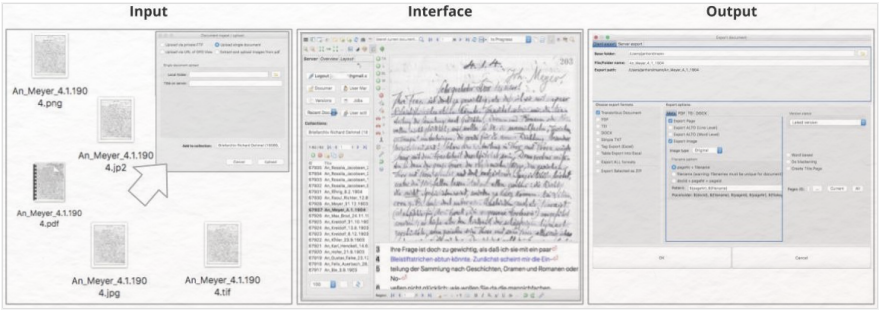
Toolbeitrag: Transkribus

Jan Horstmann [1]

1. Universität Münster

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3746
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2018-10-15 auf fortext.net
Lizenz:	CC BY SA		open ∂ access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.



Der Workflow von Transkribus: Laden Sie Ordner mit einseitigen oder mehrseitigen PDF-Dateien oder auch Bilddateien (JPEG, PNG, TIFF, JP2) hoch, lassen Sie die Linien im Manuskript bestimmen und transkribieren Sie. Das Transkript kann z. B. als PDF, TEI-konformes XML, als DOCX, TXT etc. herunter geladen und weiter verwendet werden.

Hinweis: Der folgende Tooleintrag bezieht sich auf die 2018 verfügbare Transkribus-Version. Das Tool hat sich seither stark weiterentwickelt.

- **Systemanforderungen:** Desktopbasiert, benötigt Internetverbindung für Serverzugriff, kann offline mit lokalen Daten verwendet werden, unterstützt alle Betriebssysteme, benötigt Java Runtime Environment
- **Stand der Entwicklung:** Seit 2016, wird weiter entwickelt
- **Herausgeber:** Universität Innsbruck
- **Lizenz:** Kostenfrei, aber nicht Open Source
- **Weblink:** https://www.transkribus.org (eine verschlankte Webversion eignet sich bei Bedarf für kurzfristigere Transkriptionsaufgaben)
- **Im- und Export:** Transkribus-Dokument, Excel-Datei, PDF, TEI-XML, DOCX, TXT (vgl. Reintext-Version); nur Import: JPEG, PNG, TIFF, JP2
- **Sprachen:** Niederländisch, Englisch, Finnisch, Französisch, Deutsch, Schwedisch, Polnisch, Dänisch etc. Für mehr Informationen: https://readcoop.eu/transkribus/public-models/

1. Für welche Fragestellungen kann Transkribus eingesetzt werden?

Das Kerngeschäft von Transkribus ist die Digitalisierung (Horstmann 2024a) von Handschriften, d. h. das manuelle Transkribieren und die automatisierte Handschriftenerkennung (HTR) (Horstmann 2024b). Zusätzlich wird auch eine optische Zeichenerkennung (OCR) für Druckschriften angeboten. Editionswissenschaftliche Projekte können in Transkribus ausgeführt werden, die Digitalisierung kann aber auch als Vorbereitung für eine Weiterverarbeitung der Texte mit anderen digitalen Tools dienen. Transkribus bietet grundsätzlich auch die Möglichkeit, die erstellten Transkripte nach selbst gewählten Kategorien zu annotieren (vgl. Annotation) und größere Textmengen (vgl. Korpus) nach diesen Kategorien zu durchsuchen.

2. Welche Funktionalitäten bietet Transkribus und wie zuverlässig ist das Tool?

Transkribus bietet eine Vielzahl von Funktionen, darunter:

- Archivierung von Textsammlungen und zugehöriger Scans oder Transkriptionen; Anreicherung mit Metadaten
- Automatische und manuelle Segmentierung des Textes
- Tagsetzung (vgl. Tagset), Kommentierung und Annotation
- Transkription
- Nutzung automatischer HTR-Funktionen für deutsch- und englischsprachige Texte
- Training (vgl. Machine Learning) eines eigenen HTR-Modells für eine bestimmte Schrift
- OCR (Funktionen von Abbyy FineReader (Schumacher 2024)): Einlesen von lateinischer Schrift, Fraktur und Mischformen in verschiedenen Sprachen
- Fehlerquotenmessung von HTR und OCR

Grundvoraussetzung für die Nutzung ist, dass hochwertige Scans verwendet werden. Für diesen Zweck hat Transkribus das ScanTent für perfekte Scan-Bedingungen und die Android-App DocScan (Kleber u. a. 2017) für einen einfachen Upload der Dateien in Ihren Transkribusaccount entwickelt. Die Leistung wird damit zuverlässig und auch vergleichsweise schnell. Auch komplexere Layouts (wie bspw. Tabellen oder Texte mit mehreren Spalten) können vom HTR- und OCR-Programm häufig automatisch richtig erfasst werden. Die HTR bietet zudem die Möglichkeit einer manuellen Auszeichnung der Zeilen und ihrer Abfolge.

3. Ist Transkribus für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / -
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	teilweise
Leichter Einstieg	-
Handbuch vorhanden	✓
Handbuch aktuell	teilweise
Tutorials vorhanden	✓
Erklärung von Fachbegriffen	teilweise
Gibt es eine gute Nutzerbetreuung?	✓

Transkribus' grafische Nutzeroberfläche (GUI) (vgl. GUI) ist sehr komplex und ohne Einführung nur wenig intuitiv nutzbar. Über die vielen Funktionen können Sie sich im englischen oder deutschen WIKI einen Überblick verschaffen. Fachbegriffe werden dort größtenteils kurz erklärt, allerdings mit Ausnahmen wie z. B. der Unterschied zwischen *line* und *baseline*. Ein deutsches Benutzerhandbuch erklärt die Benutzeroberfläche zwar en detail, bezieht sich jedoch auf eine ältere Version und ist daher in einigen Punkten veraltet. Anfragen per Mail beantwortet das Transkribus-Team i. d. R. zügig und ausführlich.

4. Wie etabliert ist Transkribus in den (Literatur-)Wissenschaften?

Für Transkriptionsprojekte ist Transkribus europaweit die erste Anlaufstelle und viele Editionen werden mit Transkribus-Unterstützung erstellt. Laut Aussagen von Transkribus sind unter 55 zur Zeit aktiv laufenden Projekten 16 Editionsprojekte; 10 geplante Projekte haben noch nicht begonnen und weitere 6 haben Interesse bekundet (Stand Juli 2018).

5. Unterstützt Transkribus kollaboratives Arbeiten?

Ja. Textsammlungen (Collections) können mit anderen Nutzer*innen einzeln geteilt und dann gemeinsam transkribiert und annotiert (vgl. Annotation) werden. Nach dem Speichern der Transkriptionen und sonstiger Metadaten werden diese den anderen Nutzer*innen der jeweiligen Collection automatisch zugänglich gemacht. Gemeinsam können zudem Textsammlungen erweitert und Transkriptionsrichtlinien erstellt werden.

6. Sind meine Daten bei Transkribus sicher?

Ja. Beim Erstellen eines Accounts ist die Angabe Ihres Namens, der Mailadresse und eines Passwortes nötig. Bei der Registrierung wird zudem die IP-Adresse abgerufen und geschützt gespeichert. Auch Trainingsdaten werden erhoben, dies jedoch vor allem für die Verbesserung der HTR-Funktion und ohne dass ein Zugriff auf die Dokumente selbst stattfände. Dies geschieht, um die tooleigene HTR-Funktion stetig zu verbessern und

zukünftig Handschriften digitalisieren zu können, ohne jeweils ein eigenes Training vorschalten zu müssen. Es ist kein Widerspruch möglich, die Daten werden jedoch wieder gelöscht, wenn Sie Ihren Account löschen.

Hochgeladene Texte werden auf einem Server der Universität Innsbruck gespeichert. Die Texte befinden sich in einem geschützten Login-Bereich und sind nur durch diejenigen Transkribus-Nutzer*innen einsehbar, denen Sie Zugriff geben. Alternativ können Sie offline mit lokalen Daten arbeiten.

Externe und weiterführende Links

- Transkribus: https://web.archive.org/save/https://transkribus.eu/Transkribus/ (Letzter Zugriff: 04.06.2024)
- Transkribus Benutzerhandbuch: https://web.archive.org/save/https://help.transkribus.org/de/erste-schritte (Letzter Zugriff: 04.06.2024)
- Deutsches Transkribus-Wiki: https://web.archive.org/save/https://help.transkribus.org/de (Letzter Zugriff: 04.06.2024)
- Englisches Transkribus-Wiki: https://web.archive.org/save/https://web.archive.org/save/https://help.transkribus.org (Letzter Zugriff: 04.06.2024)
- Transkribus und andere Sprachen: https://readcoop.eu/transkribus/public-models/ (Letzter Zugriff: 04.06.2024)
- Transkribus Webversion: https://web.archive.org/save/https://transkribus.eu/read/login/?next=/read/library/ (Letzter Zugriff: 04.06.2024)
- ScanTent: https://web.archive.org/save/https://scantent.cvl.tuwien.ac.at/en/ (Letzter Zugriff: 04.06.2024)

Bibliographie

Colutto, Sebastian, Günther Hackl, Philip Kahle und Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In:, 19–24. Kyoto, Japan. doi: 10.1109/ICDAR.2017.307, ttps://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8270253 (zugegriffen: 17. September 2018).

Horstmann, Jan. 2024b. Methodenbeitrag: Digitale Manuskriptanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3744, https://fortext.net/routinen/methoden/digitale-manuskriptanalyse.

———. 2024a. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

Kleber, Florian, Markus Diem, Fabian Hollaus und Stefan Fiel. 2017. Mass Digitization of Archival Documents using Mobile Phones. In: *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, 65–70. https://dl.acm.org/citation.cfm?doid=3151509.3151526 (zugegriffen: 17. September 2018).

Mühlberger, Günter und Tamara Terbul. 2018. Handschriftenerkennung für historische Schriften. Die Transkribus Plattform. *b.i.t. online* 21, Nr. 3: 218–222. https://www.b-i-t-online.de/heft/2018-03/fachbeitrag-muehlberger.pdf (zugegriffen: 17. September 2018).

Schumacher, Mareike. 2024. Toolbeitrag: Abbyy FineReader. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3742, https://fortext.net/tools/tools/abbyy-finereader.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv` , sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

HTR HTR steht für *Handwritten Text Recognition* und ist eine Form der Mustererkennung, wie auch die OCR. HTR bezeichnet die automatische Erkennung von Handschriften und die Umformung dieser in einen elektronischen Text. Die Automatisierung beruht auf einem Machine-Learning-Verfahren.

IP-Adresse Die Vernetzung von Computern wird in einem Internetprotokoll (IP) festgehalten, woraufhin jedes angebundene Gerät in diesem Computernetz eine IP-Adresse erhält. So werden die Geräte adressierbar und erreichbar gemacht. Die IP gehört zu den personenbezogenen Daten, da über sie auf Ihre Identität geschlossen werden kann.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

WIKI Ein Wiki ist eine Webseite mit einer Sammlung von Informationen zu ausgewählten Themen, die i. d. R. von mehreren Nutzer*innen zusammengestellt werden. Zu jedem Eintrag in einem Wiki gibt es eine Diskussionsseite, die auch frühere Versionen des Eintrags anzeigt.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.