READ
Recognition and Enrichment of Archival Documents

Strukturkennzeichnung und-training

Version v.1.9.1
Letzte Überarbeitung dieser Anleitung 12.12.2019

Diese Anleitung zeigt Ihnen, wie Sie Ihre Dokumente mit Strukturkennzeichnungen, wie „paragraph“, „heading“, „caption“ oder „footer“ versehen können. Diese Kennzeichnung macht es möglich, die Struktur Ihrer Dokumente zu definieren. Mittlerweile kann die Struktur in Transkribus aus trainiert werden.

Für den Fall, dass Sie nach Informationen bezüglich satzbasierter Kennzeichnungen (Tagging – z.B. Personen, Orte, etc.) suchen, konsultieren Sie bitte die Anleitungen Transkriptionen mit Auszeichnungen versehen und Transkribus Transcription Conventions.

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.

Einführung

Mit der Strukturkennzeichnung (Structural Tagging) in Transkribus können Sie die Struktur Ihrer Dokumente auszeichnen.

Zudem ist es möglich, Modelle für die automatische Strukturerkennung zu trainieren. Strukturkennzeichnungen bilden die Trainingsdaten für diesen Prozess.

Es ist nicht nötig jede einzelne Struktureigenschaft im Dokument zu kennzeichnen – konzentrieren Sie sich auf jene Abschnitte, die für Sie wichtig sind.

Die Strukturerkennung in Transkribus ermöglicht es Ihnen:

- Die Dokumente in Strukturabschnitte wie Absätze, Überschriften oder Seiten zu unterteilen
- Eigene Kennzeichnungskategorien für Ihre individuellen Bedürfnisse hinzuzufügen
- Diese Kennzeichnungen in Zukunft für das Training eines Modells zu nutzen

Interface

- Öffnen Sie als erstes das Dokument in Transkribus
- Die Strukturkennzeichnungsinterface finden Sie, indem Sie zuerst den “Metadata“ Tab und dann den „Structural“ Tab wählen.

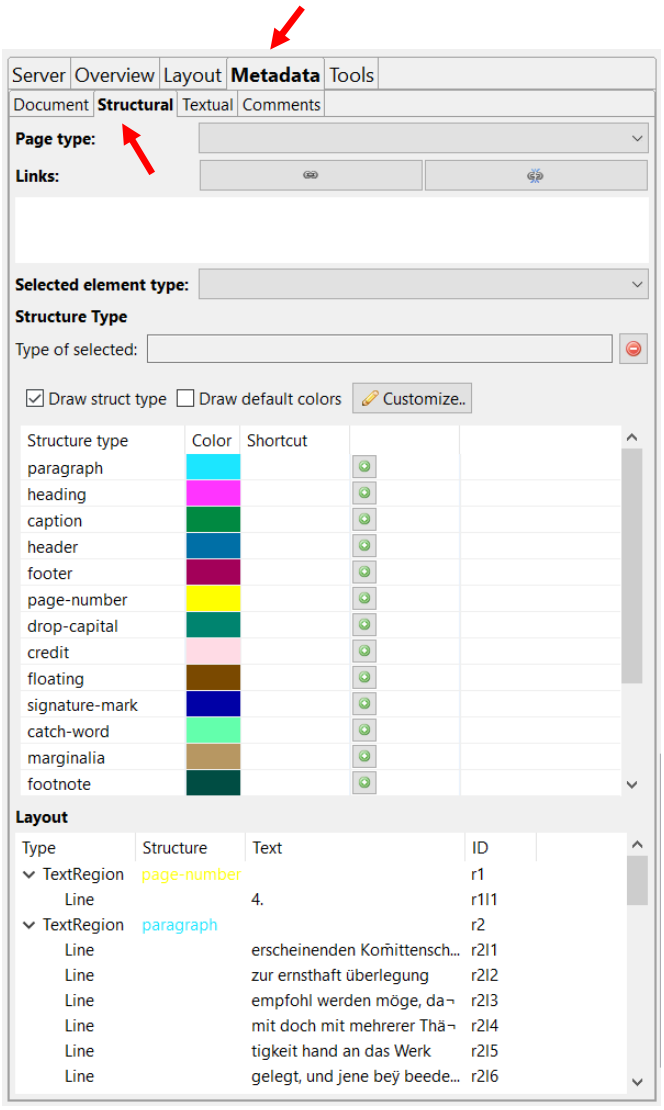


Abbildung 1 Wo Sie die Strukturerkennungsoptionen finden

- In der Mitte des Tabs werden die vordefinierten Strukturkategorien angezeigt

Eigene Kennzeichnungskategorien definieren

- Um Ihre eigenen Kategorien zu definieren, klicken Sie auf die „Customize“ Schaltfläche, damit öffnen Sie das „Tag configuration“ Fenster.

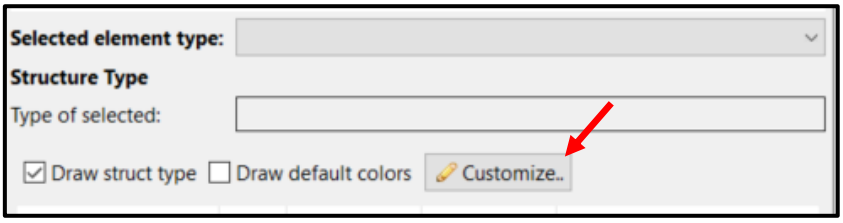


Abbildung 2 „Customize“ Schaltfläche

- Um eine neue Kennzeichnungskategorie zu definieren, tippen Sie einfach den Namen in das leere Feld am unteren Ende des Fensters, dann klicken Sie auf den grünen Kreis mit dem weißen Kreuz.

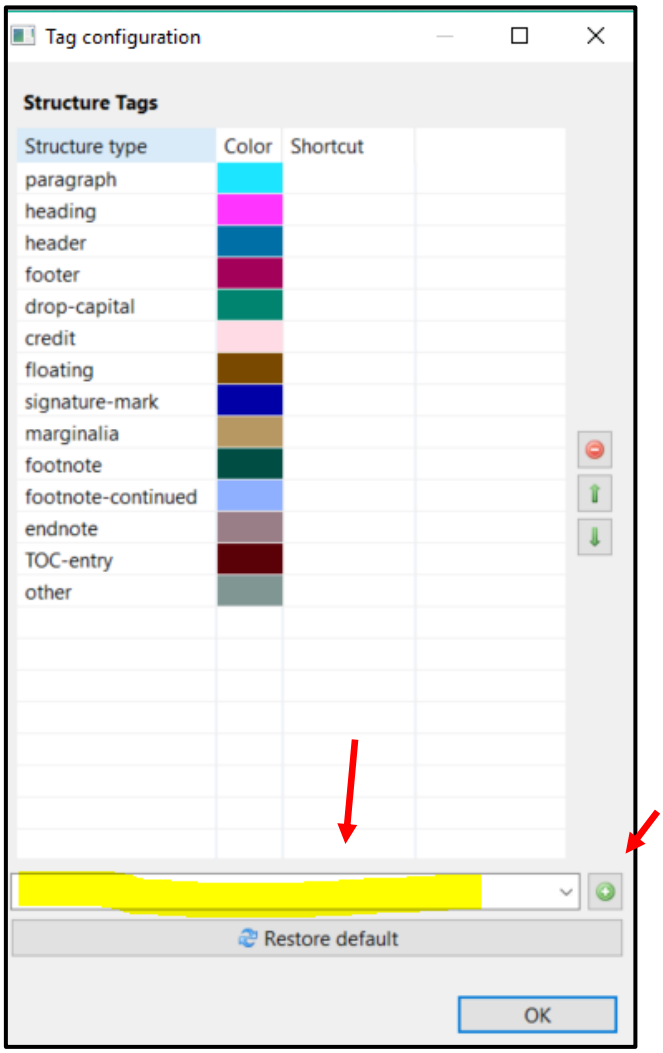


Abbildung 3 Neue Kennzeichnungskategorie definieren

- Im „Tag configuration“ Fenster können Sie die Farben der einzelnen Kategorien anpassen. Dafür klicken Sie einfach auf das Farbfeld, es öffnet sich ein Fenster, in dem Sie die eine Farbe auswählen können.

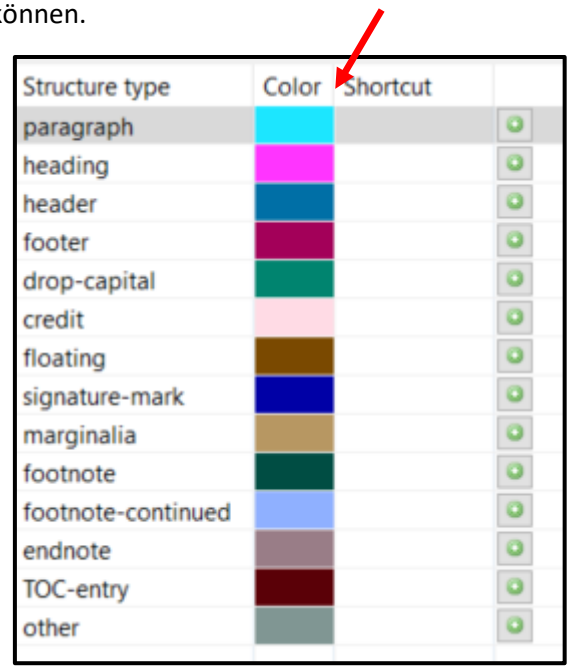


Abbildung 4 Farben anpassen

Abbildung 5 Farbe wählen

- Die neuen Kategorien, die Sie definieren sind automatisch für alle Dokumente in Ihren Kollektionen verfügbar.

Elemente mit Kennzeichnungen versehen

- Textregionen, sowohl als auch Zeilenregionen können mit Strukturkennzeichnungen versehen werden.
- Die automatische Strukturerkennung arbeitet auf Regionsebene, daher macht es Sinn, die Kennzeichnungen auch den Zeilenregionen zuzuweisen.

- **Achtung:** Sie müssen nicht jedes Element Ihrer Dokumente kennzeichnen – Ziel ist es, jene Bereiche zu markieren, die für Sie von Interesse sind.
- Um eine Markierung zu setzen, klicken Sie zuerst auf die „Item visibility“ Schaltfläche im Hauptmenu und stellen Sie sicher, dass Text- und Zeilenregionen im Dokument sichtbar sind.

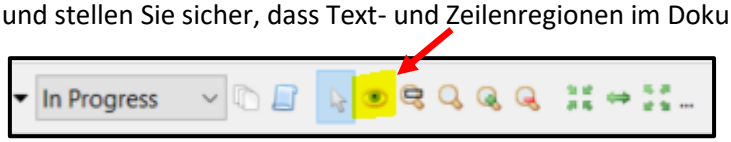


Abbildung 6 “Item visibility” Schaltfläche

- Klicken Sie auf eine Text- oder Zeilenregion im Dokument. Mehrere Regionen zugleich können Sie wählen, indem Sie währen dem Klicken die „Strg“-Taste gedrückt halten.
- Sie haben zwei Optionen für die Markierung:
  - Sie klicken auf den grünen Kreis mit dem weißen Kreuz auf der rechten Seite des „Structural“ Tab.

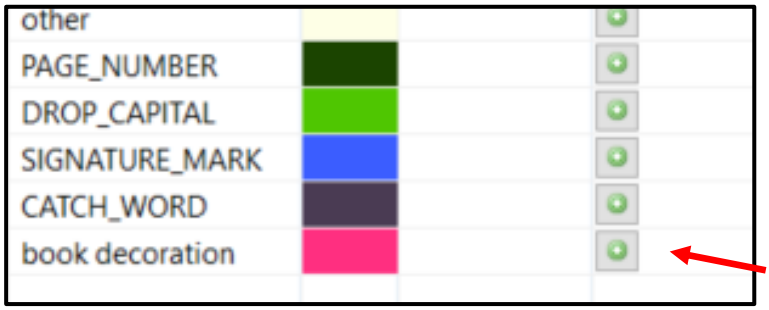


Abbildung 7 Tags zuweisen mit Klick auf den grünen Kreis

  - Oder Sie machen einen Rechtsklick auf den markierten Abschnitt im Bild und wählen dann die gewünschte Kategorie unter „Assign structure type“.

Abbildung 8 Kennzeichnung durch Rechtsklick zuweisen

Formen verbinden

- Es ist möglich zwei Strukturkennzeichnungen miteinander zu verbinden, zum Beispiel eine Verbindung zwischen einer Zeile und der damit verbundenen Fußnote. Dazu verwenden Sie die „Links“ Schaltfläche im „Structural“ Tab.
- Die Schaltfläche links dient dazu, die Verbindung herzustellen, die Schaltfläche rechts, um eine solche Verbindung zu löschen.
- Für das Training spielt die Verbindung von Formen allerdings keine Rolle.

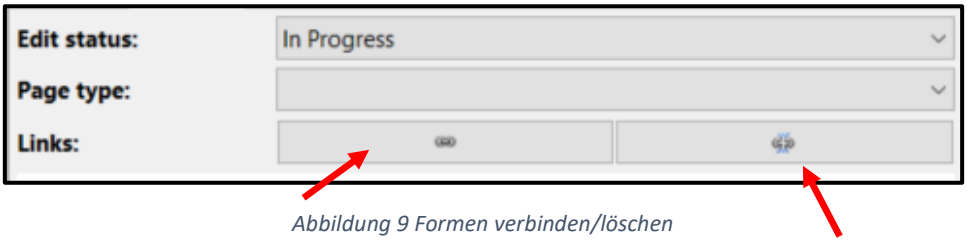


Abbildung 9 Formen verbinden/löschen

Seiten-Typ

- Sie können für jede Seite Ihres Dokuments einen Seiten-Typ (page type) wählen.
- Mögliche Typen sind:
  - Front cover
  - Back cover
  - Title
  - Table-of-contents
  - Index
  - Content
  - Blank
- Öffnen Sie dazu die gewünschte Seite und wählen Sie dann den passenden Typ, indem Sie auf die Schaltfläche neben „Page type“ klicken.
- Auch der Seiten-Typ wird beim Struktur-Training nicht mittrainiert.

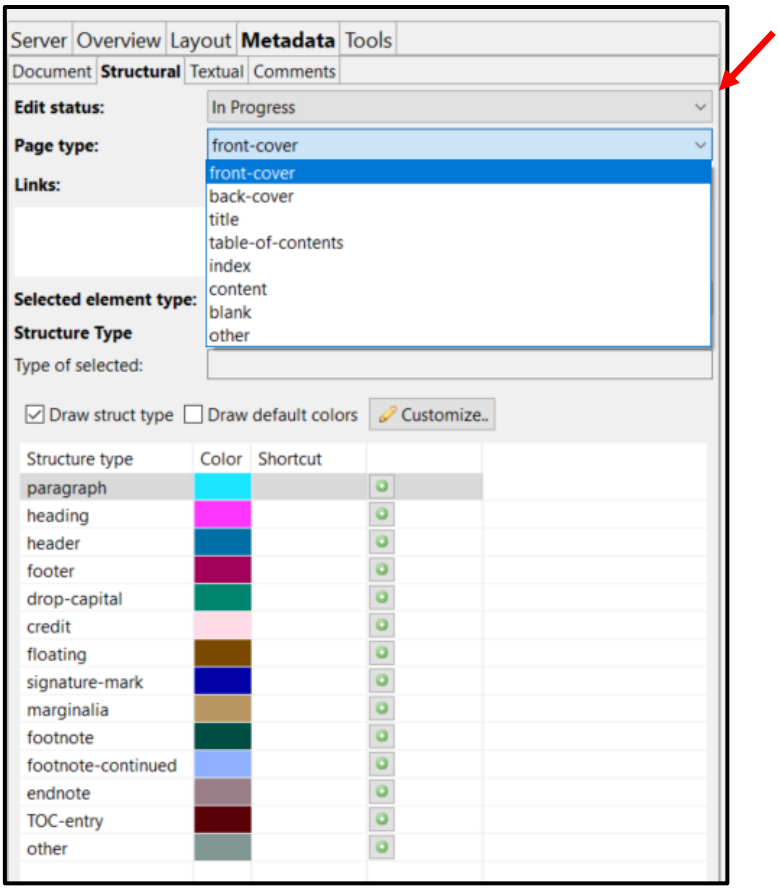


Abbildung 10 Seiten-Typ wählen

Layout Bereich

- Im „Layout“ Bereich des „Structual“ Tabs können Sie zwischen den Strukturtypen im Dokument hin- und herspringen.
- In diesem Bereich finden Sie einen Überblick über die Strukturtypen in Ihrem Dokument und einen Ausschnitt des transkribierten Texts. So kommen Sie schneller zur gewünschten Stelle.
- Um zur gewünschten Stelle zu springen, doppelklicken Sie auf die dazugehörige Zeile im „Layout“ Bereich. Transkribus springt dann automatisch auf diese Stelle im Bild.

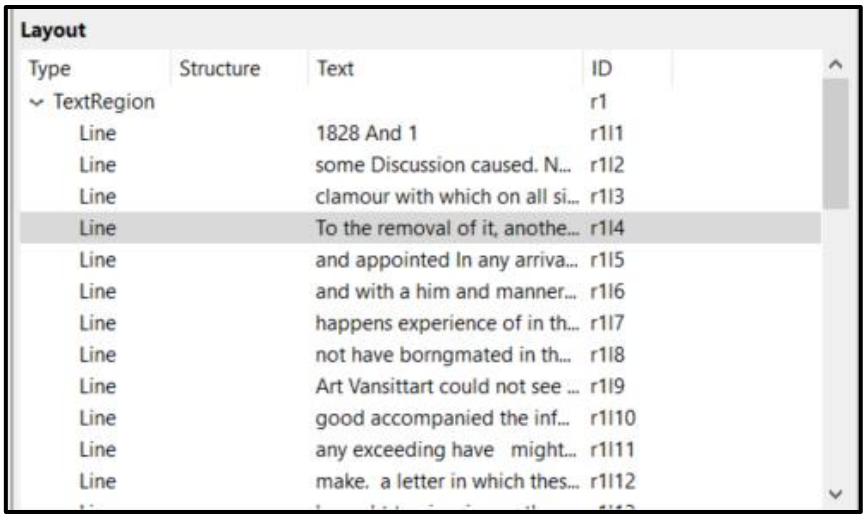


Abbildung 11 “Layout” Bereich

- Die Kennzeichnungen, die von Ihnen gesetzt wurden, erscheinen in der „Structure“ Spalte. Neben der Bezeichnung gibt es einen kleinen Pfeil. Wenn Sie auf diesen klicken, können Sie schnell und einfach den Strukturtyp ändern.

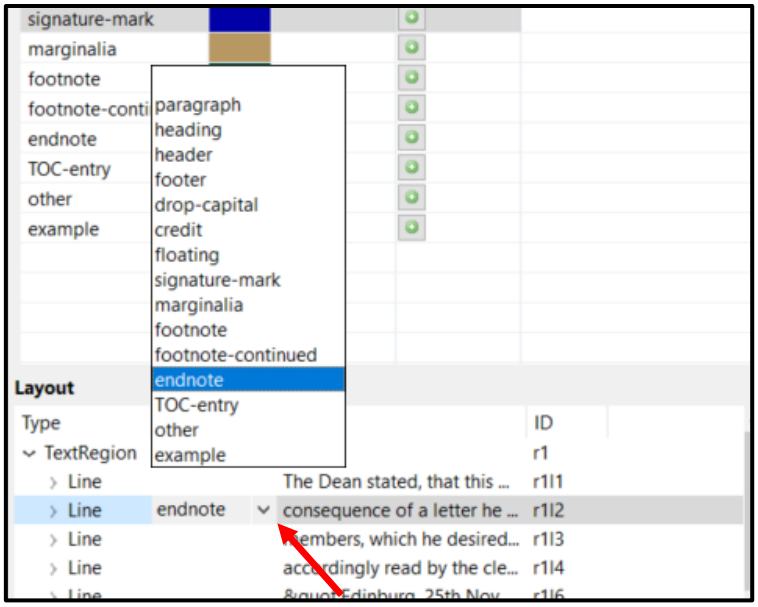


Abbildung 12 Strukturtyp im „Layout“ Bereich ändern

Weitere Optionen

Strukturmarkierungen löschen

- Um Strukturmarkierungen zu löschen, klicken Sie auf den Strukturtyp im „Layout“ Bereich und wählen Sie die „—delete—“ Option (erste Zeile).

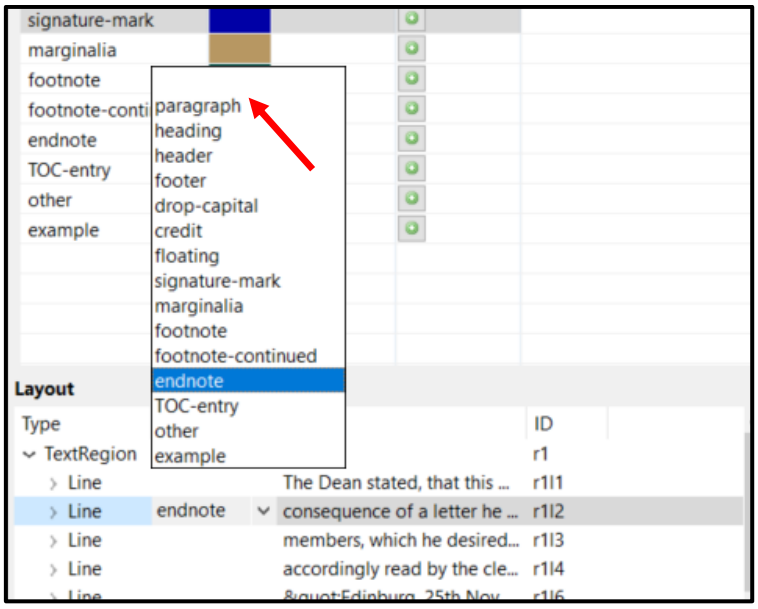


Abbildung 13 Strukturkennzeichnungen löschen

„Draw struct type“ Option

- Wenn Sie diese Option wählen, erscheint der Name des zugewiesenen Strukturtyps im Bild.
- Wenn die Option nicht ausgewählt wird, bleibt die Bezeichnung verborgen.

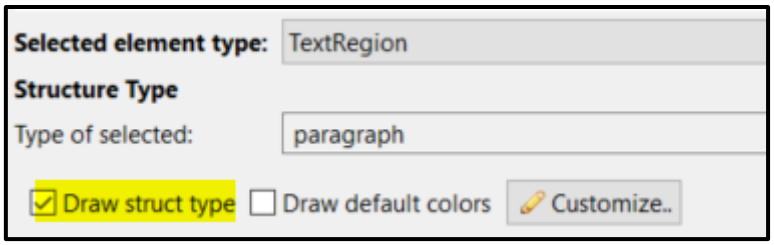


Abbildung 14 Strukturbezeichnung im Bild zeigen

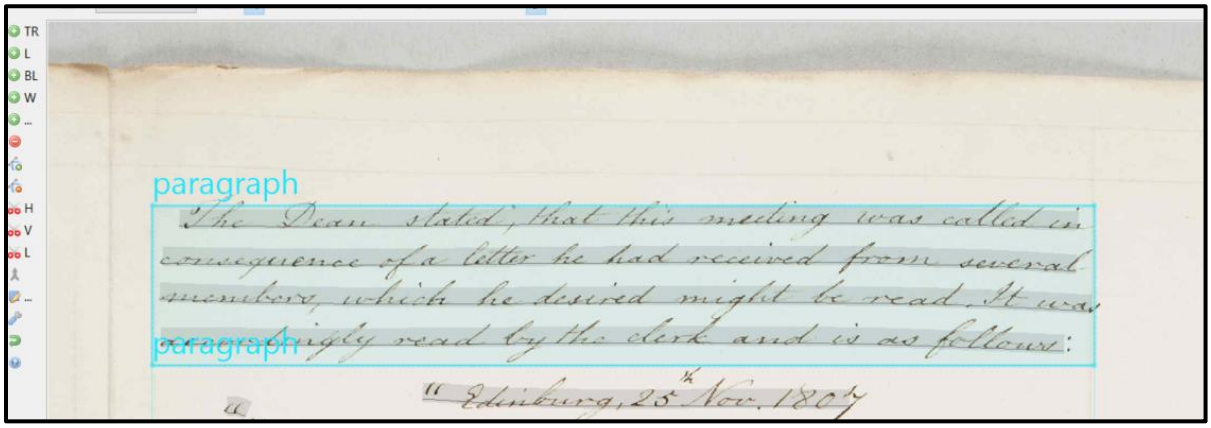


Abbildung 15 „Draw struct type“ Ansicht

„Draw default colours“ Option

- Die Strukturkennzeichnung verwendet andere Farben, als die Text- und Zeilenregionen. Wenn Sie Strukturkennzeichnungen vergeben werden sich deshalb die Farben im Bild verändern.
- Wenn Sie die Standard-Text- und Zeilenregionsfarben anzeigen möchten, wählen Sie die „Draw default colours“ Option.
- Die gesetzten Strukturkennzeichnungen werden dadurch nicht gelöscht, es werden nur die Standardfarben, anstatt der Strukturkennzeichnungsfarben angezeigt.

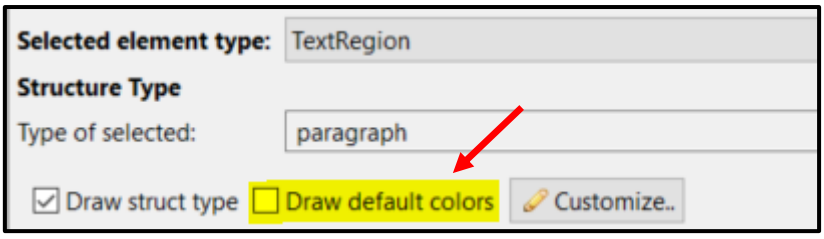


Abbildung 16 „Draw default colors“ – Standardfarben anzeigen

„Type of selected“ Anzeige

- Wenn Sie auf eine Text- oder Zeilenregion in Ihrem Dokument klicken, wird im „Type of selected“ Feld angezeigt, welcher Strukturtyp dem markierten Abschnitt zugewiesen wurde.

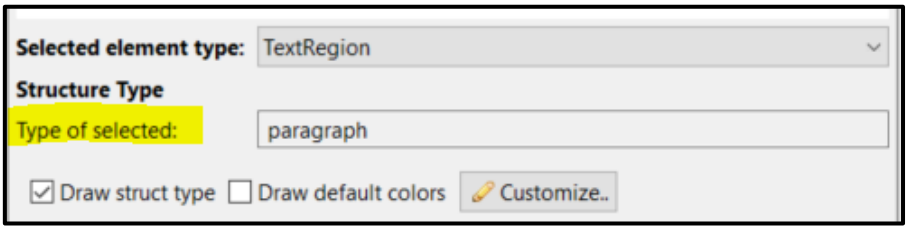


Abbildung 17 „Type of selected“ Anzeige

Struktur-Training

Mit dem Struktur-Training erhalten Sie ein Modell, dass die Struktur Ihrer Dokumente automatisch erkennen kann. Die Effizienz hängt, wie bei Texterkennung, von der Menge und der Klarheit der Daten ab. Wenn Sie in etwa 50 Beispiele jeder Struktur-Tag-Kategorie gekennzeichnet haben, macht es Sinn ein Training zu starten, 50-100 Seiten an Trainingsmaterial sollten also ein brauchbares Modell generieren können. Natürlich können Sie auch ein Training mit weniger Informationen starten. Entsprechend wird das Ergebnis dann aber eine höhere Fehlerquote aufweisen.

Nachdem Sie das Markieren mit dem Structural-Tagging abgeschlossen haben, können Sie direkt mit dem Trainieren beginnen. Öffnen Sie dazu den „Tools“-Tab. Dort finden Sie im Abschnitt „Other Tools“ den „P2PaLA“-Button. Wenn Sie darauf klicken, öffnet sich das Fenster für das Struktur-Training.

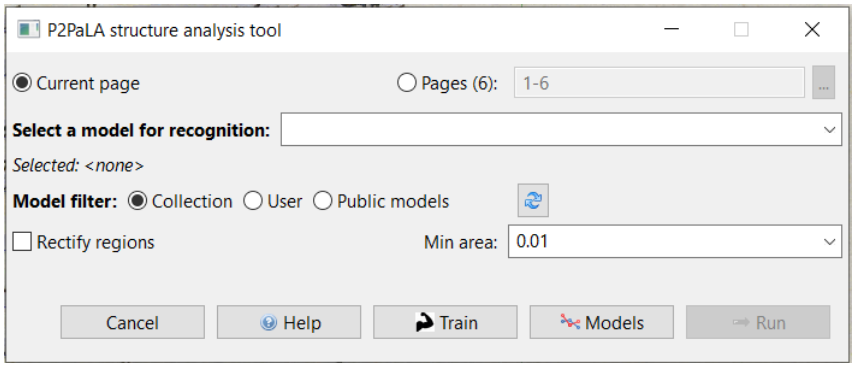


Abbildung 18 Strukturerkennung

Klicken Sie auf „Train“. Damit öffnet sich das Fenster für Trainings-Parameter.

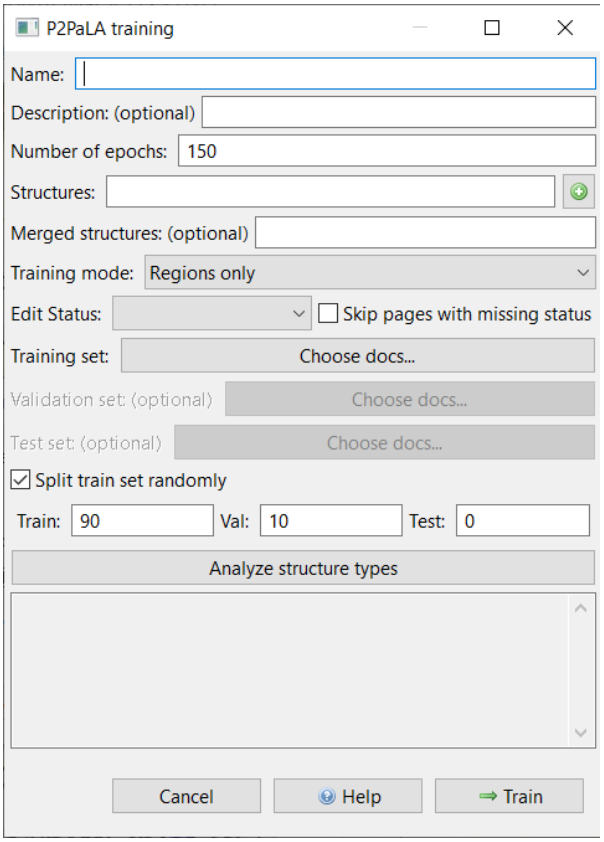


Abbildung 19 Strukturtraining

Im oberen Bereich fügen Sie Informationen zum Modell hinzu.

„Structures“: hier können Sie jene Struktur-Typen hinzufügen, die trainiert werden sollen. Bitte achten Sie beim Eingeben darauf, nicht die Leertaste zu verwenden und auf Groß- und Kleinschreibung (je nachdem wie der Tag anfangs definiert wurde). Wir empfehlen nur Kleinschreibung zu verwenden. Außerdem empfehlen wir Bindestriche (-) und Unterstriche (_) als einzige Spezialzeichen zu verwenden.

- `Beispiel: paragraph heading footnote page-number`

„Merged Structures“: dies wird verwendet, um bestimmte Struktur-Typen im Training mit anderen gleich zu behandeln (z.B. „footnote-continued“ oder „footer“ wie „footnote“). Bei der Eingabe trennen Sie die Liste der Strukturtypen, die zusammengefügt werden sollen, durch einen Beistrich.

- `Beispiel: footnote:footnote-continued,footer heading:header`

“Training mode”: hier können Sie festlegen, ob sie nur Regionen oder Baselines oder beides trainieren möchten. Bitte beachten Sie, dass das Baseline-Training nicht bedeutet, dass Strukturtypen auf Zeilenebene trainiert werden, sondern es dient zur Erkennung der Baselines selbst.

“Edit status”: wenn Sie die letzte Version des Dokuments für das Training verwenden möchten, können Sie dieses Feld frei lassen. Falls Sie auf eine andere Version zurückgreifen möchten, können Sie diese hier auswählen.

“Training Set”: hier wählen Sie die Trainingsdaten aus.

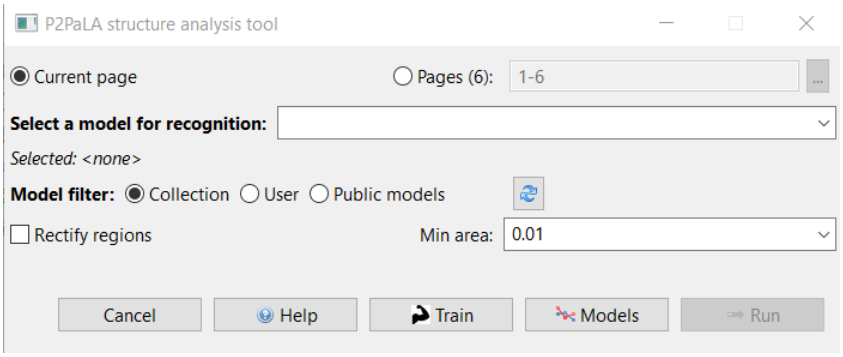
“Analyze structure types”: gibt Ihnen einen Überblick über die Anzahl der verfügbaren Tags in Ihren Dokumenten.

Um das Training zu starten, klicken Sie auf “Train”.

Nach Abschluss des Trainings steht Ihnen das Modell in Ihrer Collection zur Verfügung und kann bei Bedarf mit anderen Collections geteilt werden.

Anwendung eines Struktur-Modells

Wenn Sie ein Strukturmodell auf ein Dokument anwenden möchten, um Strukturtypen erkennen zu lassen, öffnen Sie ebenfalls die „P2PaLa“-Funktion im „Tools“-tab.



Wählen Sie oben im Fenster aus, welche Seiten Sie gerne erkennen lassen möchten.

„Model filter“:

- „Collection“: wenn das gewünschte Modell in Ihrer Collection liegt

- „User“: wenn Sie das Modell trainiert haben
- „Public models“: wenn Sie gerne ein öffentliches Modell verwenden würden.

Die verfügbaren Modelle erscheinen dann neben „Select a model for recognition“. Wählen Sie das Modell, das Sie gerne verwenden möchten. Eine Übersicht der Modelle mit mehr Details erhalten Sie, wenn Sie auf „Models“ klicken.

„Rectify regions“: Bei Aktivierung werden alle Regionen auf den Begrenzungsrahmen der erkannten Region vereinfacht.

„Min area“: Der eingestellte Wert definiert die Größe sehr kleiner (nicht relevanter) Regionen, die entfernt werden. Der Default-Wert beträgt 0.01. Das ist eine recht vorsichtige Einstellung. Der Wert kann normalerweise ohne Bedenken auf 0.1 erhöht werden, außer es befinden sich sehr kleine Regionen von Interesse in Ihren Dokumenten.

Um die Erkennung zu starten, klicken Sie auf „Run“.

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, das von der Europäsichen Kommission finanziert wird.