READ
Recognition and Enrichment of Archival Documents

Vorhandene Transkriptionen zum Trainieren eines Handschriftenerkennungs-modells verwenden

Version v1.9.1
Letzte Aktualisierung dieses Guides: 04.01.2020

In dieser Anleitung finden Sie Informationen zur Anwendung des text to image (t2i) tools. Wenn Sie bereits über Transkriptionen verfügen und diese verwenden möchten, um ein Handschriftenerkennungsmodell (HTR-Modell) in Transkribus zu trainieren, hilft Ihnen das t2i dabei, die Transkriptionen mit den Scans in Transkribus zusammenzuführen.

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

Die Transkribus Plattform macht es Nutzern möglich ein HTR Modell zu trainieren, das Dokumente automatisch erkennt. Das Modell wird trainiert, indem ihm Bilder von Dokumenten mit einer bestimmten Handschrift und deren exakte Transkription gezeigt werden.

In den letzten 20 Jahren wurden viele wissenschaftliche Transkriptionsprojekte durchgeführt. Eine große Menge von Dokumenten wurde transkribiert und ist jetzt in elektronischer Form verfügbar. Alle diese Transkriptionen können ganz einfach als Trainingsmaterial für HTR verwendet werden.

Das t2i-Tool, das vom CITlab an der Universität Rostock entwickelt wurde, erstellt Trainingsdaten basierend auf den vorhandenen Transkriptionen. Das Tool verwendet einen Algorithmus, der die Transkriptionen automatisch mit den Bildern der handschriftlichen Dokumente verknüpft. In weiterer Folge kann damit ein HTR-Modell trainiert werden. Das Tool eignet sich vor allem für all jene, die schon über 500-1000 Seiten an Transkriptionen verfügen.

Anstatt Trainingsdaten für HTR manuell in Transkribus zu erstellen, können Sie also einfach ihre vorhandenen Transkriptionen verwenden, um die Technologie auszuprobieren. So können Transkriptionen erstellt werden, ohne dass der Workflow eines Projekts verändert werden muss. Bitte beachten Sie, dass das t2i-Tool auf einem HTR-Modell basiert, das für sich eine gewisse Fehlerquote besitzt und deshalb nicht in der Lage ist ein gänzlich fehlerfreies Ergebnis zu liefern. Deshalb kommt man um einzelne händische Ausbesserungen an der t2i-Transkription nicht umhin. Für den Fall, dass Sie eine 100%ig fehlerfreie Transkription wünschen, kann es schneller sein, die vorhandenen Transkriptionen händisch in Transkribus zu kopieren.

Vorbereitungen

- Wenn Sie mit dem t2i-Tool arbeiten möchten, brauchen Sie Zugriff auf die digitalisierten Bilder und Transkripte ihrer Dokumente
- Diese Dateien müssen laut den nachfolgenden Anleitungen vorbereitet werden, bevor sie mit vom t2i-Tool verarbeitet werden können.

Seitenmenge

- Wir empfehlen Ihnen den Trainingsprozess mit mindestens 20.000 Wörtern (ca. 100 Seiten) transkribierten Materials zu starten.
- T2i ist speziell bei einer großen Menge von verfügbaren Transkripten interessant (500 oder mehr Seiten).
- Die Technologie kann schnell große Mengen von Transkripten verarbeiten (100.000 Seiten und mehr).
- Die neuralen Netzwerke in HTR lernen schnell. Je mehr Trainingsdaten verfügbar sind, desto besser werden die Resultate.

Bilddateien

- Alle Arten von Bilddateien können verarbeitet werden.
- Die Bilder sollten eine Auflösung von mindestens 200 ppi haben. Alternativ – wenn die Bilder mit einer Kamera aufgenommen wurden – gilt die Faustregel, dass ein einzelnes Zeichen mindestens 15-20 Pixel haben soll.
- Die Genauigkeit der HTR ist auch von der Bildqualität abhängig. Trotzdem können auch Daten von Mikrofilmen und Schwarzweißbildern mit genug Trainingsdaten verarbeitet werden.

Transkriptionsdateien

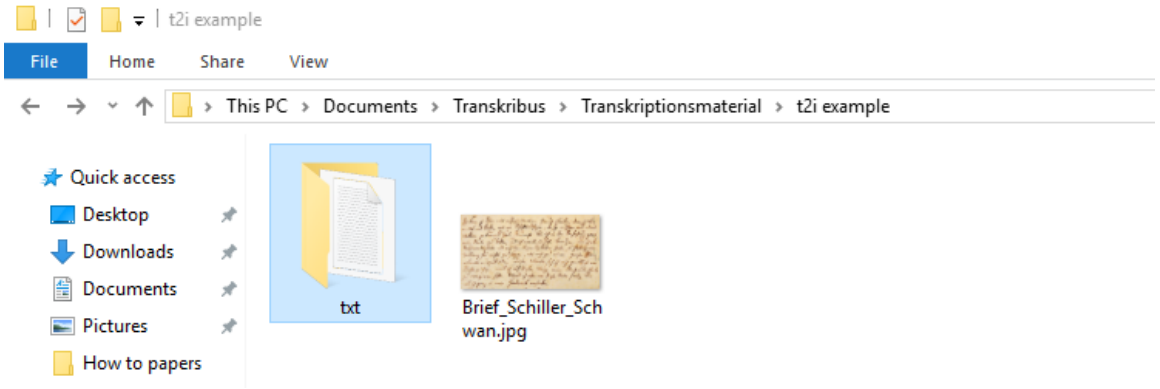
- Alle Transkripte sollten als einfache Textdateien (TXT) gespeichert werden.
- Wenn Sie über Transkriptionen im TEI (Text Encoding Initiative), Word, XML oder HTML-Format verfügen, sollten Sie diese in TXT-Dateien umwandeln. Um diesen Arbeitsschritt effizient zu gestalten, ist es sinnvoll sich (bei entsprechenden Kenntnissen) eine einfache File-Preparation-Software zu programmieren, um sich das händische Erstellen der Textdateien zu ersparen.
- Transkriptionen sollten auf Seitenlevel gespeichert werden, das heißt eine txt-Datei pro Seite.

Transkriptionen

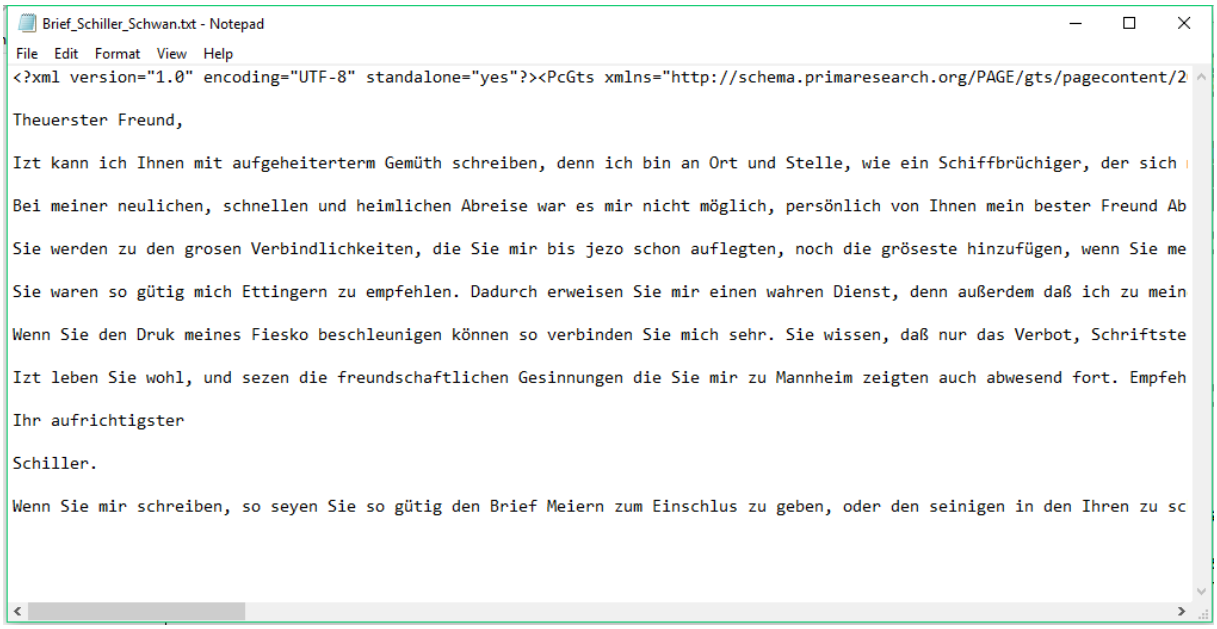
- Transkriptionen sollten keine unnötigen Formatierungen enthalten.
- Die Zeilenumbrüche in der Transkription können beibehalten werden. Es ist jedoch nicht nötig am Ende jeder Zeile einen Zeilenumbruch zu setzen.
- Das t2i-Tool meistert auch Fälle, in denen das Wort ohne Bindestrich auf zwei Zeilen aufgeteilt wurde.
- Wenn ein Wort in Ihrem Transkript unleserlich ist, macht es Sinn die ganze Zeile zu löschen, in der das Wort vorkommt. Somit wird diese Zeile nicht für das HTR-Training verwendet.
- Transkriptionen müssen nicht komplett sein. Wenn Wörter im Transkript fehlen, werden sie nicht für das HTR-Training verwendet.
- Verschiedene Unicode-Zeichen können berücksichtigt werden, unter anderem Arabisch und Hebräisch.
- Transkriptionen, in denen Abkürzungen ausgeschrieben wurden, können für das t2i und HTR-Training verwendet werden (die Abkürzungen werden automatisch ausgeschrieben).

Dateivorbereitung

- Die Dateien, die Ihre Bilder und Transkriptionen beinhalten, sollten klar zuordenbar sein. Dafür speichern Sie am besten jede Bilddatei mit dem exakt selben Namen wie die zugehörige TXT-Datei.
- Wenn Sie die Bilder und Transkripte vorbereitet haben, bitte folgendermaßen strukturieren:
    - Name des Dokuments
        - TXT
- Sie können Ihre Dateien direkt auf Transkribus hochladen. Für den Upload der TXT-Dateien sollte ein separater Ordner “Text” im Ordner mit den Bilddateien vorhanden sein.



Darstellung 1 Aufteilung der Ordner



Darstellung 2 TXT Dateil

HTR-Modell

- Für die Anwendung des t2i benötigen Sie ein HTR-Modell, das zu Ihren Dokumenten passt.
- Dazu können Sie, falls vorhanden, ein bereits existierendes eigenes oder öffentliches Modell verwenden,
- …oder ein Modell eigens für das t2i vorbereiten, indem Sie einige Seiten des Transkripts händisch in Transkribus kopieren und diese anschließend trainieren. Mehr Informationen über das Training von Modellen finden Sie in dieser Anleitung: Modell Training in Transkribus

T2i in Transkribus

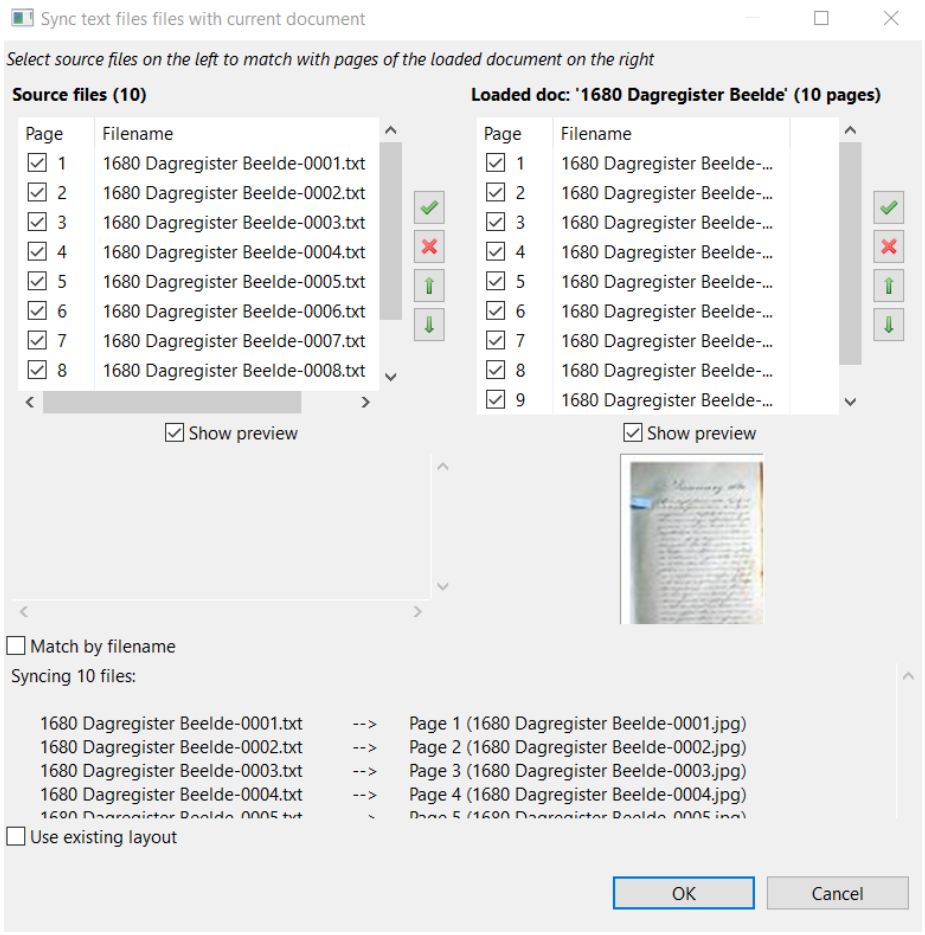
Scans und Transkriptionen gemeinsam hochladen

- Wenn Sie Scans und Transkriptionen gemeinsam hochladen, folgen Sie den oben genannten Anweisungen und nutzen Sie anschließend den gewöhnlichen Transkribus-Import. Sie finden die Funktion im Hauptmenu.

Scans und Transkriptionen getrennt hochladen

Für den Fall, dass Sie die Images schon zu einem früheren Zeitpunkt ohne die Text-Files hochgeladen haben, gehen Sie folgendermaßen vor:

- Images in Transkribus öffnen
- Text-Files in einem eigenen Ordner am PC abspeichern
- Auf „Main Menu“ im Transkribus links oben klicken.
- „Document“ wählen
- Auf „Sync local text files with doc“ klicken
- Text-Files im Directory auswählen.
- Damit öffnet sich das folgende Fenster:

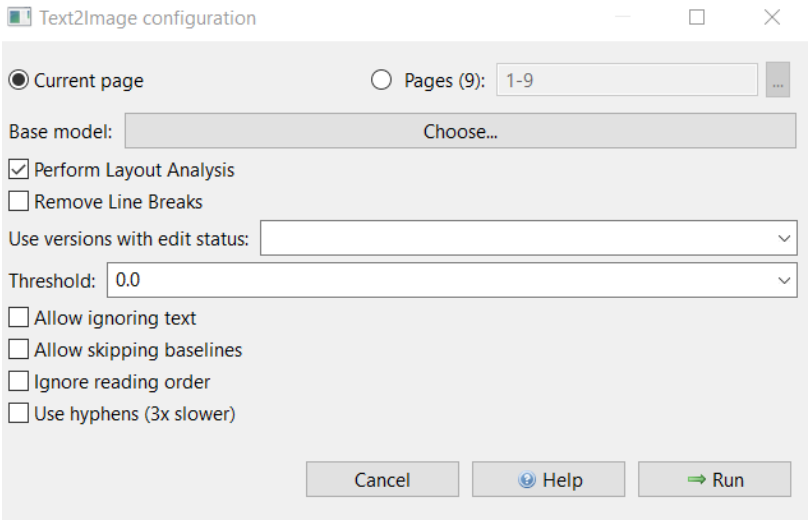


Darstellung 3 Sync text files with doc

- **„Use existing layout“:** Normalerweise startet das t2i eine neue Layout Analyse für das Dokument, sollte das nicht erwünscht sein, kann diese Option abgewählt werden.
    - Vorteile der Verwendung des bereits vorhandenen Layouts: es kann im Nachhinein einfach korrigiert werden, indem die Zeilen mit „Strg“ und „Enter“ an die richtige Stelle gebracht werden.
    - Risiko bei Layout-Erstellung direkt im Zuge des t2i: es kann passieren, dass einzelne Zeilen im Matching fehlen.
- **„Match by filename“:** anklicken um die Dateien nach Namen zu synchronisieren.
- Mit „OK“ bestätigen

Matching in Transkribus

- Importieren Sie die Dokumente mithilfe einer der oben genannten Optionen in Transkribus.
- Öffnen Sie den „Tools“-Tab in Transkribus. Im Abschnitt „Other Tools“ finden Sie die t2i-Funktion. Wenn Sie darauf klicken, öffnet sich folgendes Fenster:



Darstellung 4 t2i configuration

- Wählen Sie oben die Seiten, die zusammengeführt werden sollen.
- **„Base Model“:** Wählen Sie ein passendes Model für das Dokument, das als Basis für das t2i dienen soll.
- **„Perform Layout Analysis“:** Normalerweise startet das t2i eine neue Layout Analyse für das Dokument, sollte das nicht erwünscht sein, kann die Option abgewählt werden.
- **„Remove Line Breaks“:** wählen Sie diese Option, wenn die Zeilenumbrüche im Text-File nicht richtig gesetzt sind. Sie legen damit fest, ob Zeilenumbrüche berücksichtigt werden sollen oder nicht.
- **„Use versions with edit status“:** Für den Fall, dass Sie für das Matching nicht die letzte Version des Dokuments verwenden möchten, können Sie die gewünschte Version hier auswählen. Diese Option nimmt Bezug auf die Transkribus-Status, die dem Dokument zu zugewiesen wurde.
- **„Threshold“:** Gibt an, wie hoch die Übereinstimmung sein muss, dass ein „Match“ ausgeführt wird. Der Grundeinstellung liegt bei 0.0 weil falsche „Matches“ im Nachhinein recht einfach korrigiert werden können. Umso niedriger der „Threshold“-Wert – umso toleranter ist die Funktion in der Zuweisung.
- **„Allow ignoring text“:** wenn sich in den Text-Files Text befindet, der sich im Image nicht wiederfindet.
- **„Allow skipping baselines“:** wählen Sie diese Option, wenn im Text-File eventuell einzelne Zeilen fehlen.
- **„Ignore reading order“** lässt das t2i die Zeilenordnung, die im Rahmen der Layout Analyse definiert wird, ignorieren. Diese Funktion auszuwählen kann bei komplizierten Layouts hilfreich sein (zum Beispiel, wenn sich sowohl vertikale als auch horizontale Schrift in dem Dokument befindet) und bei Schriften, die von rechts nach links gelesen werden.
- **„Use hyphens“:** Mit dieser Option legen Sie fest, dass folgende Satzzeichen einen Zeilenumbruch auslösen: - = : ¬

Ergebnisse verbessern

- Nachdem das t2i-Matching abgeschlossen ist, können falsch zugewiesene Zeilen korrigiert werden.
- Dazu springen Sie am besten von Textregion zu Textregion und kontrollieren jeweils die erste und letzte Zeile.

- Um zu korrigieren können Sie die Transkription mit „Strg“ und „Enter“ um eine Zeile nach unten verschieben. Mit „Return“ können Sie die Transkription um eine Zeile nach oben verschieben. Selbstverständlich können sie im Text Editor auch ganz einfach Text löschen oder hinzufügen.
- Wenn Sie Zeilen oder Regionen gänzlich löschen möchten, lässt sich dies gut über den „Layout Tab“ erledigen. Dort können Sie die zu löschende Form markieren und mit dem roten Minus-Kreis-Symbol löschen.

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