Lerneinheit: Manuskriptdigitalisierung mit Transkribus

Jan Horstmann [1]

1. Universität Münster

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3745
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-03-18 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.

Eckdaten der Lerneinheit

- Anwendungsbezug: Briefmanuskript von Richard Dehmel
- Methode: Manuelle digitale Transkription eines Manuskripts
- Angewendetes Tool: Transkribus
- Lernziele: Download eines eingescannten Briefes, Installation und Nutzung des Tools, Export des erzeugten Transkripts
- Dauer der Lerneinheit: ca. 120 Minuten
- Schwierigkeitsgrad des Tools: mittel

Bausteine

- Anwendungsbeispiel
  Welchen Brief transkribieren Sie? Erstellen Sie ein digitales Transkript eines Briefes von Richard Dehmel an Rainer Maria Rilke und verknüpfen Sie den elektronischen Text mit der Handschrift.
- Vorarbeiten
  Was müssen Sie tun, bevor es losgehen kann? Lernen Sie, wie man Transkribus installiert und einen Text hochlädt.
- Funktionen
  Welche Funktionen bietet Ihnen Transkribus zur digitalen Transkription von Handschriften? Lernen Sie die einzelnen Komponenten des Tools kennen und lösen Sie Beispielaufgaben.
- Lösungen zu den Beispielaufgaben
  Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie Antworten.

1. Anwendungsbeispiel

In dieser Lerneinheit werden wir am Beispiel eines Briefes des Schriftstellers Richard Dehmel an Rainer Maria Rilke lernen, wie man handschriftliche Manuskripte digital transkribieren kann. Handschriften stellen eine besondere Herausforderung für die automatische Digitalisierung dar (vgl. Möglichkeiten der Textdigitalisierung (Horstmann 2024a)), weil handgeschriebene im Gegensatz zu maschineller Schrift individuellen Mustern folgt und nicht nur zwischen unterschiedlichen Autor*innen, sondern häufig auch innerhalb eines Oeuvres oder gar einzelner Texte variiert. Wollen Sie handschriftliche Dokumente nicht nur als Bilddateien (d. h. als Scans) digitalisieren und mit Metadaten versehen, sondern auch eine Volltexttranskription vornehmen, um den Text elektronisch les- und durchsuchbar zu machen (vgl. Digitale Manuskriptanalyse (Horstmann 2024b)), bietet sich das Tool Transkribus (Horstmann 2024c) an. In diesem Tool wird die Handschrift direkt mit dem Transkript verknüpft, sodass es bei größeren Mengen von Manuskripten der gleichen Handschrift möglich ist, diese Schrift mit maschinellem Lernen (vgl. Machine Learning) zu „trainieren“ und damit automatisch transkribierbar (vgl. HTR) zu machen. Transkribus ist eine im Rahmen des europäischen *READ*-Projektes („Recognition and Enrichment of Archival Documents“) entwickelte, in der Basisversion kostenfrei nutzbare Forschungsplattform, die momentan mehr als 10.000 Nutzer*innen das Transkribieren, Erkennen und Untersuchen von historischen Dokumenten ermöglicht, ohne dass umfangreiches technisches Vorwissen vonnöten ist.

2. Vorarbeiten

In der Staats- und Universitätsbibliothek Hamburg befindet sich ein umfangreiches Briefarchiv des zu Beginn des 20. Jahrhunderts berühmten Schriftstellers Richard Dehmel. Die Briefe liegen teilweise bereits als digitale Scans vor, wurden bislang jedoch nicht transkribiert. Wir werden in dieser Lerneinheit beispielhaft einen Brief Dehmels an Rainer Maria Rilke vom 17. Januar 1906 transkribieren. Um sich den Brief als PDF herunterzuladen, folgen Sie dem Link. Sie sehen den eingescannten Brief, den Sie sich mit einem Klick auf „Gesamtwerk als PDF“ herunterladen können, nachdem Sie den Nutzungsbedingungen zugestimmt haben. Tipp: Wollen Sie Transkribus für Ihre eigenen Manuskripte verwenden und eine Handschrift trainieren, achten Sie unbedingt darauf, dass Sie ordentliche (hochauflösende) Scans verwenden, um beste Ergebnisse zu erzielen.

Abb. 1: Ansicht des Dehmelbriefes an Rilke vom 17.01.1906, Dehmelarchiv SUB Hamburg

Da man in Transkribus PDF-Dateien direkt hochladen kann, bedarf es keiner weiteren Vorbereitung (vgl. Preprocessing) des Dokumentes. Die Umformung eines eingescannten in einen computerlesbaren Text ist eine unumgängliche Voraussetzung für Methoden der digitalen Textanalyse, sollten die Texte nicht bereits digital vorliegen. Eine manuelle (vgl. Keying) oder automatische digitale Transkription findet daher im Zuge der Digitalisierung vor der digitalen Analyse statt.

Als nächstes installieren Sie sich das Transkribus-Tool und legen sich einen Account an, damit Sie ihre Dokumente und Transkriptionen jederzeit wiederfinden und verwalten können. Folgen Sie dazu diesem Link und klicken auf den für Ihr Betriebssystem passenden Download-Button.

Abb. 2: Downloadoptionen auf der Transkribus-Webseite

Im sich öffnenden Login-Fenster können Sie sich entweder mit einem Google-Account anmelden, oder sich unter „sign up“ registrieren.

Nach dem Login laden Sie sich das Tool durch einen erneuten Klick auf den entsprechenden Download-Button herunter und extrahieren (insofern das nicht von selbst geschieht) die ZIP-Datei (vgl. ZIP) auf ihrem Rechner. Sie können das Programm nun an einem Ort speichern, an dem Sie es leicht wiederfinden können. Öffnen Sie per Doppelklick das Programm (d. h. die Datei mit dem Transkribus-Logo). Sollte Ihr Computer Transkribus nicht öffnen wollen, liegt das vermutlich an Ihren Sicherheitseinstellungen. Wie Sie eine programmspezifische Ausnahme hinzufügen können, erfahren Sie in den Videos, die wir für Sie auf Zenodo bereitstellen (forTEXT 2019a; forTEXT 2019b).

Sie haben nun das Tool Transkribus erfolgreich installiert. Nach dem Öffnen des Programms klicken Sie oben links auf den „Login"-Button, wählen Ihren Accounttyp (Google oder Transkribus) und loggen sich ein. Sie sehen die Arbeitsoberfläche (vgl. GUI) von Transkribus (vgl. Abb. 3) und eine Mitteilung, welche neuen Funktionen in letzter Zeit erschienen sind. Diese können Sie schließen. Da Ihre Arbeitsschritte auf dem Transkribus-Server gespeichert werden, ist für die Arbeit mit dem Tool eine Verbindung mit dem Internet Voraussetzung.

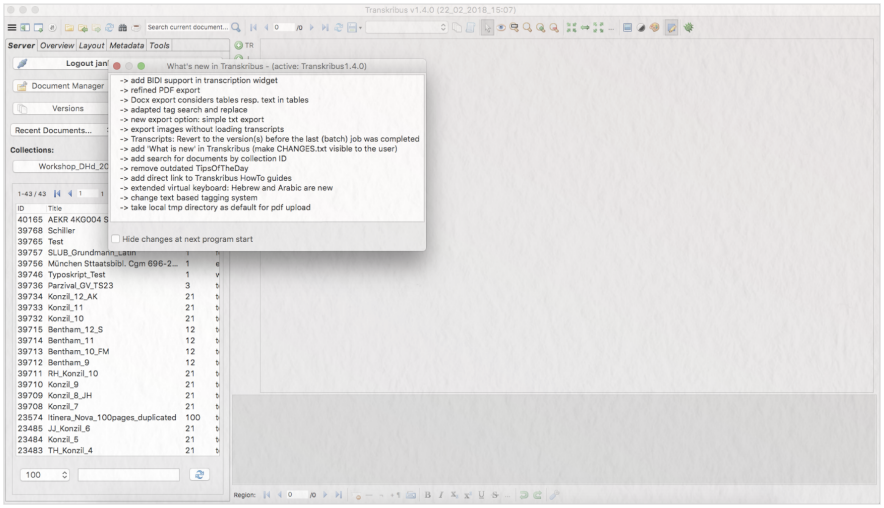


Abb. 3: Arbeitsoberfläche von Transkribus

In der oberen Leiste sehen Sie verschiedene Icons, mit denen Sie verschiedene Ansichten einstellen und diverse dokumentbezogene Aktionen durchführen können. Eine kurze Erläuterung erhalten Sie beim Hovern über das jeweilige Icon. Hier finden Sie auch das Icon für den **Dokumentimport**, auf das Sie nun klicken.

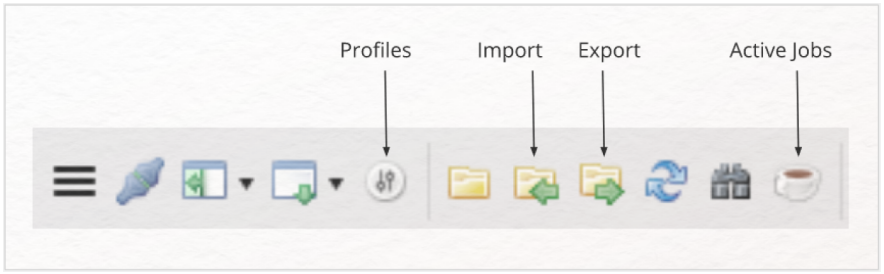


Abb. 4: Icons im Transkribus-Menü

Wählen Sie in dem sich öffnenden Fenster die Option „Extract and upload images from pdf" und suchen per Klick auf den Button rechts neben dem Eingabefeld die zuvor gespeicherte PDF-Datei des Dehmelbriefes. Bevor Sie schließlich den „Upload"-Button klicken, erstellen Sie unter „Add to collection" Ihre eigene Kollektion, in der dieser Brief gespeichert werden soll. Kollektionen (d. h. Sammlungen (vgl. Korpus)) helfen Ihnen bei der Organisation zusammengehörender Dokumente.

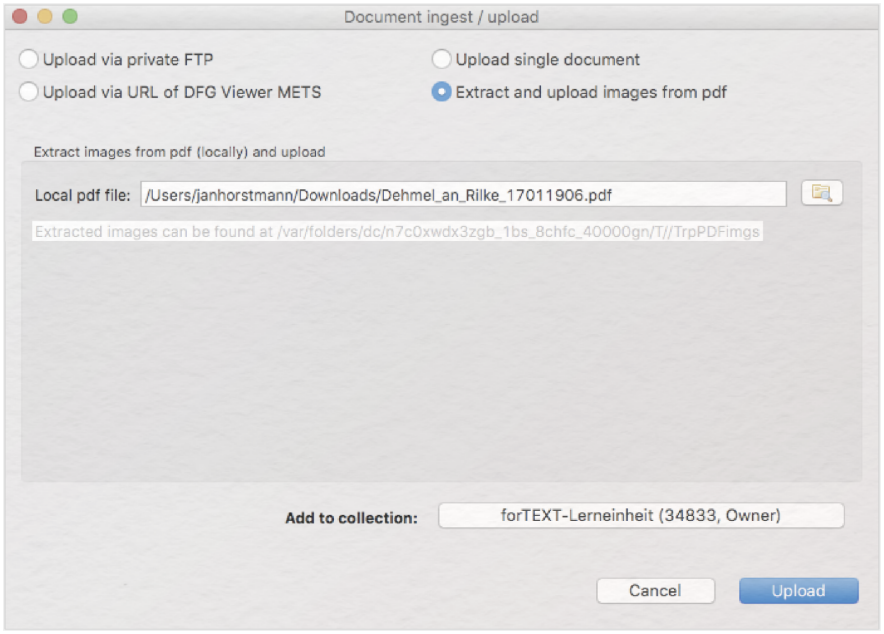


Abb. 5: Dokumentupload in Transkribus

Die Datei wird nun hochgeladen und Sie erhalten eine Meldung, wenn der Upload abgeschlossen wurde. In der linken Spalte der Benutzeroberfläche können Sie ihre Kollektion auswählen, unter der dann die enthaltenen Dokumente (in diesem Fall nur der Brief Dehmels) aufgeführt werden. Da hochgeladene Dateien zunächst vom Transkribus-Server verarbeitet werden müssen, kann es einige Minuten dauern, bis das Dokument angezeigt wird. Klicken Sie dafür hin und wieder auf den „Reload"-Button unten rechts in dieser linken Spalte (siehe Abb. 6). Tipp: Sie können in Transkribus eine Liste der bereits abgeschlossenen und noch aktiv ausgeführten Jobs mit einem Klick auf das Kaffeetassen-Symbol in der Leiste oben links einsehen (vgl. Abb. 4).

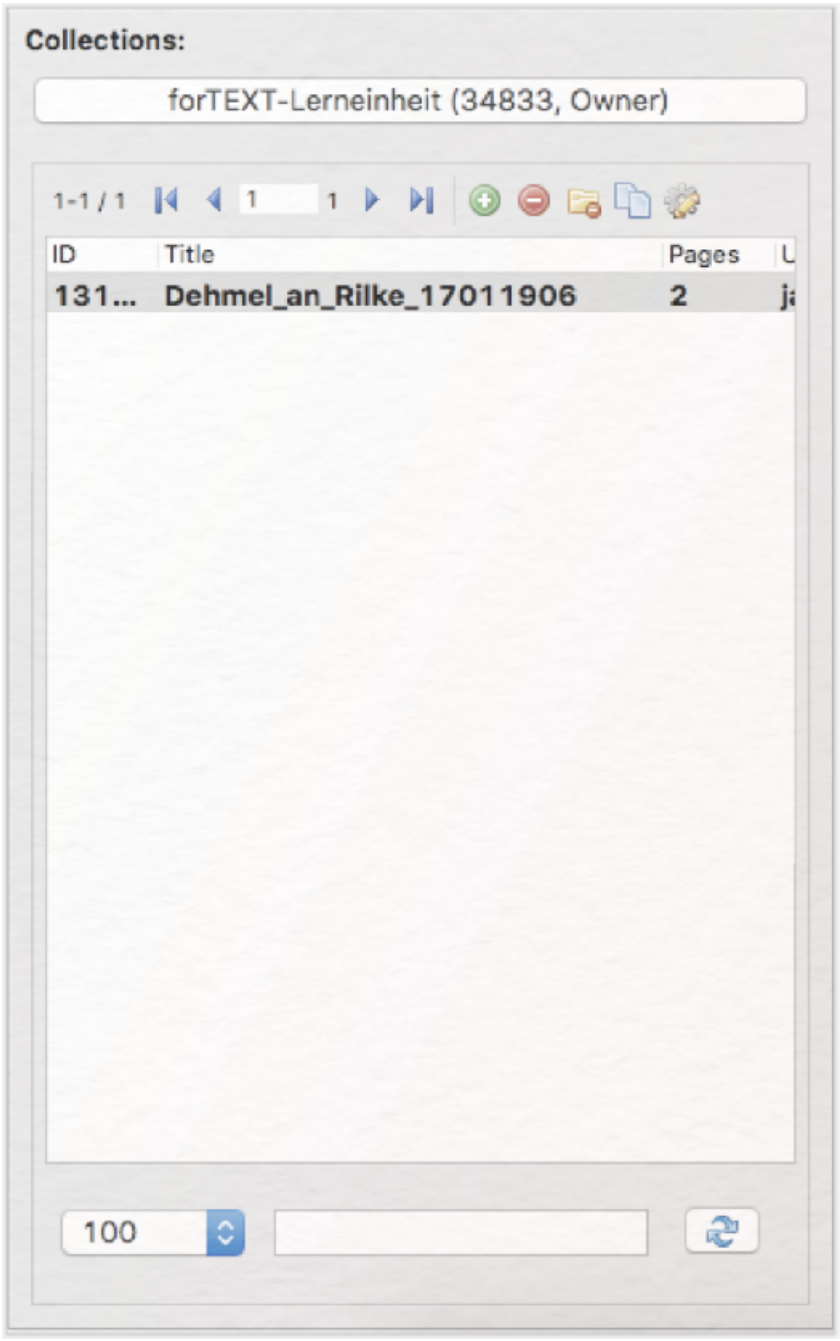


Abb. 6: Dokumente in einer Transkribus-Kollektion

Ein Doppelklick auf den schließlich erscheinenden Dokumentnamen öffnet das Manuskript auf der rechten Seite. Der Button „Document Manager“ bietet Ihnen die Möglichkeit, die Dokumente in einer Kollektion zu verwalten und bspw. einzelne Seiten aus einem Dokument zu löschen. Sollten Sie Manuskripte haben, die noch nicht als Scans vorliegen, empfiehlt sich die von Transkribus entwickelte und kostenlos nutzbare Android-App DocScan, die eingescannte Manuskripte zeitsparend direkt Ihrem Transkribus-Account hinzufügt. Sie haben nun ein Manuskript in Transkribus hochgeladen und sind damit bereit für die digitale Transkription!

3. Funktionen

Transkribus stellt eine direkte technische Verknüpfung zwischen den einzelnen Zeilen des Manuskripts (das als Bilddokument noch nicht elektronisch lesbar ist) mit der entsprechenden manuell eingegebenen Transkription her: die Grundvoraussetzung für ein maschinelles Lernen, das eine automatische Erkennung der Handschrift möglich machen soll. Um sicherzustellen, dass der transkribierte Text immer dem richtigen Text auf der jeweiligen Manuskriptseite zugeordnet werden kann, müssen auf jeder neuen Manuskriptseite zunächst immer die einzelnen Textbereiche und die Zeilen definiert werden. Das Tool bietet hierfür bei regelmäßig geschriebenen

Manuskripten eine verhältnismäßig zuverlässige automatische Unterstützung; kompliziertere Layouts (z. B. Listen oder Ergänzungen zwischen den Zeilen oder am Rand des Textes) bedürfen häufig einer manuellen Korrektur. Als Menschen können wir leicht erkennen, an welcher Stelle des eigentlichen Textes bspw. eine Ergänzung am Rand gelesen werden soll. Dem Tool müssen Sie diese Lesereihenfolge explizit beibringen, um eine korrekte Referenzierung von Bildausschnitt/Manuskriptstelle und Transkript garantieren zu können.

Für diesen, auch **Segmentierung** genannten Arbeitsschritt, wählen Sie zunächst das „Segmentation"-Profil wie in Abb. 7 gezeigt in Transkribus aus.

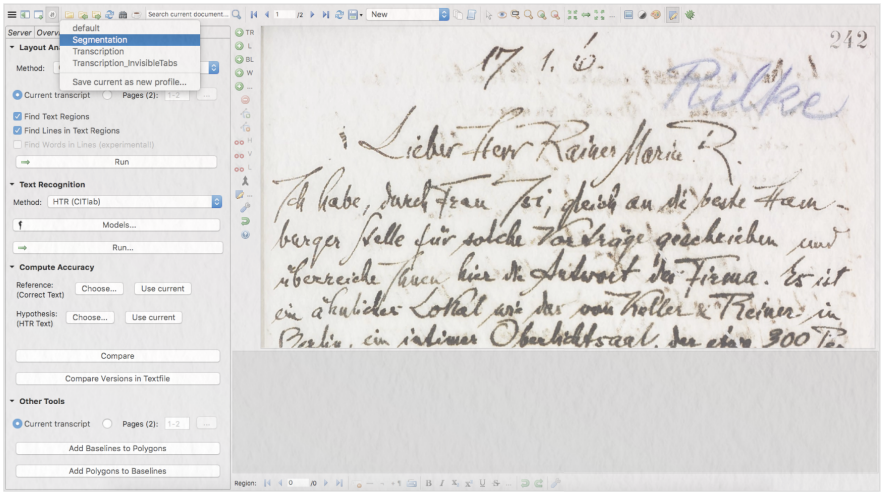


Abb. 7: Profile in Transkribus wechseln

Unter dem Profile-Icon finden Sie mehrere Tabs, von denen Sie nun „**Tools**" auswählen. Hier finden Sie zahlreiche Funktionen bspw. für die Layoutanalyse oder die Texterkennung. Uns interessiert in dieser Lerneinheit hiervon besonders die **Layoutanalyse**. Das unter diesem Punkt vorausgewählte „Current transcript" wird die Zeilen auf der aktuell aufgerufenen ersten Seite des geladenen Manuskripts als solche auszeichnen. Aktivieren Sie rechts daneben die andere Option „Pages", und begrenzen Sie den Seitenrahmen auf die beiden Briefseiten, damit das Layout auf beiden Seiten analysiert wird. Außerdem wird Transkribus in diesem Schritt Textregionen und die Zeilen innerhalb dieser Textregionen finden (automatisch vorausgewählt sind „Find Text Regions" und „Find Lines in Text Regions"). Ein Klick auf den „Run"-Button darunter startet die automatische Layoutanalyse. Sie erhalten jeweils eine Meldung, dass der Job ausgeführt wird und dass die Aufgabe abgeschlossen wurde. Transkribus fragt Sie, ob Sie die Seite nun neu laden wollen. Klicken Sie auf „Yes".

(Exkurs: Sollten Sie eingescannte Textdokumente haben, die nicht handschriftlich sondern gedruckt (aber dennoch nicht computerlesbar) sind, bietet Transkribus auch einige **OCR-Funktionen** (vgl. OCR) des sonst kostenpflichtigen Abbyy FineReaders für zahlreiche Sprachen an. Laden Sie dafür Ihr Dokument wie beschrieben in eine Kollektion. Der zweite Abschnitt im Tab „Tools" heißt „Text Recognition" und hier können Sie als Methode statt der voreingestellten HTR auf OCR umstellen. Ein anschließender Klick auf „Run" erstellt Ihnen, nachdem Sie die Sprache des Dokuments und die Schriftart (Fraktur, Latein oder gemischt) ausgewählt haben, automatisch ein Transkript, das Sie nur noch korrigieren müssen. Achtung: Frakturschriften stellen erfahrungsgemäß für OCR-Tools große Probleme dar. Der Abbyy FineReader ist hierfür zwar das am häufigsten empfohlene Tool, auch hier müssen Sie sich jedoch noch auf teilweise umfangreiche manuelle Nachkorrekturen einstellen.)

Klicken Sie nun auf das Manuskript, sehen Sie einen grün eingefärbten Rahmen über der gesamten Manuskriptseite (dies ist die **Textregion**) und unter allen Zeilen blasse rote Linien (die sog. **Baselines**).

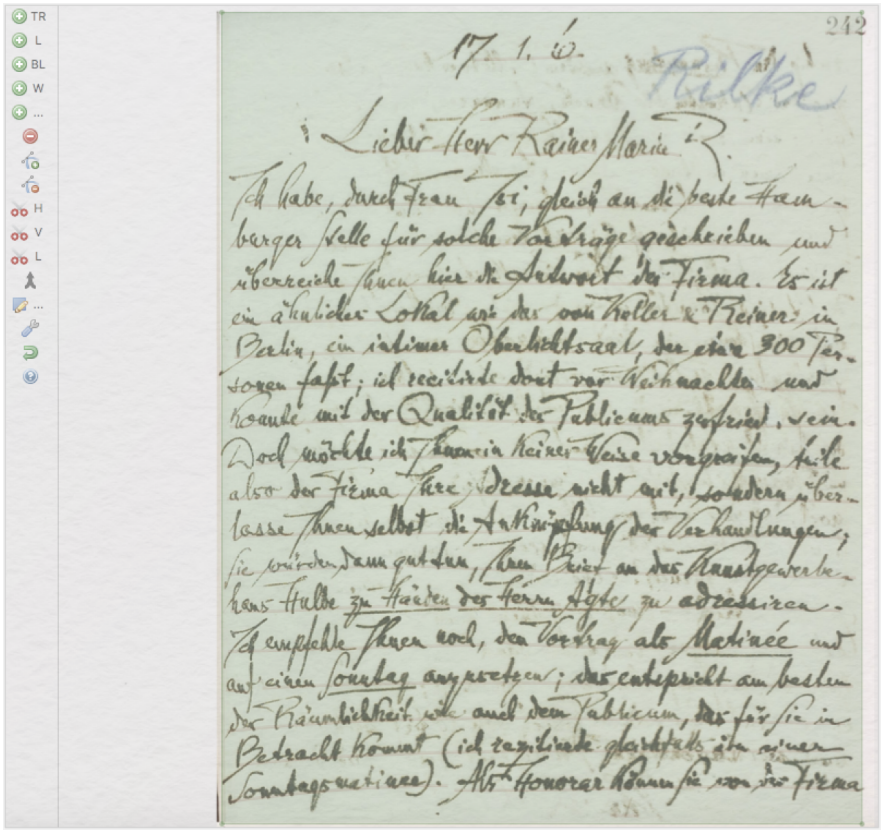


Abb. 8: Textregion und Baselines in Transkribus

Sie können das Manuskript auf der Arbeitsfläche verschieben, indem Sie außerhalb des Dokumentes klicken und ziehen (Achtung: Wenn Sie auf das Dokument klicken und ziehen, verschieben Sie die Textregion oder die Baselines). Links neben dem Manuskript finden Sie u. a. Buttons, um Textregionen und Baselines manuell hinzuzufügen (siehe Abb. 8). Wenn Sie in das Dokument hereinzoomen (mit den Funktionen Ihrer Maus bzw. Ihres Touchpads oder über die Lupensymbole oberhalb des Manuskriptes), können Sie einzelne Baselines auswählen, indem Sie sie anklicken. Dort sehen Sie dann auch, dass jede Baseline einzelne Punkte miteinander verbindet, die bei Bedarf per Drag & Drop manuell verschoben werden können, sollte die automatische Erkennung bspw. eine Zeile falsch ausgezeichnet haben.

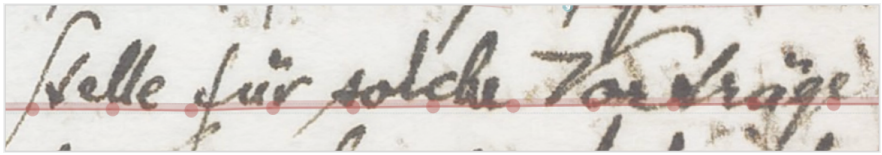


Abb. 9: Ausschnitt einer Baseline in Transkribus

Nun geht es an die manuelle Korrektur der Zeilenauszeichnungen. In der Menüleiste oberhalb des Manuskriptes sehen Sie ein Icon, das wie ein Auge aussieht. Klicken Sie auf das Auge (siehe Abb. 10), haben Sie die Möglichkeit, unterschiedliche Aspekte des Layouts (un-)sichtbar zu machen. Stellen Sie sicher, dass Baselines („Show baselines“) und die Leseabfolge der Zeilen („Show lines reading order“) aktiviert sind. Eine korrekte **Reihenfolge der Baselines** ist wichtig, wenn Sie das zu erstellende Transkript als durchgängigen korrekten Text lesen können wollen. Ein erneuter Klick auf das Auge schließt das Menü wieder. Zoomen Sie dann an den Beginn des Briefes, sodass Sie die Baselines gut sehen und die kleinen Nummerierungszahlen am Beginn jeder Baseline erkennen können.

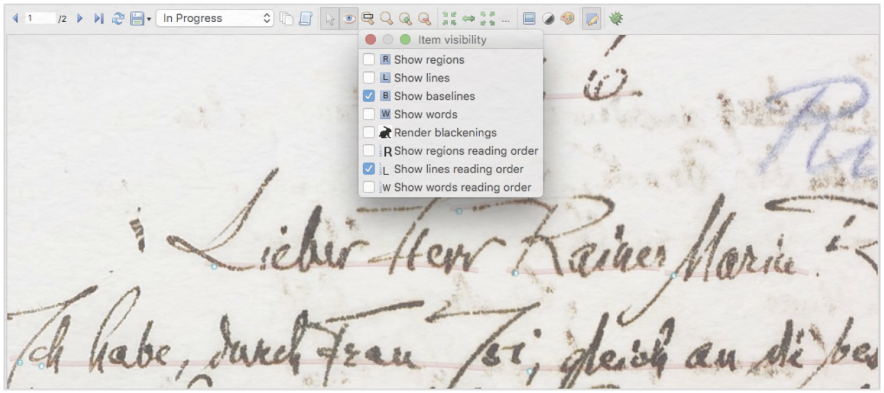


Abb. 10: Anzeigeoptionen für Textregionen und Baselines

Aufgabe 1
Untersuchen Sie die Baselines der beiden Manuskriptseiten. Welche Arten von Fehlern fallen Ihnen auf und was könnten die Gründe für die fehlerhafte Auszeichnung sein?

Manuelle Korrektur: Wenn Sie ein maschinelles Training der Handschriften anstreben, spielen die „Fehler" in der automatischen Auszeichnung keine Rolle, solange Sie während der Durchführung der Transkription die transkribierten Wörter immer der richtigen Baseline zuordnen. In dieser Lerneinheit ist es uns jedoch wichtig, dass eine Manuskriptzeile auch nur als eine Baseline ausgezeichnet wird (sodass dem fertigen Manuskript ebenfalls die korrekten Zeilenumbrüche entnommen werden können bzw. das gesamte Transkript in der natürlichen Reihenfolge gelesen werden kann). Dafür lassen sich die Baselines einzeln durch Auswahl der einzelnen Punkte einer Baseline mit der Maustaste manuell manipulieren oder mit Rechtsklick (bzw. Backspace-Taste) löschen. Ganz neue Baselines setzen Sie mithilfe des Buttons „+ BL" links neben dem Manuskript. Verfahren Sie so mit dem gesamten Brief. Bei komplexen Layouts mag die manuelle Auszeichnung (durch Hinzufügen von Baselines über „+ BL") der effizientere Weg sein, um Manuskriptzeilen und Lesereihenfolgen korrekt zu identifizieren. In diesen Fällen überspringen Sie den Schritt der automatischen Layoutanalyse und legen auch die Textregion(en) manuell fest.

Sie haben nun alle Baselines korrekt ausgezeichnet. Wechseln Sie über den „Profile"-Button oben links in die „Transcription"-Ansicht. Unter dem Manuskript erscheint ein weißes Feld mit durchnummerierten Zeilen: Hier werden Sie in Aufgabe zwei die **Transkription** vornehmen bzw. den entzifferten Text als elektronischen Text eingeben. Die Größe des weißen Feldes können Sie beliebig verändern, indem Sie es einfach weiter oder weniger weit über das Manuskript ziehen. Wenn Sie eine einzelne Zeile im Manuskript anklicken, sehen Sie, dass die entsprechende Zeile auch im Transkript hervorgehoben wird. Das funktioniert auch andersherum: Die jeweils ausgewählte Manuskriptzeile ist blau unterstrichen und die entsprechende Transkriptzeile erscheint in blauer Schrift.

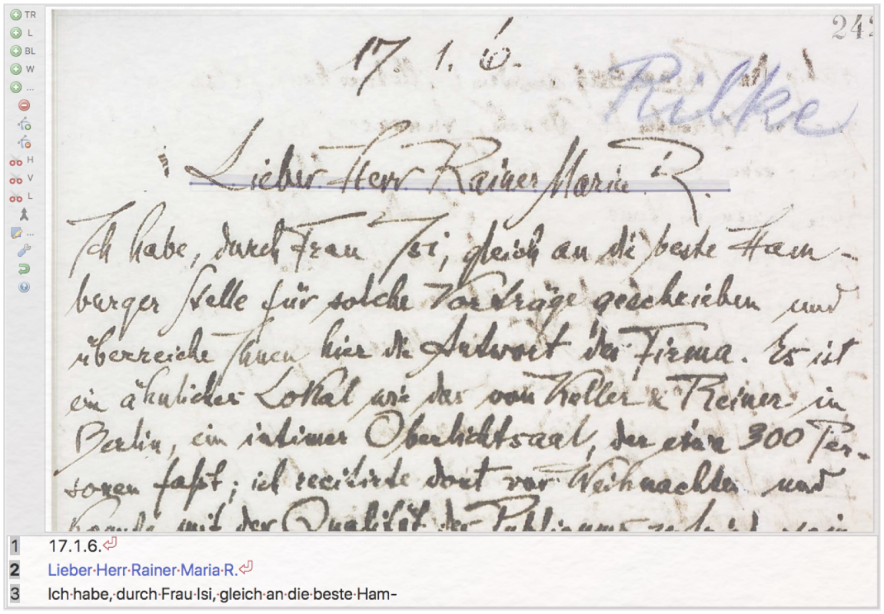
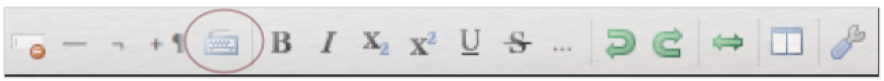


Abb. 11: Transkriptionsprofil in Transkribus mit ausgewählter Zeile

Diese Verknüpfung basiert auf der vorherigen Auszeichnung der Baselines. Der transkribierte Text muss daher immer genau hinter diejenige Nummer geschrieben werden, welche die entsprechende Baseline referenziert, um eine spätere HTR ermöglichen zu können. Hier zeigt sich auch der wesentliche Unterschied zwischen OCR und HTR: Während die automatische Erkennung von maschinell gedruckter Schrift buchstabenweise funktioniert, arbeitet die Handschriftenerkennung zeilenbasiert und vergleicht Buchstaben somit immer in ihrem Kontext (vgl. Möglichkeiten der Textdigitalisierung (Horstmann 2024a)). Vergleichen Sie bspw. einmal die beiden Es („e“) im ersten Wort „Lieber“ miteinander. Die umgebenden Buchstaben bestimmen in Handschriften immer das individuelle Aussehen eines Buchstaben und führen somit zu nicht zu unterschätzenden Abweichungen.

Aufgabe 2
Transkribieren Sie jetzt Dehmels Brief an Rainer Maria Rilke. Wie gehen Sie mit nur schwer lesbaren Wörtern oder Buchstaben um? Was hat Rilke Dehmel zu Weihnachten geschenkt? Wann und mit wem wird Dehmel auf einer Vortragsreise sein?

Sie haben nun den Brief Dehmels an Rilke vollständig transkribiert. Beim Transkribieren werden Sie bemerkt haben, dass Dehmel häufig mit Unterstreichungen arbeitet. Damit ein Machine-Learning-Algorithmus begreift, dass diese Unterstreichungen nicht zum geschriebenen Wort gehören, sondern sie als Hinzufügung interpretieren kann, sollten auch derlei **Formatierungen und Sonderzeichen** minutiös in das Transkript mit eingearbeitet werden. Dafür finden Sie in der Leiste unterhalb des Transkriptionsfeldes die entsprechenden Icons (siehe Abb. 12). Ein Tastatur-Icon bietet Ihnen zudem zahlreiche weitere Schriften (wie arabische, erweiterte lateinische, hebräische Schrift sowie alchemistische und astronomische Symbole) und auch Sonderzeichen (im Tab „General Punctuation“).



Formatierungsoptionen für das in Transkribus erstellte Transkript

Selbstverständlich hängt es von Ihrem konkreten Projektziel ab, wie minutiös das Transkript tatsächlich erstellt werden muss: Für Editionen und maschinelles Lernen ist eine große Genauigkeit unumgänglich. Da wir in dieser Lerneinheit weder eine Edition erstellen wollen noch ein maschinelles Lernen anstreben, interessiert uns lediglich der Inhalt des Manuskriptes, bei dem Unterstreichungen, Sonderzeichen, etc. nur ein bedingtes Erkenntnisinteresse haben.

Exkurs: Sollte es Ihr Ziel sein, ein **Modell** für eine spezifische Handschrift zu **trainieren** (d. h. mithilfe eines maschinellen Lernens eine automatische Erkennung der Handschrift zu ermöglichen), ist die Devise: Je mehr Text einer Handschrift Sie manuell transkribieren, desto besser wird das trainierte Modell werden (Transkribus empfielt zwischen 5000 und 15.000 Wörter bzw. 25-75 Seiten). Alternativ lassen sich im Tab „Tools“ unter „Text Recognition“ bereits trainierte Handschriftenmodelle (wie die gotische oder die Kurrentschrift) auswählen. Das langfristige Ziel des Projektes ist es, die vielen in Transkribus angesiedelten Transkriptionsprojekte zu nutzen,

um viele unterschiedliche Schreibstile zu trainieren. Auf diese Weise wird es zukünftig möglich sein, die meisten handschriftlichen Dokumente ohne vorheriges individuelles Training zu erkennen (vgl. das umfangreiche Transkribus-WIKI für weitere Informationen und Anleitungen). Da maschinelles Lernen sehr rechenaufwendig ist, ist die Möglichkeit, Modelle selbst zu trainieren, in der Standardversion nicht implementiert. Nehmen Sie dafür bitte Kontakt mit dem Transkribus-Team auf, das dieses Feature für Sie freischalten kann.

Haben Sie Ihr Manuskript erfolgreich transkribiert, bietet Ihnen Transkribus zahlreiche Möglichkeiten der **Metadatenanreicherung**. Oben links finden Sie den Tab „Metadata“, der vier weitere Untertabs für die verschiedenen Arten von Metadaten verbirgt: „Document“, „Structural“, „Textual“ und „Comments“. Alle Metadaten, die Sie hier hinzufügen, können Sie später zusammen mit dem Transkriptionsdokument als TEI-XML speichern und in anderen Tools der digitalen Textanalyse weiter verwenden. Wir werden uns hier nur beispielhaft auf die textuellen Metadaten konzentrieren.

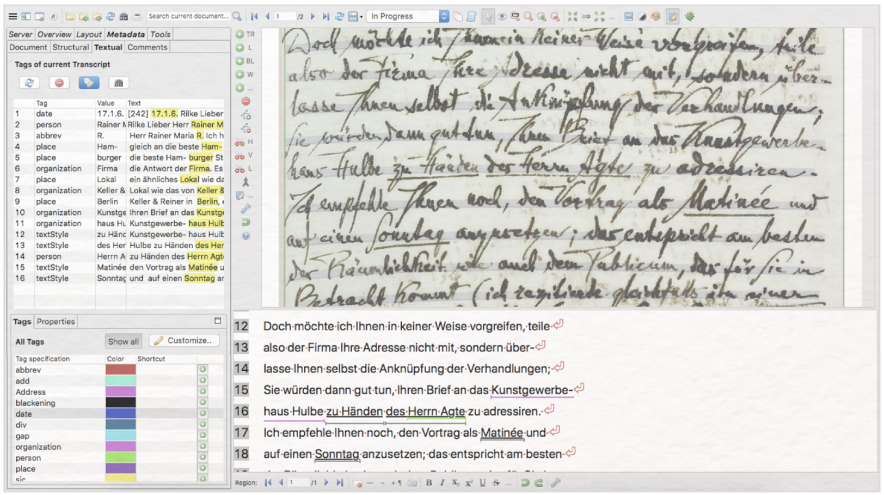


Abb. 13: Metadaten: Tags und Annotationen in Transkribus

Unten links sind Ihnen hier bereits einige Beispielkategorien (vgl. Tagset) vorgegeben, die Sie im Transkript annotieren (vgl. Annotation) können. Indem Sie ein Wort oder eine Passage markieren und dann auf das „+“-Zeichen neben der entsprechenden Tag-Kategorie klicken, belegen Sie die entsprechende Textstelle mit diesem Tag. Zu jedem Tag gibt es zudem entsprechende Properties (vgl. Property), wie z. B. die Angabe von Typen oder Ländernamen bei Orten (Tag „place“), Geburts- und Sterbedaten von Personen (Tag „person“) usw. Die Tagkategorien und Properties sind über den „Customize“-Button erweiterbar.

Aufgabe 3
Testen Sie alle Metadatenfunktionen von Transkribus. Annotieren Sie im Zuge dessen alle Organisationen, Orte und Personen. Wie viele gibt es davon, wo ergeben sich Schwierigkeiten und was kann diese Auszeichnung nützen?

Transkribierte und mit Metadaten versehene Manuskripte können schließlich in unterschiedlichen Formaten **exportiert** werden. In der Icon-Leiste oben finden Sie neben dem Import- ein Export-Icon, das Sie zu der folgenden Ansicht führt.

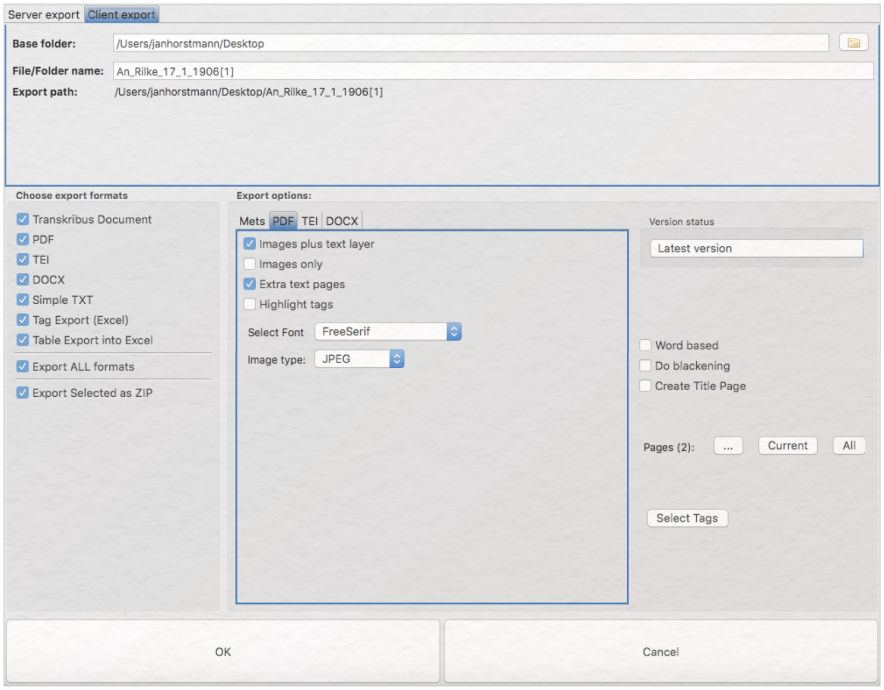


Abb. 14: Exportoptionen in Transkribus

Im oberen Bereich des Panels legen Sie unter „Client export“ den Ort fest, an dem Ihr Transkript gespeichert werden soll (alternativ können Sie den „Server export“ nutzen: Transkribus wird Ihnen dann automatisch einen Downloadlink zumailen). Auf der linken Seite können Sie individuell festlegen, in welchem Format oder welchen Formaten Sie das Transkript exportieren möchten. Unterschiedliche Formate bieten auch unterschiedliche Vor- und Nachteile. Ein PDF hat bspw. den Vorteil, dass Sie nach wie vor das handschriftliche Manuskript sehen werden, das aber nun computerlesbar sein wird. Bei DOCX- oder TXT-Dokumenten ist das nicht der Fall, dafür können Sie aber den transkribierten Text direkt lesen, kopieren, weiterverarbeiten etc. Soll ein Transkript mitsamt der hinzugefügten Metadaten für andere digitale Methoden und Tools weiter verwendet werden, ist sicher das TEI-XML-Format am vorteilhaftesten. Den standardisierten TEI-Vorgaben entsprechend kombiniert das Format Text- und Metadaten (auf Dokument-, Struktur-, Text- und Kommentarebene). So können z. B. die von Ihnen bereits in Transkribus gesetzten und im XML-Dokument gespeicherten Annotationen in einem anderen Tool als solche erkannt, erweitert, verfeinert und analysiert werden (hierfür böte sich bspw. CATMA an). Darüber hinaus besteht die Möglichkeit, sämtliche Formate gemeinsam zu exportieren und sie bei Bedarf in einer ZIP-Datei zusammenzufassen, um den Download zu beschleunigen. Zu jedem Exportformat bietet Transkribus zudem weitere Optionen in der mittleren Spalte an (vgl. Abb. 14); so können Sie z. B. ein PDF-Dokument erzeugen, das nicht nur das computerlesbare Manuskript anzeigt, sondern eine extra Textseite zu jeder Manuskriptseite hinzufügt, die besonders lesefreundlich ist und aus der Sie den transkribierten Text leichter herauskopieren können. Ein Klick auf den „OK“-Button exportiert Ihr fertiges Transkript. Zudem speichert Transkribus alle Ihre Transkripte in der entsprechenden Kollektion, sodass Sie im Tool stets dort weiterarbeiten können, wo Sie aufgehört haben.

In dieser Lerneinheit haben Sie ein Briefmanuskript Richard Dehmels digitalisiert, indem Sie den eingescannten Brief in Ihren Transkribus-Account hochgeladen, transkribiert, mit Metadaten versehen und anschließend exportiert haben. Dadurch ist das Manuskript computerlesbar geworden und kann nun mit weiteren digitalen Methoden erforscht werden. Transkribus selbst bietet Ihnen zahlreiche Informationen, Videotutorials und mit Transkribus LEARN auch Möglichkeiten, die digitale Handschriftentranskription mit vielen unterschiedlichen Handschriftenarten zu erlernen.

4. Lösungen zu den Beispielaufgaben

Aufgabe 1: Untersuchen Sie die Baselines der beiden Manuskriptseiten. Welche Arten von Fehlern fallen Ihnen auf und was könnten die Gründe für die fehlerhafte Auszeichnung sein?

Der erste „Fehler“ (Baseline 1) ist die Kennzeichnung der „242“ am oberen rechten Rand der ersten Manuskriptseite. Als menschliche Betrachtende sehen wir direkt, dass diese Zahl nicht zur Dehmel’schen Handschrift gehört, sondern eine maschinelle Nummerierung des Briefes innerhalb einer größeren Briefsammlung darstellt. Das Programm macht in dieser Hinsicht jedoch keinen Unterschied. Der zweite Fehler (Baseline 3) ist

die Identifikation des oberen Striches des Buchstabens R in der Anrede als eigene Zeile. Außerdem wird die Anrede als drei eigenständige Zeilen ausgezeichnet und das letzte „R." ist nicht mehr Teil der Baseline. Ähnliches passiert häufig im gesamten Brief. Gründe dafür können unterschiedliche Höhen der einzelnen Wörter oder auch unterschiedlich große Wortabstände sein.

Aufgabe 2: Transkribieren Sie jetzt Dehmels Brief an Rainer Maria Rilke. Wie gehen Sie mit nur schwer lesbaren Wörtern oder Buchstaben um? Was hat Rilke Dehmel zu Weihnachten geschenkt? Wann und mit wem wird Dehmel auf einer Vortragsreise sein?

In der Transkription von Handschriften spielen Zusammenhänge und Vergleiche eine große Rolle. Schwer zu entziffernde Buchstaben oder Wörter können häufig aus dem Wort- oder Satzzusammenhang erschlossen werden. Wissen Sie bspw. nicht, dass Dehmel seine zweite Frau „Isi" genannt hat (Ida Dehmel, geborene Coblenz, verheiratete Auerbach), könnte dieser Eigenname eine Herausforderung darstellen, da Dehmels großes „I" für Sie evtl. speziell aussehen könnte. Zwei Zeilen weiter unten finden Sie jedoch das alltägliche Wort „Ihnen" und ein Vergleich der Anfangsbuchstaben der beiden Wörter erschließt Ihnen auch den Namen „Isi". Zu Weihnachten hat Dehmel von Rilke übrigens dessen berühmt gewordenes *Stunden-Buch* (1905) bekommen, für das er sich inniglich bedankt. Seine Vortragsreise wird vom 13. bis 23. März 1906 stattfinden und „Frau Isi" wird ihn begleiten.

Aufgabe 3: Testen Sie alle Metadatenfunktionen von Transkribus. Annotieren Sie im Zuge dessen alle Organisationen, Orte und Personen. Wie viele gibt es davon, wo ergeben sich Schwierigkeiten und was kann diese Auszeichnung nützen?

Annotiert man sehr dicht (d. h. dass bspw. alle Erwähnungen von „ich", „wir", „Sie" etc. als Person annotiert werden) zählt das Tool schließlich sechs Organisationen, fünf Orte und 21 Personen auf der ersten, vier Organisationen, zwei Orte und 19 Personen auf der zweiten Manuskriptseite. Selbstverständlich handelt es sich dabei nicht um jeweils individuelle Organisationen, Orte und Personen; Koreferenzen können über die Properties aufgelöst werden. Eine weitere Schwierigkeit ergibt sich bei Ausdrücken wie „300 Personen", „Publicum", „Adresse" oder auch „unsres", bei denen eine Bezeichnung als Person oder Ort zumindest angezweifelt werden könnte. Schließlich erzeugt Transkribus (anders als andere Annotationstools) zwei Annotationen, wenn ein Tag über einen Zeilenwechsel hinaus gesetzt wird, z. B. bei „Ham-burger" oder „Kunstgewerbe-haus". Dies führt zu quantitativen Verzerrungen. Eine Auszeichnung dieser Konzepte in den Metadaten kann ein Transkript jedoch bereits für eine Netzwerkanalyse (Schumacher 2024a) vorbereiten und bildet gleichsam ein manuelles Pendant zur Named Entity Recognition (Schumacher 2024b).

Externe und weiterführende Links

- Abbyy FineReader: https://web.archive.org/save/https://www.abbyy.com/de-de/finereader/pricing/?msclkid=af8d631f212e14702ac83f61a8591f30&affsrc=1&CJEVENT=16ae0fec3f5411e9823200a20a180511 (Letzter Zugriff: 04.06.2024)
- Brief Dehmels an Rainer Maria Rilke vom 17.01.1906: https://web.archive.org/save/https://digitalisate.sub.uni-hamburg.de/recherche/detail?tx_dlf%5Bid%5D=20874&tx_dlf%5Bpage%5D=1&cHash=1e7ca4d4cba6f9052272f057fca11ed3 (Letzter Zugriff: 04.06.2024)
- CATMA: https://web.archive.org/save/http://catma.de (Letzter Zugriff: 04.06.2024)
- Dehmel-Archiv: https://web.archive.org/save/http://www.sub.uni-hamburg.de/sammlungen/nachlass-und-autographensammlung/dehmel-archiv.html (Letzter Zugriff: 04.06.2024)
- DocScan: https://web.archive.org/save/https://www.transkribus.org/de/docscan (Letzter Zugriff: 04.06.2024)
- Transkribus: https://web.archive.org/save/https://www.transkribus.org/de (Letzter Zugriff: 04.06.2024)
- Transkribus Kontakt: https://web.archive.org/save/https://help.transkribus.org/kb-tickets/new (Letzter Zugriff: 04.06.2024)
- Transkribus Modell-Einrichtung und Training: https://web.archive.org/save/https://help.transkribus.org/de/modell-einrichtung-und-schulung (Letzter Zugriff: 04.06.2024)
- Transkribus Videotutorials: https://web.archive.org/save/https://www.youtube.com/channel/UC-txVgM31rDTGlBnH-zpPjA (Letzter Zugriff: 04.06.2024)
- Transkribus LEARN: https://web.archive.org/save/https://transkribus.eu/r/learn/ (Letzter Zugriff: 04.06.2024)
- Transkribus Wiki: https://web.archive.org/save/https://help.transkribus.org (Letzter Zugriff: 04.06.2024)

Bibliographie

forTEXT. 2019a. Tutorial: Sicherheitsaufnahme für Internetprogramme Hinzufügen (Mac). 19. Januar. https://doi.org/10.5281/zenodo.11074232.

———. 2019b. Tutorial: Sicherheitsausnahme für Internetprogramme Hinzufügen (Windows). 25. Januar. https://doi.org/10.5281/zenodo.11074222.

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.

———. 2024c. Toolbeitrag: Transkribus. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3746, https://fortext.net/tools/tools/transkribus.

Schumacher, Mareike. 2024a. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, https://fortext.net/routinen/methoden/netzwerkanalyse.

———. 2024b. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

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.

Double-keying Double-Keying ist eine Variante des Keying, bei der zwei Personen den Inhalt eines Dokumentes abtippen. Anschließend sucht ein Computerprogramm nach Differenzen zwischen den beiden Versionen. Gefundene Tippfehler werden dann von einer dritten Person korrigiert. So entstehen nahezu fehlerfreie Textdigitalisate.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

Keying In den Bibliotheks- und Textwissenschaften beschreibt Keying das manuelle Erfassen, also das Abtippen, eines Textes im Zuge seiner Digitalisierung (siehe auch Double-Keying).

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.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

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.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

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.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.