READ
Recognition and Enrichment of Archival Documents

Transkribieren mit Transkribus für das Training der Handschriftenerkennung

Version v1.8.0
Letzte Aktualisierung dieses Guides: 24.10.2019

Transkribus ist eine Plattform für die automatische Texterkennung, Transkription und das Durchsuchen von historischen Dokumenten mit Hilfe der Handschriftenerkennung (HTR+).

Transkripte in Transkribus können unter anderem verwendet werden um:

- ein neuronales Netzwerk ("Model") zu trainieren, das automatisch gedruckte oder handschriftliche Texte erkennt.
- wissenschaftliche und standardisierte Transkripte zu erstellen, die als Grundlage einer digitalen Ausgabe dienen können.

Diese Einführung ermöglicht es Ihnen, schnell Trainingsdaten für die automatische Erkennung Ihrer Dokumente zu erstellen.

Wenn Sie bereits transkribierte Dokumente haben, konsultieren Sie bitte die Anleitung Vorhandene Transkriptionen für das Training eines Modells verwenden.

Laden Sie den Transkribus Expert Client herunter, oder stellen Sie sicher, dass Sie die neueste Version verwenden:

- https://transkribus.eu/

Besuchen Sie das Transkribus Wiki für mehr Informationen und weitere How to Guides:

- https://transkribus.eu/wiki/

Transkribus und die zugrundeliegende Technologie werden durch die folgenden Projekten und Plattformen verfügbar gemacht:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Kontakt

- Das Transkribus Team: email@transkribus.eu

Inhalt

Das READ Projekt wird durch das Horizon 2020 Forschungs- und Innovationsprogramm im Rahmen des Fördervertrags Nr. 674943 finanziert.

Einleitung

Diese Anleitung erklärt, wie Sie Transkribus verwenden können, um Transkripte zu erstellen. Diese können unter anderem genutzt werden als

- Trainingsdaten für ein HTR+ Modell, das Dokumente in weiterer Folge automatisch transkribieren kann
- Als Basis für eine digitale wissenschaftliche Edition

Um ein Dokument in Transkribus zu transkribieren, sind im Wesentlichen drei Schritte durchzuführen:

Schritt 1: Hochladen

- Laden Sie Ihr Dokument auf die Transkribus Plattform hoch

Schritt 2: Segmentieren

- Automatische Segmentierung zur Erkennung von Regionen und Zeilen in Ihrem Dokument

Schritt 3: Transkription

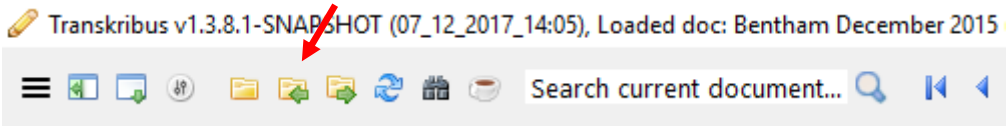
- Transkribieren Sie den Text Zeile für Zeile

Diese einfache Form der Transkription reicht für ein HTR+ Training aus. HTR+ funktioniert für Handschrift-, als auch Druckschriftdokumente.

Für jene, die an wissenschaftlichen Editionen arbeiten, kann in Transkribus auch eine anspruchsvollere Transkription angefertigt werden. Man kann die Lesereihenfolge einstellen, Tags und Metadaten hinzufügen, Abkürzungen ausschreiben und vieles mehr.

Dokumente auf Transkribus hochladen

- Um alle nötigen Werkzeuge für Ihre Dokumente verwenden zu können, müssen diese auf dem Transkribus Server liegen (das heißt, sie müssen hochgeladen werden).
    - **Achtung: Alle Collections und Dokumente auf Transkribus sind privat**. Nur Sie und Nutzer die von Ihnen autorisiert wurden, können die Dokumente sehen. Sie sind nicht öffentlich zugänglich. Das Hochladen eines Dokuments auf den Transkribus Server ist ein rein technischer Prozess.
- Um ein Dokument hochzuladen klicken Sie auf die “Import Document(s)” Schaltfläche im Main Menu.

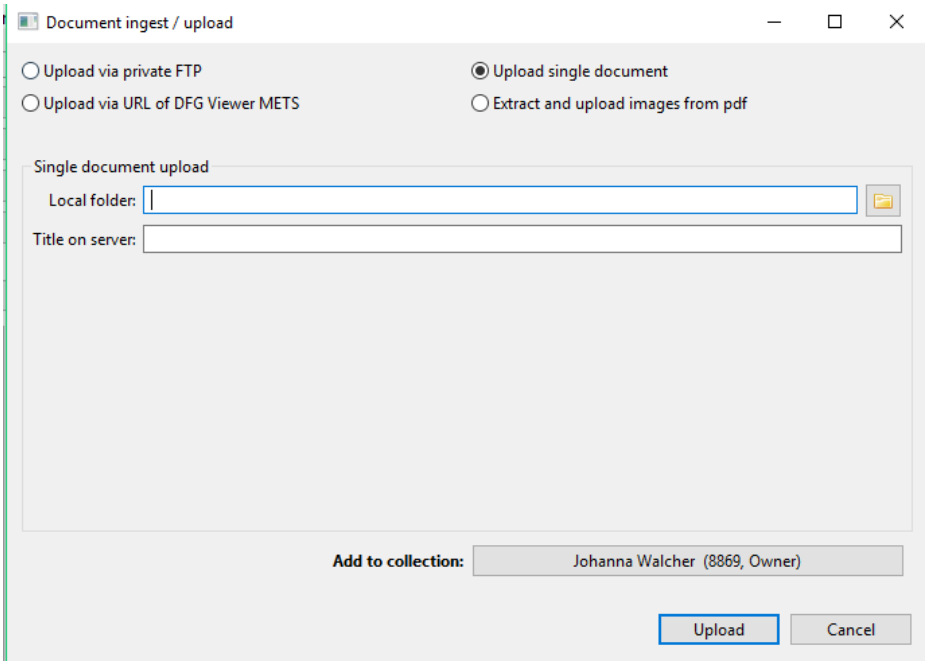


Darstellung 1 Dateien in Ihre persönliche Collection hochladen

- Sie haben drei Optionen:
    - **Einzelne Dokumente** aus einem lokalen Ordner **hochladen**:
        - mit dieser Option können Sie Dokumente bis zu 500 MB hochladen**.**
        - Dafür wählen Sie die „Upload Single Document“ option.
        - Bitte beachten Sie hier, dass sich die Images in einem extra Ordner befinden müssen. Wenn Sie die Dateien auswählen, scheint es als ob der Ordner leer wäre. Das ist nicht so, sie können die Dokumente in diesem Schritt nur nicht

sehen. Es reicht einfach den Ordner zu markieren und dann mit „OK“ zu bestätigen.

- **Upload via FTP**
  - Diese Option bietet sich führ mehrere große Dokumente an.
- **Upload via URL** des DFG Viewer METS
  - Diese Option ermöglicht den direkten Upload von Dokumenten aus Quellen, die den DFG (Deutsche Forschungsgemeinschaft – German Science Funds) Viewer unterstützen
- **Bilder aus PDF-Dateien extrahieren und hochladen**



Darstellung 2 Wählen Sie "Upload single document" für Dokumente bis zu 500 MB

Segmentierung

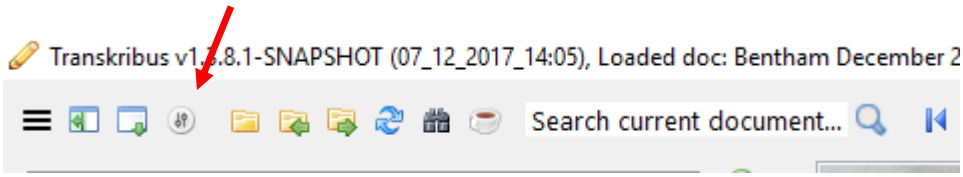
Einleitung

- Sobald Sie Ihre Dokumente auf Transkribus hochgeladen haben, können Sie mit dem Segmentieren beginnen.
- Damit die Handschriftenerkennung funktioniert, müssen **Text und Bild** in Transkribus **verbunden sein**. Das geschieht durch die Unterteilung Ihres Dokuments in „Text Regions“, „Lines“ und „Baselines“.

Ansichten

- Verschiedene Ansichten sind verfügbar, um Ihnen das Segmentieren und Transkribieren zu erleichtern.
- Sie können Ansichten für Segmentierung und Transkription auswählen, indem Sie auf die “Profiles” Schaltfläche im Hauptmenu klicken.
- Das “Segmentation” Profil stellt Baselines in Rot dar. Das macht das Erkennen von Fehlern der automatischen Segmentierung leichter.
- Das “Transcription” Profil blendet ein Textbearbeitungsfeld ein, in dem Sie ihr Dokument transkribieren können.

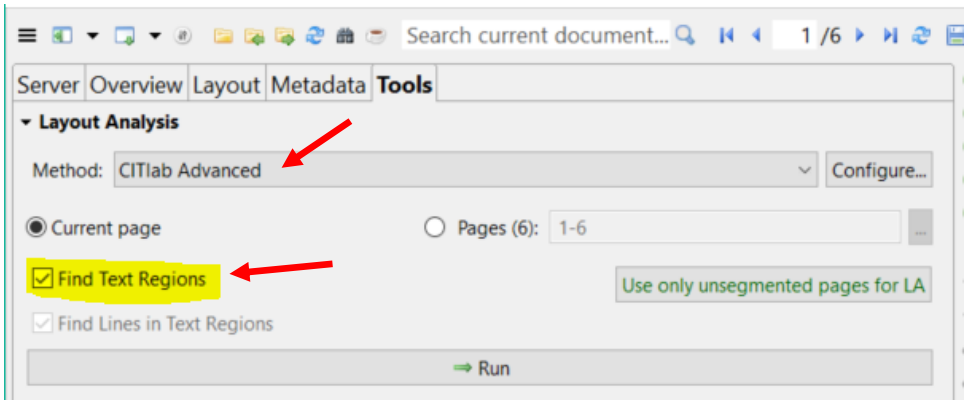
- Sie können das “default” Profil verwenden um beide Arbeiten zu erledigen.



Darstellung 3 Viewing profiles für Segmentierung und Transkription

Automatisch Text Regions, Lines und Baselines erkennen lassen

- Wählen Sie die “Segmentation” Ansicht im Main Menu aus.
- Klicken Sie auf den Reiter “Tools” auf der linken Seite Ihres Bildschirms und gehen Sie zum Bereich “Layout Analysis”.
- Wählen Sie bei “Method:” die Option “CITlab Advanced”.
- Geben Sie an, ob Sie die Layoutanalyse für die aktuelle Seite, bestimmte Seiten oder das gesamte Dokument durchführen möchten.
- Überprüfen Sie, dass “Find Text Regions” ausgewählt ist.
- Klicken Sie auf “Run“.



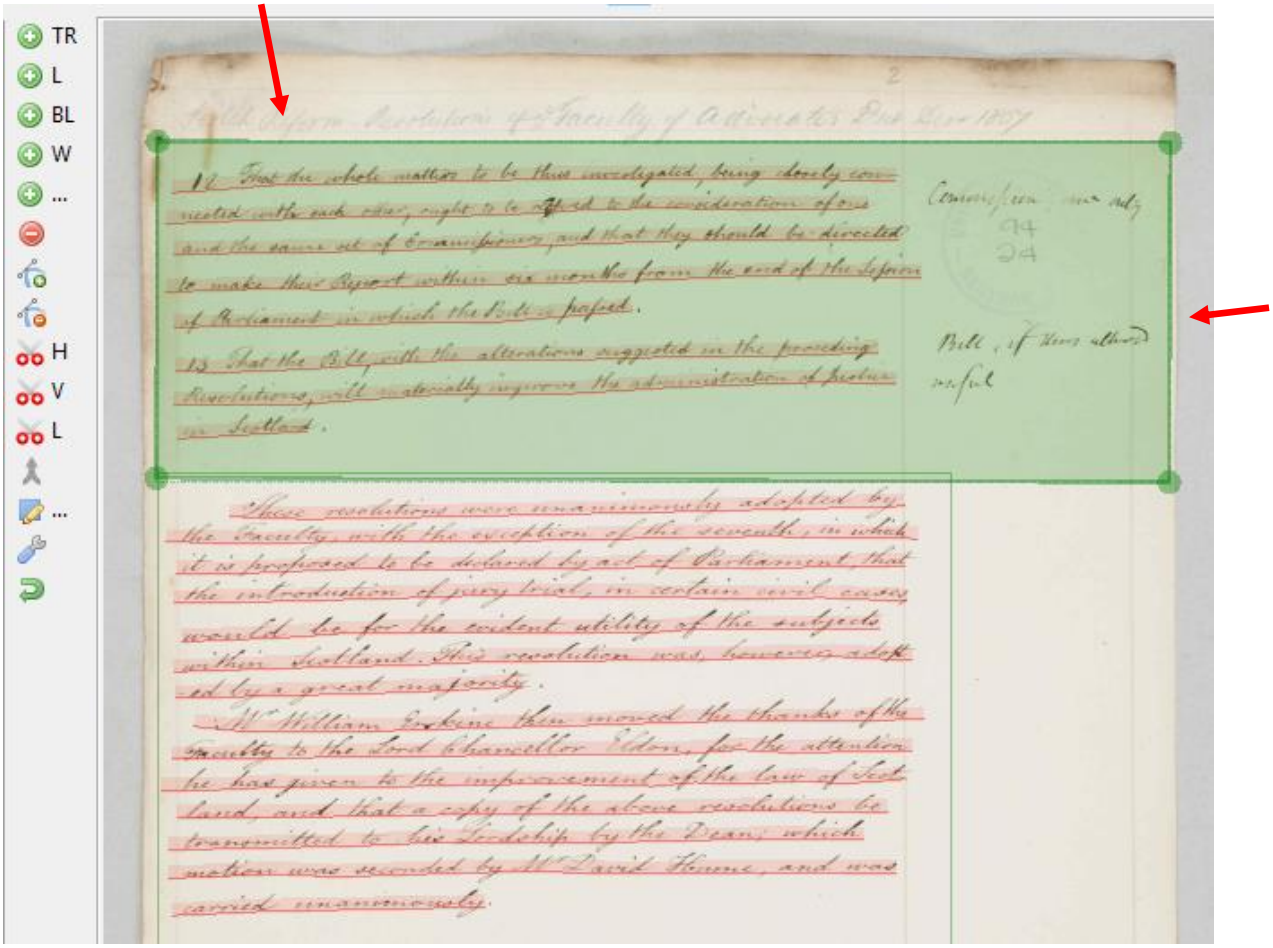
Darstellung 4 Automatische Segmentierung im Reiter “Tools” durchführen

- Das **Programm erkennt automatisch** die Text Regions, Lines und Baselines in Ihrem Dokument.
- In den meisten Fällen sind die Resultate sehr genau.
- Wenn Dokumente ein sehr komplexes Layout haben, können manuelle Korrekturen nötig sein.

Das Ergebnis der automatischen Segmentierung korrigieren

- **Achtung:** Wenn Sie ein Handschriftenerkennungsmodell trainieren, muss die Position der Text Regions nicht 100 % genau sein und die Lesereihenfolge (reading order) ist nicht relevant.
- Wenn Sie an einer wissenschaftlichen Ausgabe arbeiten, die einen höheren Grad an Genauigkeit verlangt, kann der Text manuell korrigiert werden.

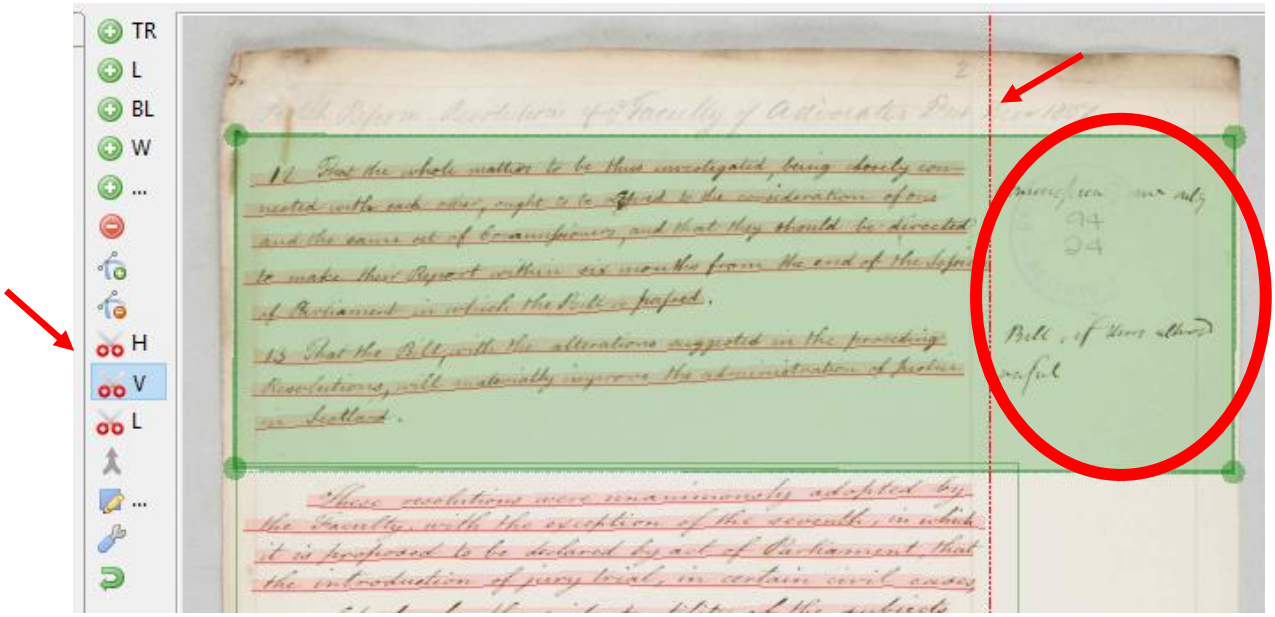
Eine zusätzliche Line wurde irrtümlich hinzugefügt oder ausgelassen



Darstellung 5 Eine Line in eine bestehende Text Region einfügen

- Im obigen Beispiel wurde die erste Line vom Programm ausgelassen. Wenn Sie diese zur bestehenden Text Region hinzufügen wollen, gehen Sie wie folgt vor:
  - Klicken Sie in die Region, sodass diese hervorgehoben wird.
  - Verschieben Sie die Ränder der Text Region, so wie nötig.

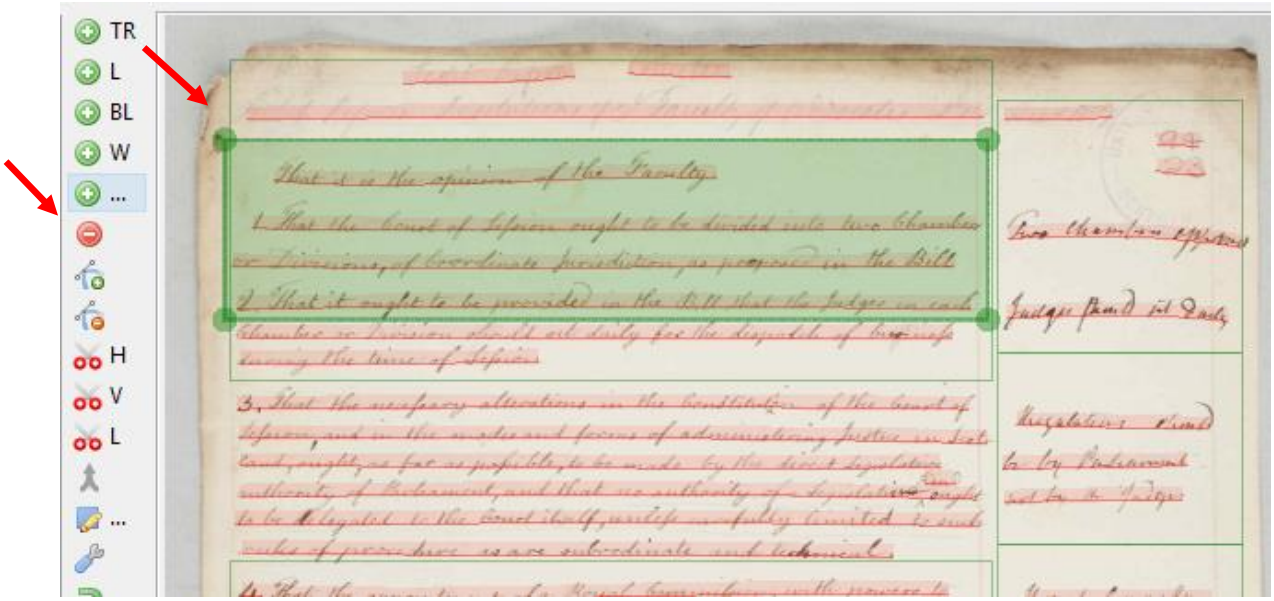
Eine Marginalie soll sich in einer eigenen Textregion befinden



Darstellung 6 Eine Text Region aufteilen

- Wenn Sie eine Text Region in zwei aufteilen wollen, funktioniert das mit den Werkzeugen im Canvas Menu.
- Darstellung 6 zeigt: der “V-button” teilt eine Text Region vertikal.
- Mit dem “L-button” kann man die Text Region mit einer selbst gezeichneten Linie teilen.

Eine nicht benötigte Region entfernen

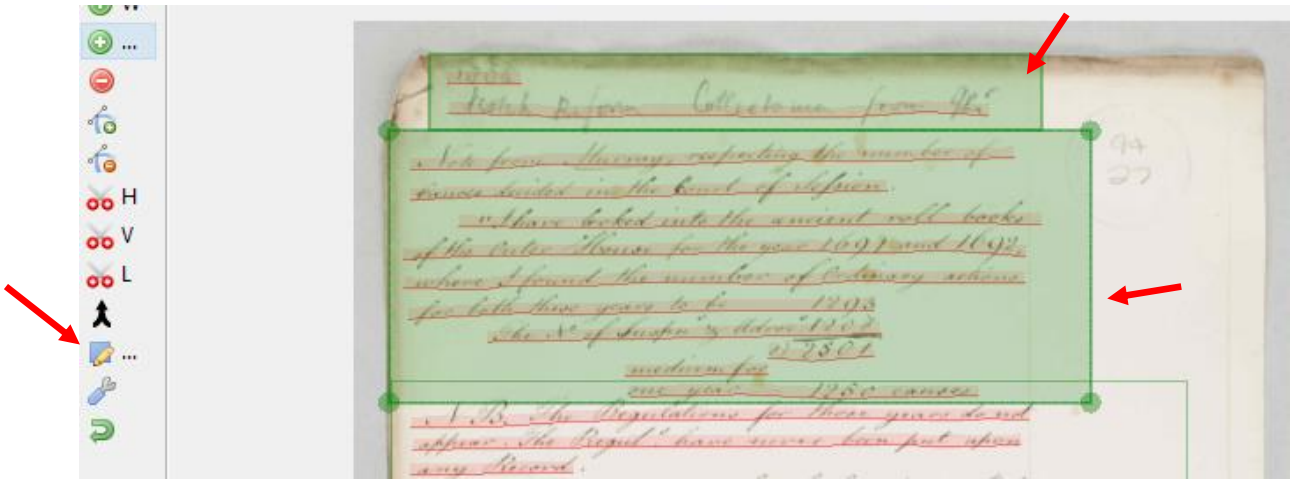


Darstellung 7 Eine Region entfernen

- Im oberen Beispiel überlappen sich zwei Regionen, eine der beiden kann entfernt werden.
- Klicken Sie auf die Region die Sie entfernen möchten und klicken Sie dann auf die rote “Remove a shape” Schaltfläche.

Zwei Regions verbinden

- Manchmal kann es passieren, dass das Programm zwei Text Regionen erstellt, wo nur eine nötig wäre. In diesem Fall können Sie die beiden ganz einfach verbinden.
    - Halten Sie die “STRG” Taste auf Ihrer Tastatur gedrückt und klicken Sie nacheinander auf beide Regionen.
    - Klicken Sie auf die “Merges the selected shapes” Schaltfläche im Canvas Menu.



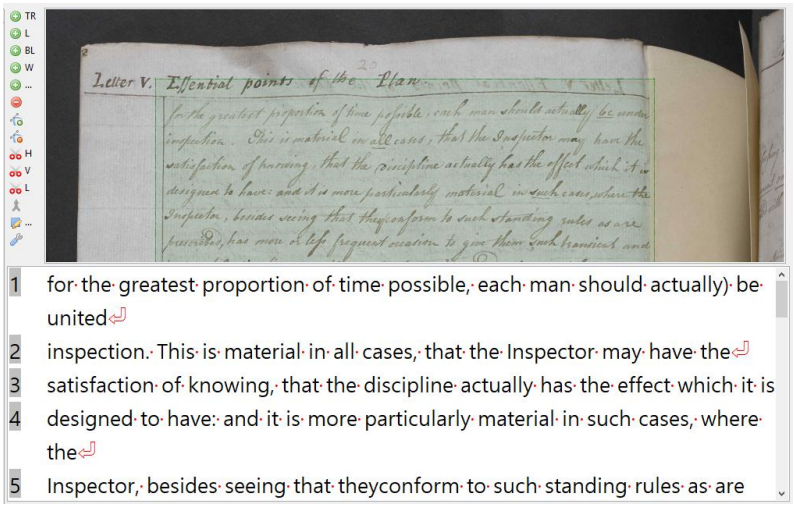
Darstellung 8 Zwei Text Regions verbinden

Baselines korrigieren

- Natürlich können auch die Baselines verändert werden.
- Klicken Sie auf die gewünschte Baseline und verschieben Sie Teile der Line, teilen Sie sie auf oder verbinden Sie zwei Baselines.
- Sie können auch Baselines löschen und eine neue von Hand zeichnen. Um eine neue Baseline zu kreieren, klicken Sie auf die „+BL“ Schaltfläche im Canvas Menu. Klicken Sie einmal um mit dem Zeichnen zu beginnen und doppelklicken Sie um die Line abzuschließen.
- Achtung: Baselines sind essentiell für die Handschriftenerkennung; Line Regions müssen nicht korrigiert werden.

Einfache Transkription – für das HTR+ Training

- Wählen Sie das “Transcription” Profil im Hauptmenu.
- Im Text Editor unter dem Bild gibt es jeweils **eine Zeile für jede Line/Baseline im Bild**. Das Bild und der Text sind so verbunden.



Darstellung 9 Dokument transkribieren

- Transkribieren Sie den Text in der Sprache Ihres Quelldokuments. Verwenden Sie das Zeichensystem Ihrer Tastatur.
- Es können mehrere Personen gleichzeitig an einem Dokument arbeiten, um Probleme zu vermeiden, besser nicht zugleich auf einer Seite. Sie können Ihr Dokument anderen Nutzern zugänglich machen, indem Sie auf die “User Manager” Schaltfläche im “Server” Tab klicken.

Ein Modell trainieren

- Wenn Sie ein HTR+ Modell trainieren möchten, ist diese einfache Transkription ausreichend.
- Wie empfehlen das Modell-Training zu starten, wenn Sie zwischen 5 000 und 15 000 Wörter (ca. 25-75 Seiten) transkribiert haben. Wenn Sie mit gedrucktem Material arbeiten, können Sie auch schon mit weniger Transkriptionsmaterial starten.
- Wenn Sie bereit sind, senden Sie uns bitte eine kurze Nachricht (email@transkribus.eu), dann schalten wir Sie für die Trainingsfunktion frei, der nicht automatisch im Expert-Client inkludiert ist.
- Wie man ein Modell in Transkribus trainiert, lesen Sie in dieser Anleitung: Modell Training in Transkribus

Danksagung

Wir möchten den vielen Nutzern danken, die mit ihrem Feedback zur Verbesserung der Transkribussoftware beigetragen haben.

Transkribus wird der Öffentlichkeit im Rahmen des H2020e Infrastrukturprojekts READ (Recognition and Enrichment of Archival Documents) zugänglich gemacht. Dieses Projekt wird von der Europäischen Kommission im Rahmen des Fördervertrags Nr. 674943 finanziert