READ
Recognition and Enrichment of Archival Documents

Modell Training in Transkribus

Version v 1.8.0
Letztes Update dieses Guides: 24.10.2019

Dieser Guide erklärt, wie Sie Transkribus verwenden können, um ein Handschriftenerkennungsmodell (HTR+ Modell) zu trainieren. Nachdem Sie das Modell trainiert haben, kann es automatische Transkriptionen erstellen und Ihnen helfen, Ihre Dokumente nach Wörtern zu durchsuchen.

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

Einleitung 3
Vorbereitung 3
Training 3
HTR+ Training einrichten 4
Base Model 6
Training Set 6
Test Set 7
Fortschritt verfolgen 7
Nach dem Training 8
Statistiken 9
HTR Transkripte erstellen 10
Ein Modell teilen 11
Vorteile des Modelltrainings 13
Credits 13

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

Einleitung

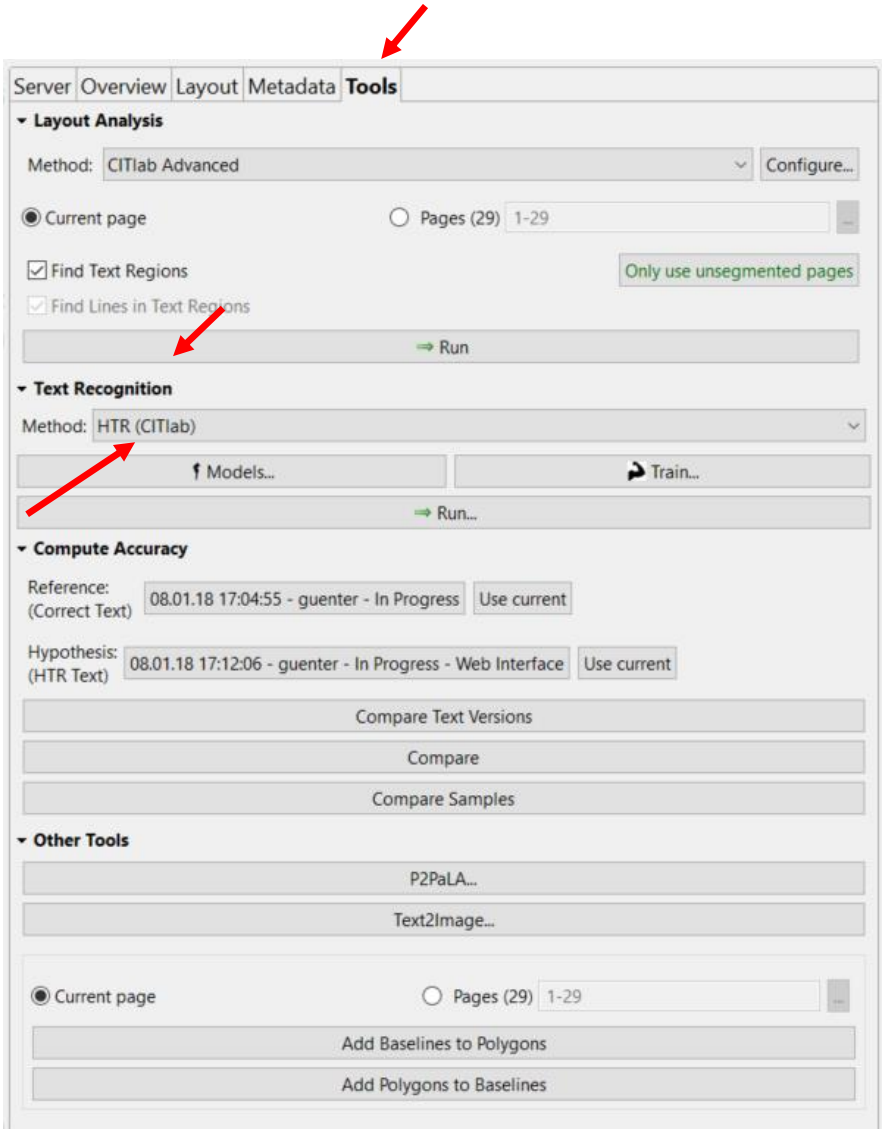
- Die Transkribus Plattform erlaubt es Nutzern ein HTR+ Modell zu trainieren, um Dokumente automatisch erkennen zu lassen. Das Modell muss trainiert werden, um einen speziellen Schreibstil zu erkennen. Das geschieht durch das „zeigen“ von Bildern und den dazugehörigen genauen Transkriptionen.
- Für das Training eines Modells sind zwischen 5.000 und 15.000 Wörter (ca. 25-75 Seiten) an transkribiertem Material nötig. Für gedruckte Texte sind im Normalfalld weniger Trainingsdaten nötig, als für handgeschriebene Texte.
- Die Trainingsfunktion ist in der Standardversion der Transkribus Plattform nicht inkludiert. Wenn Sie ein Modell trainieren möchten kontaktieren Sie bitte das Transkribus Team (email@transkribus.eu). Sie erhalten dann Zugriff auf das Feature.

Vorbereitung

- Wir empfehlen, dass Sie den Trainingsprozess, abhängig davon, ob Sie mit gedruckten oder handgeschriebenen Dokumenten arbeiten, mit 5.000-15.000 Wörtern transkribiertem Material beginnen.
- Das neuronale Netzwerk im Hintergrund lernt schnell und je mehr Trainingsdaten vorhanden sind, desto besser das Ergebnis.
- Sie können Trainingsdaten für Ihr HTR+ Modell erstellen, indem Sie Bilder hochladen und den Text transkribieren. Eine vollständige Anleitung dazu finden Sie hier: Transkribieren mit Transkribus.
- Wenn Sie bereits vorhandene Transkripte haben, können Sie diese verwenden, um Ihr Modell zu trainieren. Mehr Informationen dazu, finden Sie in dieser Anleitung: Vorhandene Transkriptionen zum Training eines Modells verwenden.

Training

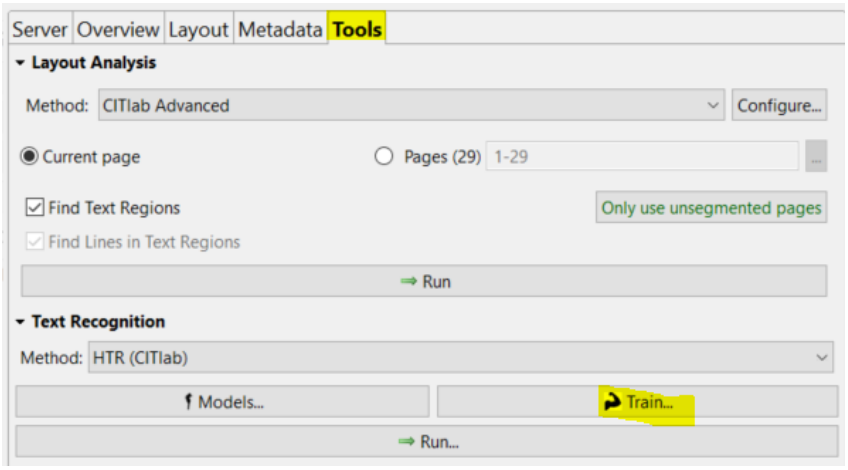
- Die Grundeinstellungen für das Training eines Modells finden Sie im Reiter **“Tools”** im Bereich **“Text Recognition“**.
- Die **“Method”**, “HTR (CITlab)” ist die aktuell effektivste Trainingsmethode.
- Wenn Sie auf die Schaltfläche “Models” klicken, können Sie sehen welche Modelle verfügbar sind und mit welchen Dokumenten Sie trainiert wurden.
- Mit der Schaltfläche **“Train”** gelangen Sie zu den Optionen für das Training eines Modells.



Darstellung 1 Modelltraining Bereich

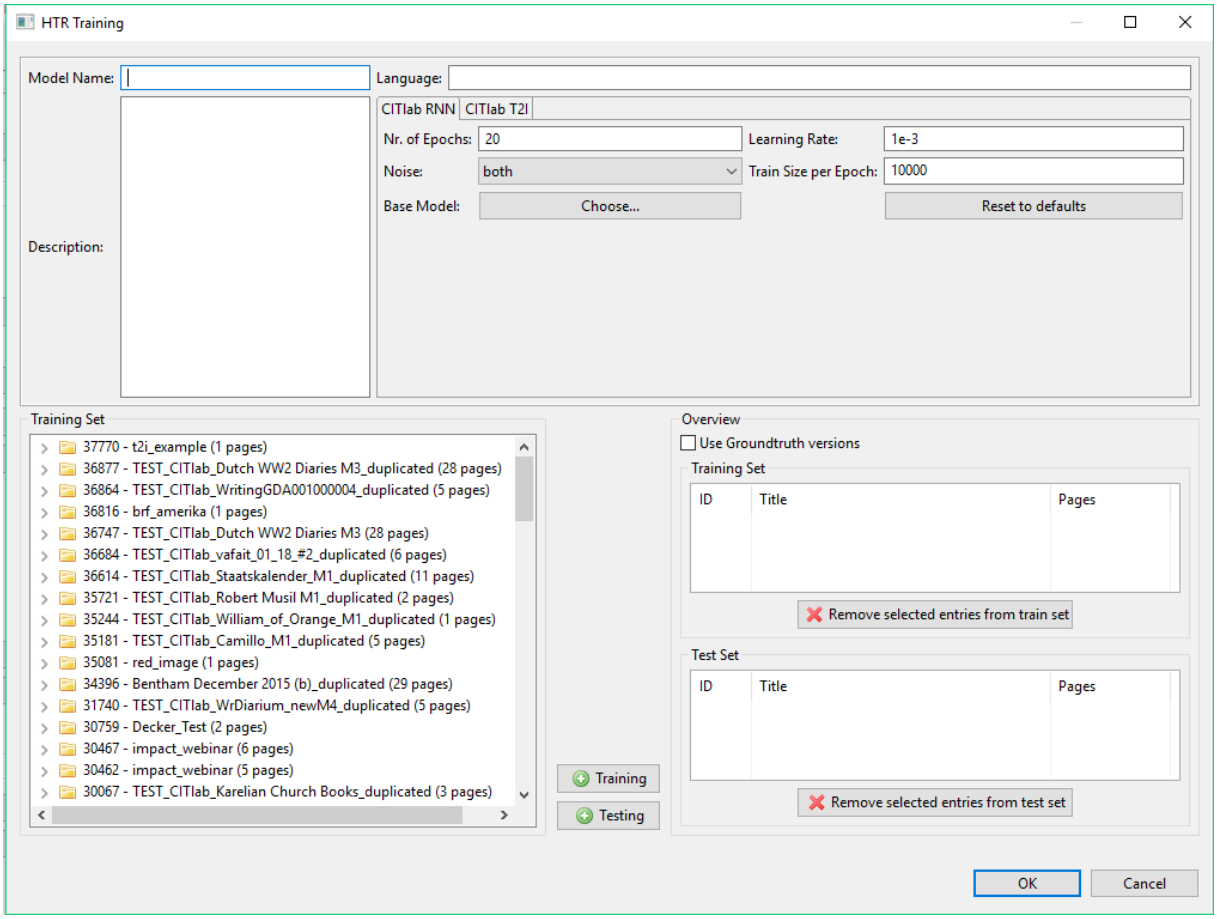
HTR+ Training einrichten

- Um zum “HTR+ Training” Fenster zu gelangen, klicken Sie auf die Schaltfläche “Train” im Reiter “Tools”.



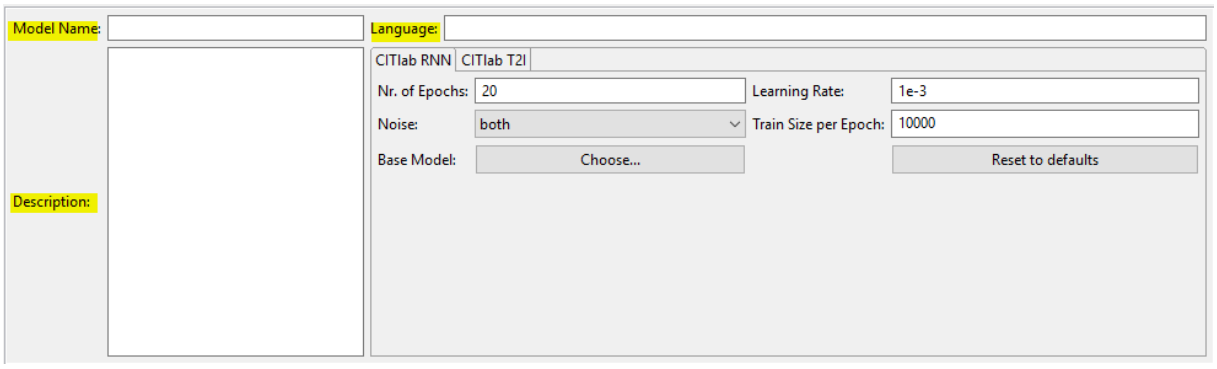
Darstellung 2 Das “HTR Training” Fenster öffnen.

- Das folgende Fenster öffnet sich:



Darstellung 3 "HTR Training" Fenster

- Im oberen Bereich geben Sie die Angaben über Ihr Modell ein.



Darstellung 4 Angaben zum Modell einfügen

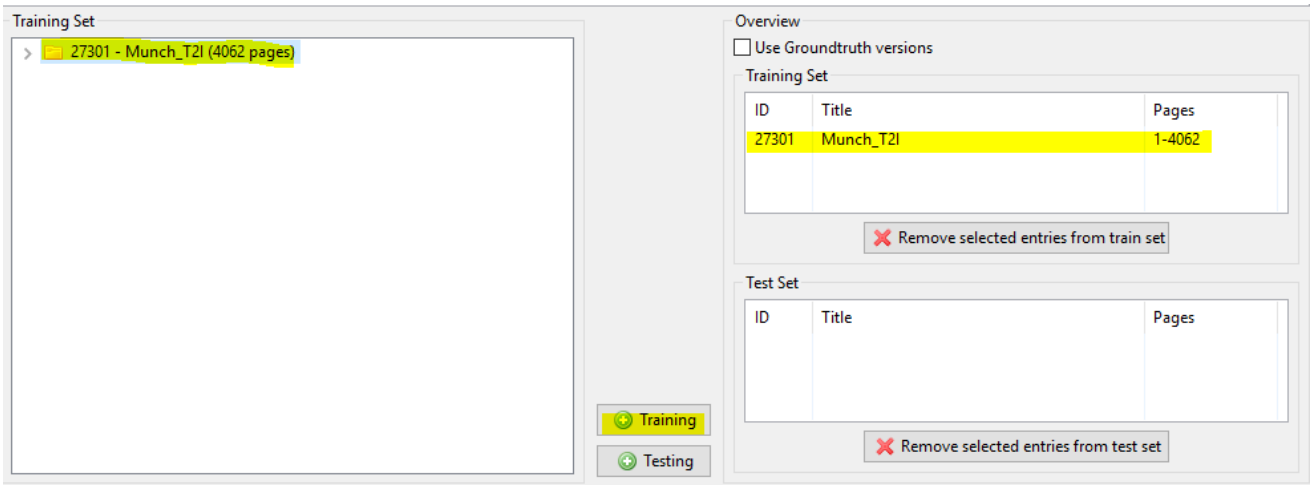
- Fügen Sie bitte Folgendes hinzu
  - Name des Modells (frei wählbar)
  - Sprache (der Schrift im Dokument)
  - Beschreibung (der Dokumente und der Seiten die als Trainings- und Testdaten dienen)
- Achtung: "Nr. of Epochs" bezieht sich darauf, wie oft die Trainingsdaten evaluiert werden. Wenn Sie die Anzahl der Epochen erhöhen dauert der Trainingsprozess länger.

Base Model

- Wenn Sie nicht von ganz vorne beginnen möchten und bereits ein Modell existiert, auf das Sie Ihr Training aufbauen können, gibt es in Transkribus die Möglichkeit, dem Training ein Base-Model hinzuzufügen. Die Daten, die das Base-Model enthält, werden so zum neuen Model dazutrainiert.
- Um von diesem Vorteil zu profitieren, müssen die Daten des Base-Models dem neuen Model ähnlich sein.
- Mit Hilfe eines Base-Models kann der Trainingsprozess verschnellert werden. Eine Verbesserung muss im individuellen Fall getestet werden und kann nicht immer garantiert werden.
- Ein großer Vorteil der Nutzung eines Base-Models ist, dass man schon mit einer kleineren Menge an Trainingsdaten starten kann und so den zeitlichen Aufwand der händischen Transkription verringert.
- Um ein Base Model zu verwenden, wählen Sie das gewünschte mit „Choose…“ neben „Base Model:“.

Training Set

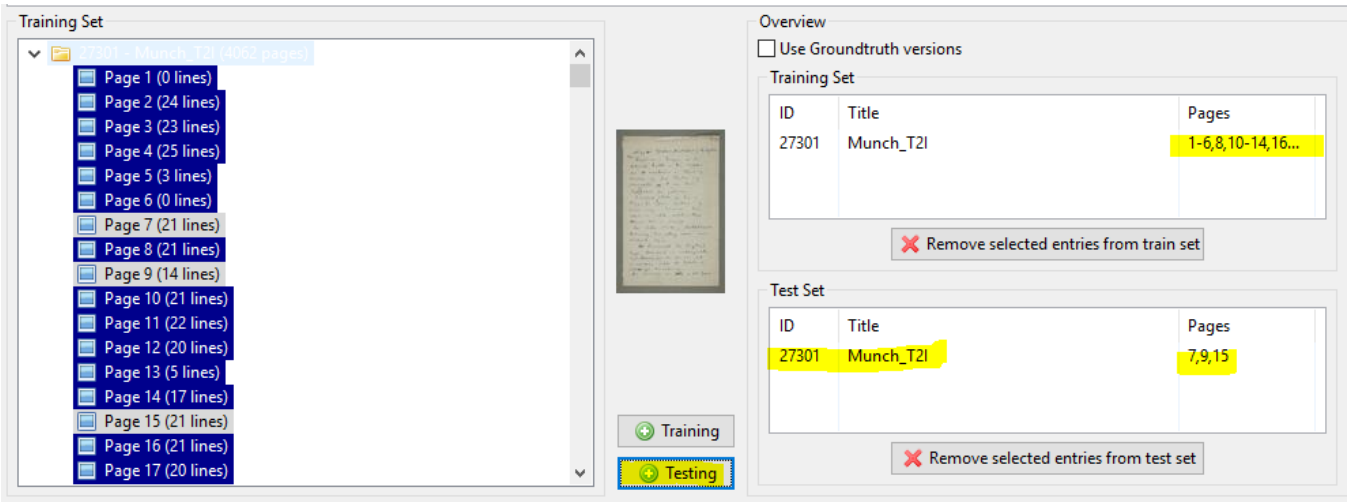
- Als nächstes wählen Sie die Seiten aus, die Sie als Trainingsdatenset verwenden möchten.
- Um alle Seiten Ihres Dokuments zum Trainingsset hinzuzufügen, klicken Sie auf den Ordner und dann auf “+Training”.
- Um eine Reihe von Seiten Ihres Dokuments zum Trainingsset hinzuzufügen, doppelklicken Sie auf den Ordner, dann klicken Sie auf die erste Seite, die Sie hinzufügen möchten, halten Sie die Umschalttaste gedrückt und klicken Sie auf die letzte hinzuzufügende Seite. Dann klicken Sie auf “+Training”.
- Um einzelne Seiten Ihres Dokuments zum Trainingsset hinzuzufügen, doppelklicken Sie auf den Ordner, dann halten Sie die „STRG“-Taste gedrückt und klicken auf die Seiten, die Sie als Trainingsdaten verwenden möchten. Wenn Sie alle ausgewählt haben, klicken Sie auf “+Training”.
- Die ausgewählten Seiten erscheinen dann im Bereich “Training Set”.



Darstellung 5 Alle Seiten zum Trainingsset hinzufügen

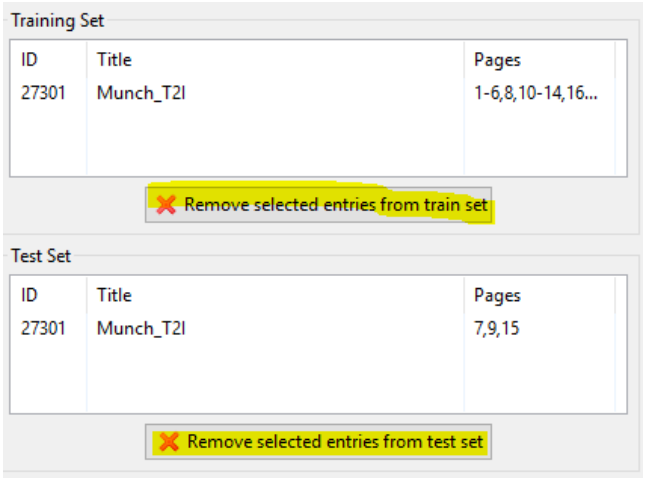
Test Set

- Während des Trainingsvorgangs werden einige Seiten als Test Set zur Seite gelegt. Diese werden nicht für das Training des HTR+ Modells verwendet. Stattdessen dienen Sie dazu, die Performance Ihres Modells zu testen.
- Wir empfehlen zumindest ein Test Set Seite für jede 50-100 Seiten im Training Set.
- Die Seiten in Ihrem Test Set sollten den Stil der Seiten des Training Sets widerspiegeln.
- Je mehr Seiten Ihr Test Set enthält, desto länger dauert das Training.
- Um Seiten zum Test Set hinzuzufügen folgen Sie demselben Ablauf wie beim Training Set aber klicken Sie auf die Schaltfläche „+Testing“.



Darstellung 6 Seiten zum Test Set hinzufügen

- Um Seiten aus dem “Training Set” oder dem “Test Set” zu entfernen, klicken Sie auf die Seite und dann auf das rote „X“.

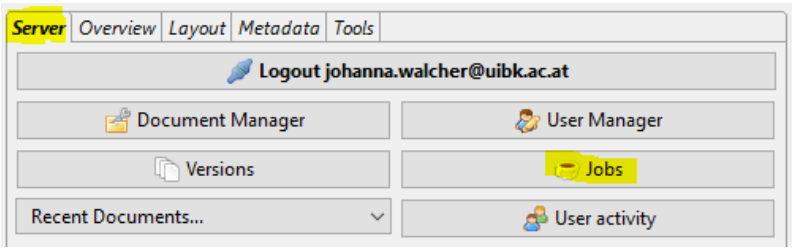


Darstellung 7 Seiten entfernen

- Sie können die Seiten, die im Test Set verwendet werden, in der Modellbeschreibung anführen.
- Sie können das Training starten, indem sie auf „OK“ klicken.

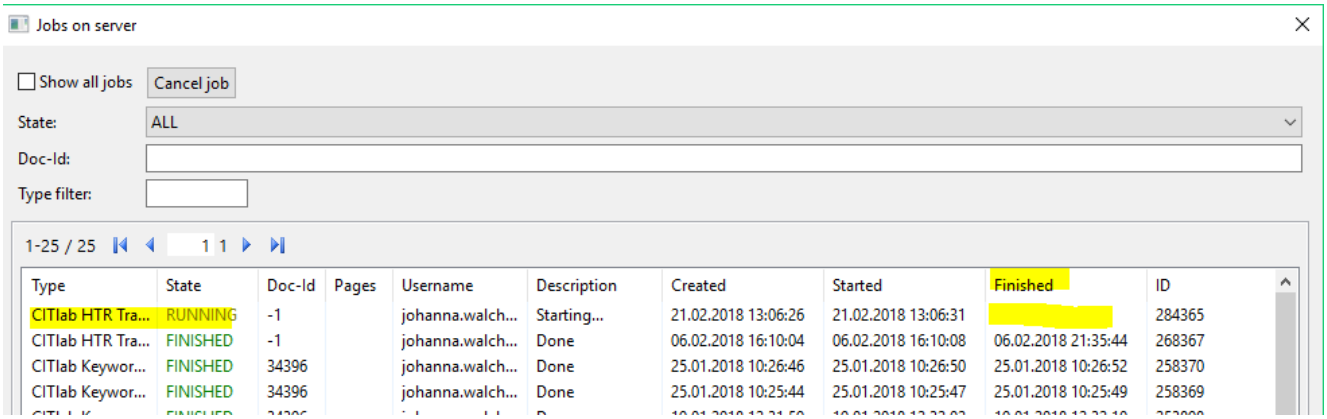
Fortschritt verfolgen

- Sie können den Fortschritt des Trainings verfolgen, indem Sie auf die Schaltfläche „Jobs“ im Reiter „Server“ klicken.



Darstellung 8 Fortschritt des Trainings mit der Schaltfläche "Jobs" überprüfen

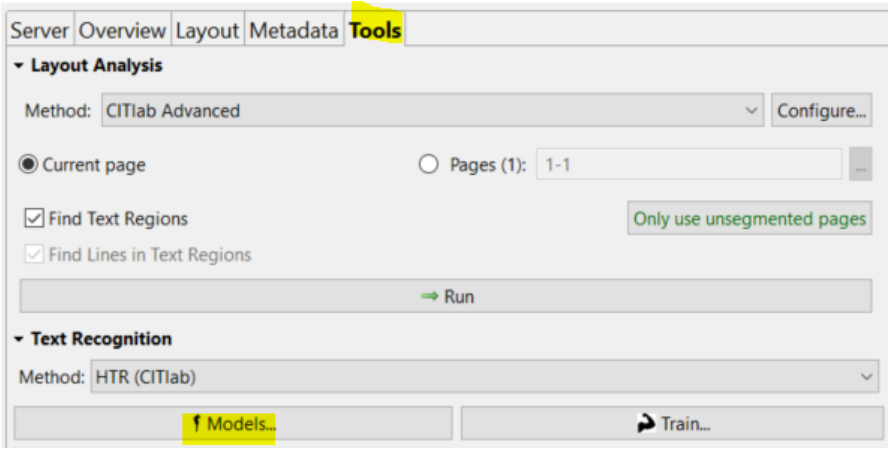
- Der Abschluss jeder Epoche und auch der Abschluss des Trainings wird im Fenster „jobs on server" angezeigt.
- Das Training eines HTR+ Modells nimmt etwas Zeit in Anspruch. Sie können währenddessen andere Arbeiten in Transkribus erledigen oder die Plattform schließen, ohne das Training zu unterbrechen.



Darstellung 9 "Jobs on server" Überblick

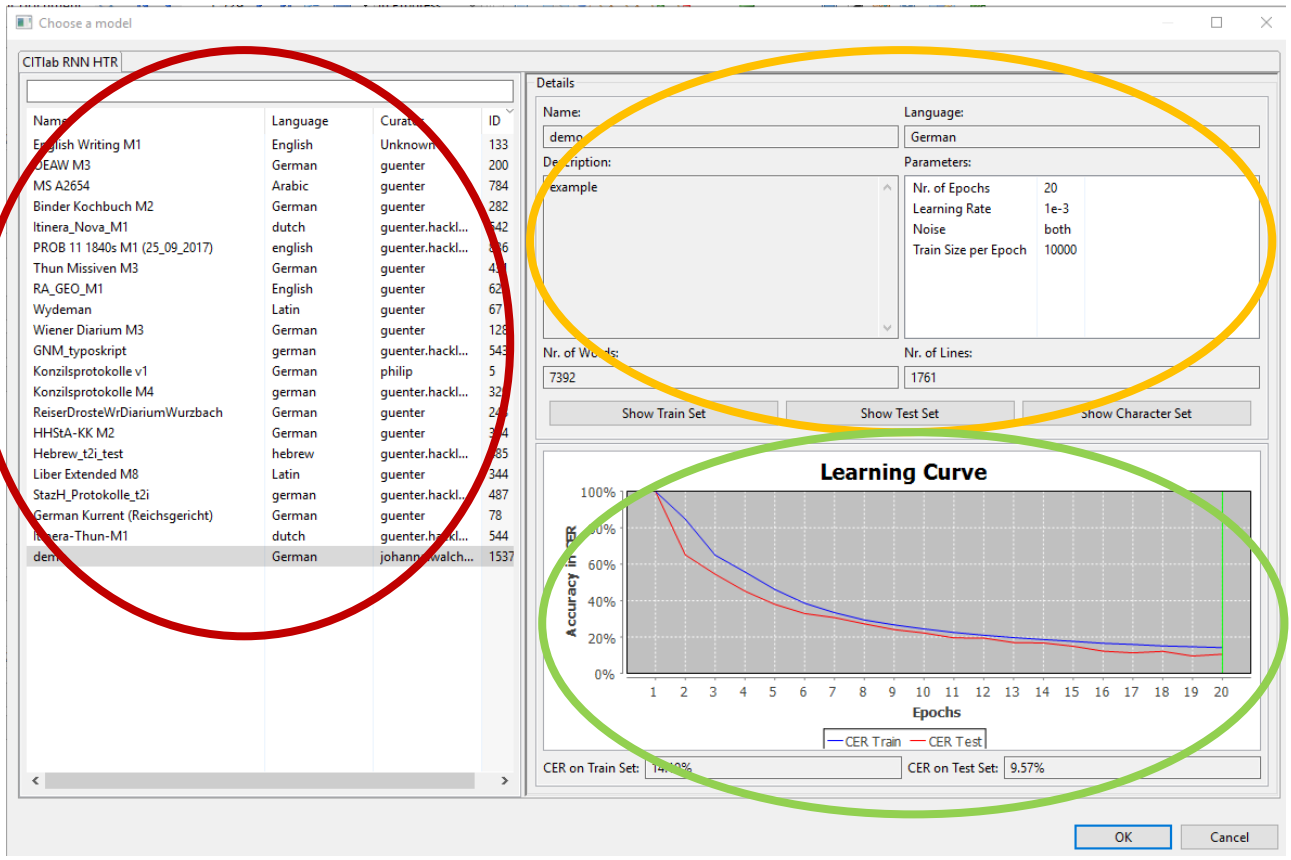
Nach dem Training

- Nach dem Abschluss des Trainings steht das Modell in Ihrer Collection zur Verfügung.
- Um das Modell aufzurufen klicken Sie auf "Models" im Reiter "Tools".



Darstellung 10 Das "Choose a model" Fenster öffnen

- Es öffnet sich das folgende Fenster:

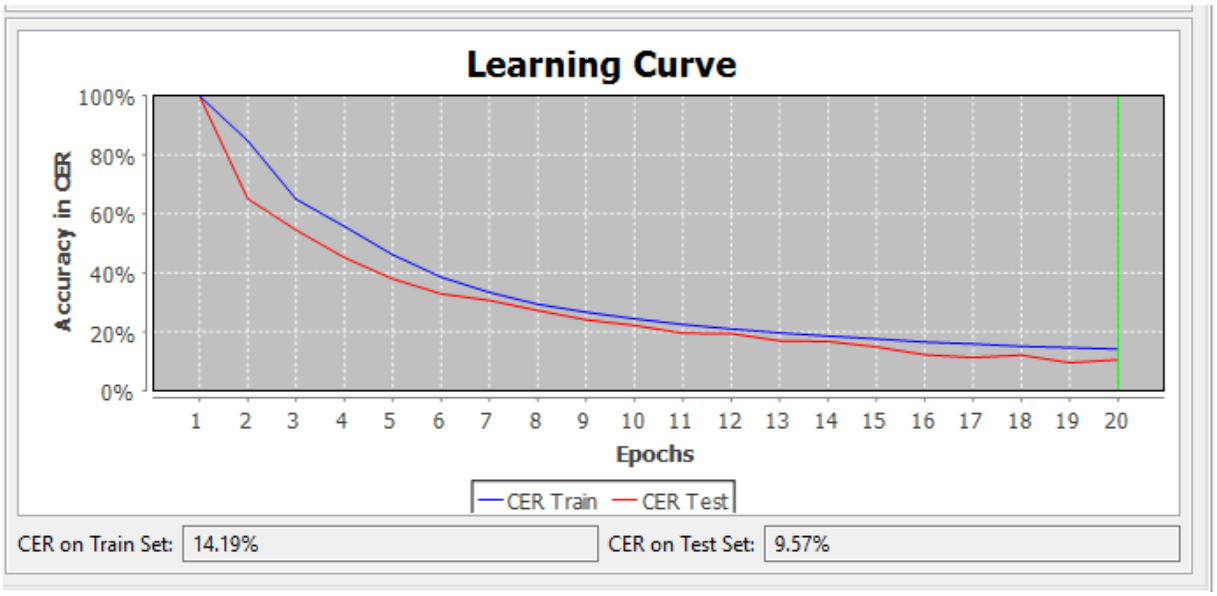


Darstellung 11 "Choose a model" Fenster

- Auf der linken Seite sehen Sie einen Überblick aller verfügbaren Modelle.
- Oben rechts sehen Sie die Angaben zum Modell.
- Rechts unten sehen Sie die Lernkurve des Modells. Mehr Informationen zu diesen Statistiken finden Sie weiter unten.

Statistiken

- Die "Learning Curve" Grafik verbildlicht die Genauigkeit Ihres Modells.



Darstellung 12 "Learning Curve" Ihres Modells

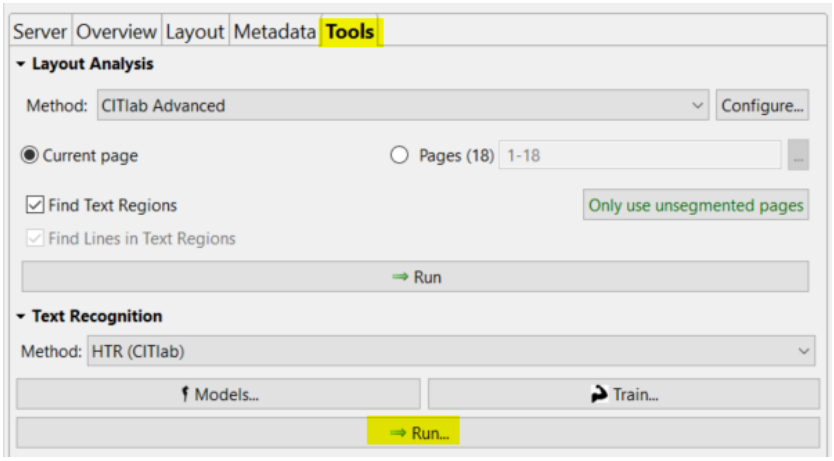
- Die y-Achse ist als “Accuracy in CER” definiert (siehe Darstellung 12).
- “CER“ steht für **Character Error Rate**, also der Prozentsatz der Zeichen, die von der HTR nicht richtig transkribiert wurden.
- “**Accuracy in CER**” wird in Prozent auf der y-Achse angezeigt. Die Kurve startet immer bei 100% und bewegt sich mit Trainingsfortschritt und Verbesserung des Modells nach unten.

- Die X-Achse ist als “**Epochs**” definiert.
- Während des Trainingsprozesses führt Transkribus nach jeder Epoche eine Evaluation durch. In Darstellung 12 wurde das „Training Set“ in 20 Epochen unterteilt.
- Wenn Sie ein Modell trainieren, können Sie angeben in wie viele Epochen das Training Set unterteilt werden soll. Je mehr Epochen es gibt desto länger dauert das Training.

- Die **Grafik** zeigt eine rote und eine blaue Linie.
- Die **blaue Linie** zeigt den Fortschritt des Trainings.
- Die **rote Linie** zeigt den Fortschritt der Evaluierungen im Test Set.
- Zuerst trainiert sich das Programm im **Training Set**, und testet sich danach mit Hilfe von den Seiten im **Test Set**.
- Unter der Grafik sind zwei Prozentzahlen sichtbar, die sich auf die CER für das Training Set und das Test Set beziehen.
- In Darstellung 12 hat das Modell eine CER von 14.19 % für das Training Set und eine CER von 9.57 % für das Test Set.
- Der Wert für das Test Set ist aussagekräftiger, weil er zeigt, wie gut das Modell auf Seiten abschneidet, auf die es nicht trainiert wurde.
- Resultate mit einer CER von 10 % oder weniger können für automatisierte Transkriptionen verwendet werden.
- Resultate mit einer CER von 20 -30 % reichen aus, um mit Keyword Spotting zu arbeiten. Mehr Informationen dazu gibt es hier: Suche in Dokumenten mit Keyword Spotting.

HTR Transkripte erstellen

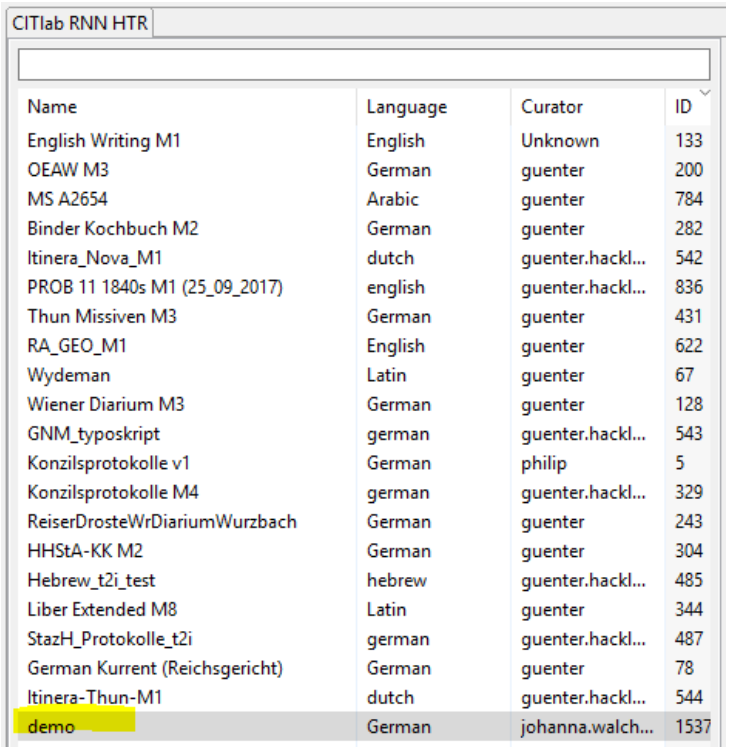
- Mit Ihrem Modell können Sie jetzt automatisch Transkripte von Dokumenten in Ihrer Kollektion generieren.
- Zuerst laden Sie Ihr Dokument auf Transkribus hoch.
- Dann segmentieren Sie Ihr Dokument in Textregionen, Zeilen und Baselines.
- Für mehr Informationen über das Hochladen und Segmentieren, lesen Sie bitte die Anleitung Transkribieren mit Transkribus.
- Um auf Ihr Modell zuzugreifen, klicken Sie auf den Reiter “Tools” und gehen Sie zum “Text Recognition“ Bereich.
- Klicken Sie auf “Run” und dann auf “Configure”. Wählen Sie Ihr HTR Modell aus der Liste auf der linken Bildschirmseite aus und klicken Sie auf OK.
- Wählen Sie aus, ob Sie ein HTR-Transkript von einer oder mehreren Seiten erstellen möchten.
- Klicken Sie auf “Run” um den Texterkennungsprozess zu starten.
- Sobald die Texterkennung fertig ist, müssen die Seiten neu geladen werden und es erscheint die automatische Transkription im Textfeld.



Darstellung 13 Modell starten

Ein Modell teilen

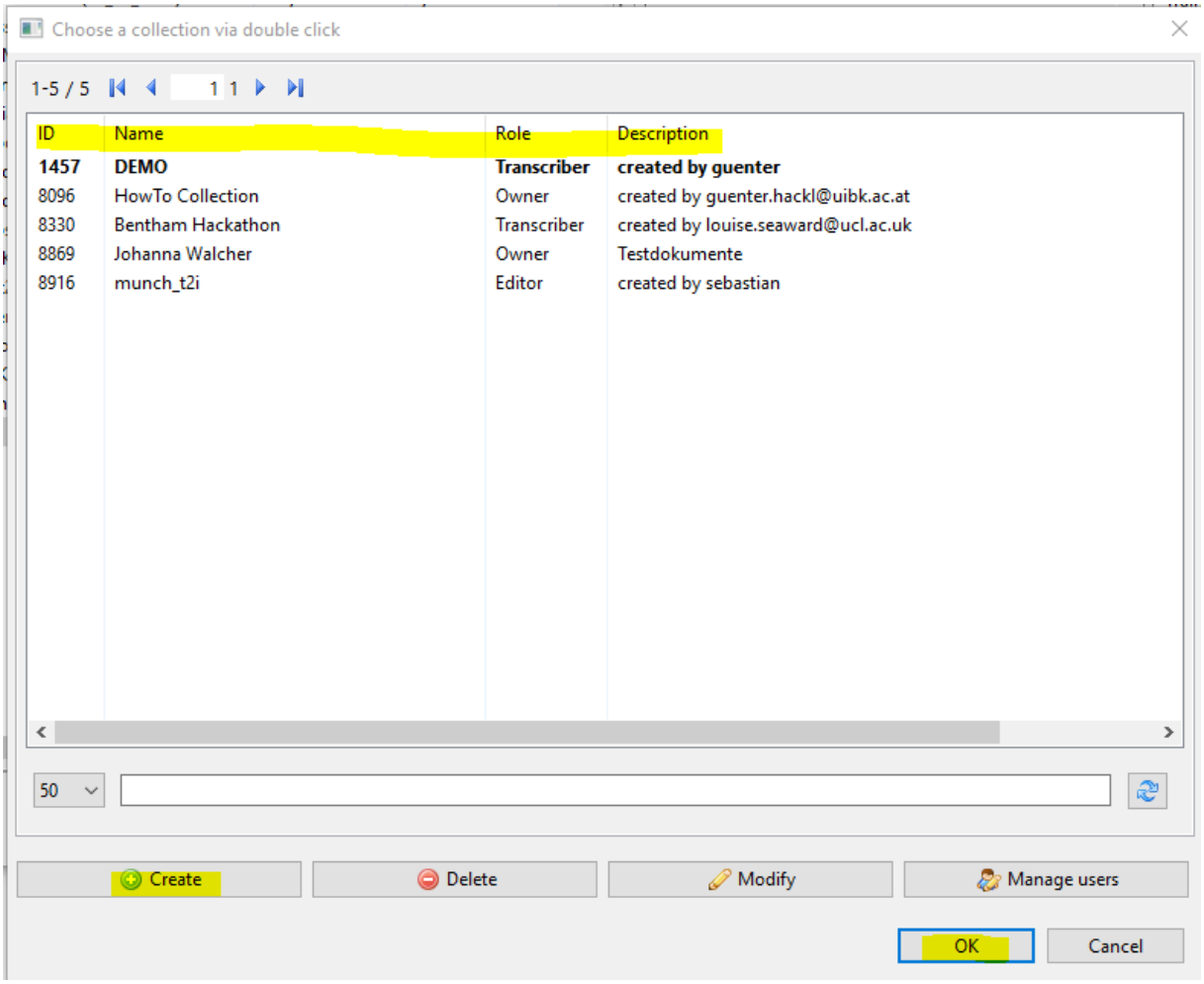
- Sie können Ihr HTR-Modell mit anderen Kollektionen in Transkribus teilen, sowohl mit Ihren eigenen, als auch mit anderen Nutzern.
- Wenn Sie Ihr Modell mit einer anderen Kollektion teilen möchten, müssen Sie Zugriff auf diese Kollektion haben.
- Rechtsklicken Sie auf den Namen Ihres Modells (auf der linken Seite des “Choose a model” Fensters).



Darstellung 16 Ein Modell durch Rechtsklick auf das Modell teilen

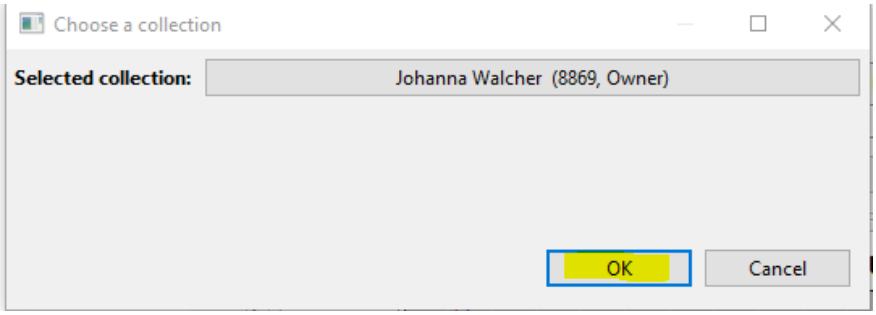
- Dann wählen Sie “Share model…”
- Das “Choose a collection via double click” Fenster öffnet sich.
- Im nächsten Fenster klicken Sie auf die Kollektion, mit der Sie das Modell teilen möchten und dann auf “OK”.

- In diesem Fenster können Sie auch mit der Schaltfläche “Create” eine neue Collection für das Modell erstellen.
- Klicken Sie “OK” um zu bestätigen.

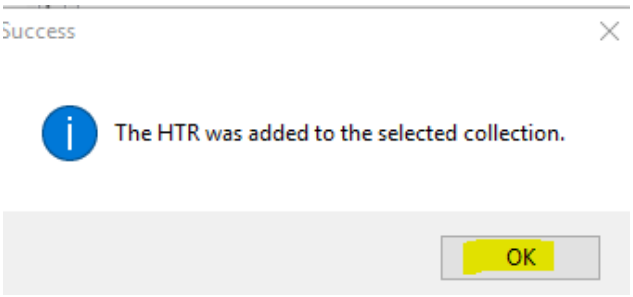


Darstellung 17 Modell teilen

- Wenn Sie die Collection ausgewählt haben, klicken Sie noch einmal auf “OK” und das Modell wird geteilt.



Darstellung 18 Das Teilen des Modells bestätigen



Darstellung 19 Das Modell wurde geteilt

Vorteile des Modelltrainings

- Nach Abschluss des Trainings können Sie Ihr Modell auf jedem anderen historischen Dokument mit ähnlicher Schrift ausprobieren.
- Sie können Ihr Dokument mit anderen teilen, die auch davon profitieren können.
- Sie können das Training mit mehr Daten wiederholen um noch bessere Resultate zu erzielen.
- Sie können die Genauigkeit Ihres Modells mit der “Compute Accuracy” Funktion messen. Die Resultate der HTR hängen davon ab, wie ähnlich und wie klar die Schrift im historischen Dokument ist.
- Das Transkribus Team arbeitet an einem Algorithmus, der es ermöglicht, automatisch jede Art von Dokument zu transkribieren, ohne vorher Trainingsdaten vorbereiten zu müssen. Die Technologie lernt vor allem von Trainingsdaten die in Transkribus verarbeitet werden.
- Je mehr Daten zur Verfügung stehen, desto effizienter wird die Technologie. Trainieren Sie Ihr eigenes Modell und seien Sie Teil davon! 😊

Credits

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, das von der Europäsichen Kommission finanziert wird.