Retro | KILDER TIL DANSKE KOMMUNERS HISTORIE

VEJLEDNING I MASKINEL TEKST-GENKENDELSE I TRANSKRIBUS

- **KOM GODT I GANG**
- **TRÆN MODELLER**

UDARBEJDET AF

JAN MATTIAS JONSSON AGGER & KRISTIAN PINDSTRUP

TIDL. OG NUV. PROJEKTKOORDINATOR

Aarhus Stadsarkiv

December 2019

AARHUS STADSARKIV

Indhold

1. Introduktion

Udover at være et fint transskriberingsprogram har Transkribus den klare fordel, at programmet også er ganske langt i forhold til at lære computeren at læse håndskrifter. Med hjælp af det hidtil transskriberede og korrekturlæste materiale er der lavet en dansk model der kan deles og benyttes af enhver til transskriberingen eller som udgangspunkt i træningen af en model der passer mere specifikt til den enkelte protokol. Denne hedder Danish 1870-1950 v3.5 og kan anvendes af alle Transkribus-brugere. Dog er datamaterialet ikke tilgængeligt for offentligheden.

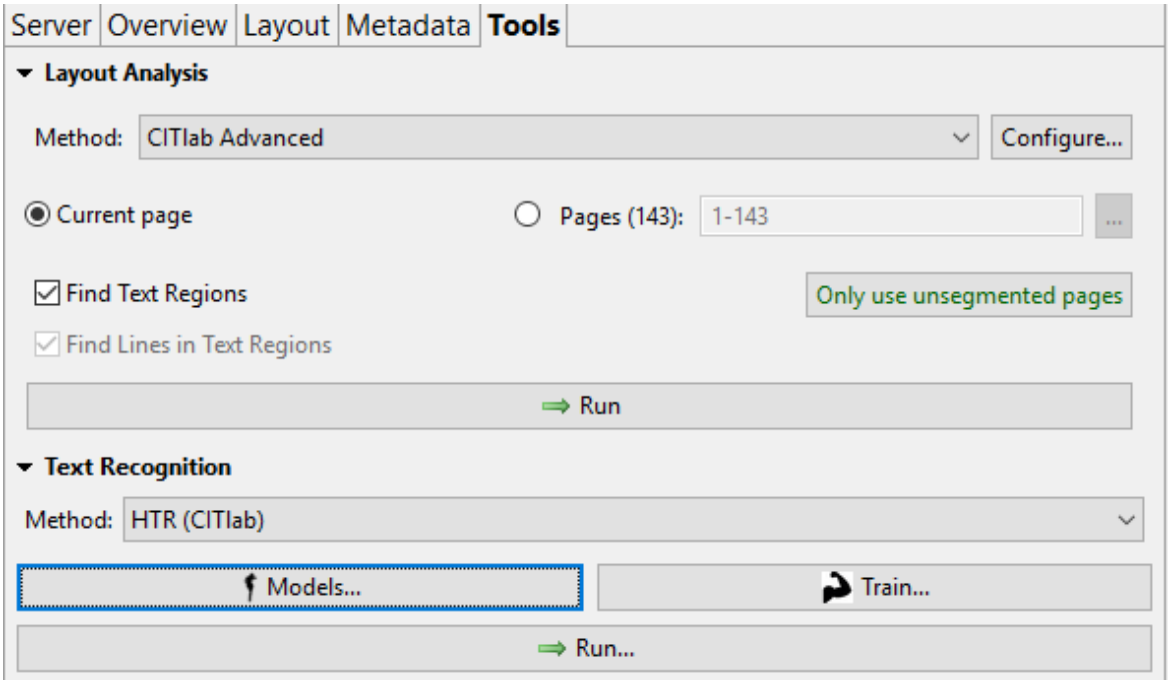
Denne vejledning vil give en introduktion til brugen og træningen af modeller.

Det anbefales at du først går i gang med arbejdet med modeller når du har lært Transkribus ordentligt at kende, da der stadig kræves et forarbejde med segmenteringen ligesom modellen forbedres af at kunne lære fra nogle allerede transskriberede sider.

2. Modeller

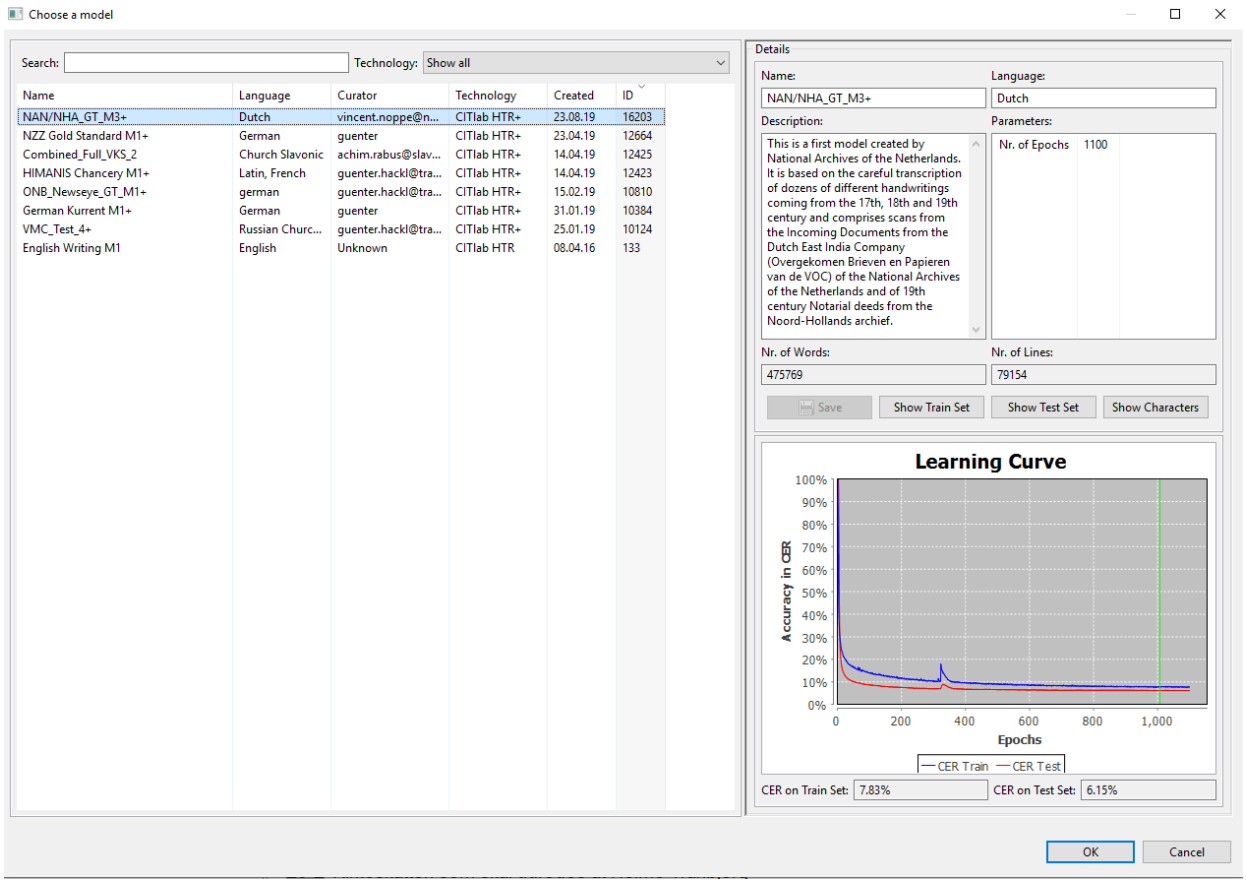
Til at starte med er det en god idé at få et overblik over hvilke modeller der er adgang til. Transkribus har udgivet en række forskellige modeller der kan bruges til forskellige sprog og skrifttyper. Der er lavet en række modeller på dansk, hvoraf Aarhus Stadsarkiv er ansvarlig for en række af dem. Den senest udgivne fra Aarhus Stadsarkiv er Danish 1870-1950 v3.5, men der kan være trænet nyere modeller, som ikke er tilgængelig. Hvis du er interesseret i at få adgang til den seneste model, kan du kontakte projektkoordinatoren for RETRO.

Du finder de tilgængelige modeller under Tools → Models



Figur 1 – Bemærk at man skal anmode Transkribus om tilladelse til træning før knappen "Train…" viser sig, hvorfor mange kun har "Models…" og "Run…"

2.1 Modeloversigt



Figur 2 – Oversigten over de modeller der er tilgang til. Forskellige samlinger kan have adgang til forskellige modeller hvis de er brugt til træning eller modeller er delt med specifikke samlinger.

Fladen i oversigten er opdelt i 3 dele.
Først og fremmest er der den egentlige oversigt, hvor titel, sprog mm er skrevet. Som standard indeholder den et begrænset antal modeller. Det er de modeller som er blevet udgivet igennem Transkribus.

Når du trykker på en af modellerne, kommer der flere detaljer øverst til højre. Dette inkluderer en beskrivelse, hvor eksempelvis materialet bag modellen og hvem der står bag den kan forklares.
Lige derunder er antallet af transskriberede ord og linjer, der kan bruges til at få en fornemmelse for hvor stort et materiale den er trænet med. Flere ord betyder oftest at sandsynligheden er større for at den er brugbar på lige præcis din protokol.
Nederst til højre vises modellens læringskurve. Selve kurven er ikke så interessant, hvis ikke man selv har trænet den, men det er derimod modellens fejlrate, Character Error Rate (CER). Den

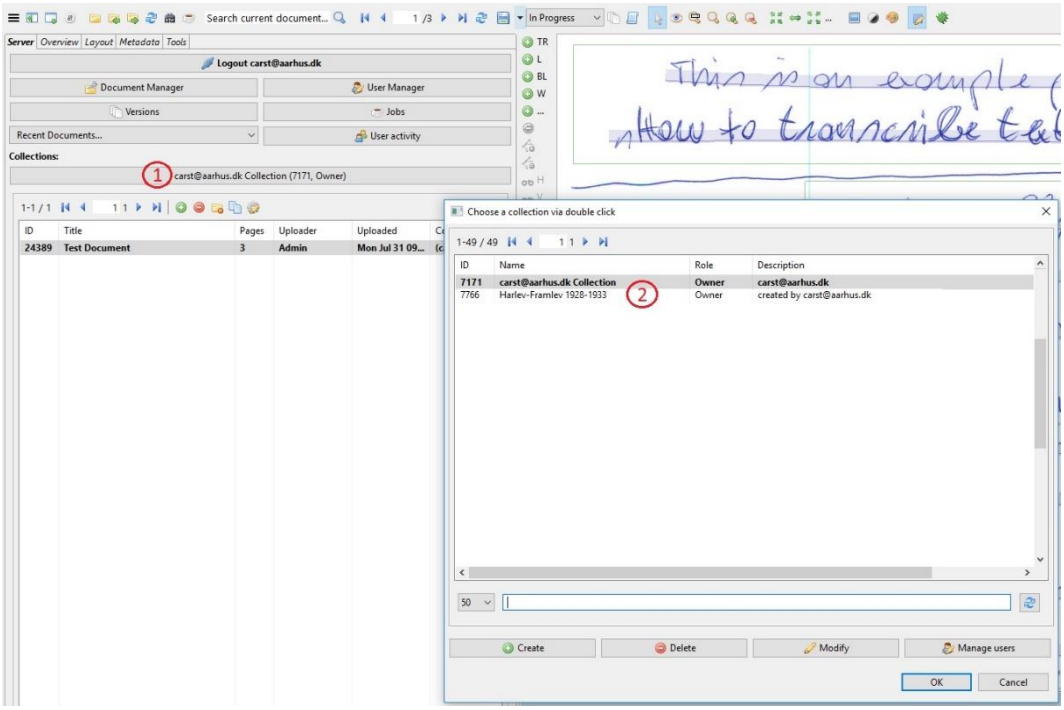
fortæller hvor præcis modellen er på lige præcis det materiale der er brugt til henholdsvis træning og test.

Som udgangspunkt er modeller med fejlprocenter under 10 brugbare i førstetransskriberingen, men det kræver selvfølgelig at protokollens håndskrift minder om den brugt i modellen.

3. Brug modeller til transskribering

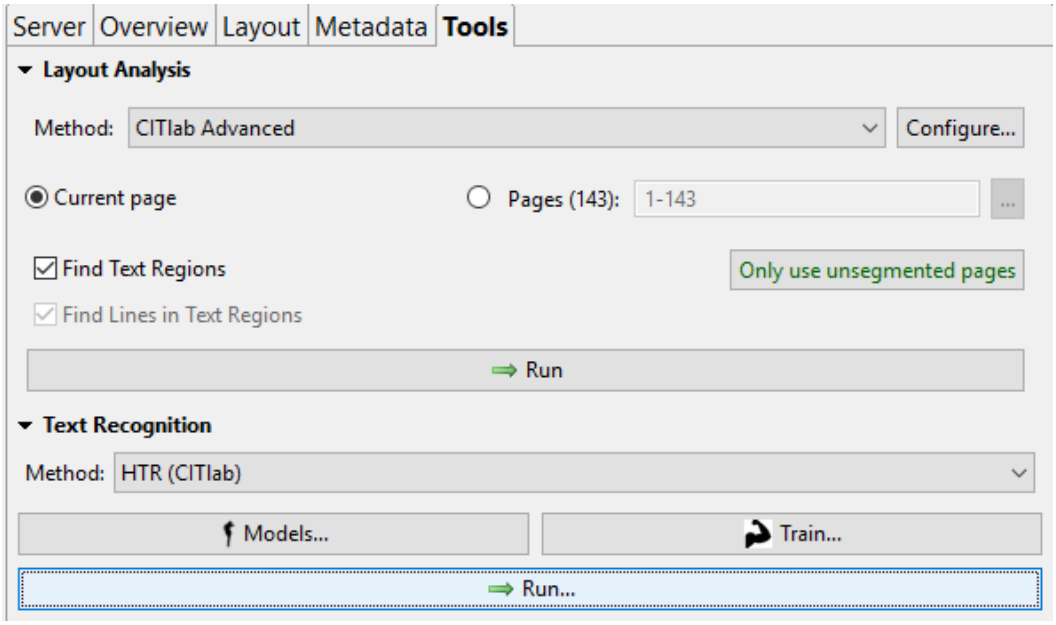
Når du har en passende model, er det bare om at komme i gang med at bruge den til den første omgang transskribering.

Først skal du åbne samling og dokument ligesom i brugervejledningens punkt 6.1.



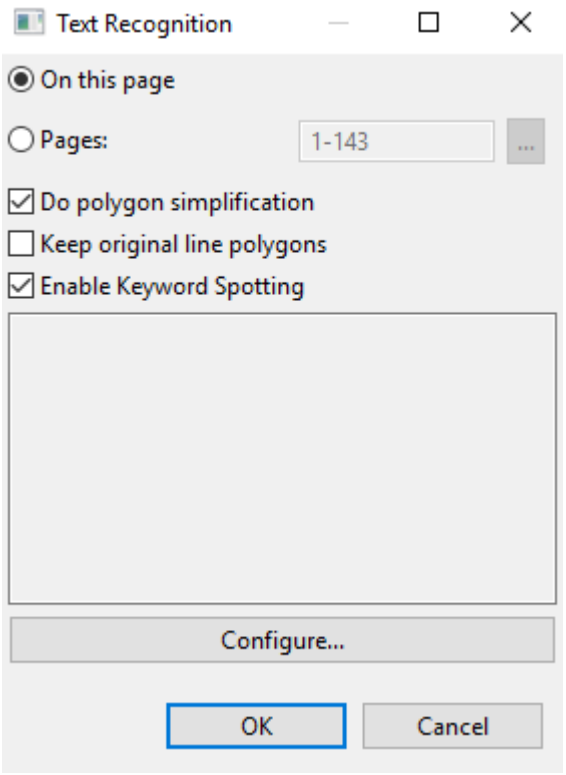
Figur 1 - Find dokument, 1) Klik på bjælken under "Collections", 2) Marker den protokol du vil åbne og klik "OK".

Når dokumentet er åbent, klikker du på Tools og derefter "Run…" under Text Recognition



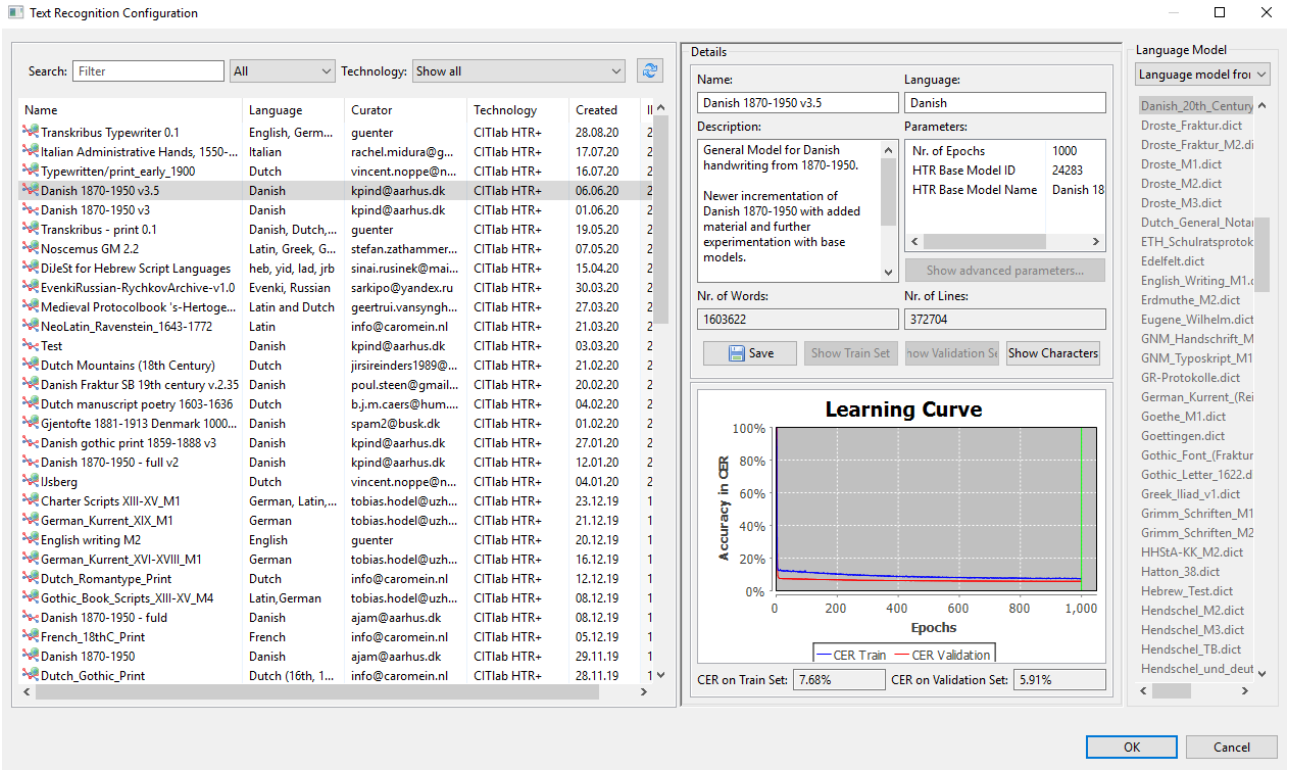
Figur 2 – Vær opmærksom på, at det er den nederste af de to knapper med "Run" der bruges her.

Herefter får du et nyt pop-up vindue, hvor du har mulighed for at vælge siderne der skal transskriberes, hvilken model der skal bruges og nogle øvrige indstillinger.



Figur 3

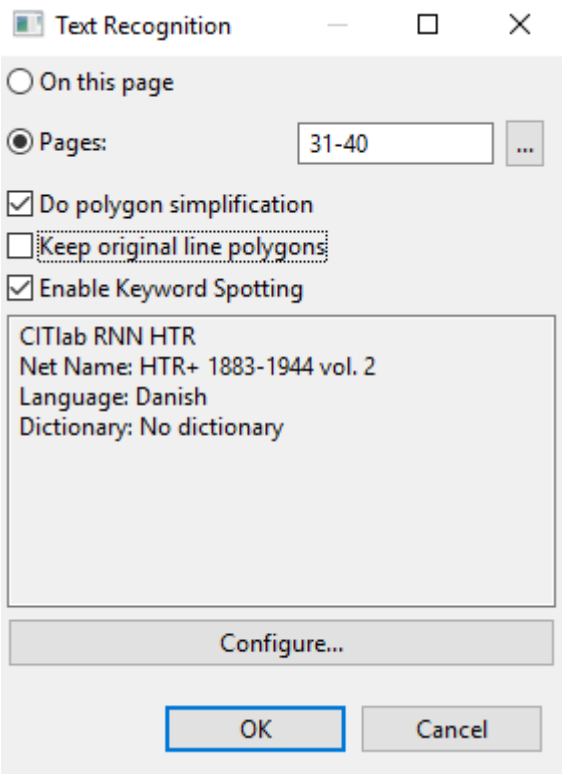
Her kan du først vælge de sider der skal transskriberes og derefter trykke på Configure… for at vælge model



Figur 4 – Her kan du vælge den ønskede model og eventuelt en ordbog

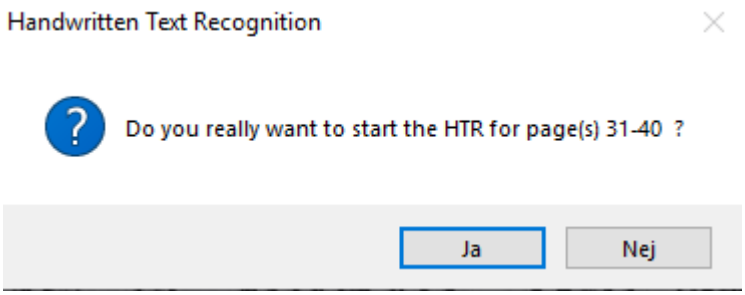
Billedet minder meget om det du får ved at trykke på Models. En oversigt, beskrivelse og læringskurve. Til højre er der dog også muligheden for at tilvælge en ordbog. Det kan være værd at eksperimentere med forskellige ordbøger, for i nogle situationer kan det øge modellens præcision. Til modellen Danish 1870-1950 v3.5 anbefales det at bruge ”Language model from dictionary”, da det i de fleste tilfælde vil give et bedre resultat.

Når du har trykket på den model du ønsker at bruge, trykker du på OK, hvorefter det forrige vindue nu ser sådan ud:



Figur 5

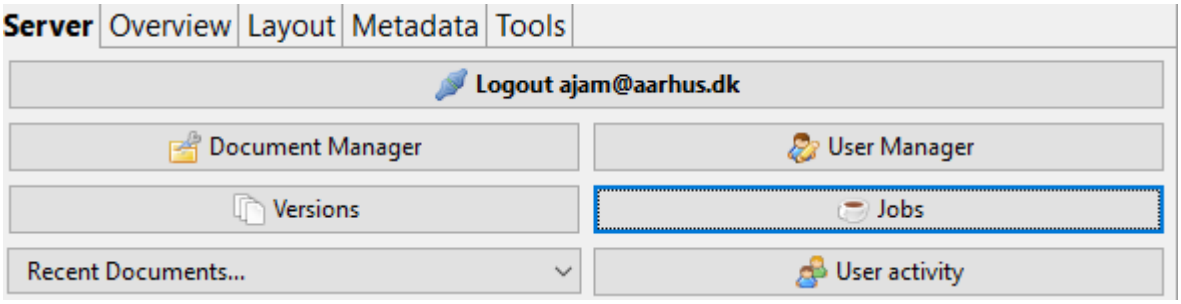
Her trykker du igen OK, hvorefter følgende popper op:



Figur 6

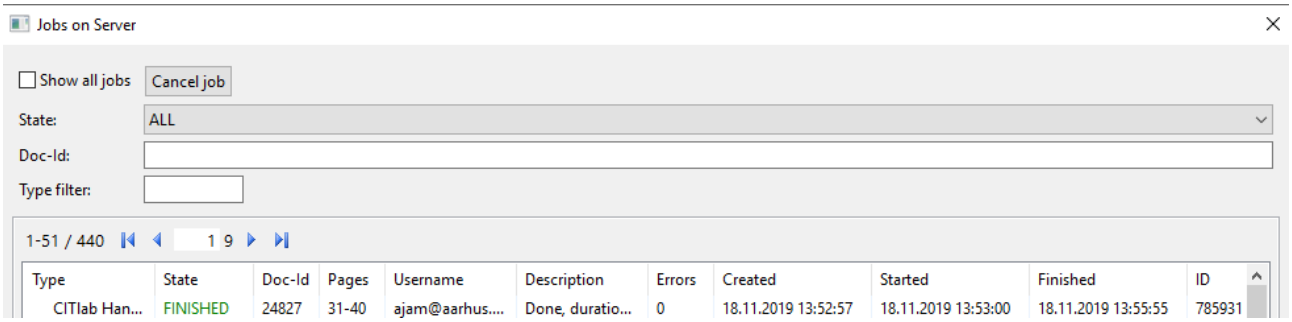
Ved at trykke Ja starter serveren med at transskribere, en proces der for 10 sider tager et par minutter. Hvis mange har igangsat jobs, kan det dog tage længere tid at igangsætte jobbet.

Du kan følge med i processen med at læse siderne ved at trykke på Jobs under Server



Figur 7 – I vinduet hvor du normalt vælger dit dokument kan du også tjekke jobs.

Her kan du se de ”jobs” du har haft gang i eller som serveren i Innsbruck laver for dig i øjeblikket. Når opgaven er færdig, står det med grønt under State.



Figur 8 – Oversigten viser her at det seneste job af typen CITlab Handwritten Text Recognition (HTR) er afsluttet

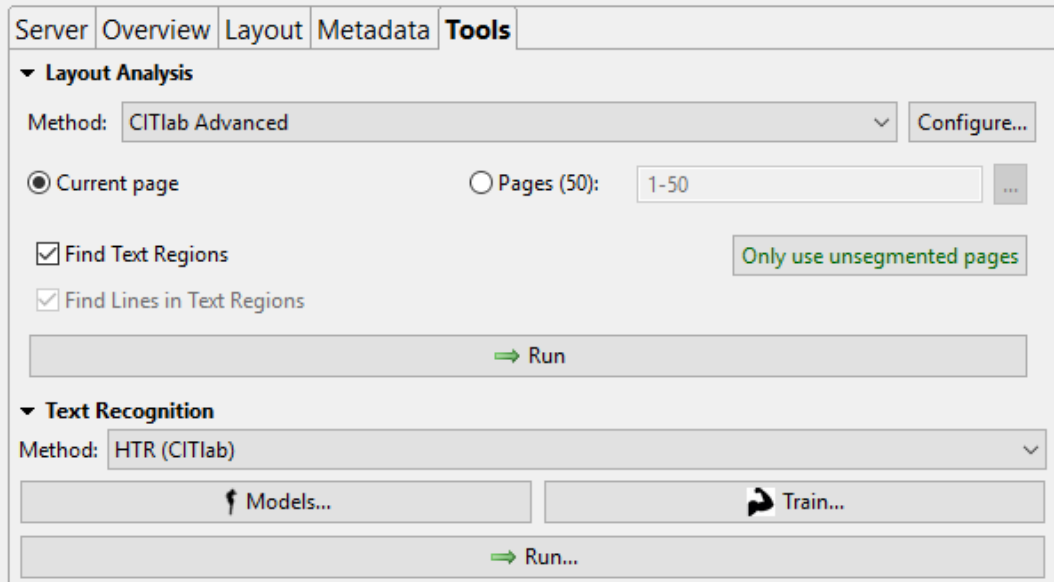
4. Træning af modeller

Hvis det viser sig ikke at være til den store hjælp at bruge allerede trænede modeller, kan løsningen være at træne en ny. Dette er særligt relevant ved gotisk håndskrift eller ved protokoller skrevet med mere besværlig håndskrift. Ligeså kan det bruges til at styrke en allerede eksisterende model, så den fungerer bedre med lige præcis din protokol.

4.1 Før du træner

For træne en model kræves det, at du allerede har transskriberet et antal sider af din protokol. Det kan være svært at vurdere helt præcist hvor mange der skal til, men som udgangspunkt bør det minimum være 50. Disse skal idéelt set være korrekturlæst, som minimum ved en ekstra gennemlæsning.

Herudover skal du også have adgang til at træne modeller. Dette har man ikke som standard, men ved at skrive til Transkribusteamet på email@transkribus.eu er det muligt at få adgang. Dette sker ikke med det samme, hvorfor der bør anmodes om adgang i god tid.

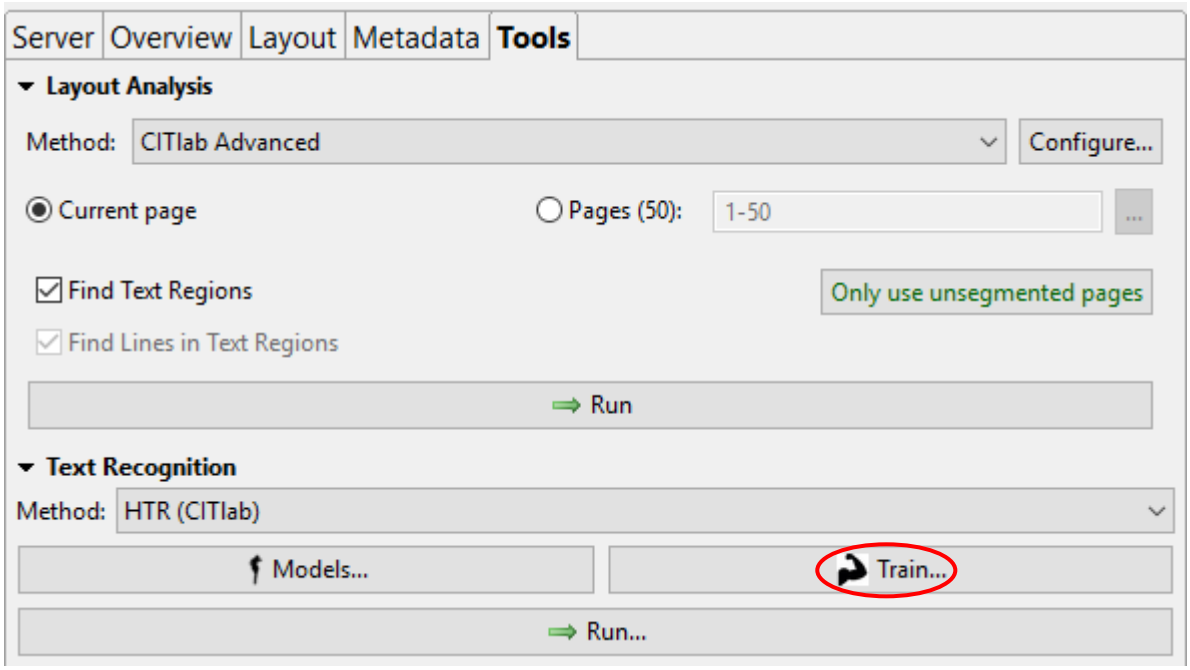


Figur 1 – Oversigten under Tools → Text Recognition ændrer sig når du har adgang til træning

4.2 Træning af modeller

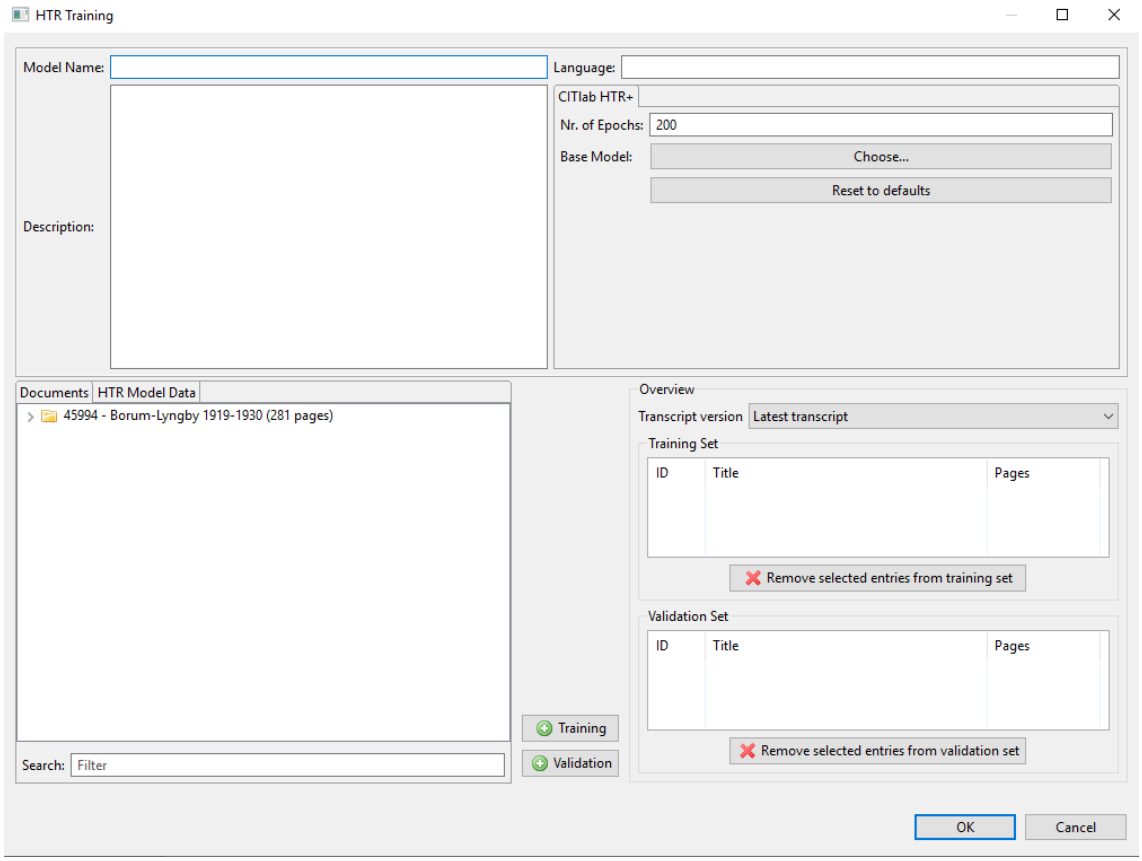
Når du har fået adgang til træningsværktøjet og har et brugbart antal transskriberede sider er det tid til at træne modellen.

Her er det vigtigt at du er inde i den samling som indeholder de sider som skal bruges til modellen. Dette gør du på samme måde som når du skal bruge maskinlæsningen eller selv skal transskribere.



Figur 2 – For at træne en ny model trykker du på Train…

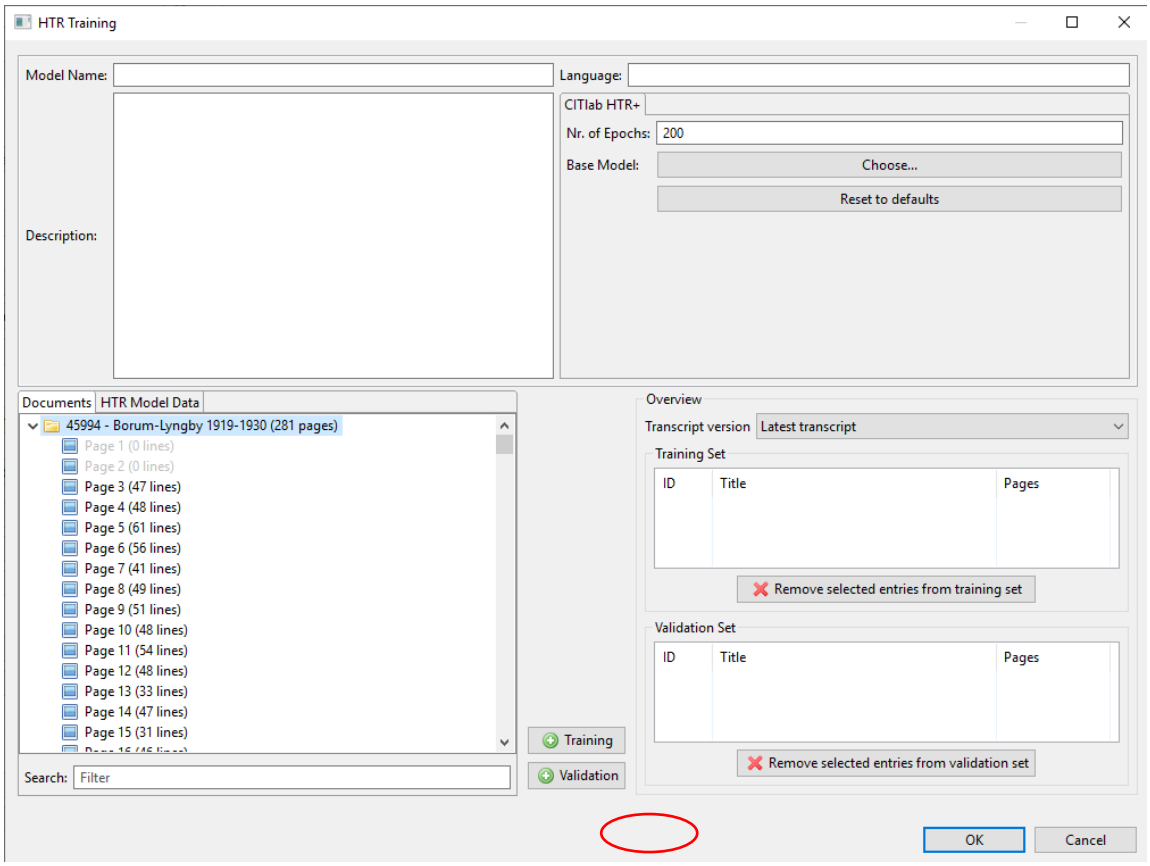
Træningen af modeller igangsættes igennem det nedenstående vindue hvor du kan navngive og beskrive modellen og vælge de relevante sider.



Figur 3 – Under Documents finder du de dokumenter der er en del af den samling/protokol du arbejder med. Det er således kun den "Collection" du har valgt, der kan ses.

Her kan du starte med at give modellen en titel, beskrivelse og notere at den er på dansk. Titlen skal være så det kan ses hvad den kan bruges til. Når der er tale om modeller som ikke skal deles, er den præcise titel og beskrivelse dog ikke så relevant.
Når du har udfyldt de tre bliver det noget mere teknisk.
”Nr. of Epochs” handler om hvor mange omgange maskinen skal gennemgå og lære af den transskriberede tekst. Det vil ofte være nok med 200 epochs medmindre der arbejdes med et større materiale. Lige herunder kan du vælge at bruge en ”base model”. Dermed får modellen et udgangspunkt i hvad den har lært i en anden model. Det vil som regel være en fordel, men kræver at du har adgang til en. Det kan du dog få af projektkoordinatoren.

Til sidst skal du vælge de sider der skal bruges til træningen. Ved at dobbeltklikke på mappen med de sider du vil bruge (eller trykke på pilen ved siden af mappeikonet) får du en oversigt over mappens sider.

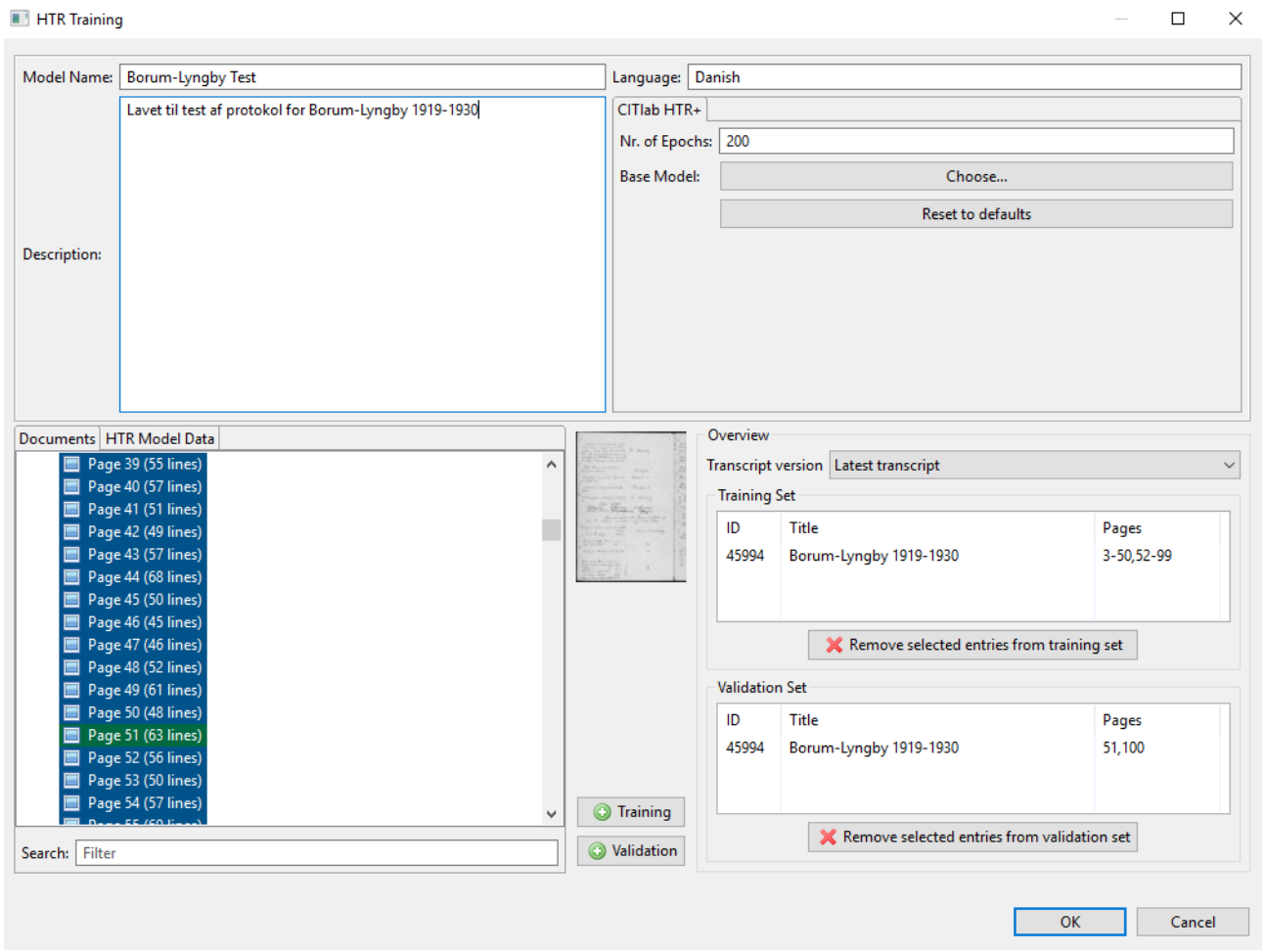


Figur 4 – Mappen er nu åben med en oversigt over siderne. Bemærk at det er noteret hvor mange linjer der er på siden. De første to sider er trykte standardsider og derfor ikke transskriberet.

Her markerer du de sider du vil bruge til træningen (ved at holde shift nede kan du markere en række sider) og trykker så på Training, hvormed siderne tilføjes under Training Set.

På samme tid skal der også være nogle sider i "Validation Set" der ikke bruges til selve træningen men fungerer som en kontrol hvor det lærte bruges til at sammenligne med den menneskelige transskribering. Den bruges således til at fortælle hvor præcis modellen bliver.
En god tommelfingerregel er at bruge én side pr. 50 sider, men der skal altid være minimum én side og denne skal have linjer.
Du tilføjer sider hertil ved at markere dem og trykke på Validation.

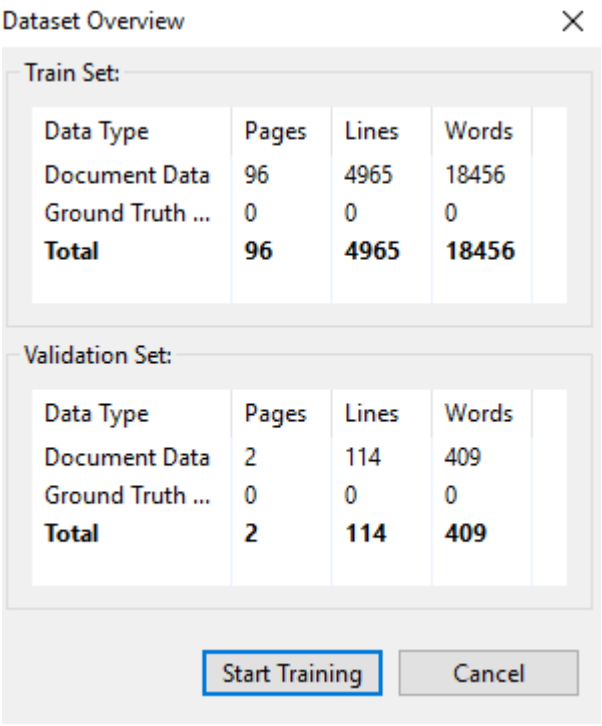
Dermed skulle du gerne ende med et vindue der minder om nedenstående:



Figur 5 – Siderne 3-100 er her brugt til enten træning eller "Validation". Hvis flere mapper bruges, ses det i tabellerne.

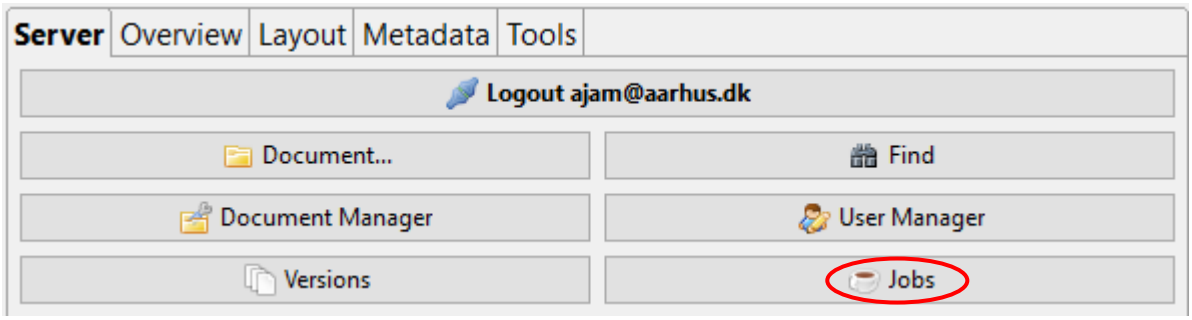
Når du har valgt de sider du ønsker at bruge til henholdsvis træning og validering trykker du på OK.

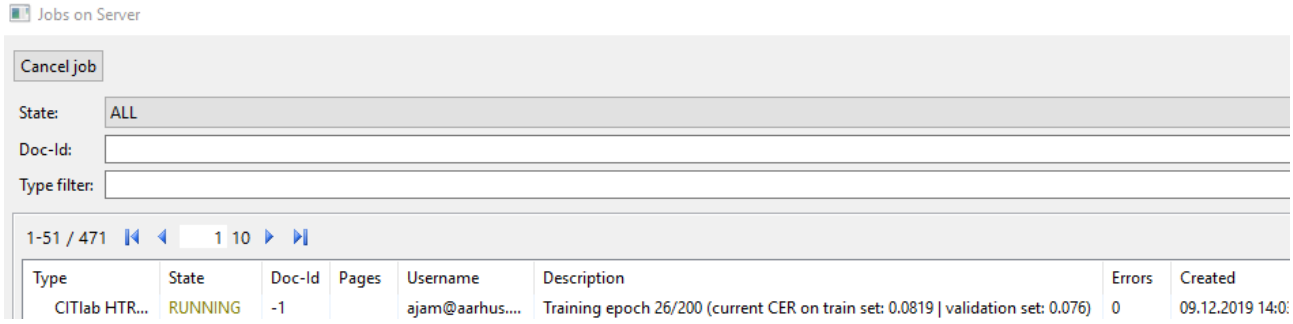
Herefter dukker en oversigt over det valgte materiale op. Tryk på Start Training for at sætte det i gang.



Figur 6 – Oversigten kan give et overblik over omfanget på det data der benyttes til træningen.

Når du har startet træningen, kan den følges ved at trykke på Jobs





Figur 8 - Her kan du følge med i hvor langt træningen eller andre jobs er kommet

Træningen tager sin tid (modellen brugt som eksempel tog 4 timer og 40 min) og du modtager en mail når den er afsluttet. Da den foregår på en server i Innsbruck, kan du dog sagtens lægge programmet til side eller arbejde videre med transskriberingen mens den træner.

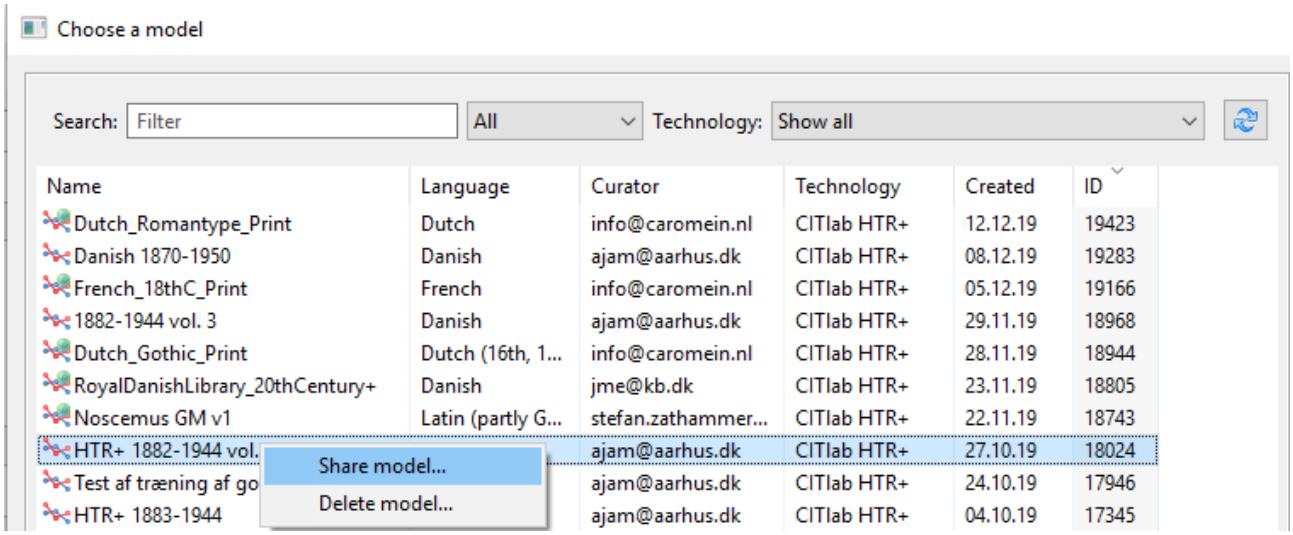
Når træningen er afsluttet, skulle du gerne kunne finde den i modeloversigten og således begynde at bruge den. Det kan dog være en idé at få den vurderet af en med erfaring med brugen af modellerne, hvis du vil bruge en model på andre protokoller end de er lavet på baggrund af.

5. Deling af modeller

Som udgangspunkt er modeller private i Transkribus, Hvis ikke den er gjort offentlig, kan den kun bruges i den samling som modellen er trænet på, ligesom den kun kan ses af brugere der har adgang til samlingen. Det er dog muligt at dele modeller med andre samlinger. Dette kan være relevant hvis du har lavet en god model som du vurderer kan bruges til andre protokoller. Det kan for eksempel være brugbart hvis du arbejder med to protokoller i forlængelse med hinanden, hvor det kan ses at der er ligheder i håndskriften således at en model kan bruges til begge. Det kan også være hvis du vil prøve at bruge din model som base model til at træne en ny baseret på en anden protokol.

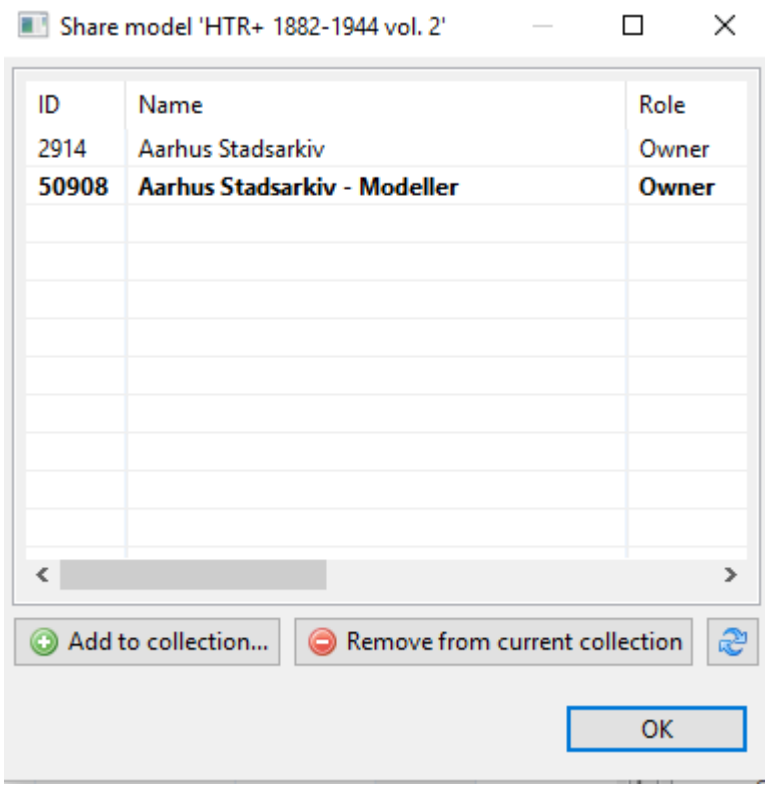
Det er som udgangspunkt ikke noget problem at dele en model med din egen samling, men hvis modellen er trænet med protokoller der er under 75 år gamle, må modellen ikke deles med andre, da det vil være muligt for modtageren at læse det materiale som modellen bygger på, hvormed personfølsomme oplysninger kan blive delt. Spørg eventuelt dit arkiv, hvis du er i tvivl om hvorvidt dit materiale kan deles uden problemer.

Selve delingen af en model er ganske simpel. Først går du ind i modeloversigten som det er forklaret i punkt 2. Her finder du den model som du vil dele og højreklikker på den og vælger Share model… .



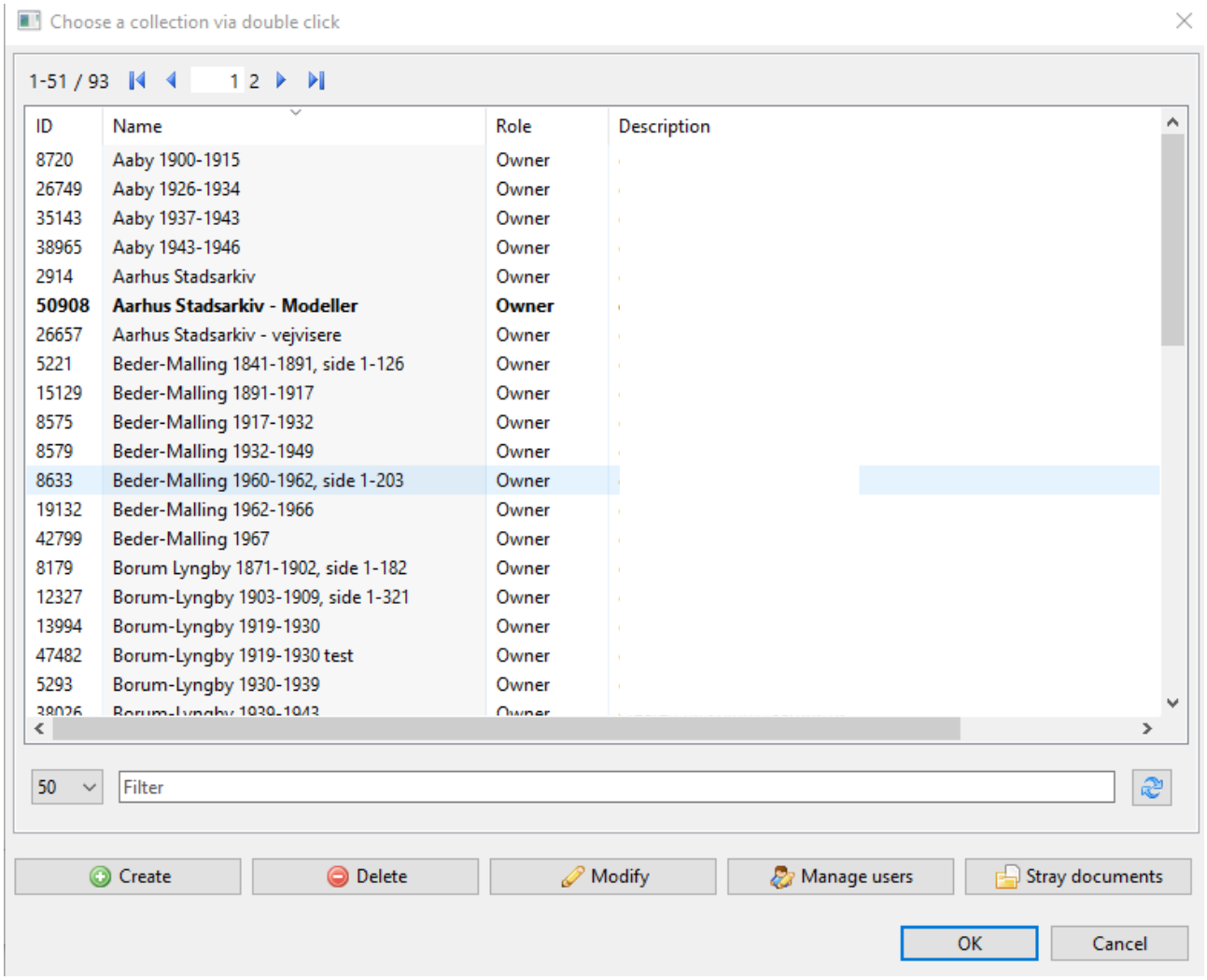
Figur 1 – Når du har fundet den ønskede model vælger du "Share model…"

Herefter kommer der så en oversigt over de samlinger som den enkelte model ligger i.



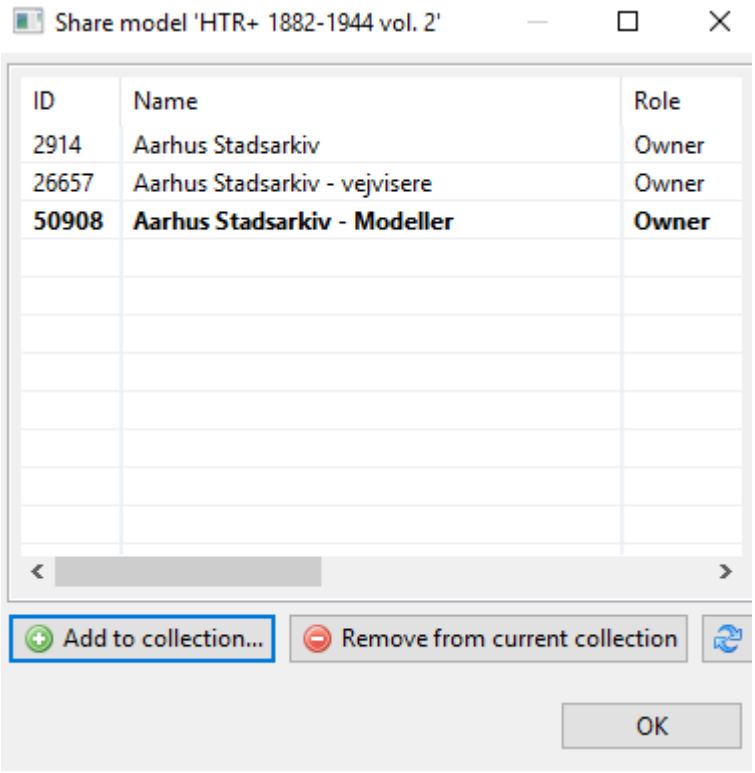
Figur 2 – I oversigten vælger du Add to collection… for at finde den samling som du vil føje modellen til

Ved at trykke på Add to collection… får du en oversigt over dine samlinger. Her markerer du den samling du vil dele med og trykker på OK.



Figur 3 – Hvis du vil dele en model med andre, kan du fx lave en ny samling og tilføje modellen hertil

Herefter dukker den nye samling op i oversigten



Figur 4 – Ved at trykke på samlingens ID og trykke på Remove from current collection er det også muligt at fjerne modellen fra en samling

Dermed skulle modellen også gerne blive synlig i den nye samling, hvormed du kan bruge den på materialet heri.