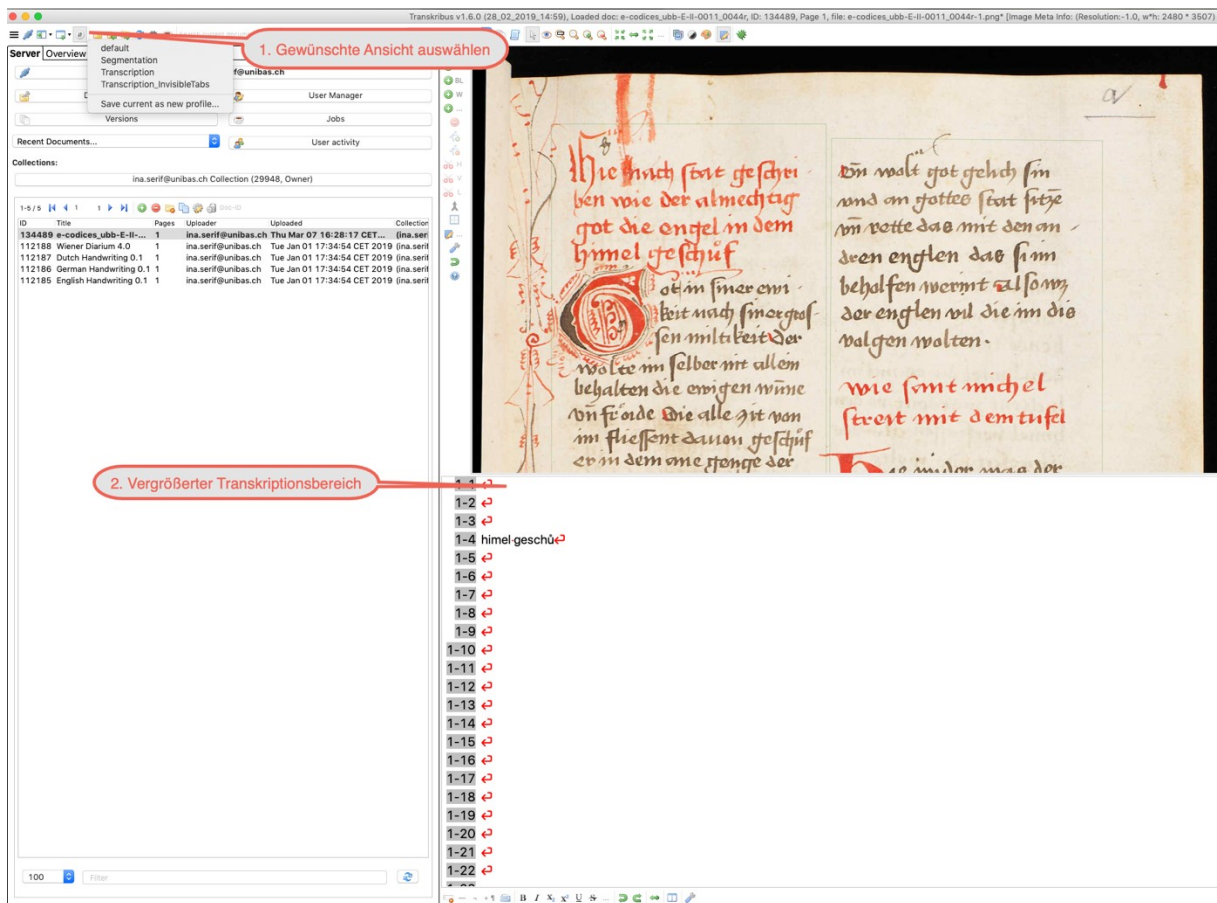


Abb. 17: Ansicht mit vergrößertem Transkriptionsbereich (eigener Screenshot, bearbeitet).

In der Menüleiste des Transkriptionsbereichs lassen sich weitere Anpassungen machen, wie beispielsweise Vergrößerung der Schrift:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 2* (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

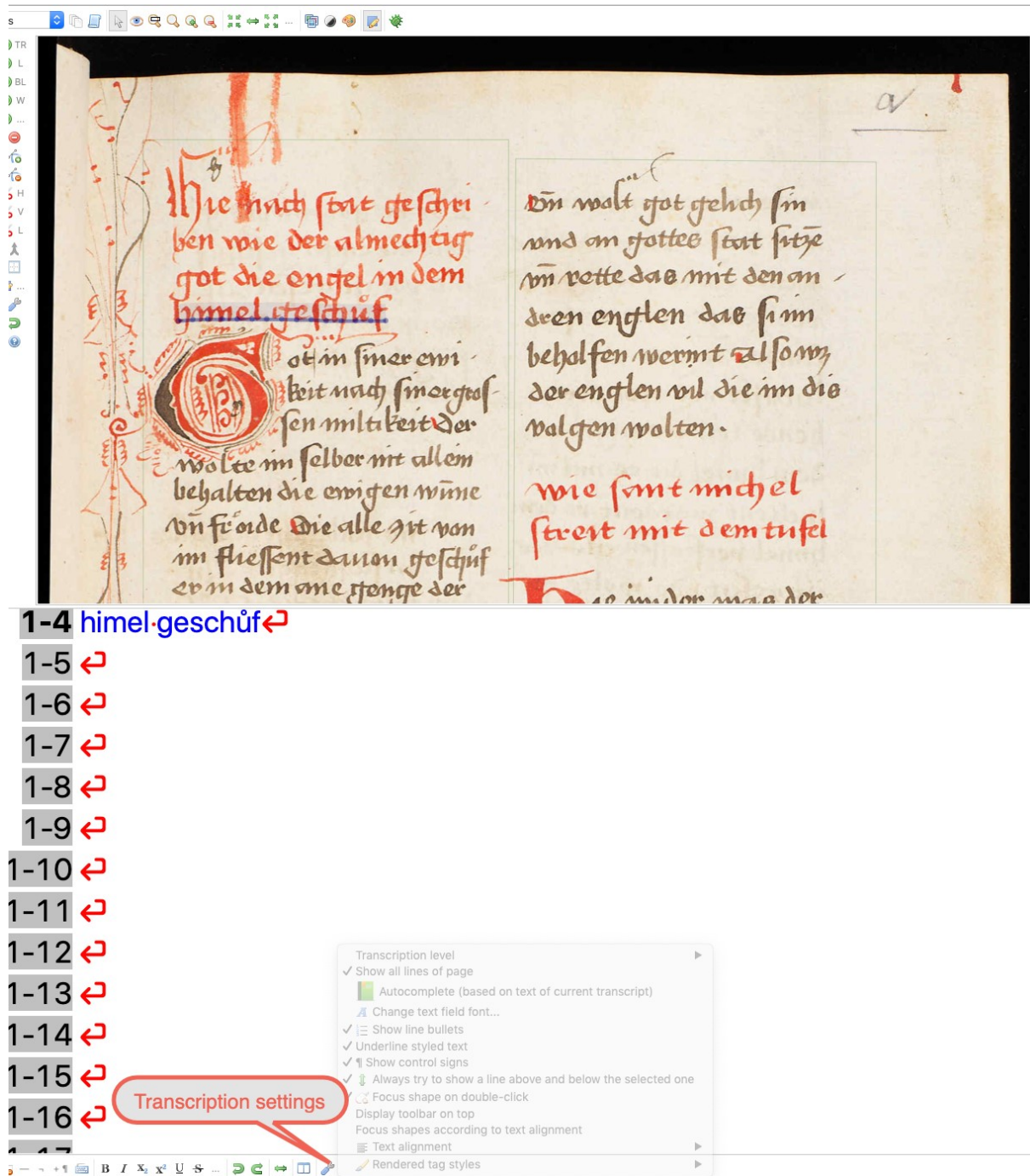


Abb. 18: Ansicht im Transkriptionsmodus mit weiteren Optionen (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 2* (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Dort stehen auch mehrere Textbearbeitungswerkzeuge und, für eine diplomatische Transkription, zahlreiche Sonderzeichen über das „Virtual Keyboard“ zur Verfügung:



Abb. 19: Textbearbeitungsleiste für die Transkription (eigener Screenshot, bearbeitet).

Die Symbolpalette des „Keyboards“ kann dabei beliebig erweitert und ein eigener Bereich mit häufig benötigten Sonderzeichen bestückt werden (wobei die bisherige Zeichensammlung recht beachtlich ist). Hierzu muss ein neues Zeichen als Unicode eingegeben werden.²¹

²¹ Eine Unicode-Tabelle, in der Zeichen in der benötigten Form „U+XXXX“ dargestellt werden, findet sich zum Beispiel hier: <https://unicode-table.com/>.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

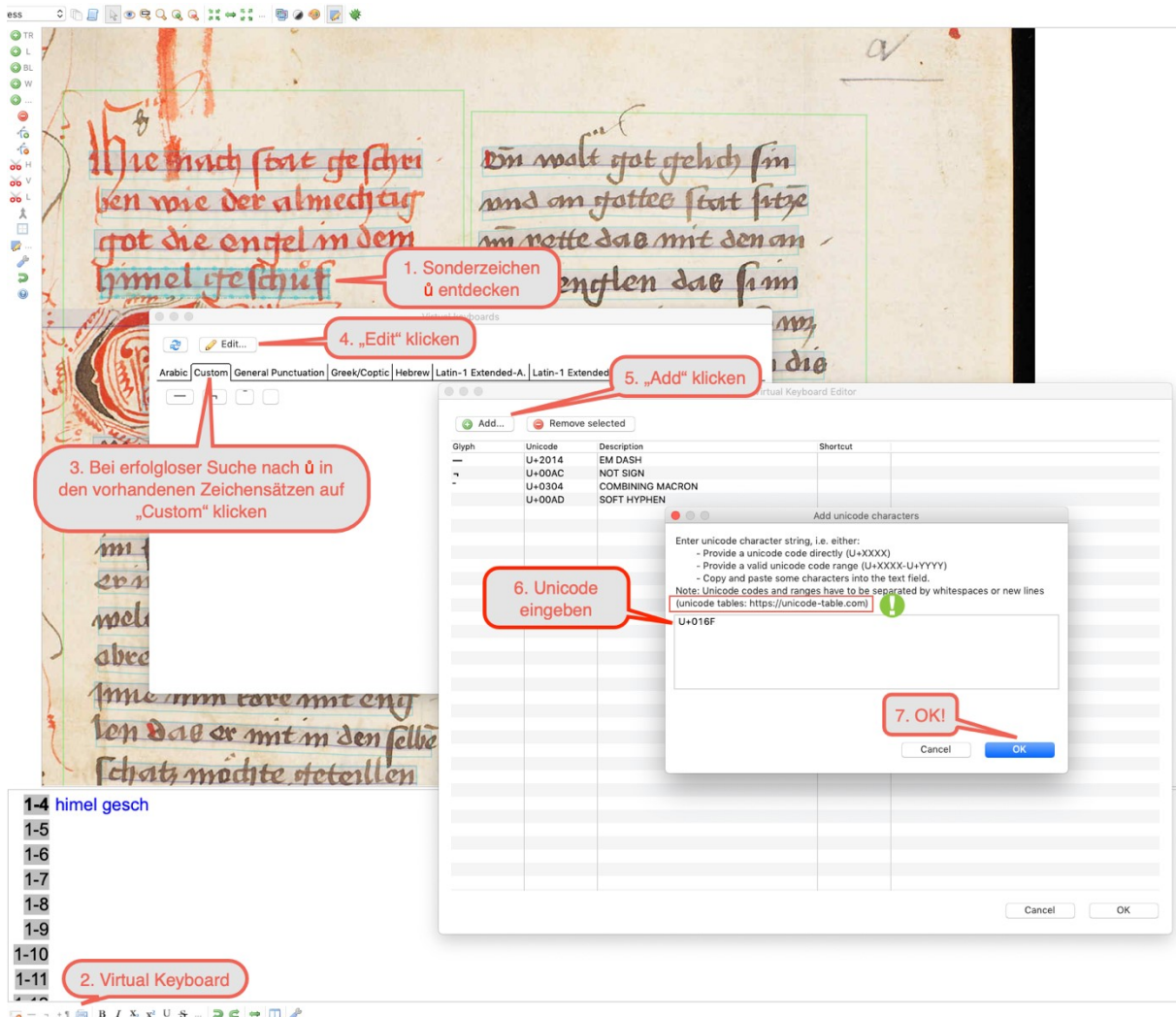


Abb. 20: Vorgehen zum Hinzufügen eines neuen Sonderzeichens (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

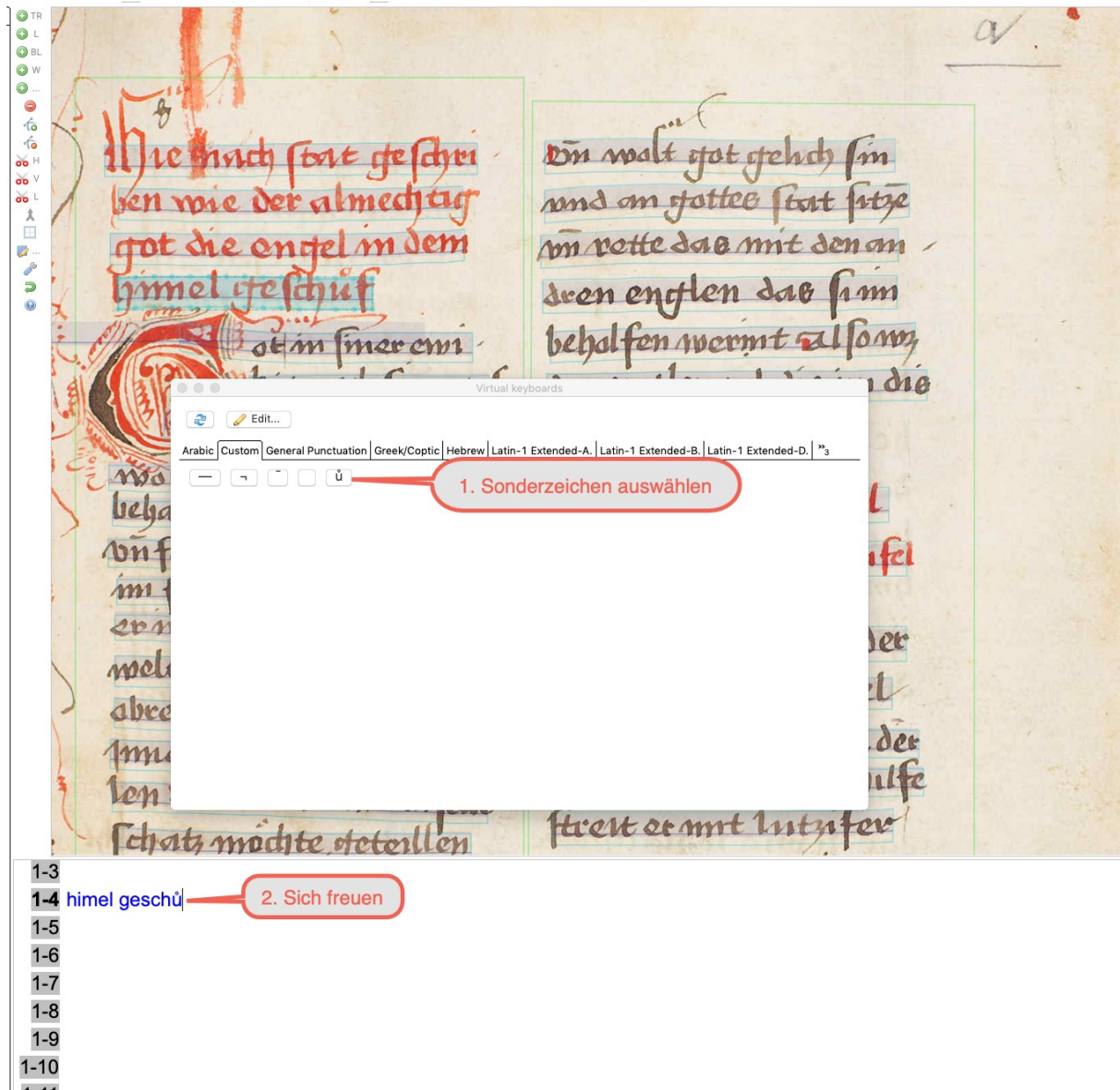


Abb. 21: Auswahl eines Sonderzeichens (eigener Screenshot, bearbeitet).

Eine so erstellte Transkription kann zum einen in unterschiedliche Formate exportiert werden:²²

²² Soll ein Lesetext erstellt werden, ist, wie oben erwähnt, die Reihenfolge der Zeilen wichtig. Soll beispielsweise eine PDF-Datei mit hinterlegtem Text entstehen, kann über die Suchfunktion eines *PDF-Viewers* natürlich auch nach Wörtern gesucht werden. Wortfolgen oder ganze Sätze werden durch falsche Umbrüche aber womöglich nicht gefunden.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

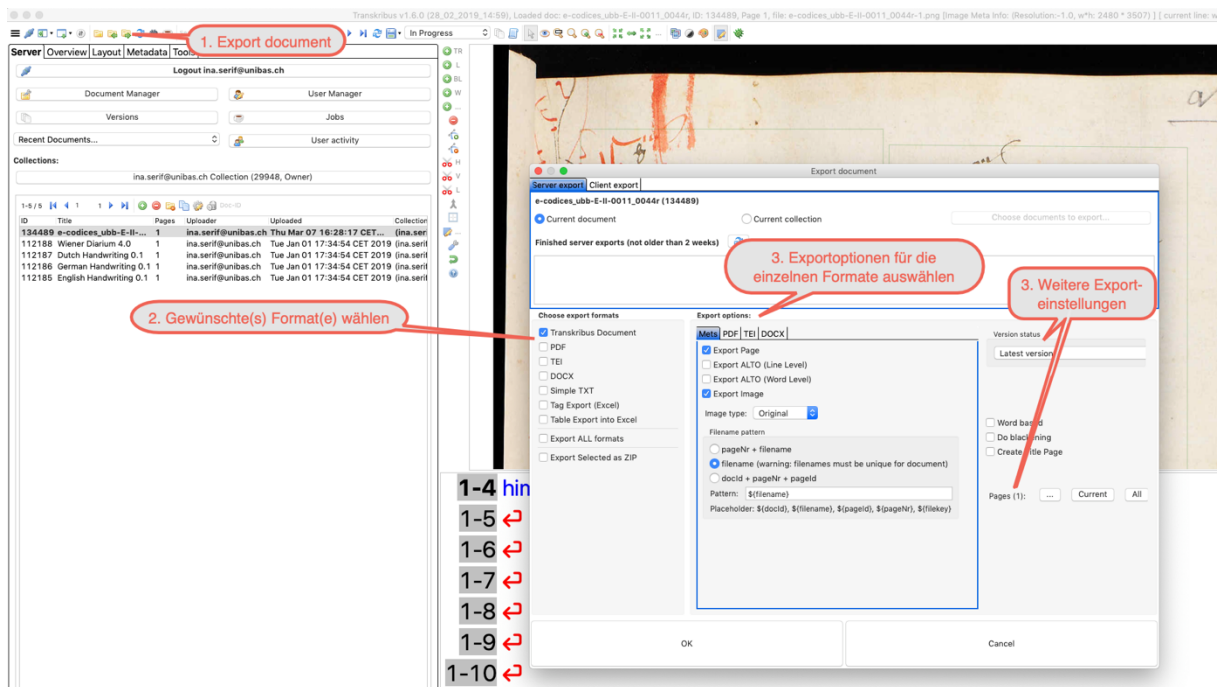


Abb. 22: Auswahl des gewünschten Exportformats und der zu exportierenden Seiten (eigener Screenshot, bearbeitet).

Zum anderen kann die Transkription als Grundlage für das Training eines Texterkennungsmodells genutzt werden, eines *Handwritten-Text-Recognition-Modells* (HTR).

Mittlerweile existieren für die unterschiedlichsten Schriften aus verschiedenen Epochen allgemeine HTR-/HTR+-Modelle,²³ die bereits mehr oder weniger gewinnbringend für eigene Dokumente genutzt werden können. In den meisten Fällen ist es weniger aufwendig bzw. zeitsparender, bereits existierende Modelle mit der eigenen korrekten Transkription, der sogenannten *Ground Truth*, zu verfeinern und zu trainieren, statt ein ganz neues Modell zu erstellen.

Über die jedem*r Benutzer*in zur Verfügung gestellte Startkollektion stehen, wie oben erwähnt, vier Modelle aus unterschiedlichen Epochen für erste Versuche zur Verfügung. Die den Modellen zugrundeliegenden Quellen können in der Modellübersicht eingesehen werden:

²³ HTR+ ist die transkribusinterne Weiterentwicklung von HTR; bereits bestehende HTR-Modelle können auf den Innsbrucker Servern ohne zusätzliches Trainingsmaterial zu HTR+-Modellen weitertrainiert werden, was etwa zur Halbierung der Fehlerrate des ursprünglichen HTR-Modells führt.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



The screenshot shows the Transkribus v1.6.0 interface. The 'Tools' tab is active, displaying a list of models. A red callout '1. Tools' points to the 'Tools' tab. Another red callout '2. Models' points to the 'Models' section. A third red callout '3. Bei Bedarf Textsample ansehen' points to the 'Show Train Set' button. A fourth red callout '4. Modell durch Auswahl bestätigen' points to the 'OK' button. The 'Learning Curve' graph shows accuracy over epochs, with a red line for 'CER Train' and a blue line for 'CER Test'. The 'CER on Train Set' is 11.91% and the 'CER on Test Set' is 9.84%.

Abb. 23: Übersicht über die vorhandenen Texterkennungsmodelle in der Beispielsammlung (eigener Screenshot, bearbeitet).

Für die Beispielseite in diesem Tutorial wurde das Modell „Comb_Gothic_Script“ genutzt, das sechs verschiedene gotische Buchschriften aus dem 13.–15. Jahrhundert kombiniert.²⁴

Die Texterkennung auf einer Seite/einem Dokument wird ähnlich wie die Layoutanalyse gestartet:

²⁴ Hierfür wurden Transkriptionen aus dem Parzival-Projekt der Universität Bern, siehe <http://www.parzival.unibe.ch/home.html>, und dem Editionsprojekt Königsfelden an der Universität Zürich, siehe <https://www.hist.uzh.ch/de/fachbereiche/mittelalter/lehrstuehle/teuscher/forschung/projekte/koenigsfelden.html> genutzt. Für Einzelheiten zum Modell siehe die Beschreibung in der Modellübersicht.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 2* (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

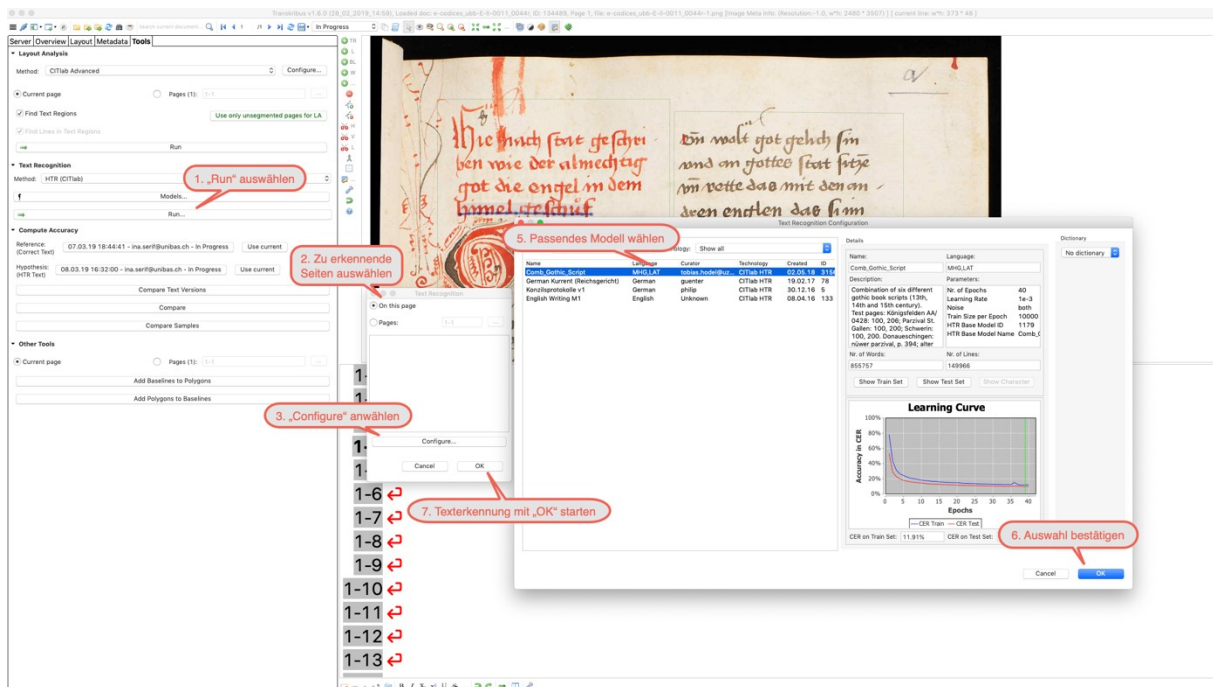


Abb. 24: Auswahl der Seiten und des Modells für die Texterkennung (eigener Screenshot, bearbeitet).

Dann den Start bestätigen und, wie bei der Layoutanalyse, erst einmal zurücklehnen.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

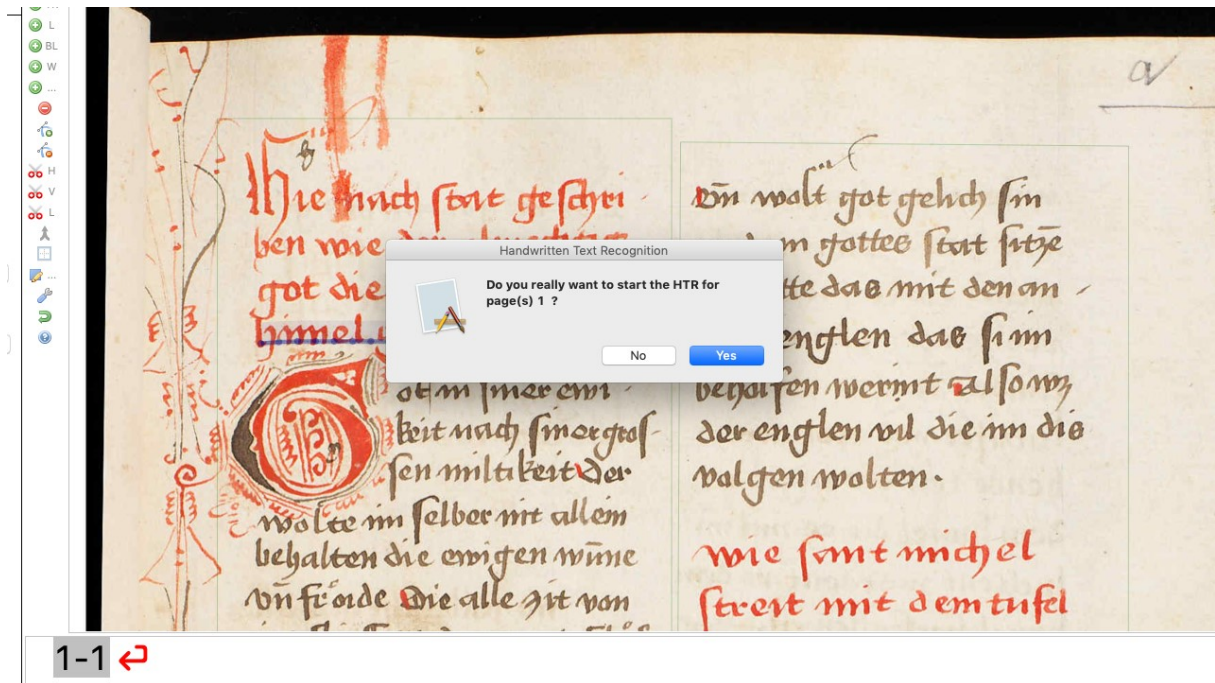


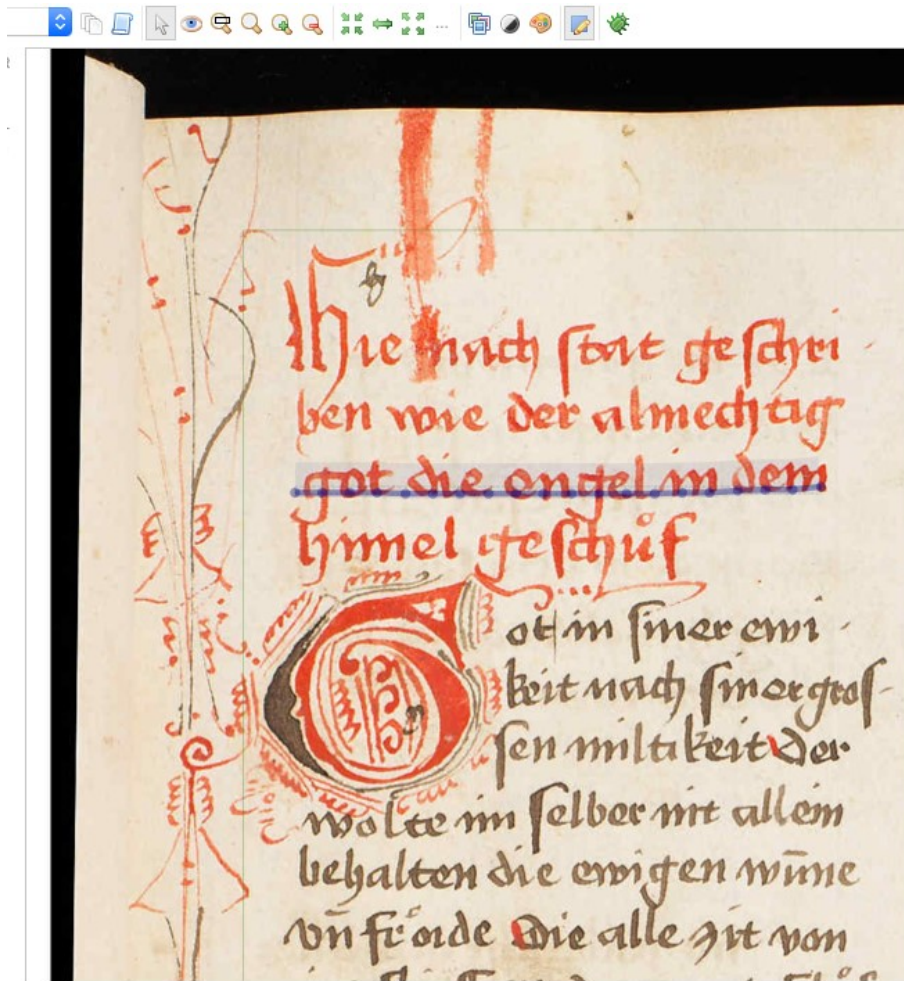
Abb. 25: Start der Texterkennung (eigener Screenshot, bearbeitet).

Das Ergebnis einer solchen Erkennung hängt, da es sich um sehr allgemeine Modelle handelt, sehr vom einzelnen Dokument bzw. der Schrift ab.²⁵ Die Beispielseite wurde folgendermaßen erkannt:

²⁵ Auf die Einbindung eines Wörterbuchs (in der Übersicht über die HTR-Modelle rechts oben unter „Dictionary“ auswählbar) wurde verzichtet, weil sich so möglicherweise durch Hyperkorrektur false positives, also gänzlich falsch erkannte (d.h. im Text gar nicht vorhandene) Wörter in einen Text einschleichen können, die nicht so schnell als solche identifiziert werden können wie einzelne falsch erkannte Zeichen.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



- 1 ie-huach-sta-ge-schzn↵
- 2 ben-wie-der-almechttag↵
- 3 got-die-engel-in-dem↵
- 4 umel-ige-schuf↵
- 5 Dsl-otz-um-umn-wi↵
- 6 it-nach-sinee-qto↵
- 7 sa-mttereen↵
- 8 wölte-nn-selber-nrt-allem↵
- 9 behalten-die-ewigen-wume↵
- 10 Vn-fr-oide-die-alle-qit-von↵

Abb. 26: Ergebnis der Texterkennung auf der Beispielseite (eigener Screenshot, bearbeitet).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Zeile eins liefert keine vernünftigen Ergebnisse, Zeile zwei hat ein falsch erkanntes „a“ im letzten Wort, Zeile drei wurde komplett korrekt erkannt, Zeile vier bis sieben hingegen scheinen nur mäßig brauchbar. Aber im Großen und Ganzen ist man mit diesem Modell offenbar auf einem guten Weg. Das weitere Vorgehen bestünde nun darin, den ausgegebenen Text selbst zu korrigieren – eine *Ground Truth* zu erstellen – und diese Korrektur für ein Training des Modells zu nutzen. Das Training kann noch nicht selbst gestartet werden; bisher besteht das Vorgehen darin, dass man sich nach einer Korrektur von einer gewissen Anzahl Seiten per Mail²⁶ mit der Bitte um ein erneutes Training meldet. Das Trainieren eines vorhandenen Modells ist ab etwa dreißig Seiten sinnvoll,²⁷ bringt dann aber bereits erhebliche Verbesserungen. Das Training eines solchen Modells dauert etwa dreieinhalb bis vier Stunden.²⁸

Bei der Erstellung der *Ground Truth* kann übrigens auch festgelegt werden, wie mit Sonderzeichen, Abbreviationen etc. umgegangen werden soll. Will man beispielsweise, dass ein „û“ in der Quelle als „uo“ in der Transkription widergegeben wird, so muss man es entsprechend transkribieren.²⁹ Bei Wörtern mit Zeilenumbruch gibt es bislang noch keinen Weg, eine Trennung zwar in der Transkription darzustellen, das Wort aber als Ganzes zu speichern; dies soll demnächst angegangen werden.

Ein neues/verfeinertes Modell muss dann nochmals auf das Dokument angewendet werden, damit die Ergebnisse sichtbar sind und die Fehlerquote („Character Error Rate“, CER) bei der Erkennung mit dem trainierten Modell abgefragt werden können, sodass man einen ersten Anhaltspunkt hat:

²⁶ Email an: email@transkribus.eu senden.

²⁷ Für ein völlig neues Modell sind mehr Seiten nötig, etwa 100 sind hier die Richtlinie. Zum Training generischer Modelle, die auf dieselbe Schriftart von unterschiedlichen Schreiber*innen angewandt werden können, vgl. den Beitrag von Achim Rabus, *Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus*, 2019 (*Draft Paper* online publiziert unter <https://www.academia.edu/s/40d211b367/recognizing-handwritten-text-in-slavic-manuscripts-a-neural-network-approach-using-transkribus-1-achim-rabus?source=work>).

²⁸ Momentan laufen relativ viele Trainings, weswegen es zu Verzögerungen kommen kann – auch hier gilt: Verlieren Sie keine Zeit!

²⁹ Man könnte auch beschließen, ein „û“ als „ü“ oder als „č“ oder als „š“ widerzugeben, oder auch ein bestimmtes Abkürzungszeichen stets aufzulösen bzw. durch etwas anderes zu ersetzen. Fehler in der Schreibung hingegen sollten keinesfalls korrigiert werden, ein Zeichen im Dokument sollte stets einem einzigen, gleichbleibenden Zeichen bzw. einer gleichbleibenden Zeichenkombination in der Transkription entsprechen.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

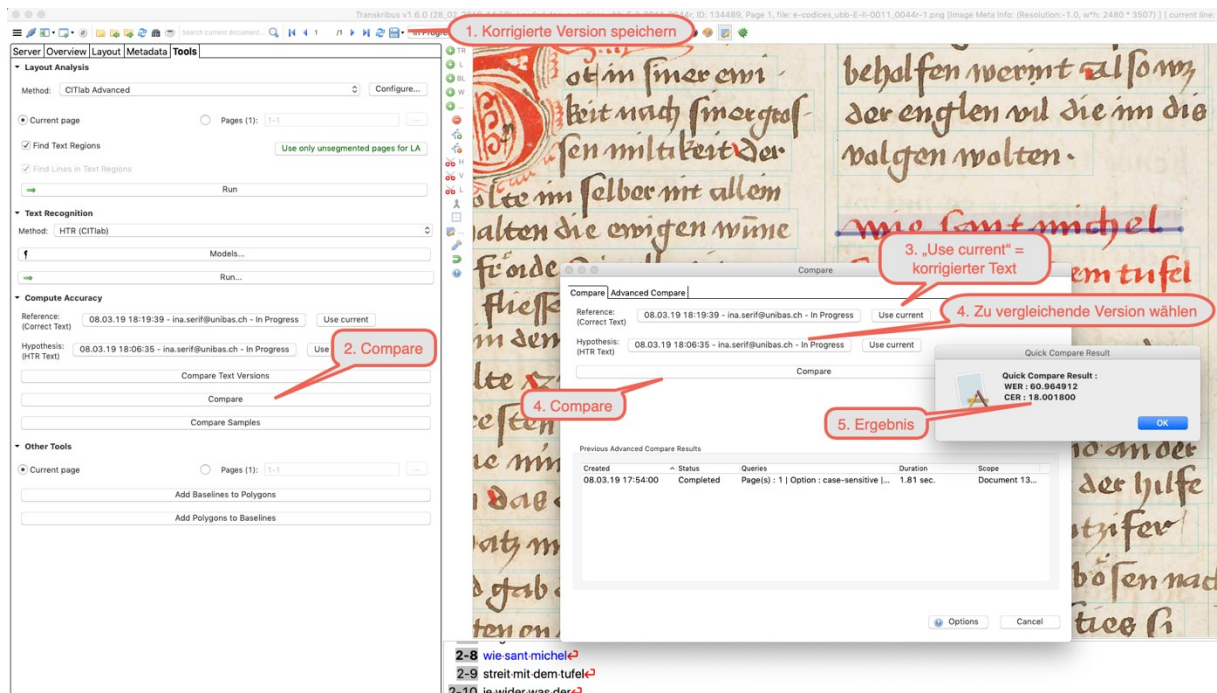


Abb. 27: Ergebnis der Berechnung der Fehlerquote (CER) eines Modells (eigener Screenshot, bearbeitet).

In diesem Dokument beträgt die CER 18% – eine fehlerfreie Transkription ist das nicht, aber das Training des Modells dürfte eine CER von unter 10% liefern. Ab einer Fehlerquote von weniger als 15% wird für die Korrektur der fehlerhaften Transkription weniger Zeit benötigt wird als für eine gänzlich neue Transkription; Spaß macht es aber eher erst ab unter 10% – zuvor hat man meist den Eindruck, die Korrektur mache mehr Arbeit als eine komplett neue Transkription.³⁰ Allein aber dadurch, dass der vorhandene Text, auch wenn er sehr fehlerhaft ist, doch einen gewissen inhaltlichen Kontext liefert, wird die Erfassung des Geschriebenen m.E. erheblich erleichtert, die Transkription also beschleunigt.

Unter „Compute Accuracy“ kann man sich mittels der Funktionen „Compare Text Versions“ und „Compare“ einen Eindruck über die Art/Häufigkeit der Fehler machen:

³⁰ Die Zahlen folgen aus Beobachtungen, die an der Universität Innsbruck in mehreren Anwendungsübungen des Programms gemacht wurden, und sind als grobe Orientierung zu verstehen.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



1. „Use current“ = korrigierter Text

2. Zu vergleichende Version wählen

3. Compare Text Versions

Abb. 28: Übersicht über die Art der Fehler bei der Texterkennung (eigener Screenshot, bearbeitet).

Bei einem Vergleich ist zum einen darauf zu achten, die richtigen Versionen auszuwählen – für die „Compare“-Variante und die „Compare-Text-Versions“-Variante müssen die Versionen jeweils wieder neu definiert werden („Use current“ und fehlerhafte erste Version). Zum anderen ist zu bedenken, dass das Ersetzen ganzer Zeilen bei der Korrektur der Transkription statt das ausschließliche Korrigieren von einzelnen Fehlern höhere Fehlerraten anzeigt: In Zeilen 1–19 bis 1–21 beispielsweise wurden keineswegs alle Buchstaben falsch erkannt, jedoch schien das

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



schnelle Löschen der ganzen Zeile und das Einfügen des korrekten Textes weniger Zeit zu beanspruchen als das mühselige Ersetzen einzelner falscher Buchstaben. Die CER dieses Modells fällt tatsächlich also etwas niedriger als die errechneten 18% aus.

Für all diejenigen, die bereits über (Teil-)Transkriptionen von Dokumenten verfügen, sei auf die Möglichkeit hingewiesen, vorhandenen Text mit dem zugehörigen Digitalisat zu verknüpfen (*text to image*), also im Nachhinein aus schon geleisteter Transkriptionsarbeit ein Modell zu erstellen und für weitere Dokumente zu nutzen.³¹

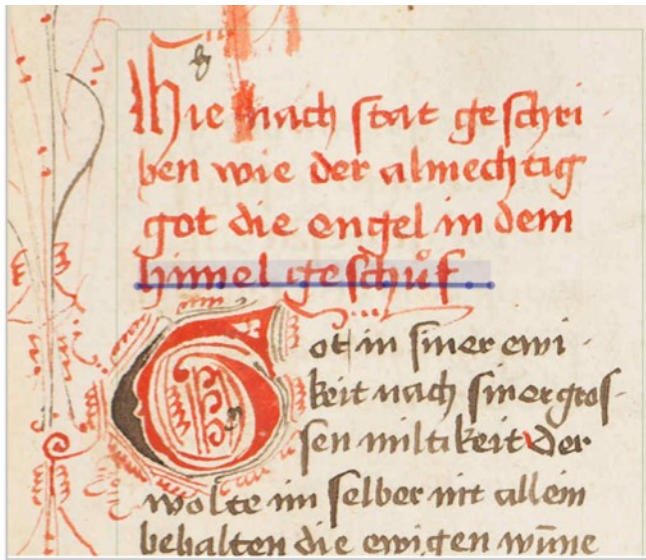
4. Durchsuchbarkeit von Quellen mit *Keywordspotting*

Neben der Aussicht, nach einem Training des Modells eine weitaus niedrigere Fehlerrate zu haben, kann mit der aktuell erzielten Rate bereits sinnvoll mit dem Text gearbeitet werden. Grundlage hierfür ist die (nicht sichtbare) Speicherung von möglichen Buchstabenvarianten bei der Textanalyse auf den Programmservern, auf die mittels einer Suchanfrage zurückgegriffen werden kann – die Texterkennung zeigt dem/der Nutzer*in im Textfeld immer diejenige Transkription an, die vom Modell als am wahrscheinlichsten eingestuft wird; mögliche Varianten werden aber im Hintergrund abgelegt und gespeichert. Das erste Wort in Zeile 4, „himel“, wurde von dem Modell als „umel“ wiedergegeben – man will es ihm nicht verübeln:

³¹ Auch Annotationen in TEI können so weitergenutzt werden, vgl. Maximilian Bryan, Tobias Hodel und Nathanael Philipp, Generierung von Trainingsdaten für die Handschrifterkennung aus TEI annotierten Dokumenten – Ein Erfahrungsbericht aus dem EU-Projekt READ, in: *INF-DH 2018 – Workshopband*, hrsg. von Manuel Burghardt und Claudia Müller-Birn, Bonn 2018, DOI: [10.18420/infdh2018-11](https://doi.org/10.18420/infdh2018-11).

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



- 1 ie·huach·sta·ge·schzn↵
- 2 ben·wie·der·almechtig↵
- 3 got·die·engel·in·dem↵
- 4 umel·ige·schuf↵

Abb. 29: Erkannter Text nach Anwendung des Beispielmotells „Comb_Gothic_Script“ (eigener Screenshot, bearbeitet).

Weitere Schreibweisen – lesbar wären neben „himmel“ vielleicht auch „lzimel“, „hunel“ oder „hinuel“ – sind zwar nicht sichtbar, können aber über das sogenannte *Keywordspotting* im Programm gesucht werden. Denn bei einer Suche über *Keywordspotting* wird nicht nur der Transkriptionsoutput, also der angezeigte Text, durchsucht, sondern auch die hinterlegten möglichen Varianten bzw. Kombinationen, eine *fuzzy search*, die aber auf tatsächlich erkannten Zeichen beruht.

Über das *Keywordspotting* findet sich auf dieser Seite auch der als „umel“ wiedergegebene „himmel“:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.

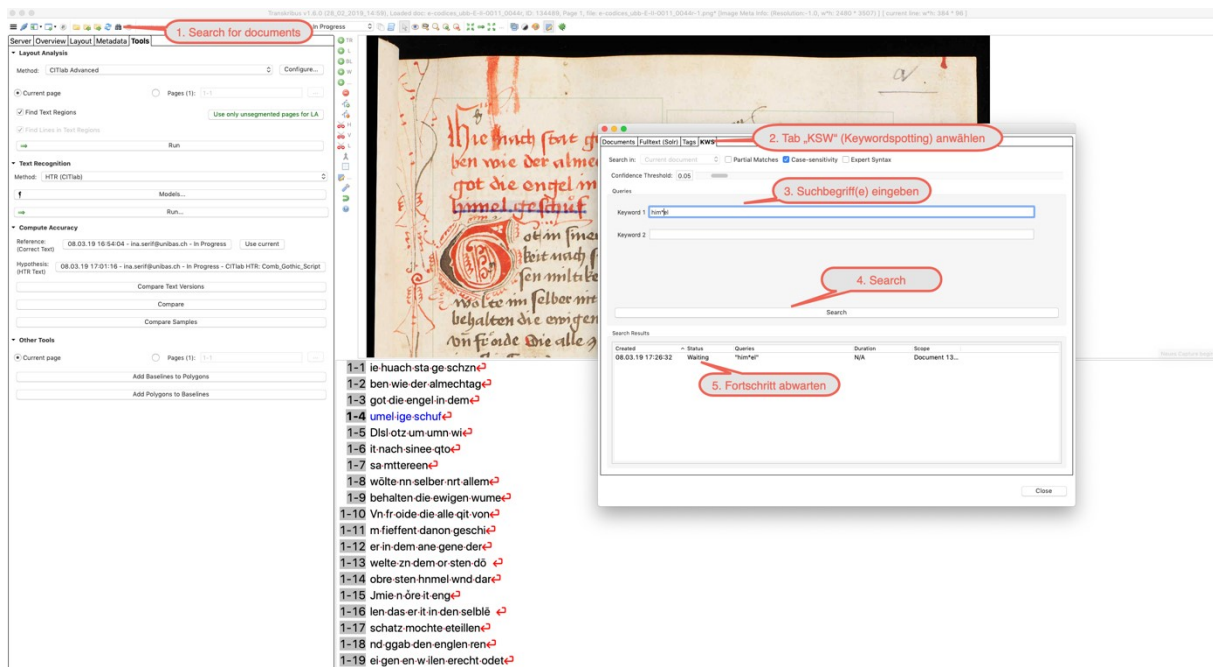


Abb. 30: Start einer Suchanfrage mit Keywordspotting (eigener Screenshot, bearbeitet).

„himel“ wurde bei der Suchanfrage trunkiert („him*el“), damit auch Schreibungen wie „him-mel“ gefunden werden.

Die Ergebnisse aus der Suche kann man sich nach Abschluss durch Doppelklick auf den Such-eintrag anzeigen lassen:

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



1. „Completed“ doppelklicken

2. Fundstelle mit Kontext durch Klick anzeigen lassen

Confidence	Page Nr.	Line transcription	Preview
0.1897	1	obre sten hnmel wnd dar	
0.0963	1	Her cib von dem humelt	
0.0511	1	umel ige schuf	

Search Results

Created 08.03.19 17:26:32 Status Completed

Preview

Close

Abb. 31: Anzeige der Suchergebnisse des Keywordspotting (eigener Screenshot, bearbeitet).

Der Text unter „Line transcription“ zeigt in der dritten Zeile weiterhin die fehlerhafte Lesart „umel“ an; die Kombination der einzelnen Zeichen zu „himel“ ist aber als Lesart vorhanden, auch wenn die „Confidence“, also die berechnete Wahrscheinlichkeit, sehr niedrig ist.³²

Da die Rechenleistung des *Keywordspotting* auf den Projektservern erfolgt, kann eine Suchabfrage je nach Länge des Dokuments etwas Zeit in Anspruch nehmen. Die Suche nach „der“ in einer anderen Handschrift mit 500 erkannten Textseiten dauerte knapp zwei Minuten für rund 18.000 Treffer; die Suche nach „himel“ benötigte für 247 Treffer rund eine Minute. Die Zeit kann je nach Serverauslastung variieren.

³² Die abgeschlossenen Suchanfragen werden in der Ergebnisanzeige gespeichert, man erhält also nebenbei auch eine Versionierung der eigenen Suchvorgänge.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Treffer im Dokument

Ergebnisse

Mit Doppelklick auf den Eintrag Sprung zur Seite

Suchdauer

Confidence	Page Nr.	Line transcription	Preview
0.5169	168	gett vonr hune, das er	
0.4603	73	dem kay ser is dz bimmel.	
0.4319	87	nidie lase der humel end	
0.3592	330	von himmel bracht dem sel	
0.3506	87	dar hinel und der eed	
0.3504	321	der aunem kostlichen hüel	
0.3379	400	ich zú Hinel Der bsch	
0.3344	125	dich got her von hiel	
0.3292	168	or hünel her gett hu	
0.3097	122	tes stnmd von Hünel zu	
0.3016	122	von imnel einbäet dir dz	
0.2945	125	gott von Hel ics bet	
0.2940	95	hoffert onemeh zechümel	
0.2893	151	won hine schlig, dz dier	

Suchtreffer „himmel“ im Transkriptionsoutput

4-25 z.was spechen oder ant
4-26 sünen mocht begeben
4-27 wereden dz hant ech
4-28 gestens als mnsinen buchen
4-29 geschriben br flog mnit
4-30 saner ranst als ein odler
4-31 nidie lase der humel end
4-32 beschrift, die geligen hrit

Abb. 32: Anzeige der Ergebnisse und Dauer des Keywordspotting in einem größeren Dokument (eigener Screenshot, bearbeitet; Handschrift im Beispiel: ms. Stuttgart, Württembergische Landesbibliothek, HB V 22, fol. 42rb, Digitalisat: <http://digital.wlb-stuttgart.de/purl/bsz393684989>, Lizenz: keine (Public Domain).

Für eine Trefferquote von über 99,5% für ein Suchwort – das bedeutet, 99,5% der tatsächlichen Fundstellen eines Wortes werden erkannt – sollte die CER maximal 15% betragen; eine CER unter 25% liefert immer noch eine Trefferquote von über 99%. Mit oben verwendetem Modell ließe sich die Beispielhandschrift also bereits sehr gut nach Worten/Wortfolgen durchsuchen.

Zwar ist der Schritt vom *distant* zum *close reading* für große Textmengen für vormoderne Texte auch mit *Keywordspotting* noch nicht ganz getan – die fehlende Standardisierung in der Schreibung erfordert weitere Schritte. Aber bereits die Nutzung gezielter Suchanfragen eröffnet ganz

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



neue Möglichkeiten für die Arbeit mit unedierten Quellen, ebenso wie das unproblematische kollaborative Arbeiten, sei es in geschlossenen oder in öffentlichen Sammlungen.³³

5. *Transkribus* – eine eierlegende Wollmilchsau?

Die in diesem Tutorial aufgezeigten Funktionen von *Transkribus* schöpfen die tatsächlichen Möglichkeiten des Programms bei Weitem nicht aus. So stehen beispielsweise unter dem Tab „Tagging“ vordefinierte und anpassbare Tags zur Auszeichnung von Text mit zusätzlichen Informationen zur Verfügung, was nicht nur für die Suche innerhalb eines Dokuments genutzt werden kann (Suche nach Tags wie „Ort“, „Person“, „Abkürzung“, „Tier“, ..., ebenso wie nach Inhalten von Tags wie „Köln“, „Peter“, „etc.“, „Pferd“), sondern auch für die Normalisierung unterschiedlicher Schreibweisen.³⁴

Neben *Transkribus* existieren weitere Texterkennungsprogramme, die ihre eigenen Vorteile haben.³⁵ Einer der größten Vorzüge von *Transkribus* ist sicher die Verbindung unterschiedlicher Funktionen in einem einzigen Programm, ohne dass spezielle technische Voraussetzungen nötig wären. Neben reiner Transkriptionsarbeit, die durch die Verknüpfung von Text und Bild erheblich erleichtert wird, trägt das Anwenden und Trainieren von Texterkennungsmodellen dazu bei, diese für die gesamte Nutzer*innengemeinschaft stetig zu verbessern: Jedes verfeinerte bzw. neue Modell erhöht dessen Nutzern respektive verringert den Trainingsaufwand für andere Dokumente. Die Durchsuchbarkeit von größeren Quellenbeständen über das *Keywordspotting* erweitert die Zugriffs- und Forschungsmöglichkeiten auf historische Dokumente. Die Verwaltung von Digitalisaten in Sammlungen, die auch mit anderen geteilt und bearbeitet werden können, ist ein weiterer Vorzug, ebenso wie die verschiedenen Exportmöglichkeiten, wodurch keine langfristige Abhängigkeit zu dem Programm entsteht und die Ergebnisse auch in anderen Kontexten genutzt werden können. Der Fortbestand und die Weiterentwicklung des

³³ Öffentliche Sammlungen sind über den Webclient zugänglich.

³⁴ Auch hierzu gibt es einen Abschnitt in der Benutzeranleitung, siehe <https://transkribus.eu/wiki/De/index.php/Benutzeranleitung> unter „Tags finden“.

³⁵ Im Bereich von *free-* und *open-source*-Entwicklungen seien hier beispielsweise *OCROpus* (siehe <https://github.com/tmbdev/ocropy>), *Calamari* (siehe <https://github.com/Calamari-OCR/calamari>), *Tesseract* (siehe <https://github.com/tesseract-ocr/tesseract>), *OCR4all* (siehe <https://github.com/OCR4all>) sowie *OCR-D – Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition* (siehe <http://ocr-d.de/>) genannt, im Bereich der Handschriftenerkennung sollte *Kraken* (siehe <http://kraken.re/>) nicht unbeachtet bleiben. Einen Vergleich einiger dieser Programme gibt es in Ted Han und Amanda Hickmann, Our search for the best OCR tool, and what we found, in: *Source*, 19. Februar 2019, <https://source.opennews.org/articles/so-many-ocr-options/>.

Zitation:

Ina Serif, Ein Wolpertinger für die Vormoderne: Zu Nutzungs- und Forschungsmöglichkeiten von Transkribus bei der Arbeit mit mittelalterlichen und frühneuzeitlichen Handschriften und Drucken, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* 2 (2019), S. 125–166, <https://mittelalter.hypotheses.org/21828>.



Programms nach Ende der Förderungsdauer in diesem Jahr dürfte durch die Überführung in eine Kooperative gesichert sein. Wie sich die finanzielle Beteiligung von teilnahme willigen Institutionen gestaltet, wird sich zeigen; es wäre sehr erfreulich, wenn die Absicht, die Nutzung für Projekte bis zu einer bestimmten Seitenzahl weiterhin kostenlos anzubieten, garantiert werden könnte.

Alle angegebenen Links wurden am 1. Mai 2019 geprüft.

Verwendete Handschriften:

ms. Basel, Universitätsbibliothek, E II 11, fol. 44r (verwendet in Abb. 6–9, 11–18, 20–31). Digitalisat via [DOI: 10.5076/e-codices-ubb-E-II-0011](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63862-p0011-9). Lizenz: [CC0 1.0](https://creativecommons.org/licenses/by-sa/4.0/).

Beschreibung unter https://aleph.unibas.ch/F/?local_base=DSV05&con_lng=GER&func=find-b&find_code=SYS&request=000117859. Die Handschrift enthält die Straßburger Chronik Jakob Twingers von Königshofen, fol. 44r den Beginn des ersten Kapitels.

Frankfurt, Universitätsbibliothek, Ms. germ. qu. 23, fol. 30v (verwendet in Abb. 10); Digitalisat via URN: [urn:nbn:de:hbz:30:2-14660](https://nbn-resolving.org/urn:nbn:de:hbz:30:2-14660); Lizenz: keine (Public Domain).

Beschreibung unter [urn:nbn:de:hbz:30:2-282-p0074-3](https://nbn-resolving.org/urn:nbn:de:hbz:30:2-282-p0074-3). Die Handschrift enthält ebenfalls die Twinger-Chronik, fol. 30v den Beginn des zweiten Kapitels.

ms. Stuttgart, Württembergische Landesbibliothek, HB V 22, fol. 42rb (verwendet in Abb. 32), Digitalisat via: <http://digital.wlb-stuttgart.de/purl/bsz393684989>; Lizenz: keine (Public Domain).

Beschreibung unter http://bilder.manuscripta-mediaevalia.de/hs//katalogseiten/HSK0069_b037.JPG.htm. Die Handschrift enthält u.a. Auszüge aus der Twinger-Chronik, fol. 42r enthält eine Stelle aus dem zweiten Kapitel.