

# Der Bündner Landbote:

→ Anpassung eines HTR+ Modells mit

Transkribus — Durchgeführt von Sarah Geiger und Jan Mittler an der Albert-

Ludwigs-Universität Freiburg, online präsentiert unter: <https://pieckh.github.io/dig-manu-studies/Studierendenprojekte/>

Im Rahmen der Übung „Digital Manuscript Studies“ wurde eine Vielzahl an Möglichkeiten aufgezeigt, wie in unserem digitalen Zeitalter mit verschiedensten Formen der Überlieferung, seien es Handschriften, Drucke oder Bildmaterial umgegangen werden kann, um eben nicht auf die Materialien im *real life* angewiesen zu sein, sondern einen förderlichen digitalen Zugang zu diesen zu schaffen.

Besonders interessant war in dieser Übung für uns die Vorstellung des Java-Scriptprogramms „Transkribus“. Transkribus wurde an der Universität Innsbruck in Zusammenarbeit mit führenden Forschungsgruppen aus Europa entwickelt. Aufgrund des großen internationalen Interesses wird das Projekt als *European Cooperative* weitergeführt. Transkribus ist eine frei zugängliche Plattform, zur Texterkennung, Layout-Analyse und Strukturerkennung von historischen Überlieferungen. Das bedeutet basal gesagt, dass Transkribus die digitale Arbeit mit alten Schriftstücken erleichtert, indem es den aufwendigen Weg der händischen Transkription verkürzt.

Dies gelingt anhand verschiedener „Modelle“, die in einem *machine learning process* auf unterschiedliche Arten von Texten angewendet werden können (Handschriften, Drucke etc.).

Um aber an den Punkt des *machine learnings* zu kommen, müssen vorher einige Schritte beachtet werden:

## 1. Auswahl des Textmaterials<sup>1</sup>

Die Komplexität und der Umfang der Arbeitsschritte ist abhängig von der Auswahl des Textmaterials. Wir haben uns für die Drucke des „Bündner Landboten“ entschieden, eine der ältesten, digital verfügbaren deutschsprachigen Zeitungen aus Graubünden (Schweiz), einfach unter [e-newspaperarchives.ch/?a=cl&cl=CL2.1846.04&sp=BLB&](http://e-newspaperarchives.ch/?a=cl&cl=CL2.1846.04&sp=BLB&) abrufbar. Die Form der Zeitung, also ein Druck, ist deutlich einfacher zu bearbeiten als Handschriften. Dies liegt daran, dass wir wenig Vorerfahrung zur Transkription in diese Arbeit mit einbringen und somit der Prozess der händischen Transkription sehr zeitintensiv geworden wäre. Andererseits fällt es auch dem Transkribus-Algorithmus leichter, mit gedrucktem, als mit handschriftlichem Material zu arbeiten.

## 2. Layouten und Überprüfung der vorläufigen *textrecognition*

Erstellt man sich nun eine *collection* bei Transkribus, sieht das Rohmaterial erstmal wie folgt aus:



Abbildung 1: Ansicht des Bündner Landboten in Transkribus. // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

Um den nächsten Schritt in Richtung des Erstellens eines eigenen HTR+ Modells zu gehen, muss der Text gelayoutet werden. Dies kann entweder händisch geschehen oder mithilfe der

<sup>1</sup> Das E-Newspaperarchiv stellt seine Dokumente nach einer kurzen Anmeldung zur Verfügung. Im Rahmen unseres Projekts zum „Bündner Landbote“ arbeiten wir ausschließlich mit Materialien, die als *public domain* gekennzeichnet sind.

Im Folgenden werden außerdem selbstangefertigte Screenshots des Programms „Transkribus“ zur Veranschaulichung verwendet.

von Transkribus frei zur Verfügung gestellten automatischen Layoutfunktion.

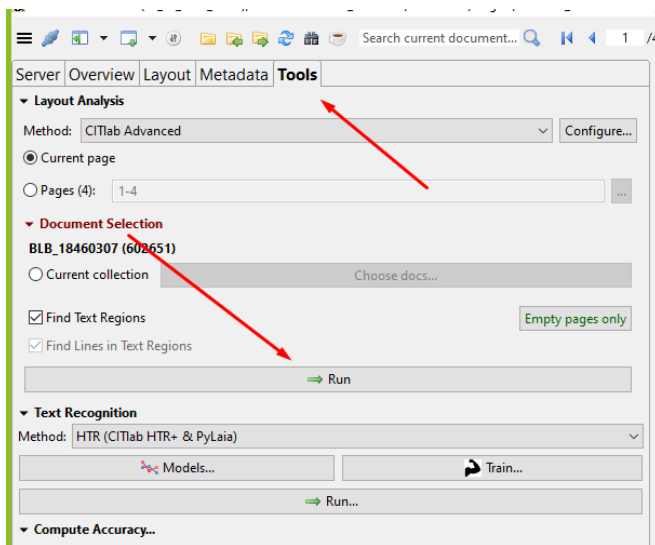


Abbildung 2: tools Ansicht in Transkribus // Screenshot angefertigt von Jan Mittler, Lizenz: public domain.

Diese Funktion sequenziert den Text in Zeilen und Wörter, was die Voraussetzung für eine erfolgreiche Transkription und Erstellung eines Modells ist.

Grundsätzlich sei hier gesagt, dass bei jedem Schritt, den man das Programm „automatisch“ machen lässt, ein *doublecheck* stattfinden sollte. Das bedeutet, dass man die generierte Sequenzierung auf Fehler überprüfen sollte und diese beheben muss. Im Laufe unserer Arbeit hat sich auch gezeigt, dass diese Vorgehensweise des *doublechecking*, also erst der Computer, dann die menschliche Analyse, deutlich zeiteffizienter ist als entweder andersherum oder nur händisch. Besonders in diesen Basisfunktionen ist Transkribus recht akkurat.

Die gelayoutete Seite sieht dann folgendermaßen aus:

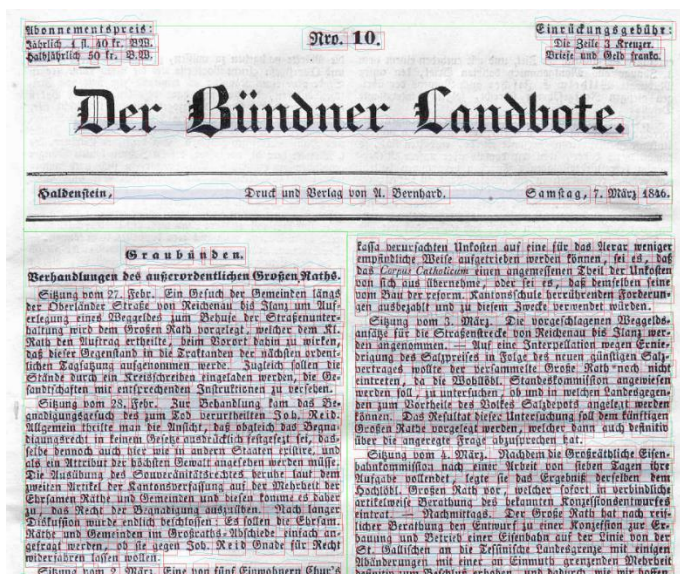


Abbildung 3: textregions in Transkribus // Screenshot angefertigt von Jan Mittler, Lizenz: public domain.

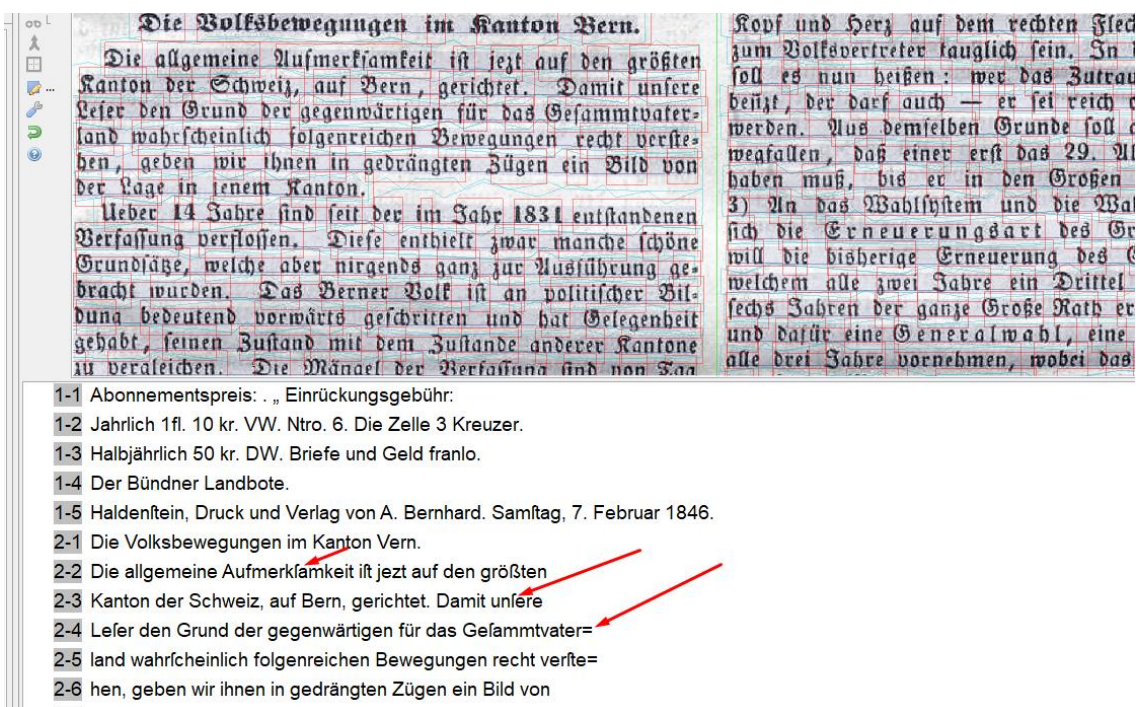
Zu sehen ist, wie einerseits die *regions* von den „leeren“ Bereichen abgetrennt werden. Hat man nun das *doublechecking*, also zum Beispiel die korrekte Nummerierung der *textregions* (Textblöcke) oder die korrekte Anlage der *baselines* (Zeilen) erfolgreich abgeschlossen, kann schon zum nächsten Schritt, dem Transkribieren, übergegangen werden.

Auch hier empfiehlt es sich wiederum, ein zur Verfügung gestelltes Basismodell einer/s anderen Benutzer/in die Vorarbeit leisten zu lassen und dieses auf Fehler zu überprüfen. Wir haben hierfür das Modell „Transkribus German Kurrent M2“ ausprobiert. Die aktuellen *public models* sind unter diesem Link: <https://readcoop.eu/transkribus/public-models/> abrufbar.

 Transkribus German Kurrent M2	German	guenter	CITab HTR+	28.01.21	3209690	9.88%	6.12%
---	--------	---------	------------	----------	---------	-------	-------

Abbildung 4: Das verwendete Modell in Transkribus. // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

In dieser Ausarbeitung finden leider nicht alle durchgeführten Zwischenschritte Platz, diese sind aber im Transkribus Wiki ([https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/#elementor-toc\\_heading-anchor-9](https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/#elementor-toc_heading-anchor-9)) nachvollziehbar. Dies sieht dann wie folgt aus:



**Die Volksbewegungen im Kanton Bern.**

Die allgemeine Aufmerksamkeit ist jetzt auf den größten Kanton der Schweiz, auf Bern, gerichtet. Damit unsere Leser den Grund der gegenwärtigen für das Gesamtvaterland wahrscheinlich folgenreichen Bewegungen recht verstehen, geben wir ihnen in gedrängten Zügen ein Bild von der Lage in jenem Kanton.

Ueber 14 Jahre sind seit der im Jahr 1831 entstandenen Verfassung verlossen. Diese enthielt zwar manche schöne Grundzüge, welche aber nirgends ganz zur Ausführung gebracht wurden. Das Berner Volk ist an politischer Bildung bedeutend vorwärts geschritten und hat Gelegenheit gehabt, seinen Zustand mit dem Zustande anderer Kantone zu vergleichen. Die Mängel der Verfassung sind nun Sanft und Herz auf dem rechten Fleck zum Volkvertreter tauglich sein. In soll es nun heißen: wer das Zutrauen besitzt, der darf auch — er sei reich werden. Aus demselben Grunde soll es wegfallen, daß einer erst das 29. Alter haben muß, bis er in den Großen 3) An das Wahlssystem und die Wahl sich die Erneuerungsgart des Er will die bisherige Erneuerung des welchem alle zwei Jahre ein Drittel sechs Jahren der ganze Große Rath er und dafür eine Generalwahl, eine alle drei Jahre vornehmen, wobei das

1-1 Abonnementspreis: ., Einrückungsgebühr:  
 1-2 Jährlich 1fl. 10 kr. VW. Ntro. 6. Die Zelle 3 Kreuzer.  
 1-3 Halbjährlich 50 kr. DW. Briefe und Geld franco.  
 1-4 Der Bündner Landbote.  
 1-5 Haldenstein, Druck und Verlag von A. Bernhard. Samstag, 7. Februar 1846.  
 2-1 Die Volksbewegungen im Kanton Bern.  
 2-2 Die allgemeine Aufmerksamkeit ist jetzt auf den größten  
 2-3 Kanton der Schweiz, auf Bern, gerichtet. Damit unsere  
 2-4 Leser den Grund der gegenwärtigen für das Gesamtvater=  
 2-5 land wahrscheinlich folgenreichen Bewegungen recht verste=  
 2-6 hen, geben wir ihnen in gedrängten Zügen ein Bild von

Abbildung 5: Transkription einer Beispielsseite durch das Basismodell. // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*

Beim *doublechecking* fällt schnell auf, dass das „Transkribus German Kurrent M2“ Modell das heute im schriftgebrauch unübliche lange-s als solches wiedergibt. Außerdem kämpft das Basismodell mit Umlauten und Satzzeichen am Ende einer Zeile. Kleine Fehler müssen dann händisch überarbeitet werden, um ein möglichst sauberes Modelltraining zu ermöglichen. Ist dies geschehen, kann in den nächsten Schritt übergegangen werden.

### 3. Modelltraining

In der *tools* Übersicht von Transkribus kann nun der vorher freigeschaltete *train button* ausgewählt werden. Das Fenster sieht dann wie folgt aus:

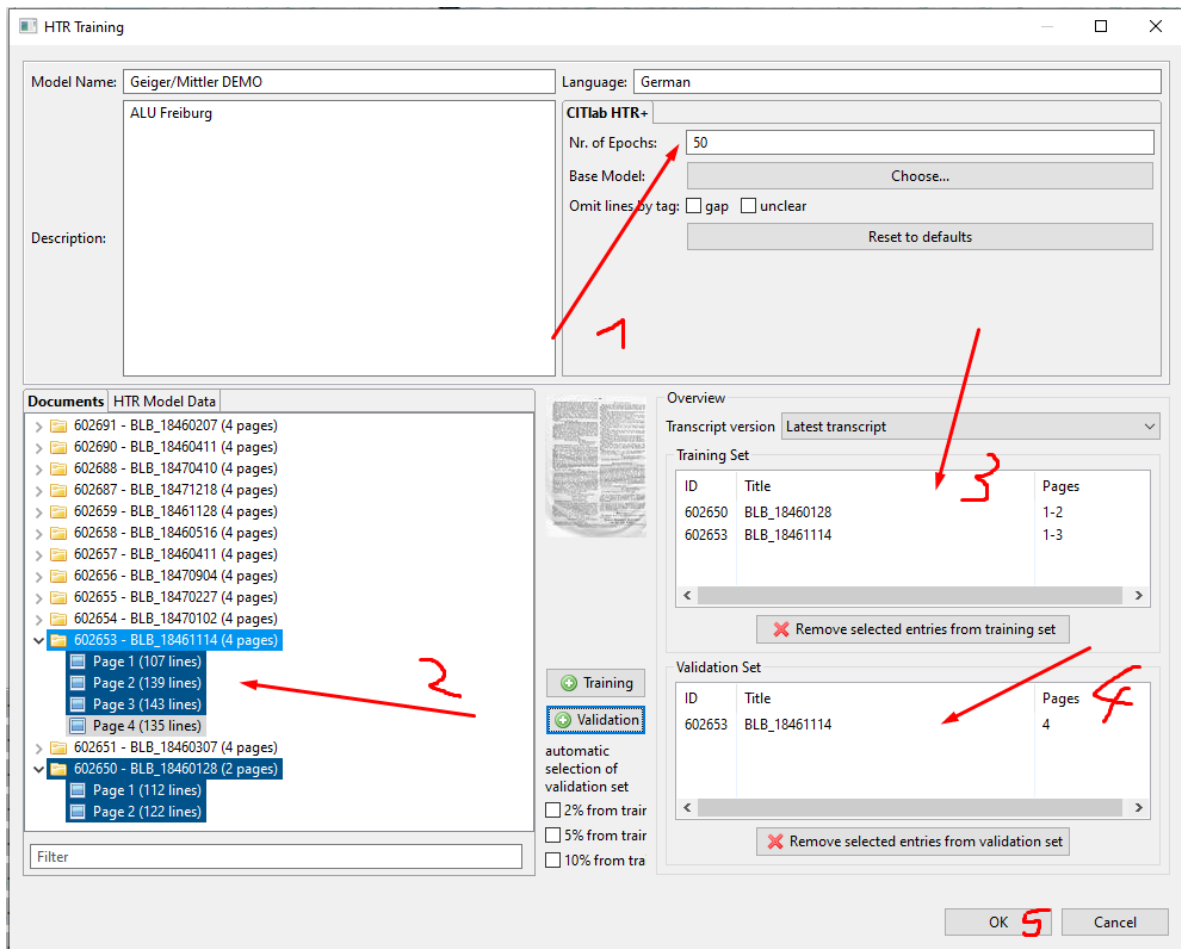


Abbildung 6: Schritt-für-Schritt Modelltraining in Transkribus. // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

Neben der Namensgebung, der Sprache und der Beschreibung des Modells ist der erste wichtige Schritt die Auswahl der Epochenanzahl des Modelltrainings. Je höher diese Zahl ist, umso öfter evaluiert Transkribus die zur Verfügung gestellte Informationen. Dies sorgt für einen besseren Lernprozess, das Problem bei einer zu hohen Epochenzahl ist allerdings, dass Transkribus immer länger für den anstehenden Rechenprozess braucht. Wir haben zum Beispiel mit 50 Epochen gearbeitet, auch hier ist es schon einigermaßen zeitintensiv, Transkribus 90 Minuten beim Rechnen zuzuschauen. Dieser Rechenvorgang hat nichts mit der eigenen Hardwarevorraussetzung zu tun, dieser Prozess findet auf den Transkribus-Servern statt, der PC könnte in dieser Phase also auch ausgeschaltet werden.

Außerdem sollte in diesem ersten Schritt ein Basismodell ausgesucht werden, welches Transkribus dann als Vorlage für das neue Modell benutzen kann, und „nur noch“ die Abweichungen angleichen muss. Auch hier haben wir wieder das „Kurrent“ Modell genutzt. Im nächsten Schritt werden dann die gewünschten Seiten für das Modelltraining eingefügt. Im vierten Schritt wird eine *validation* Seite hinzugefügt, diese dient zum Vergleich im Abschluss. Transkribus empfiehlt pro 50 Seiten Trainingsmaterial eine Seite *validation*, soviel haben auch wir benutzt.

Ist nun alles richtig eingestellt, kann das Modelltraining durch Betätigen des *OK buttons* begonnen werden. Den Status dieses Trainings kann man in der *jobs* Übersicht nachverfolgen.

The screenshot shows the Transkribus interface with the 'Jobs on Server' window open. The window displays a table of training jobs. A red arrow points to the 'RUNNING' status of the current training job.

Type	State	Doc-id	Pages	Username	Description	Errors	Created	Started	Finished
Create Docum...	FINISHED	619314		jan-mittler@t...	Done, duration: 13s 719ms	0	27.02.2021 15:44:07	27.02.2021 15:44:08	27.02.2021
Create Docum...	FINISHED	619320		jan-mittler@t...	Done, duration: 20s 418ms	0	27.02.2021 15:44:38	27.02.2021 15:44:38	27.02.2021
Layout analysis...	FINISHED	619314	1-4	jan-mittler@t...	Done, duration: 46s 264ms	0	27.02.2021 15:49:55	27.02.2021 15:49:58	27.02.2021
CITab Handwri...	FINISHED	619314	1	jan-mittler@t...	Done, duration: 54s 808ms	0	27.02.2021 15:53:02	27.02.2021 15:53:05	27.02.2021
Create Docum...	FINISHED	595920		jan-mittler@t...	Done, duration: 9m 19s 584ms	0	28.01.2021 13:57:46	28.01.2021 13:57:46	28.01.2021
CITab HTR+ Tr...	RUNNING	-1		jan-mittler@t...	Training epoch 39/50 (current CER o...	0	06.03.2021 14:49:34	06.03.2021 14:49:37	
CITab HTR+ Tr...	PENDING	-1		jan-mittler@t...		0	06.03.2021 14:56:15		

Abbildung 7: *jobs* Übersicht zur Nachverfolgung des Fortschritts // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

#### 4. Das Landbote-Modell

Nach 90 Minuten beendete Transkribus das Training des 50-Epochen Modells erfolgreich. Das Ergebnis ist besonders im Vergleich zum 5-Epochen Modell interessant.

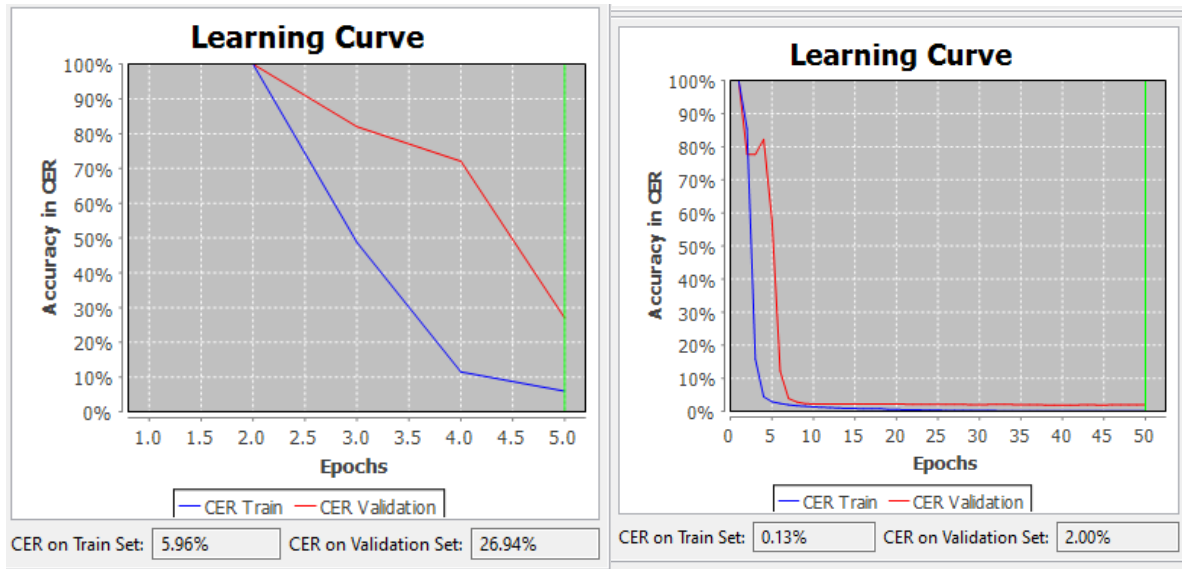


Abbildung 8+9: *learning curves* des 5/50 Epochendurchgangs. // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

Links ist die Lernkurve des Modells mit nur fünf Durchläufen zu sehen, rechts die Lernkurve des Modells mit 50 Durchläufen. Die y-Achse beschreibt die Genauigkeit in CER, also Fehler pro 100 Zeichen, die x-Achse beschreibt die vergangenen Epochen. Links ist von einem nahezu linearen Graph zu sprechen, der im Durchschnitt und nach fünf Epochen nicht unter eine Zeichenfehlerquote von 5% fällt. Rechts hingegen ist zu beobachten, dass das Modelltraining anscheinend schon nach ca. 12 Epochen auf eine Fehlerquote von zwei Prozent gefallen ist.

An der oben erklärten *validation* Seite lässt sich das anschaulich darstellen:

---

1 und nicht verlautet davon, daß die sardinische Regierung  
2 diese patriotische Feier zu verhindern gesucht habe  
3 Toskana und Lukka stehen einander feindselig ge-  
4 gegenüber! Aus Livorno wird der A. A. Ztg. unterm 16.  
5 Dez. geschrieben: „Großes Aufsehen hat das in diesen Tagen  
6 hier und in andern Städten Toskana's an den Straßen-  
7 ecken angeheftete und in den Zeitungen veröffentlichte Edikt  
8 des Großherzogs gemacht, in welchem, mit Hinweisung auf  
9 die Wiener Schlußakte und die garantierte künftige Succes-

---

1 und nichts verlautet davon, daß die sardinische Regierung  
2 diese patriotische Feier zu verhindern gesucht habe-  
3 Toskana und Lukka stehen einander feindselig ge-  
4 gegenüber! Aus Livorno wird der A. A. Ztg. unterm 16.  
5 Dez. geschrieben: „Großes Aufsehen hat das in diesen Tagen  
6 hier und in andern Städten Toskana's an den Straßen-  
7 ecken angeheftete und in den Zeitungen veröffentlichte Edikt  
8 des Großherzogs gemacht, in welchem, mit Hinweisung auf  
9 die Wiener Schlußakte und die garantierte künftige Succes-

Abbildung 10+11: oben: Das 5-Epochen-Modell, unten: Das 50-Epochen-Modell // Screenshot angefertigt von Jan Mittler, Lizenz: *public domain*.

Es ist klar zu erkennen, dass das obere Modell, welches weniger Epochendurchläufe hatte, mehr Fehler produziert als das untere. Das untere Modell hingegen scheint deutlich akkurater transkribiert zu haben, es sind kaum noch Fehler zu erkennen. Besonders spannend zu sehen ist, dass sämtliche S Buchstaben, die in der Zeitung anders als heute üblich geschrieben waren, nun, wie wir es durch unser Training intendierten, entsprechend in das moderne s transkribiert wurden. Hieran ist erkennbar, wie sich mit wenigen Schritten ein öffentlich verfügbares Modell an die eigenen Bedürfnisse anpassen lässt.



Fazit:

Das Anpassen eines HTR+ Modells ist nicht so komplex, wie wir uns das vorgestellt haben. Der Nutzen eines angepassten Modells kann in unseren Augen allerdings enorm sein. Mit einigen Stunden akribischer Arbeit, dies beinhaltet bei Drucken das Layouten und Transkribieren von wahrscheinlich ca. 15-20 Seiten, lässt sich ein gutes Modell entwickeln, das einem das restliche Transkribieren abnehmen.

Kleine Schwierigkeiten bietet Transkribus in der Transparenz seines Credit-Systems: Das Anwenden eines HTR+ Modells kostet eine gewisse Anzahl an Credits, wieviel und was diese Kosten ist nur nach größerer Recherche herauszufinden. Jeder User erhält nach dem Erstellen eines Accounts 500 kostenlose Credits, die gleiche Anzahl ist im Moment für 66 Euro unter diesem Link: <https://readcoop.eu/de/transkribus/credits/> zu erwerben. Für unsere Arbeit haben wir lediglich drei Credits für 10 Seiten Material ausgeben müssen.

Abschließend lässt sich sagen, dass wir trotz einiger weniger Umständlichkeiten sehr von der Nützlichkeit des Transkribus-Programms überzeugt waren und sind. Sollte Interesse an Einsicht in unsere Collection in Transkribus bestehen, darf gerne die angegebenen E-Mail-Adresse kontaktiert werden, wir können Sie dann hierfür freischalten.

projekt.transkribus.alu@gmail.com