Webinar (Nederlandstalig):
TRANSKRIBUS – geavanceerd.

Info: Jeroen Vandommele en Annemieke Romein
Techniek: Levien den Boer

KB nationale bibliotheek

Deze presentatie (dia’s) en het webinar zullen beschikbaar worden gesteld via de website van de KB.
U kunt de informatie dus altijd op een later moment raadplegen.

Video: KB YouTube kanaal
DOI: 10.5281/zenodo.3555097

Powerpoint: 10.5281/zenodo.3558864

KB nationale bibliotheek

Transkribus

INHOUD WEBINAR - geavanceerd

Ultra korte samenvatting basiswebinar

Geavanceerd

- Coordinatie BeNeLux
- Scantent/ DocScan
- Volgorde TR en regels aanpassen
- OCR
- Modellen: maken, beoordelen, toepassen
- KeyWordSpotting (KWS)
- Woordenboek
- Text2Image (3 methoden)
- Labelen (structuur en woorden)
- Layout: P2PaLA en NLE Document Understanding (Tabellen)

Samenvatting vorig webinar #Transkribus #basiswebinarNL

Je ‘importeert’ je bestanden naar de server in Innsbruck.
Je voert een Lay-out analyse uit (handmatig, of automatisch).
*wees je bewust van het doel dat je hebt met je tekst, misschien wil je meer geavanceerdere LA, dan is nu het moment!
Je gaat transcriberen, wees accuraat: verander de tekst niet (m.u.v. afkortingen).
Als je een aantal pagina’s gedaan hebt, kan je een model gaan maken.

Basiswebinar PowerPoint:
10.5281/zenodo.3558860

Basiswebinar:
10.5281/zenodo.3555092
https://www.youtube.com/watch?v=o6BRXq1S-b8

KB nationale bibliotheek

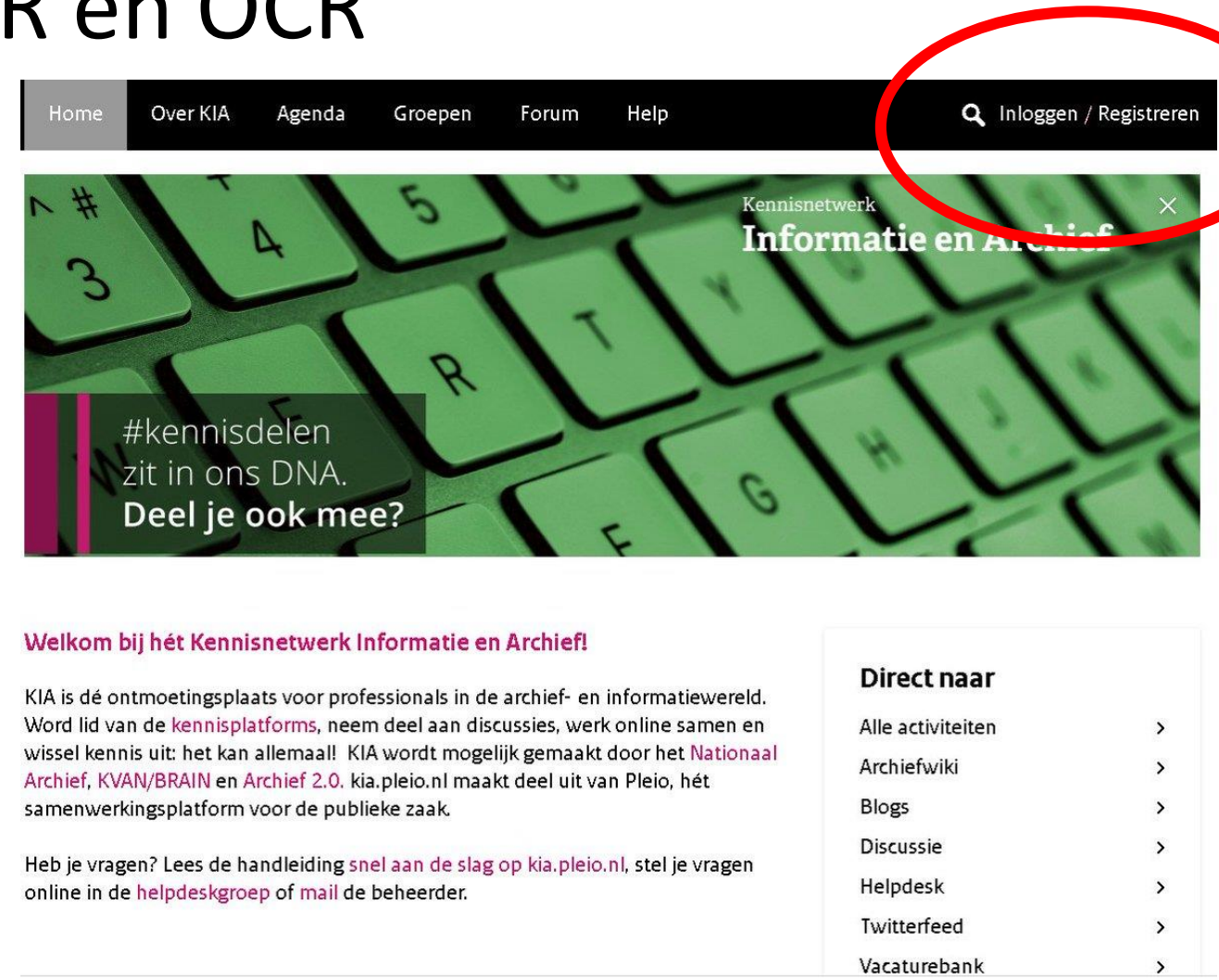
BeNeLux – Coordinatie Transkribus

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Platform Transkribus HTR en OCR

- https://kia.pleio.nl/ → registreer (gratis).
- Ga dan naar: https://kia.pleio.nl/groups/view/55812425/htr-en-ocr
- Platform voor vragen, uitwisselen van informatie, updates en elkaar op de hoogte houden van projecten, etc.



KB nationale bibliotheek

Contactinfo

- Milo van de Pol:
  E-mail: milo.vandepol@nationaalarchief.nl
  Twitter: @milo_vandepol
  (modellen inventariseren; erfgoedsector; archiefwezen)
- Annemieke Romein:
  E-mail: transkribus@caromein.nl
  Twitter: @CARomein
  (modellen inventariseren; academische wereld)

Focus op BeNeLux! (Niet strikt op de Nederlandse taal!)

“Gadgets”: Scantent & DocScan

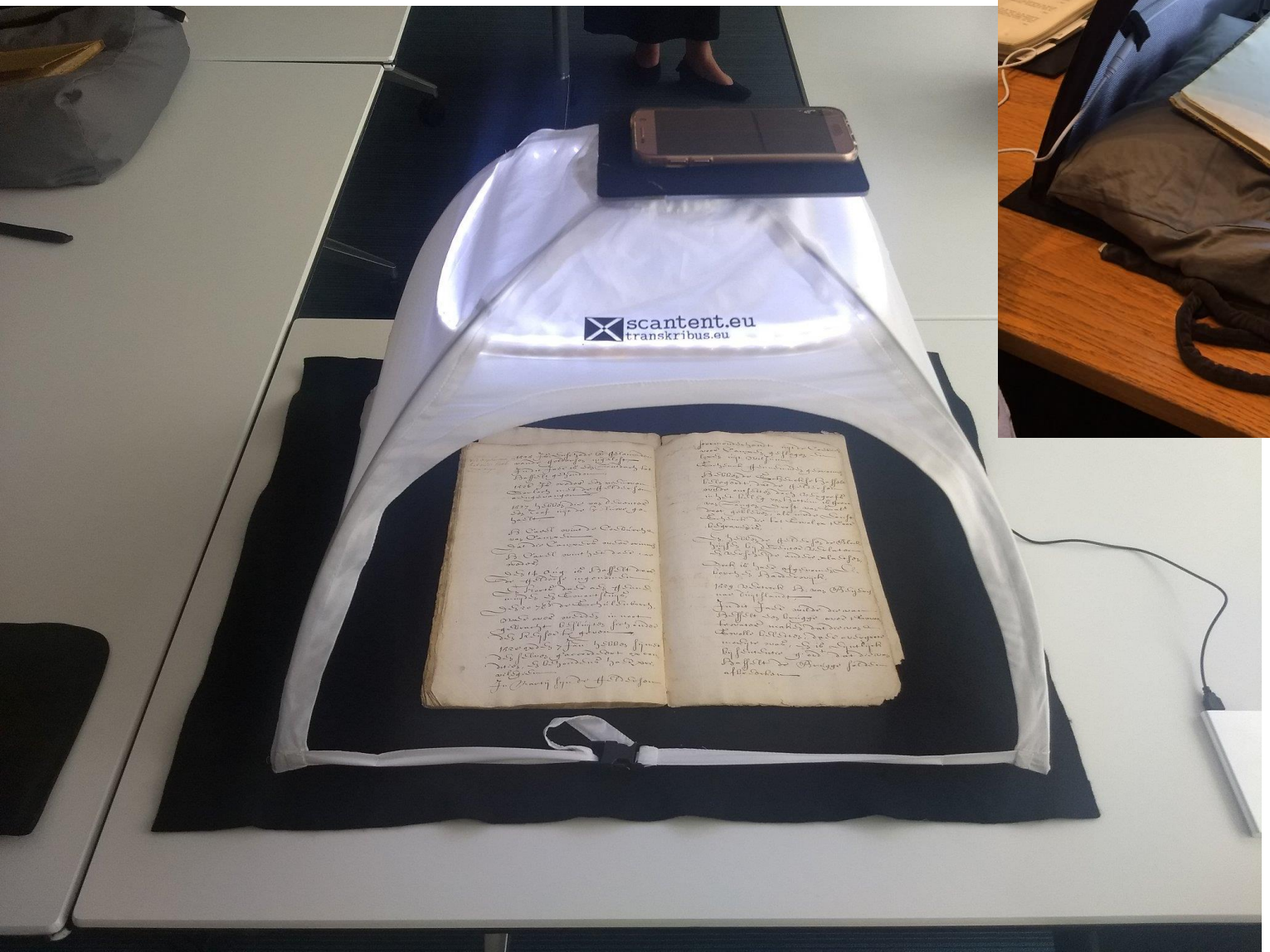
#Transkribus
#webinargeavanceerdNL

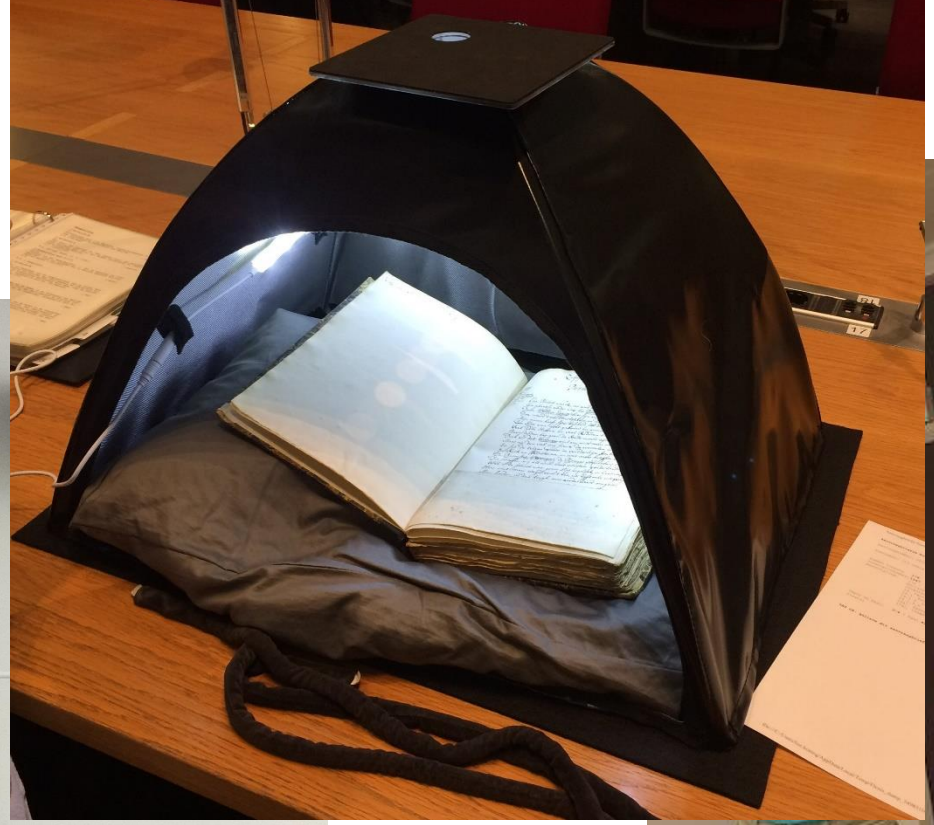
KB nationale bibliotheek

ScanTent And DocScan

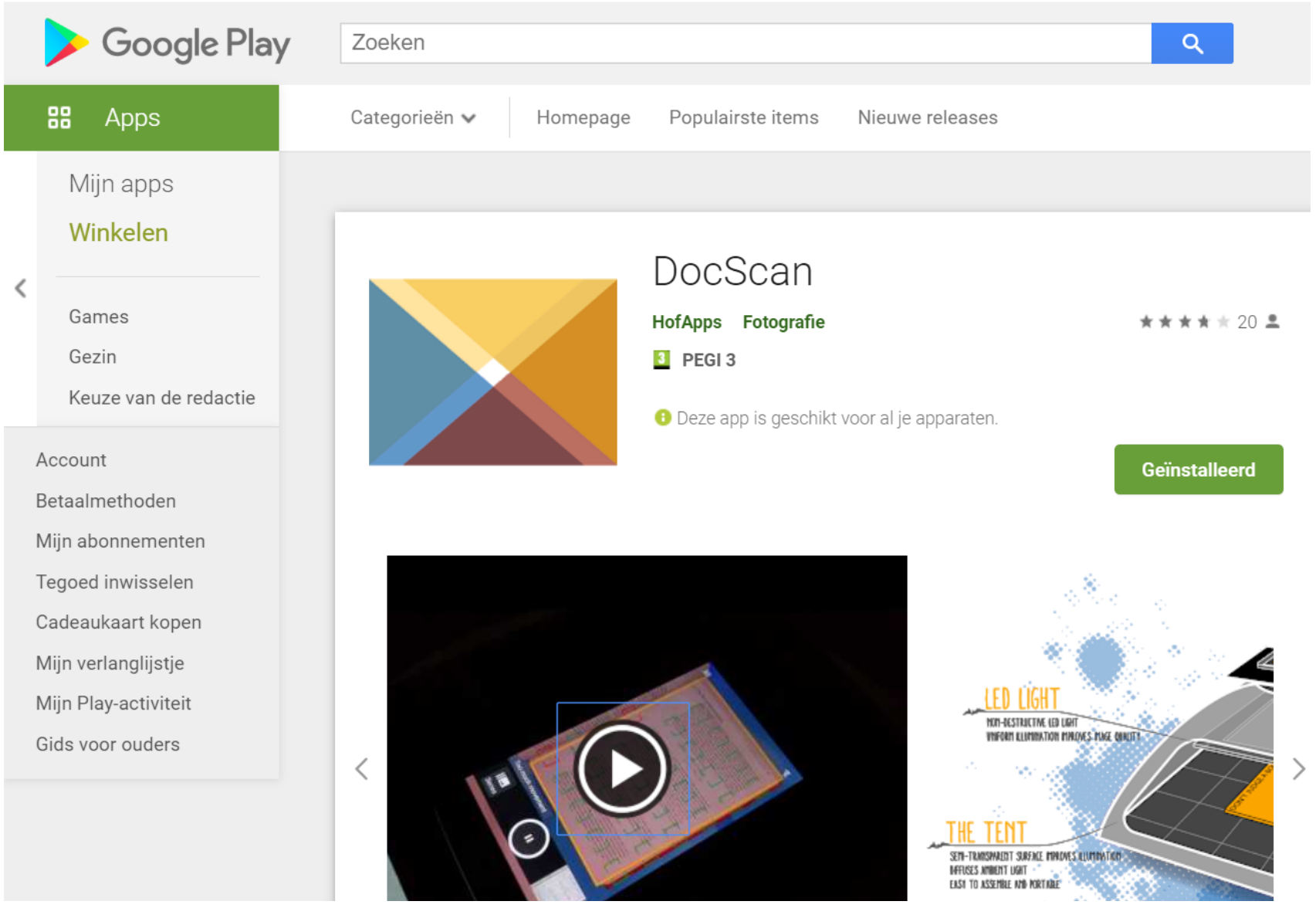
https://scantent.cvl.tuwien.ac.at/en/

Scantent





DocScan



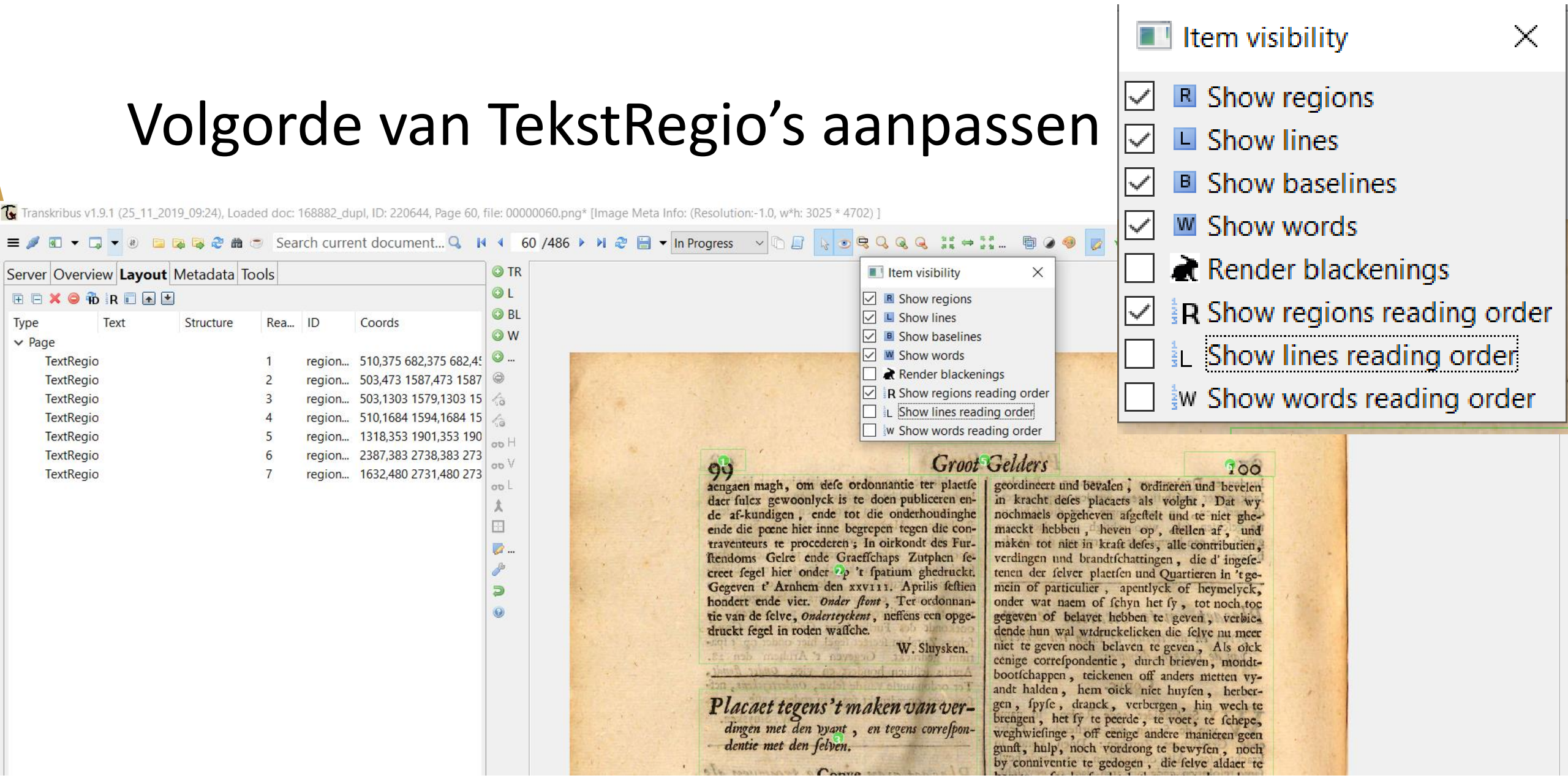
KB nationale bibliotheek

Veel voorkomend probleem: volgorde van TR of regels klopt niet…

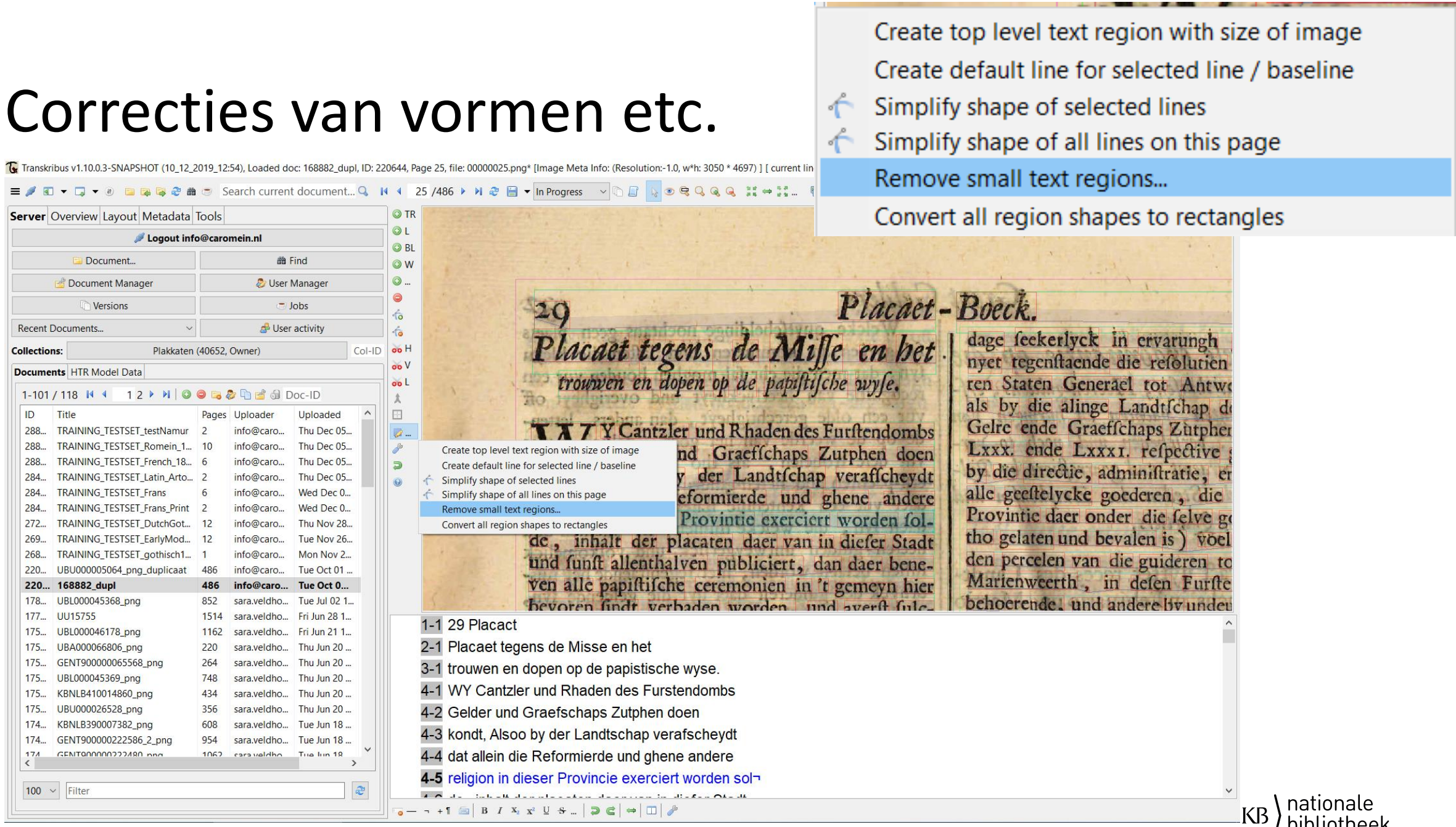
#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Volgorde van TekstRegio's aanpassen



Correcties van vormen etc.



KB nationale bibliotheek

OCR in Transkribus (ABBYY Finereader v.11)

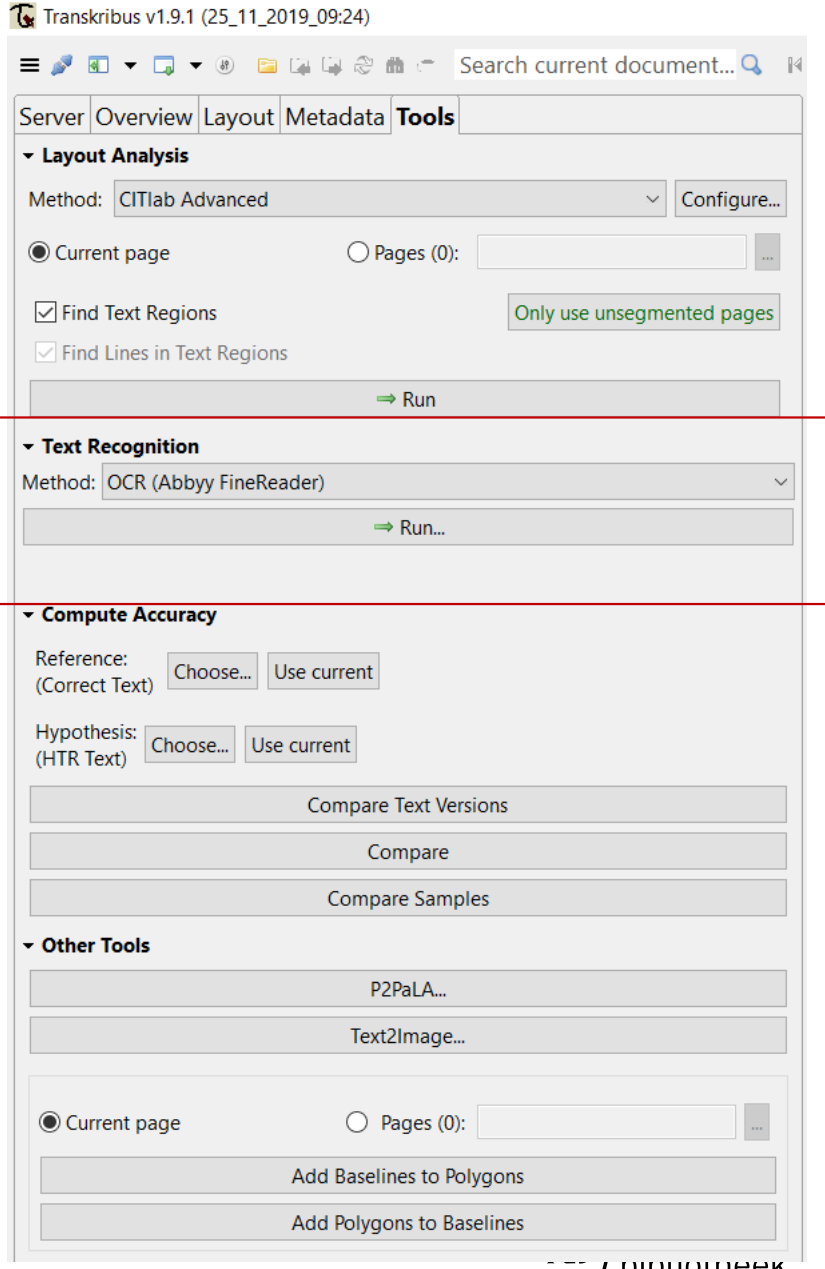
#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Optical Character Recognition

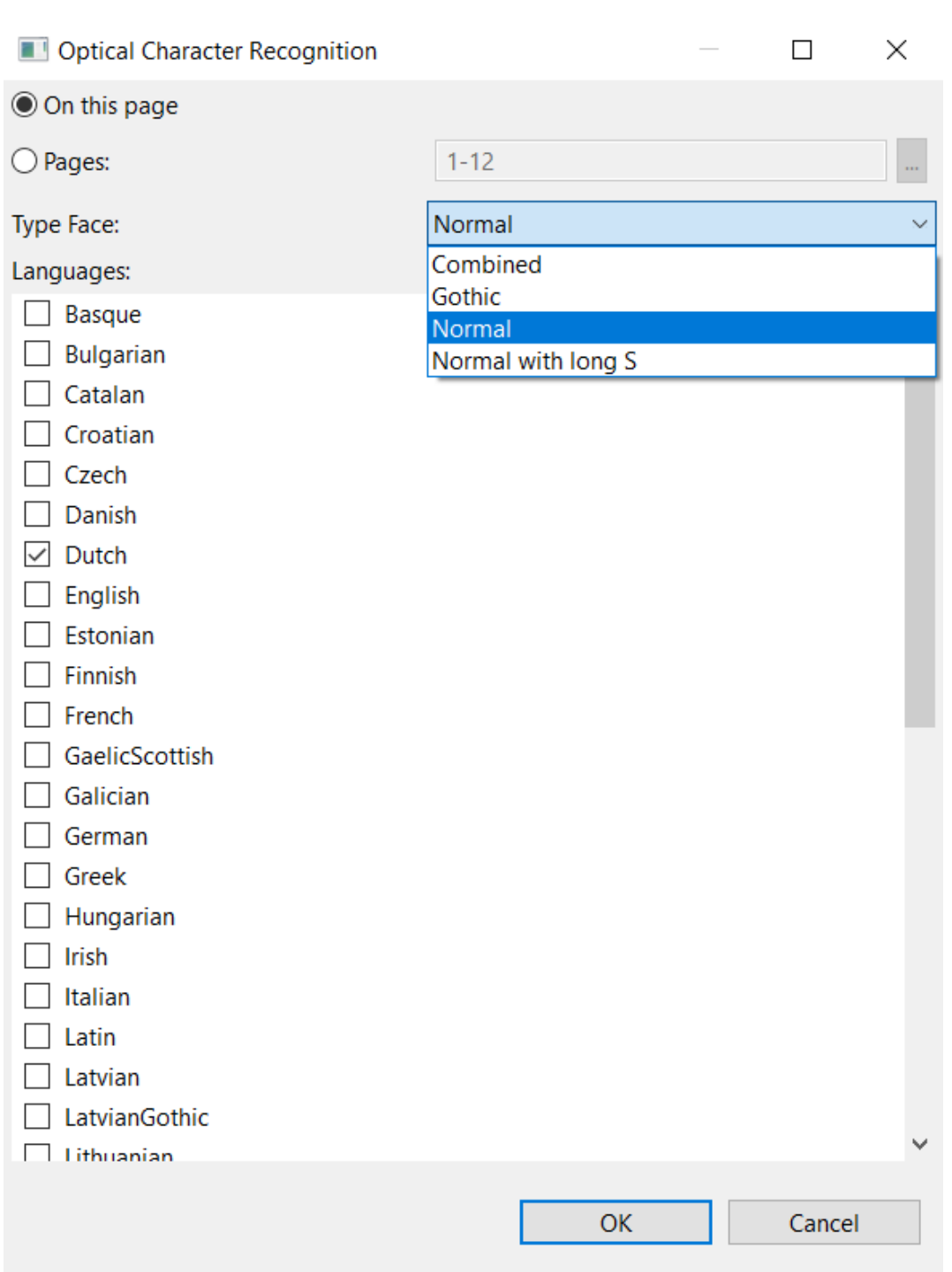
- Tab-menu, klik op Tools.
- Ga naar Text Recognition
- **Kies bij "Method" voor:** OCR (Abbyy FineReader v.11)
- Je kunt dan keuzes maken m.b.t. de taal en lettertype wanneer je op Run klikt.

- N.B. Alleen geschikt voor drukwerk.



ABBYY FineReader v.11

- Wanneer je OCR wilt toepassen moet je kiezen →
  - Taal
  - Lettertype
    - N.B. Gothic is Duits (Fraktur)!
    - NL Gothisch HTR model komt spoedig publiek (drukwerk).



Modellen:
maken, beoordelen, toepassen

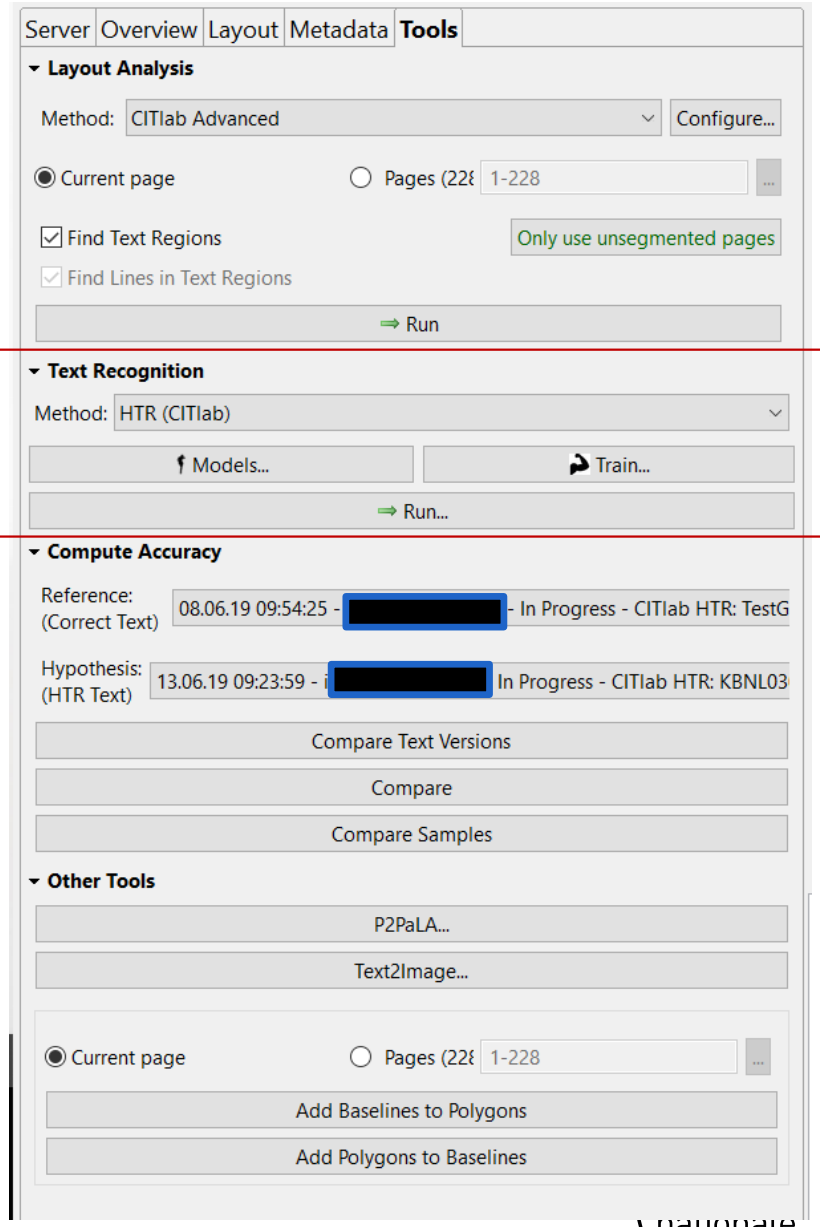
#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

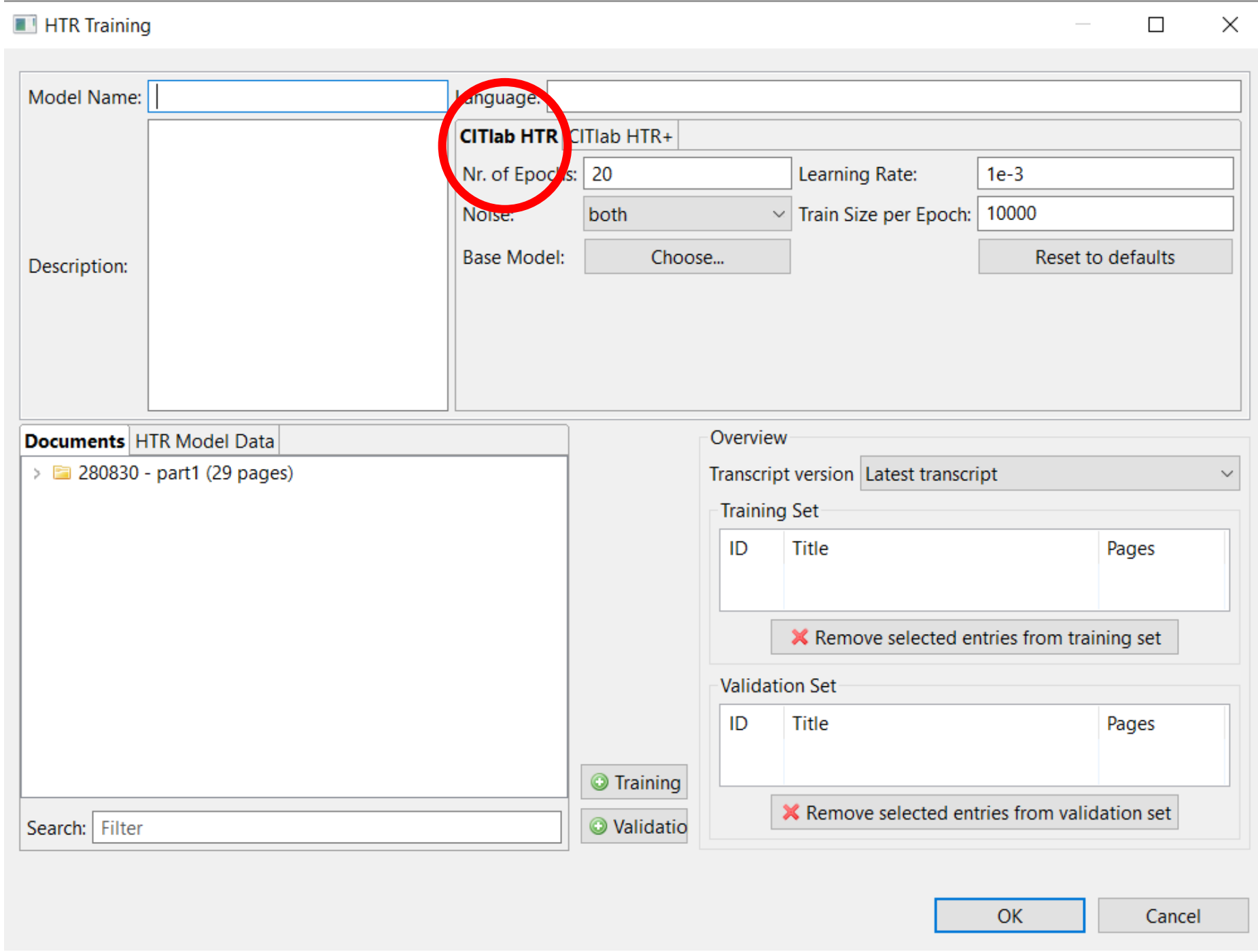
Train een eigen model

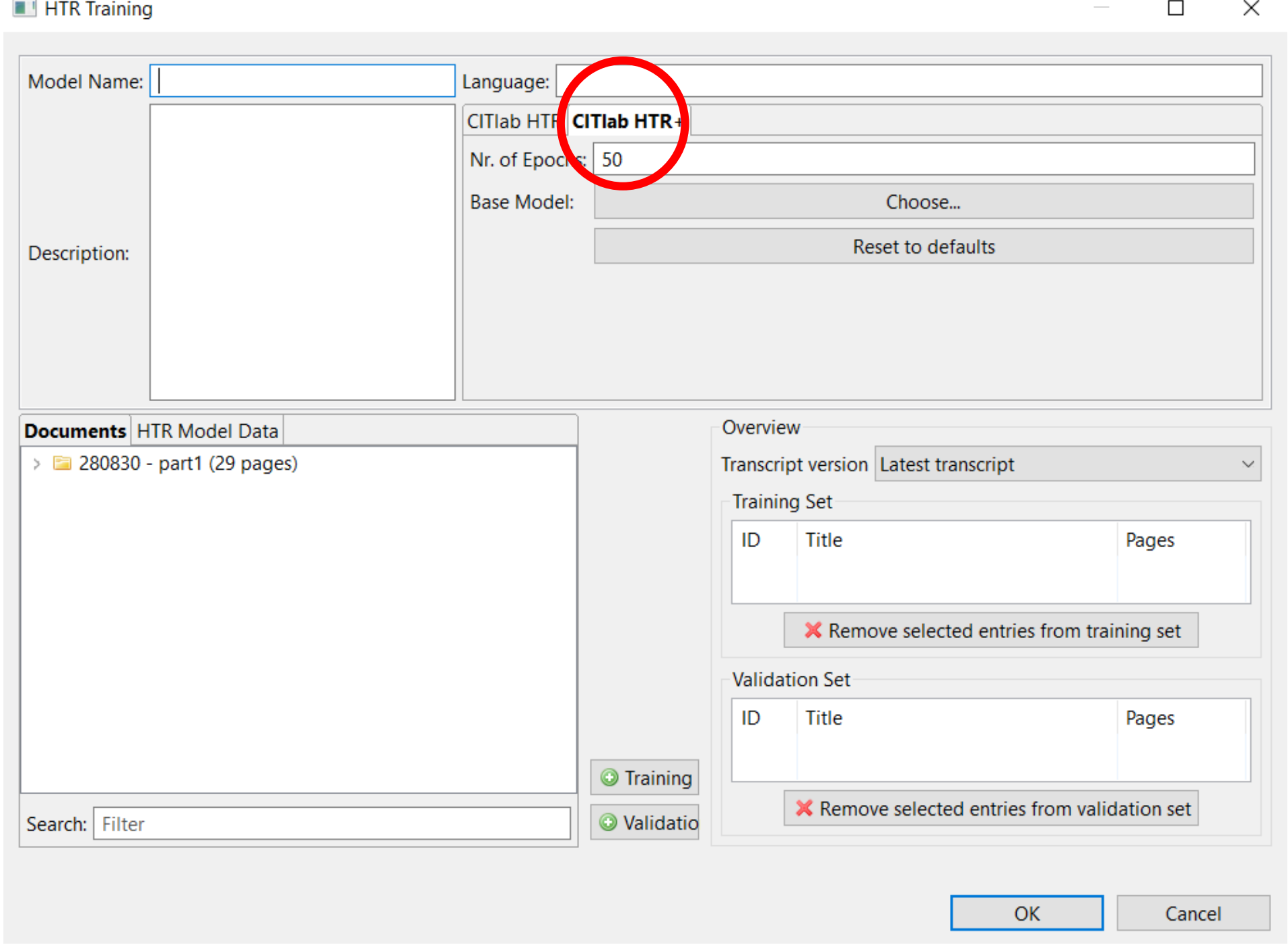
- Tab-menu, klik op Tools.
- Ga naar Tekst Recognition.
- **Kies bij “Method” voor: <u>HTR.</u>**

- Knop “Models” toont je de bestaande modellen (publieke en persoonlijke). <info>
- Knop “Train” is nodig om een eigen model kunnen trainen.
- Knop ‘Run’ is wanneer je een model wilt uit-voeren: dan pas selecteer je het model.



HTR


HTR Training
Model Name:
Language:
CITlab HTR
CITlab HTR+
Nr. of Epochs:
20
Learning Rate:
1e-3
Noise:
both
Train Size per Epoch:
10000
Base Model:
Choose...
Reset to defaults
Description:
Documents
HTR Model Data
280830 - part1 (29 pages)
Search:
Filter
Training
Validatio
Overview
Transcript version
Latest transcript
Training Set
ID
Title
Pages
Remove selected entries from training set
Validation Set
ID
Title
Pages
Remove selected entries from validation set
OK
Cancel


HTR Training
Model Name:
Language:
Description:
CITlab HTR
CITlab HTR+
Nr. of Epochs:
50
Base Model:
Choose...
Reset to defaults
Documents
HTR Model Data
280830 - part1 (29 pages)
Search:
Filter
Training
Validatio
Overview
Transcript version
Latest transcript
Training Set
ID
Title
Pages
Remove selected entries from training set
Validation Set
ID
Title
Pages
Remove selected entries from validation set
OK
Cancel

Publieke modellen (10-12-2019)

Choose a model

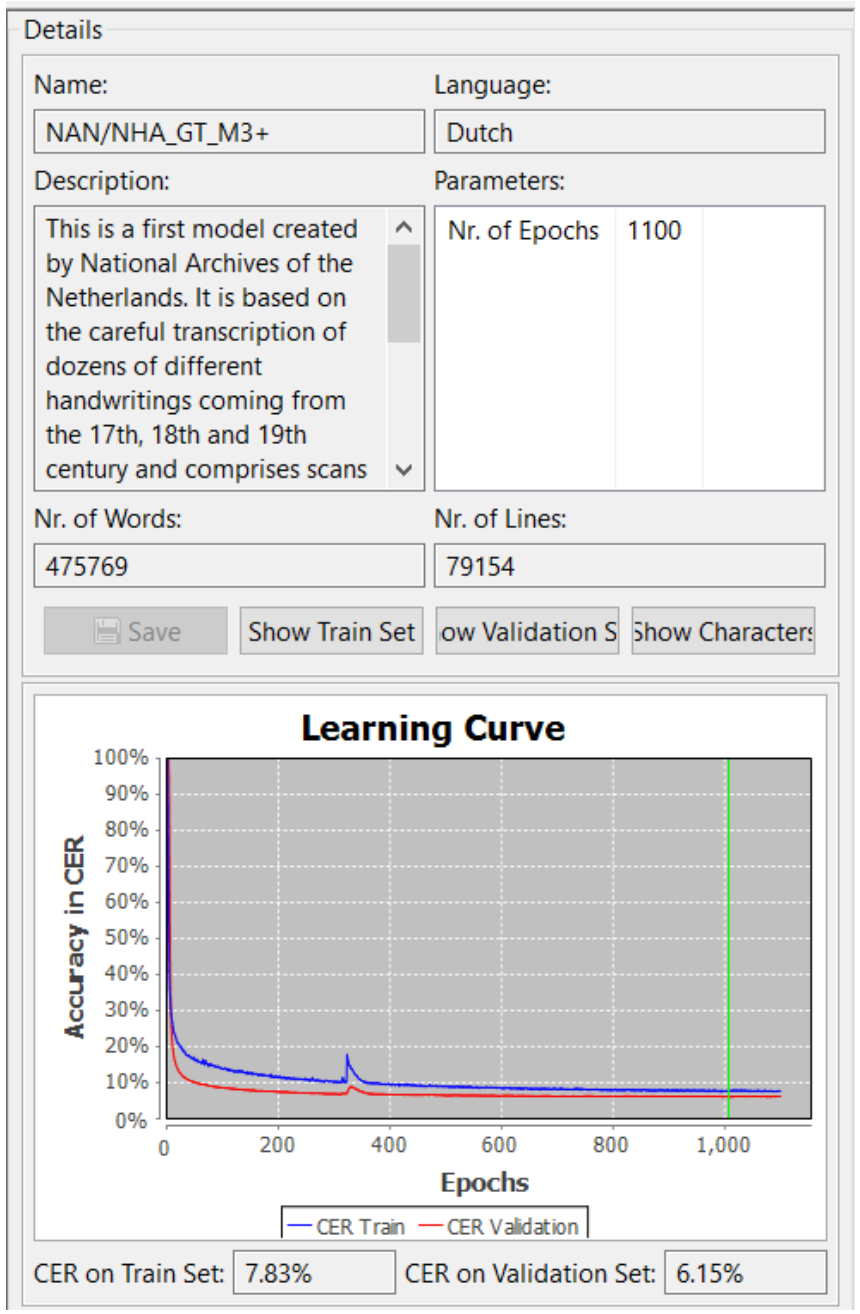
Search: | Technology: Show all

Name	Language	Curator	Technology	Created	ID
French_18thC_Print	French	info@caromei...	CITlab HTR+	05.12.19	191...
Dutch_Gothic_Print	Dutch (16th...	info@caromei...	CITlab HTR+	28.11.19	189...
Noscemus GM v1	Latin (partl...	stefan.zatham...	CITlab HTR+	22.11.19	187...
NAN/NHA_GT_M3+	Dutch	vincent.noppe...	CITlab HTR+	23.08.19	162...
Dutch Notarial Model 18th ...	Dutch	jirsireinders19...	CITlab HTR+	17.07.19	157...
NZZ Gold Standard M1+	German	guenter	CITlab HTR+	23.04.19	126...
Combined_Full_VKS_2	Church Slav...	achim.rabus@...	CITlab HTR+	14.04.19	124...
HIMANIS Chancery M1+	Latin, French	guenter.hackl...	CITlab HTR+	14.04.19	124...
ONB_Newseye_GT_M1+	german	guenter.hackl...	CITlab HTR+	15.02.19	108...
German Kurrent M1+	German	guenter	CITlab HTR+	31.01.19	103...
VMC_Test_4+	Russian Ch...	achim.rabus@...	CITlab HTR+	25.01.19	101...
English Writing M1	English	Unknown	CITlab HTR	08.04.16	133

Advies

- Modellen voor politieke, economische, sociale, culturele, literaire teksten (etc.) kunnen enorm verschillen en minder goed werken op teksten van een ander type;
- Modellen kunnen regioafhankelijk zijn (dialect);
- Periodisering is belangrijk

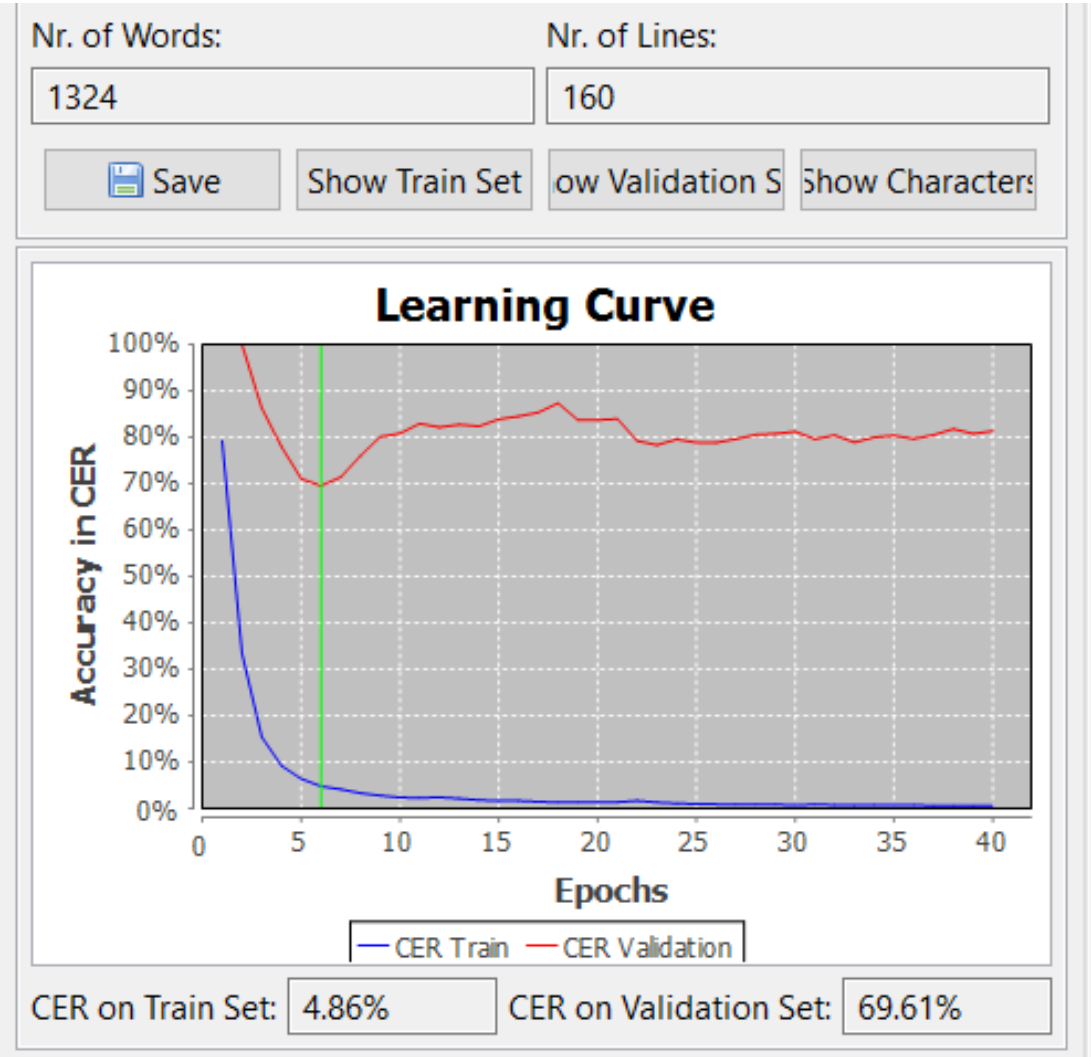
Vermeldt deze informatie aub in het ‘description’ field!



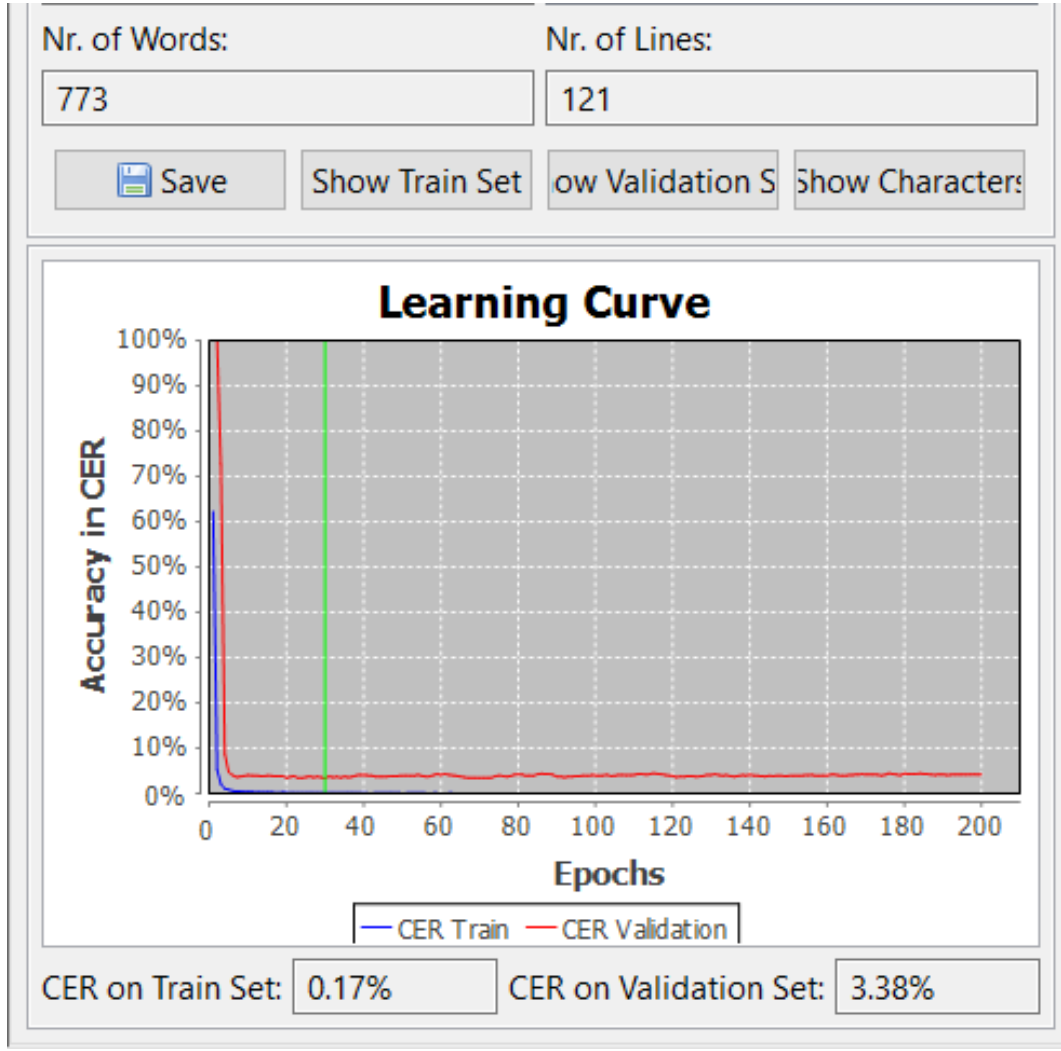
- CER= Character Error Rate (op de Trainingsset of Validatieset)
- Ook zichtbaar: hoeveel woorden en hoeveel regels (min >1000) er zijn gebruikt.
- Aantal Epochs.
- Show Characterset

KB nationale bibliotheek

Grafieken zijn niet altijd zo mooi…



Leigh Hunt (o.a. te weinig training/ validatie)

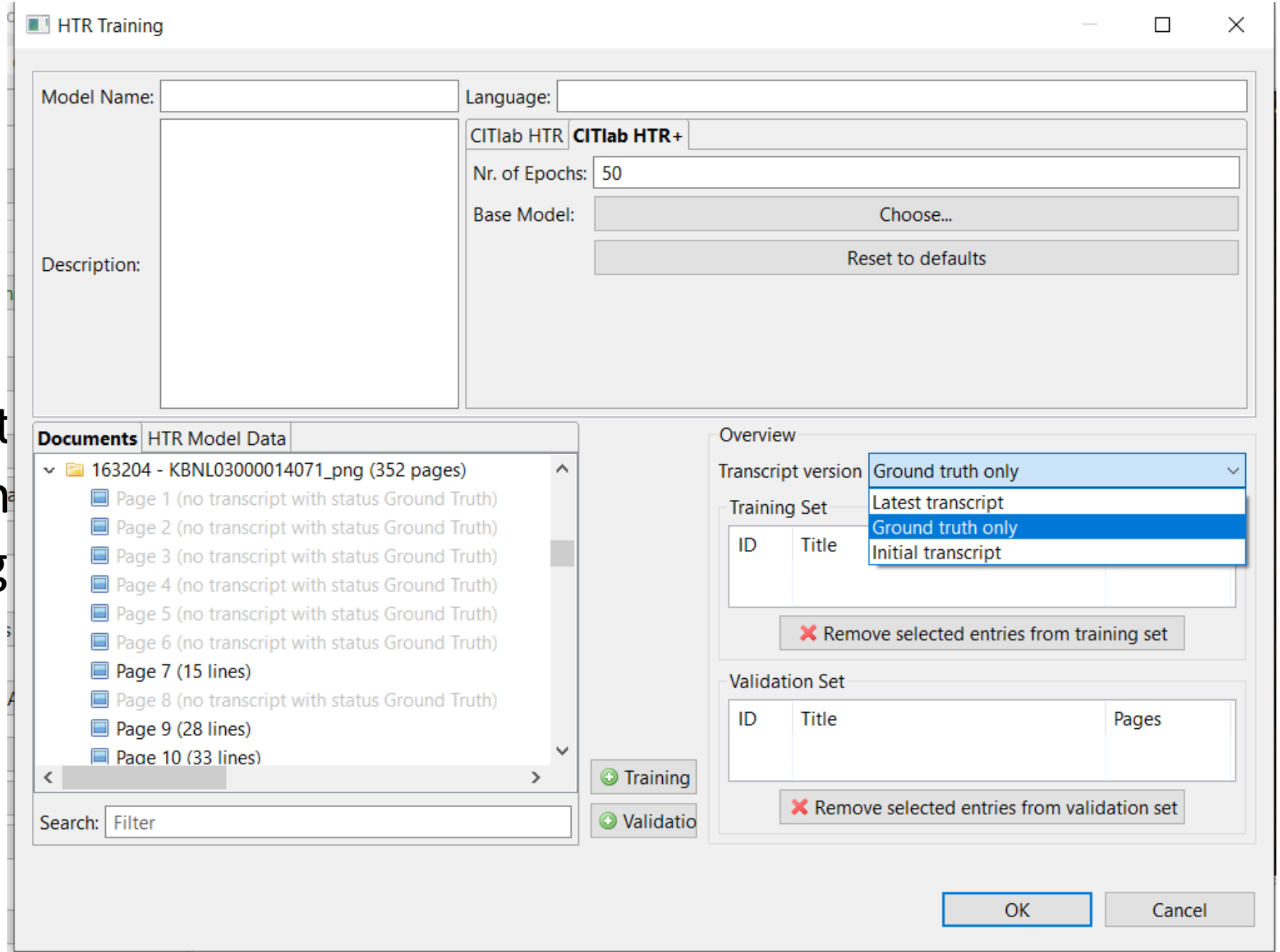


Latijn drukwerk (overfitting)

KB nationale bibliotheek

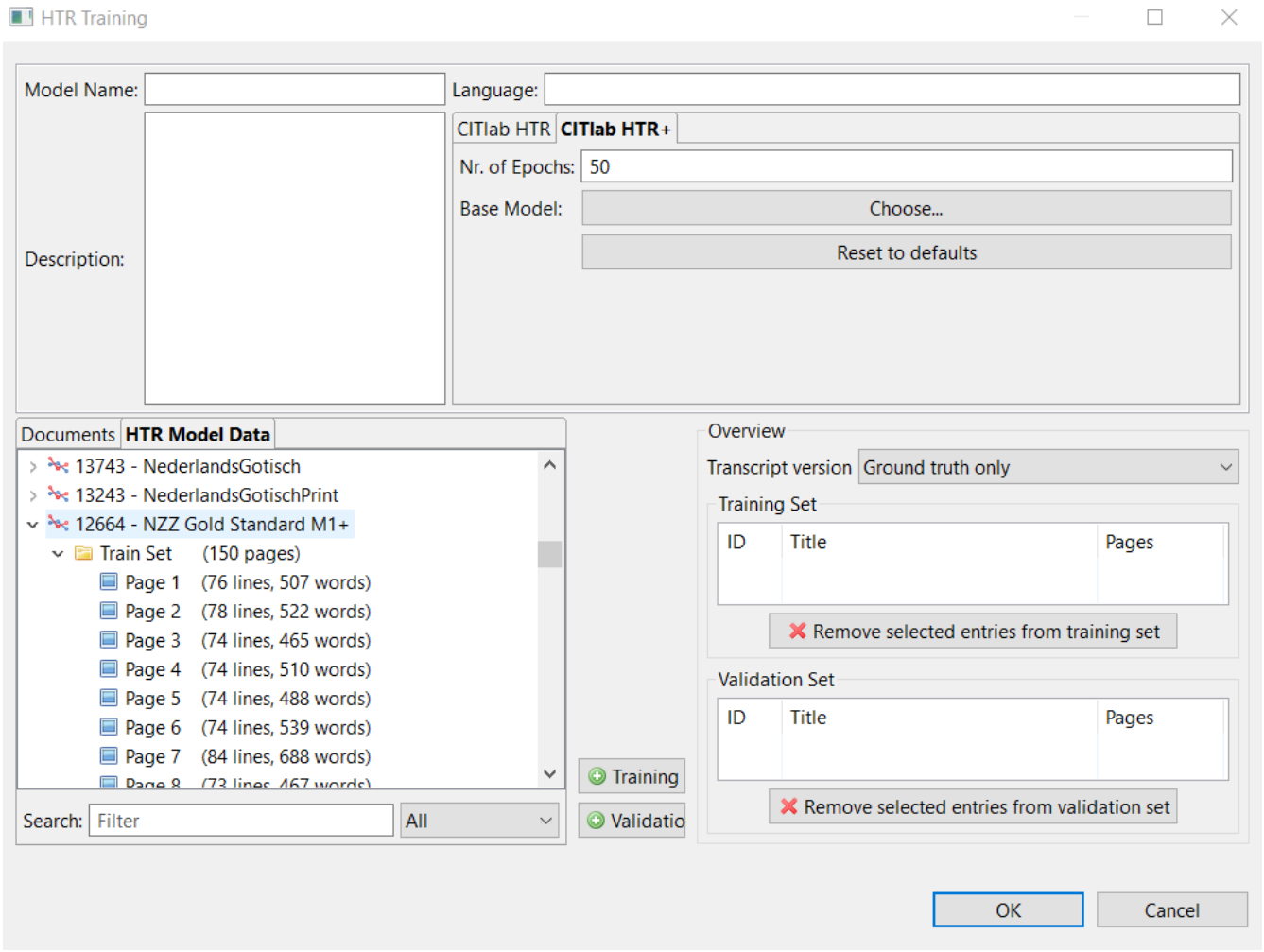
Model trainen

- Kies voor transcriptieversie “GT”
- Dan per document die pagina’s kiezen die van toepassing zijn



Trainen op basis van een bestaand model

- Methode 1. Invoeren als zijnde een basemodel.
- Methode 2. Vrijgegeven trainingsdata invoeren als trainingsset/validatieset. (Tab HTR Model Data)

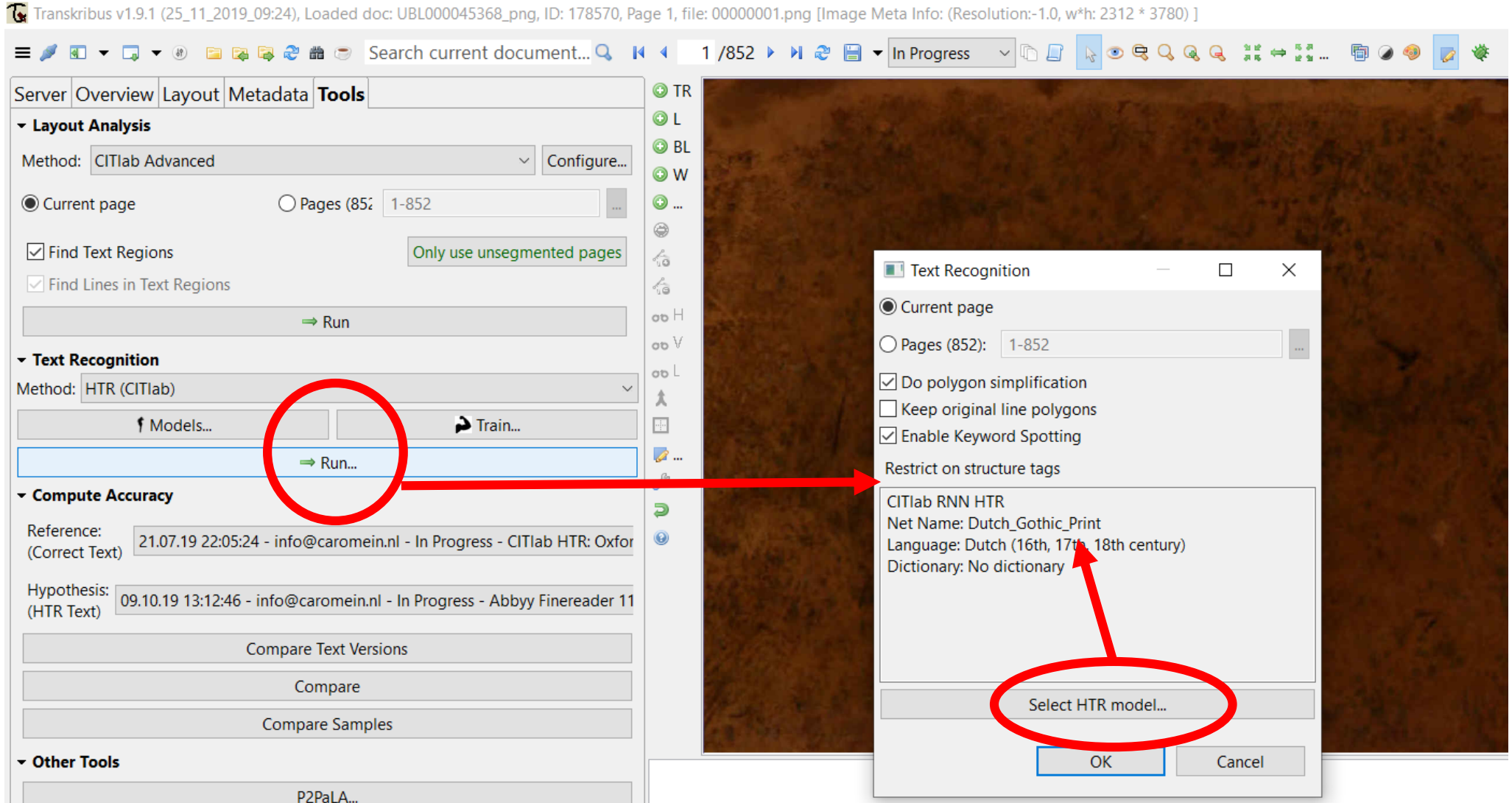


KB nationale bibliotheek

Modellen:
Werkt het ok?

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek



KB nationale bibliotheek

Transkribus v1.9.1 (25_11_2019_09:24), Loaded doc: KBNL03000014071_png, ID: 163204, Page 50, file: 00000050.png [Image Meta Info: (Resolution:-1.0, w*h: 1757 * 2216)]

Search current document... 50 /352 In Progress

Server | Overview | Layout | Metadata | **Tools**

Layout Analysis

Method: CITlab Advanced Configure...

Current page Pages (352): 1-352

Find Text Regions Only use unsegmented pages

Find Lines in Text Regions

Run

Text Recognition

Method: HTR (CITlab)

Models... Train...

Run...

Compute Accuracy

Reference: (Correct Text) 27.11.19 10:20:13 - info@caromein.nl - Ground Truth Use current

Hypothesis: (HTR Text) 29.11.19 20:11:12 - info@caromein.nl - In Progress - Abbyy Finereader 11

Compare Text Versions

Compare

Compare Samples

Other Tools

P2PaLA...

Text2Image...

Current page Pages (352) 1-352

Add Baselines to Polygons

Add Polygons to Baselines

TR

Version Comparator

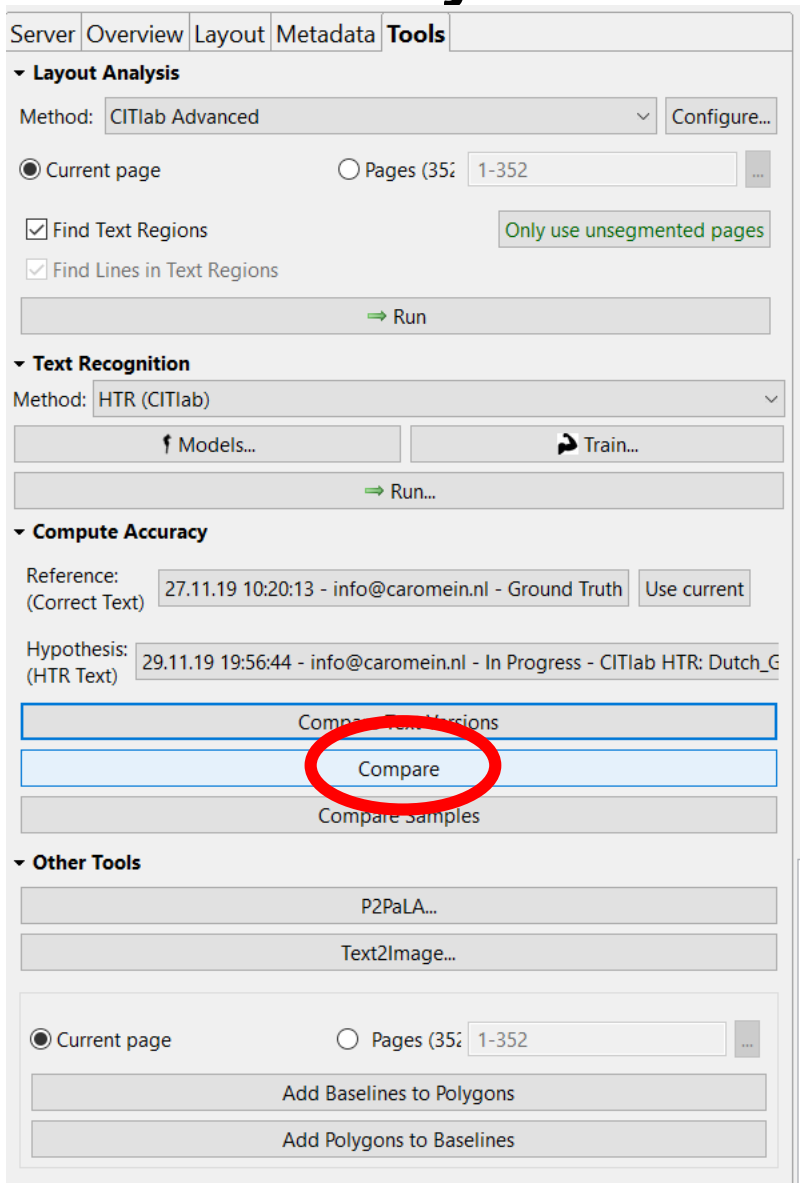
Show line numbers

42 Generale
~~fthappm~~ schappen de ~~volgende't~~ volgende 't Placcaet ofte ~~LM~~ Lijste confiscatie
subject ~~zhn /~~ zijn/ ende in ~~vier~~ dier voegen ~~angehaelt~~ aengehaest worden/ ~~dre~~ die
selve sullen in Sequestro ~~bllven /~~ blijven/ ende ~~verwant wozden~~ verwaert worden
~~vp Ve restiectlve~~ by de respective Collecteurs ofte ~~Gerievrs dienaers~~ Gerichts-dienaers in
Stadt ende ~~Lande /~~ Lande/ ter ~~thdt~~ tijdt daer over verclaert ~~olie~~ ofte voor
de waerde van dien ~~genoechfame~~ genoechsame cautie ~~geftellet fal~~ gestellet sal zijn/
~~Dp~~ Ende by aldien die selve verklaert worden van goeden ~~pry~~ prij-
se/ sullen ~~dieselde~~ dieselve ter instantie van den ~~PachterV?euck~~ Pachter/ Breuck-
~~tachter~~ Pachter ofte ~~Mvocaet Pxovinciacl doo? o?d?e~~ Advocaet- Provinciael door ordre van die
~~eerm~~ Heeren Gedeputeerden in 't ~~opmbaer dp uptmijninge~~ openbaer by uytmijninge
~~berk o ft mde Die~~ verkoft/ ende die penningen daer van ~~procederende~~ procederende/ ge-
~~distnwleert wo?den naemholrvan~~ distribueert worden nae inholt van den ~~Placcate~~ Placcate/ in dier
~~voegen /~~ voegen/ dat des Pachters ende ~~Breuck pachttrs~~ Breuck- pachters anpar-
ten ontfangen sullen ~~wordm bp~~ worden by de particuliere ~~Gntfan~~ Ontfan-
ger van 't eerste ~~Gomtop? / fonder~~ Comtoyr/ sonder dat de Pachters ~~ver~~ ver-
mogen ~~susten /~~ sullen/ hen dese voorschreven goederen aen te me-
tigen ~~ineenigrrhande~~ in eenigerhande manieren ~~alS voozsz~~ als voorsz. is ~~bp~~ by verlies
van haer anpart in ~~dieselde /~~ dieselve/ ende ~~50~~ zo. guld. ~~daerrnvoven~~ daerenboven
te verbeuren.
~~ť H 1 3 3~~ XXXIII.
Die respective Pachters sullen ~~vermogm~~ vermogen soo ~~dich~~ dick-
~~wylS alg~~ wijls als 't haer gelieft aen die ~~Lomtopren comparerm/~~ Comtoyren compareren/
~~Boecken /~~ Boecken/ en Journalen concernerende haren pacht
ende geen anderen examineren/ ende op de frauden in-
~~gmreren~~ quireren/ daer toe de Collecteurs haer ~~susten admitterm~~ sullen admitteren
ende haer ~~behooMK amwyfinge doen / susten~~ behoorlijck aenwijfinge doen/ sullen niet te
~~Min~~ min die ~~H eerm~~ Heeren Staten van die Provincie ofte derselver
Gedeputeerden visie van de ~~Doecken~~ Boecken mogen ~~nemm /~~ nemen/ ter

2-5 Stadt ende Lande / ter thdt daer over verclaert olie voor
3-1 de waerde van dien genoechfame cautie geftellet fal zijn/
4-1 <Ln Dp aldien die selve verklaert worden van goeden pry-

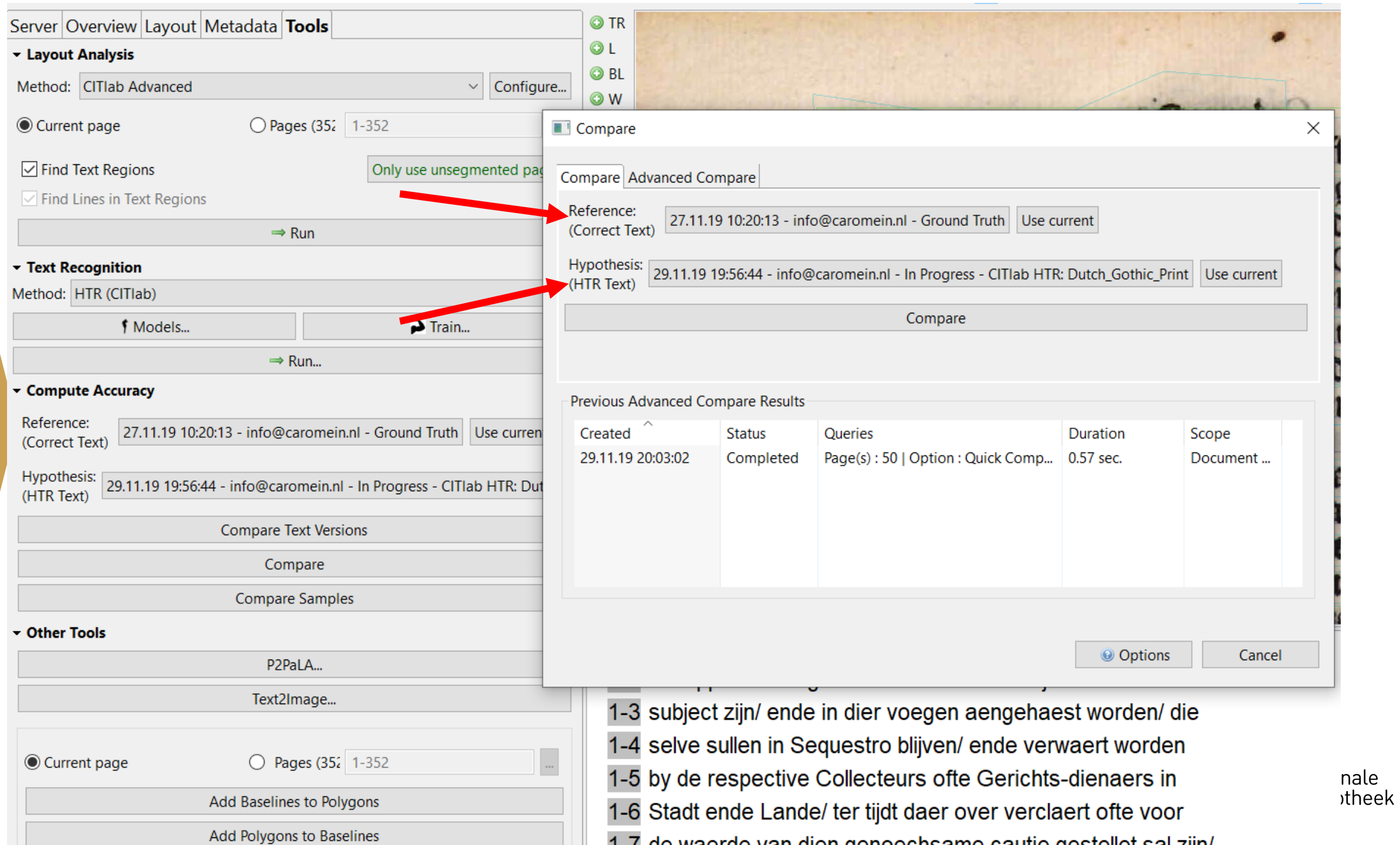
Goede Tijden, Slech... Transkribus_NL_gea... Postvak IN - Annem... Transkribus v1.9.1 (2... Error NLD 20:17 29-11-19

Hoe weet je of een model passend is voor jouw tekst?



Werkt mijn model goed?
Maak een pagina GT
Draai een model over diezelfde pagina
“Compare”

KB nationale bibliotheek


Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method:
CITlab Advanced
Configure...
Current page
Pages (352
1-352
Find Text Regions
Only use unsegmented pag
Find Lines in Text Regions
Run
Text Recognition
Method:
HTR (CITlab)
Models...
Train...
Run...
Compute Accuracy
Reference:
(Correct Text)
27.11.19 10:20:13 - info@caromein.nl - Ground Truth
Use curren
Hypothesis:
(HTR Text)
29.11.19 19:56:44 - info@caromein.nl - In Progress - CITlab HTR: Dut
Compare Text Versions
Compare
Compare Samples
Other Tools
P2PaLA...
Text2Image...
Current page
Pages (352
1-352
Add Baselines to Polygons
Add Polygons to Baselines
TR
L
BL
W
Compare
Compare
Advanced Compare
Reference:
(Correct Text)
27.11.19 10:20:13 - info@caromein.nl - Ground Truth
Use current
Hypothesis:
(HTR Text)
29.11.19 19:56:44 - info@caromein.nl - In Progress - CITlab HTR: Dutch_Gothic_Print
Use current
Compare
Previous Advanced Compare Results
Created
Status
Queries
Duration
Scope
29.11.19 20:03:02
Completed
Page(s) : 50 | Option : Quick Comp...
0.57 sec.
Document ...
Options
Cancel
1-3 subject zijn/ ende in dier voegen aengehaest worden/ die
1-4 selve sullen in Sequestro blijven/ ende verwaert worden
1-5 by de respective Collecteurs ofte Gerichts-dienaers in
1-6 Stadt ende Lande/ ter tijdt daer over verclaert ofte voor

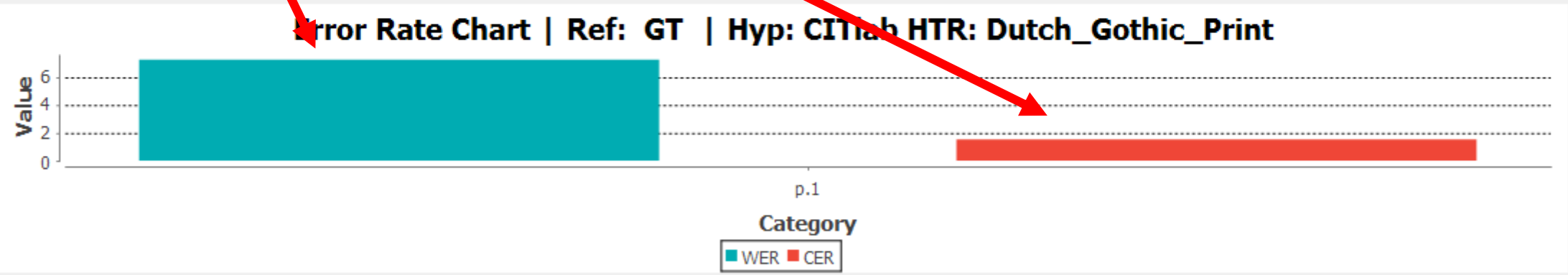
nale
otheek

Advanced Statistics

Page	WER	CER	Word Acc	Char Acc	Bag Tokens...	BT Recall	BT F1-Score
Overall	7.27 %	1.54 %	86.86 %	97.5 %	0.93	0.928	0.93

Page	WER	CER	Word Acc	Char Acc	Bag Tokens...	BT Recall	BT F1-Score
Page 1	7.27 %	1.54 %	86.86 %	97.5 %	0.92760000...	0.93180000...	0.92970521...

Compare Text Versions for Page ..



Base folder: C:\Users\jva010

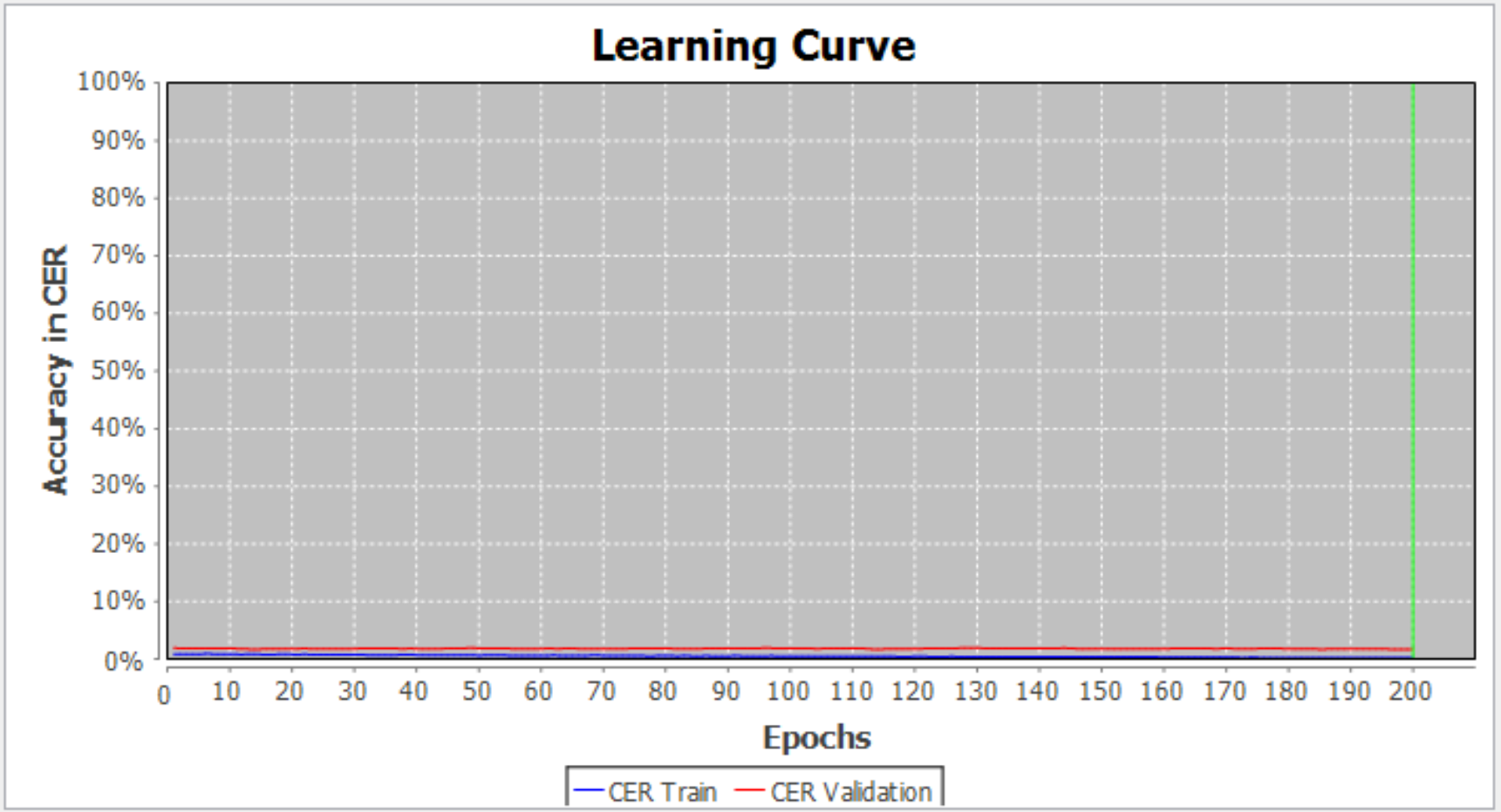
File/Folder name: DocId_276619

Export path: C:\Users\jva010\DocId_276619.xls

Nr. of Words: 51143

Nr. of Lines: 7143

Save | Show Train Set | Show Validation Set | Show Characters

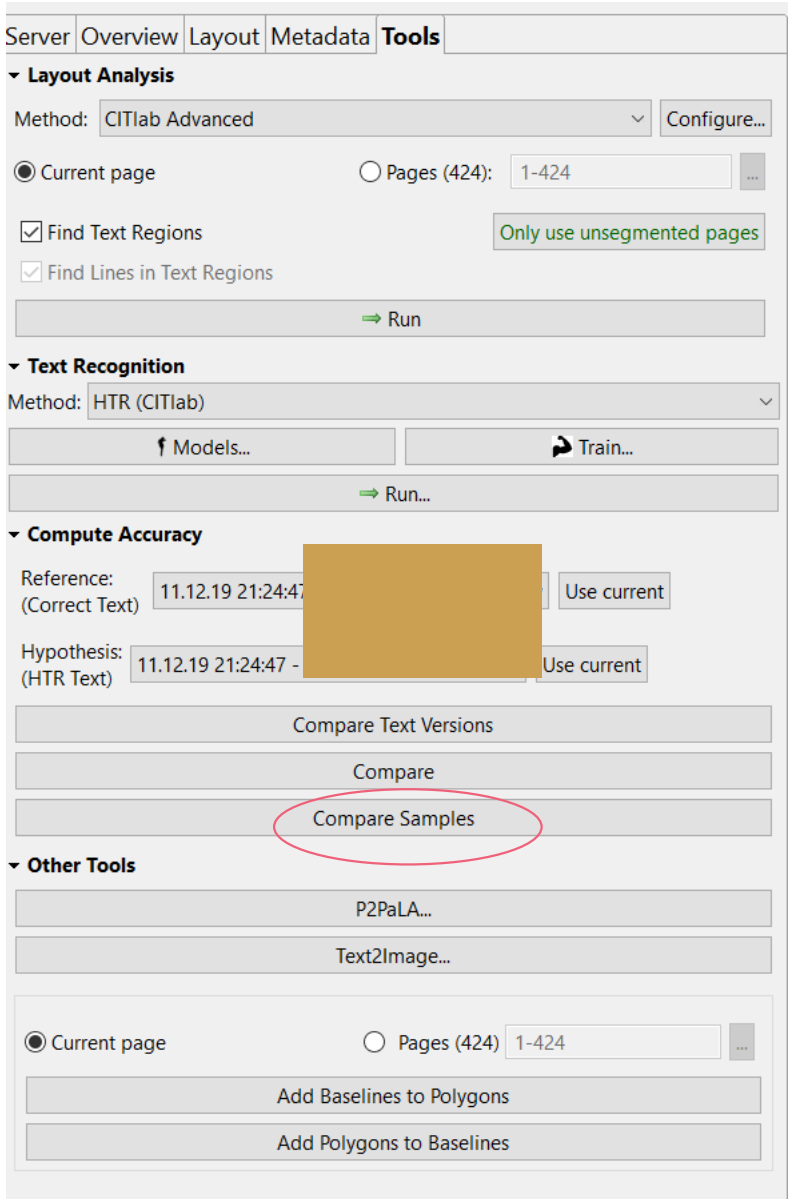


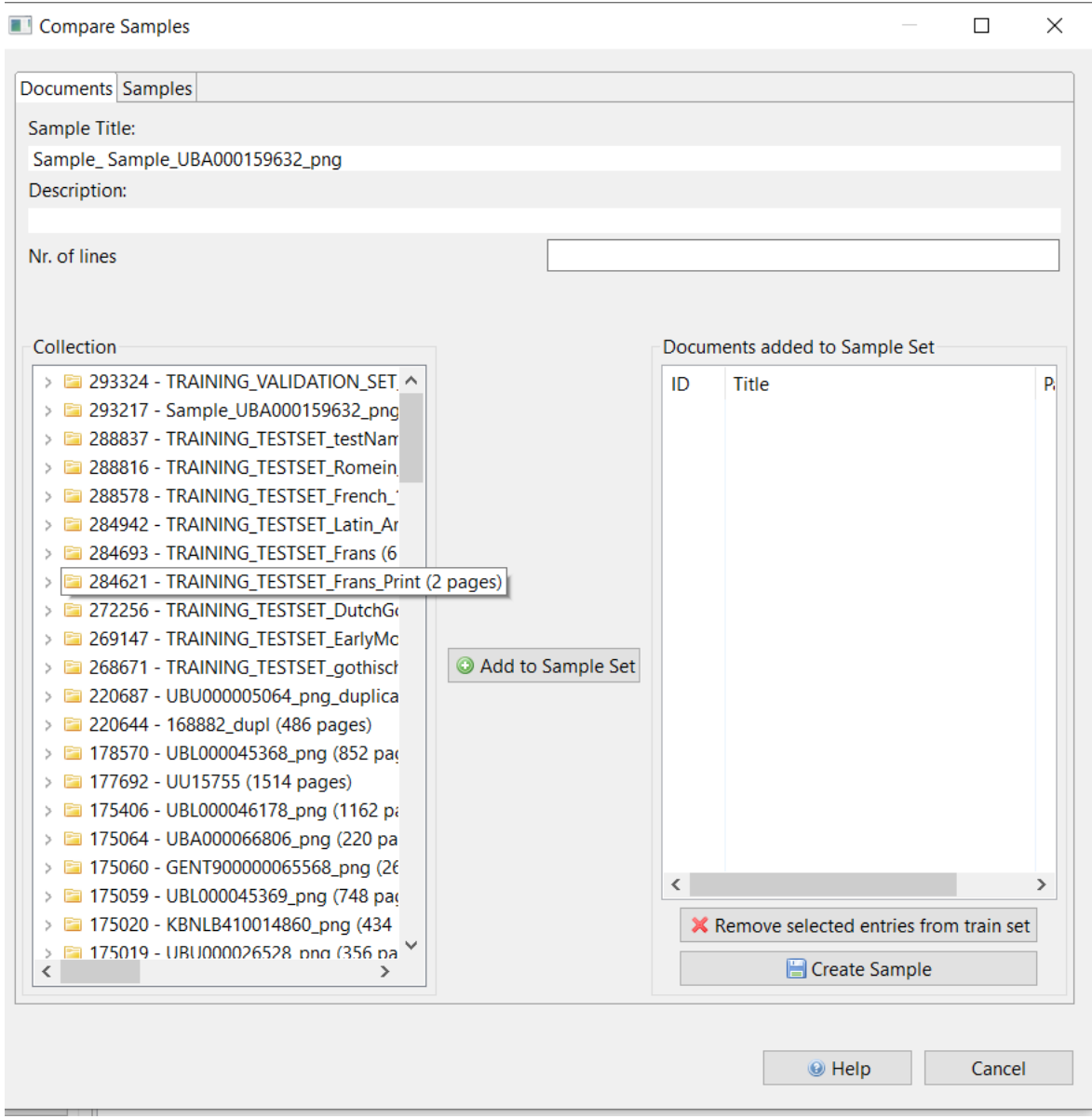
CER on Train Set: 0.22%

CER on Validation Set: 1.71%

KB nationale bibliotheek

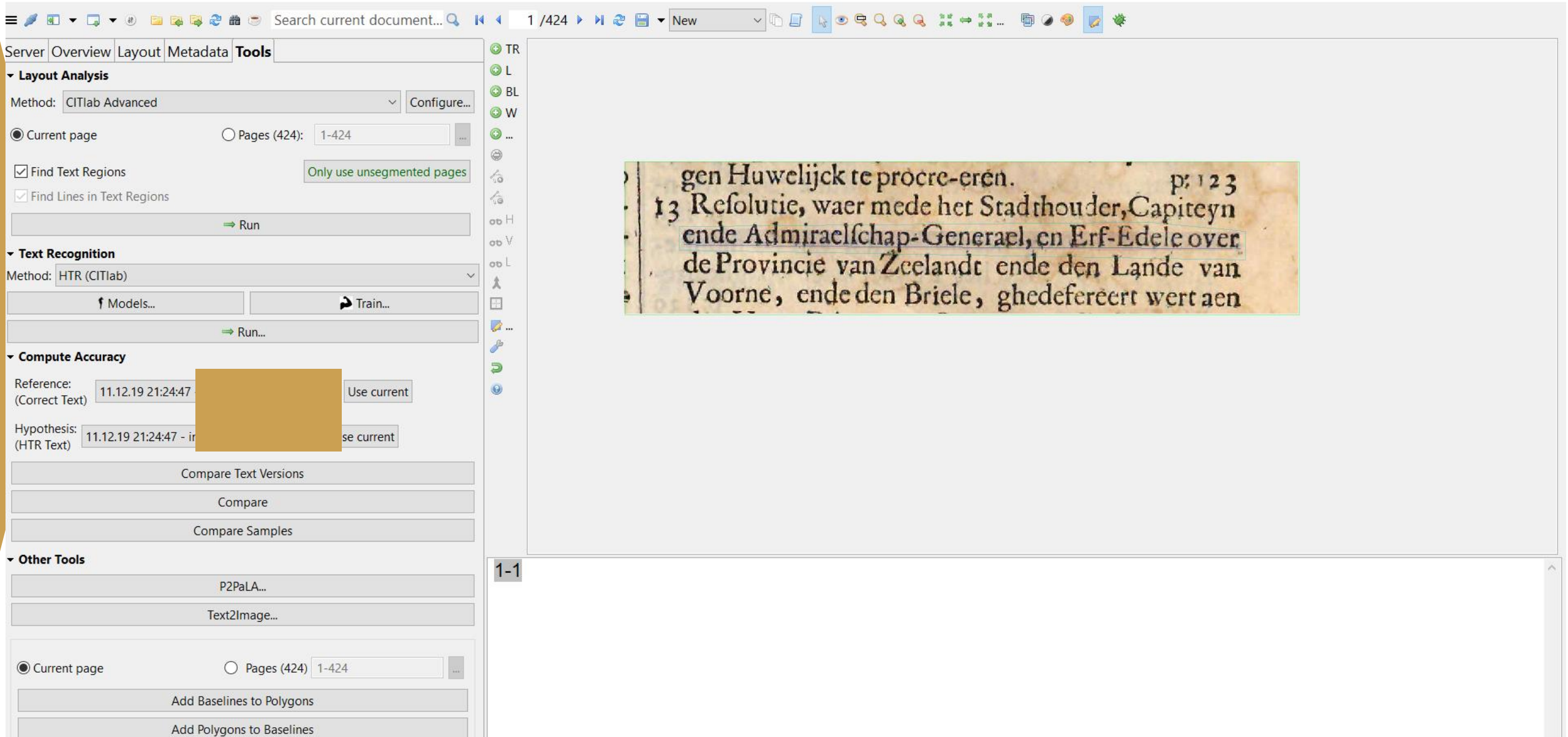
Sample





KB nationale bibliotheek

Sample


Search current document...
1 /424
New
Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method:
CITlab Advanced
Configure...
Current page
Pages (424):
1-424
Find Text Regions
Only use unsegmented pages
Find Lines in Text Regions
Run
Text Recognition
Method:
HTR (CITlab)
Models...
Train...
Run...
Compute Accuracy
Reference:
(Correct Text)
11.12.19 21:24:47
Use current
Hypothesis:
(HTR Text)
11.12.19 21:24:47 - in
se current
Compare Text Versions
Compare
Compare Samples
Other Tools
P2PaLA...
Text2Image...
Current page
Pages (424)
1-424
Add Baselines to Polygons
Add Polygons to Baselines
TR
L
BL
W
H
V
L
gen Huwelijck te procre-eren. p: 123
13 Resolutie, waer mede het Stadthouder, Capiteyn
ende Admiraelschap-Generael, en Erf-Edele over
de Provincie van Zeelandt ende den Lande van
Voorne, ende den Briele, ghedefereert wert aen
1-1

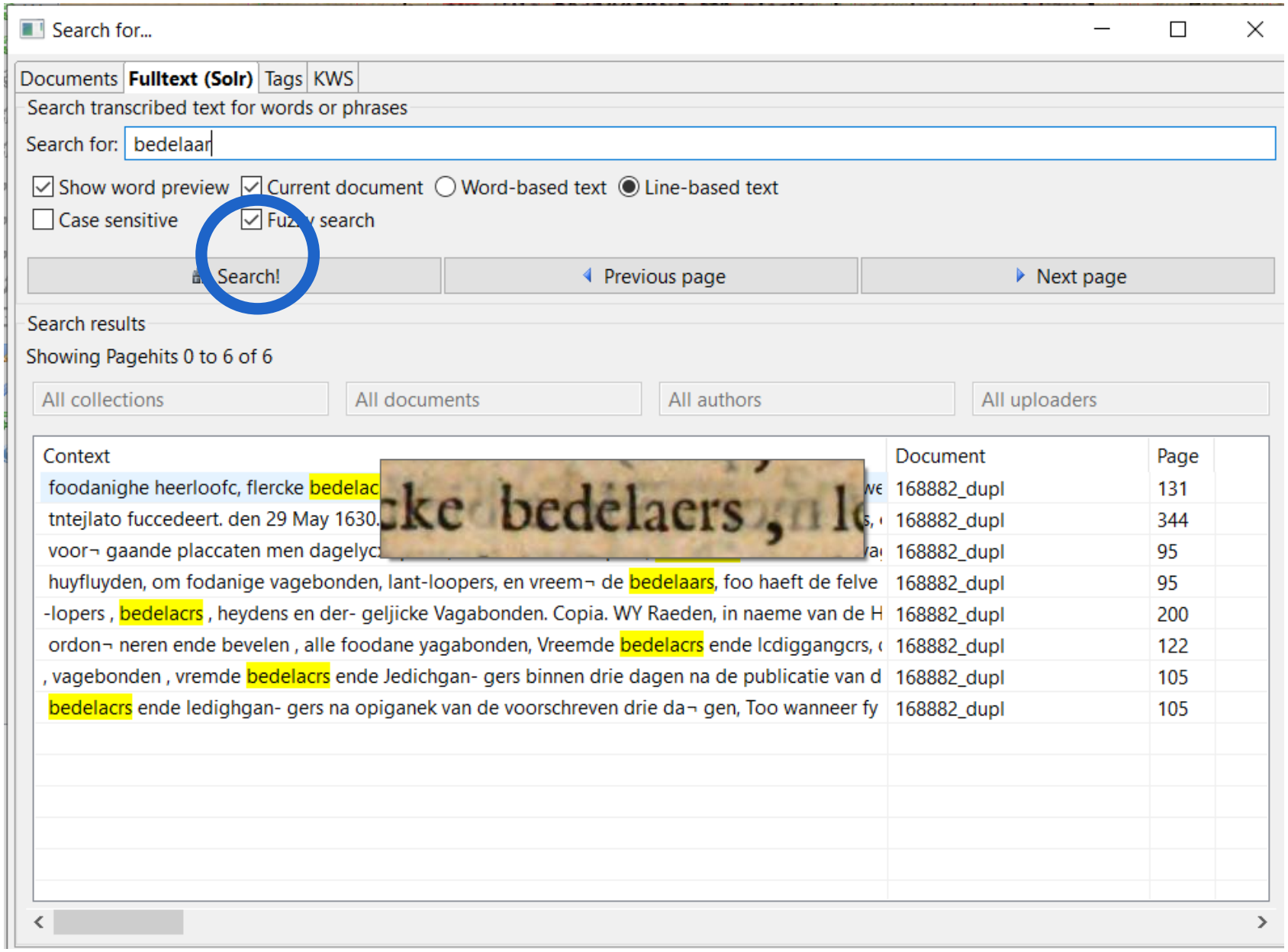
KeyWord Spotting (KWS)

#Transkribus
#webinargeavanceerdNL

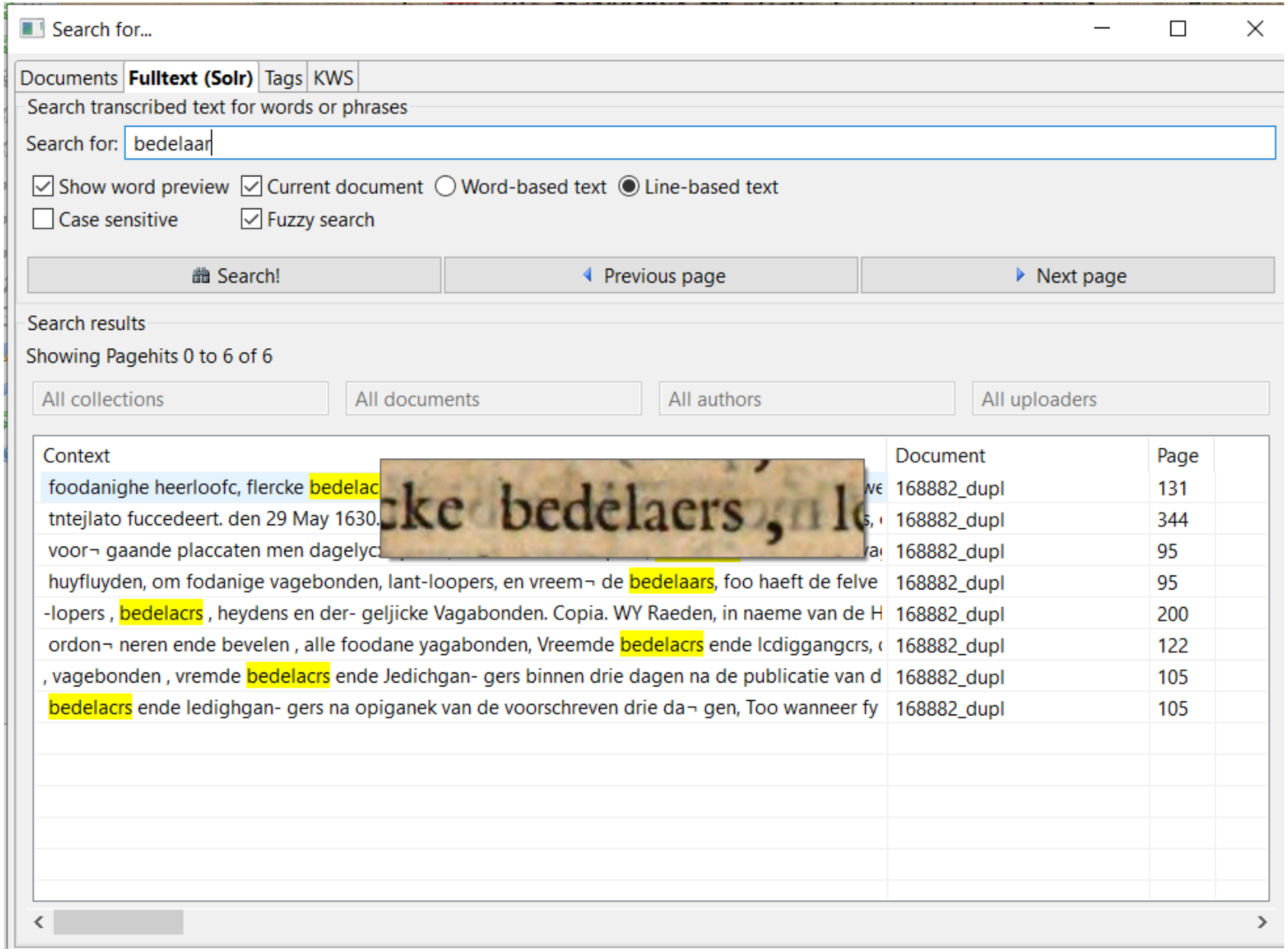
KB nationale bibliotheek

Zoeken middels KeyWord Spotting(KWS) en full text search

- Technologie om in de afbeeldingen snel naar woorden te zoeken.
- Gebaseerd op de HTR Transcripties.
- Functioneert tot ±30% CER
- Toont op basis van:
  - Confidence Value (probability)
  - Kan altijd worden getoont

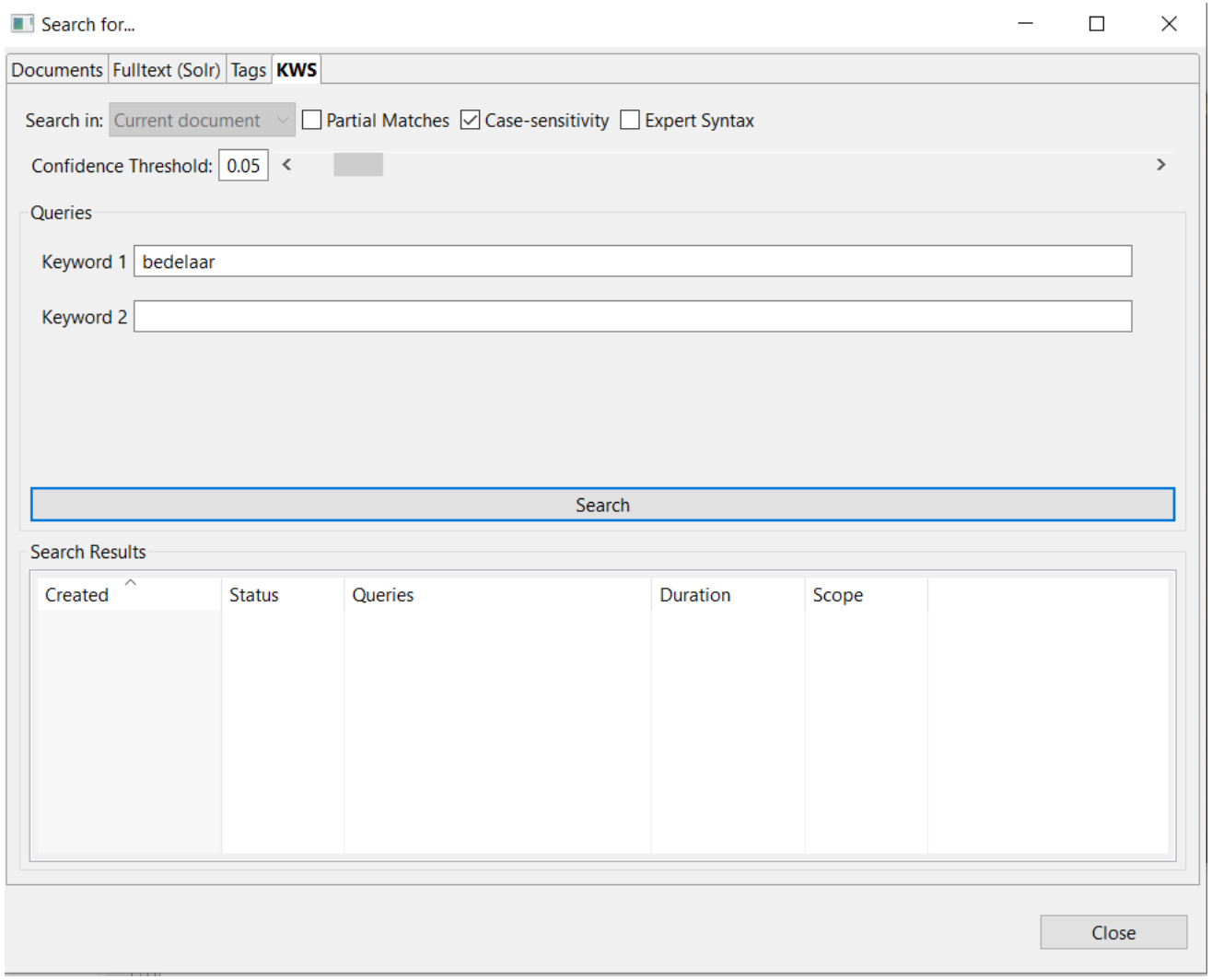


- Full tekst search (met fuzzy optie)



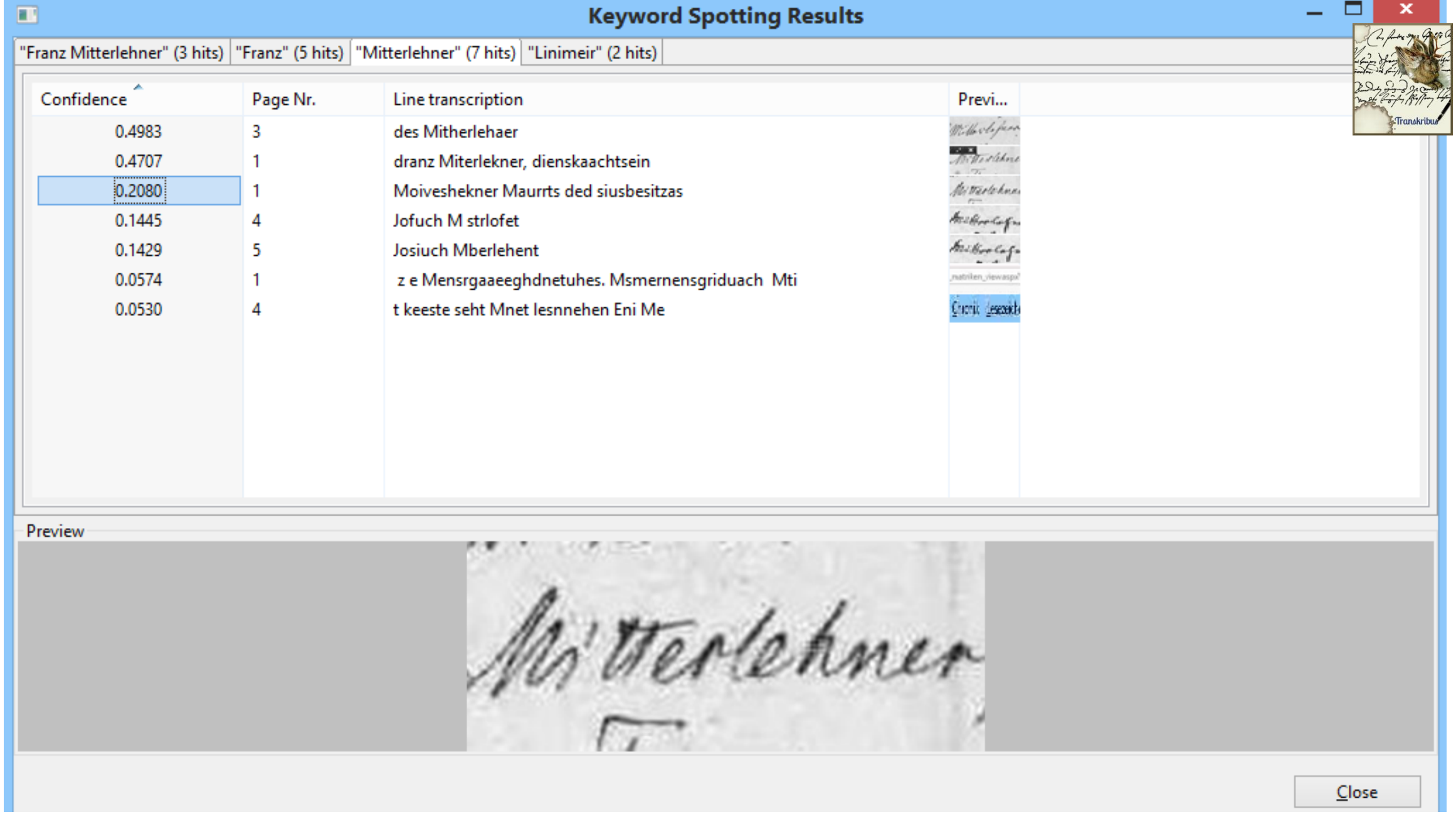
KB nationale bibliotheek

KWS



- Zoekt gedeeltelijk visueel maar deels ook transcripties (niet geheel duidelijk wat onderliggende techniek is)

KB nationale bibliotheek



KB nationale bibliotheek

KWS Interface

- Amsterdam:
- Finland: https://transkribus.eu/r/kws/

Woordenboek

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Woordenboek (dictionary)

- Eigenlijk geen woordenboek, maar een woordenlijst!
- Transkribus heeft de optie om een woordenboek te gebruiken. Dit verlaagt de WER (word error rate) maar kan het proces ook enorm vertragen.
  - Aanleveren aan Innsbruck.
  - Elk woord op een regel
  - Spellingvariatie *kan* een probleem zijn (dat wordt gecorrigeerd)

Interesse: neem contact op met het READ-team (email@transkribus.eu)

Text Recognition Configuration

Search: Filter | All | Technology: Show all

Name	Language	Curator	Technology	Created
French_18thC_Print	French	info@caromei...	CITlab HTR+	05.12.19
Dutch_Gothic_Print	Dutch (16th...	info@caromei...	CITlab HTR+	28.11.19
Noscemus GM v1	Latin (partl...	stefan.zatham...	CITlab HTR+	22.11.19
NAN/NHA_GT_M3+	Dutch	vincent.noppe...	CITlab HTR+	23.08.19
Dutch Notarial Model 18th ...	Dutch	jirsireinders19...	CITlab HTR+	17.07.19
NZZ Gold Standard M1+	German	guenter	CITlab HTR+	23.04.19
Combined_Full_VKS_2	Church Slav...	achim.rabus@...	CITlab HTR+	14.04.19
HIMANIS Chancery M1+	Latin, French	guenter.hackl...	CITlab HTR+	14.04.19
ONB_Newseye_GT_M1+	german	guenter.hackl...	CITlab HTR+	15.02.19
German Kurrent M1+	German	guenter	CITlab HTR+	31.01.19
VMC_Test_4+	Russian Ch...	achim.rabus@...	CITlab HTR+	25.01.19
English Writing M1	English	Unknown	CITlab HTR	08.04.16

Details

Name: Dutch Notarial Model 18th Century

Language: Dutch

Description:

This is the first 18th Century general model created by the City Archives of Amsterdam. It is based on thousands of scans from in total 15 different notaries who worked in Amsterdam during the 18th Century.

All notaries (except Van Hoorn and Van Esterwege) have 10 scans validation included (2671 scans training, 130 validation)):
Maten de Jonge (200)
Van Homrigh (127)

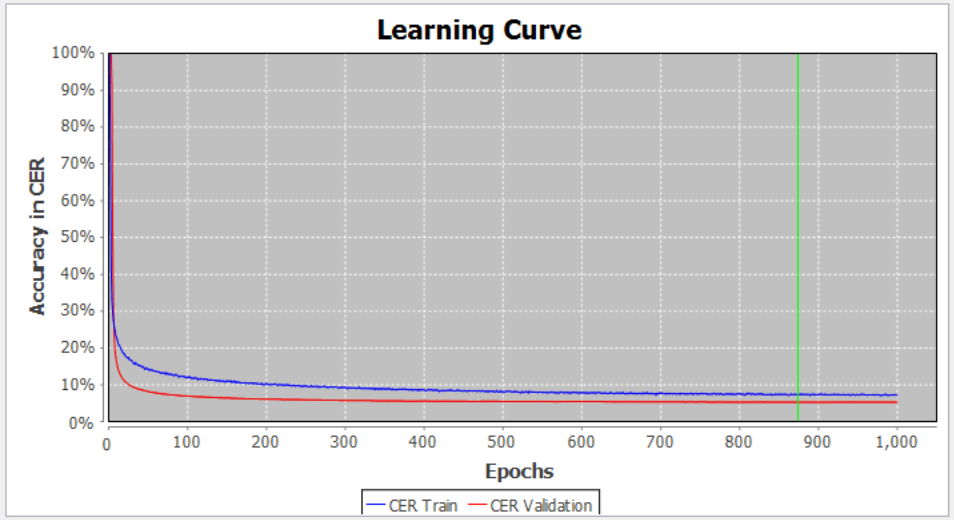
Parameters:

Nr. of Epochs	1000

Nr. of Words: 622904

Nr. of Lines: 116698

Save | Show Train Set | Show Validation Set | Show Characters



CER on Train Set: 7.42%

CER on Validation Set: 5.27%

Dictionary

No dictionary

No dictionary
Dictionary from training data
Language model from training data
Custom dictionary

A28918-M3.dict
A28918-Test.dict
A28918-test2.dict
ABP_19c_4hands_1000e.dict
ABP_7047.dict
ABP_KWS_Test.dict
ABP_M1_(19c.).dict
ABP_M2_(19c.).dict
ABP_M3_(19c.).dict
ABP_OA_0_1.dict
ABP_S_1847-1878_M3.dict
AHB_M2.dict
AHB_M3.dict
AT-HHStA-KK_M1.dict
Ambraser_Heldenbuch.dict
Anton_Reiser_M1.dict
Archivio_Ricordi_M1.dict
Archivio_Ricordi_M2.dict
Barlach_M1.dict
Becket_V2.dict
Binder_Kochbuch.dict
Binder_Kochbuch_M2.dict
Bozen+Konzilsprotokolle.dict
Bozen_Ratsprotokolle_M2.dict
Bozen_v1.dict
Brandshagen_4.dict
CCeH_Itinera-Nova.dict
Church_Slavonic_Dict_sup_lav_ser.d...
Church_Slavonic_VMC_Apostolos_F...
Church_Slavonic_VMC_Apostolos_...
Cochin_Court_Records_M1.dict

OK | Cancel

Text2Image

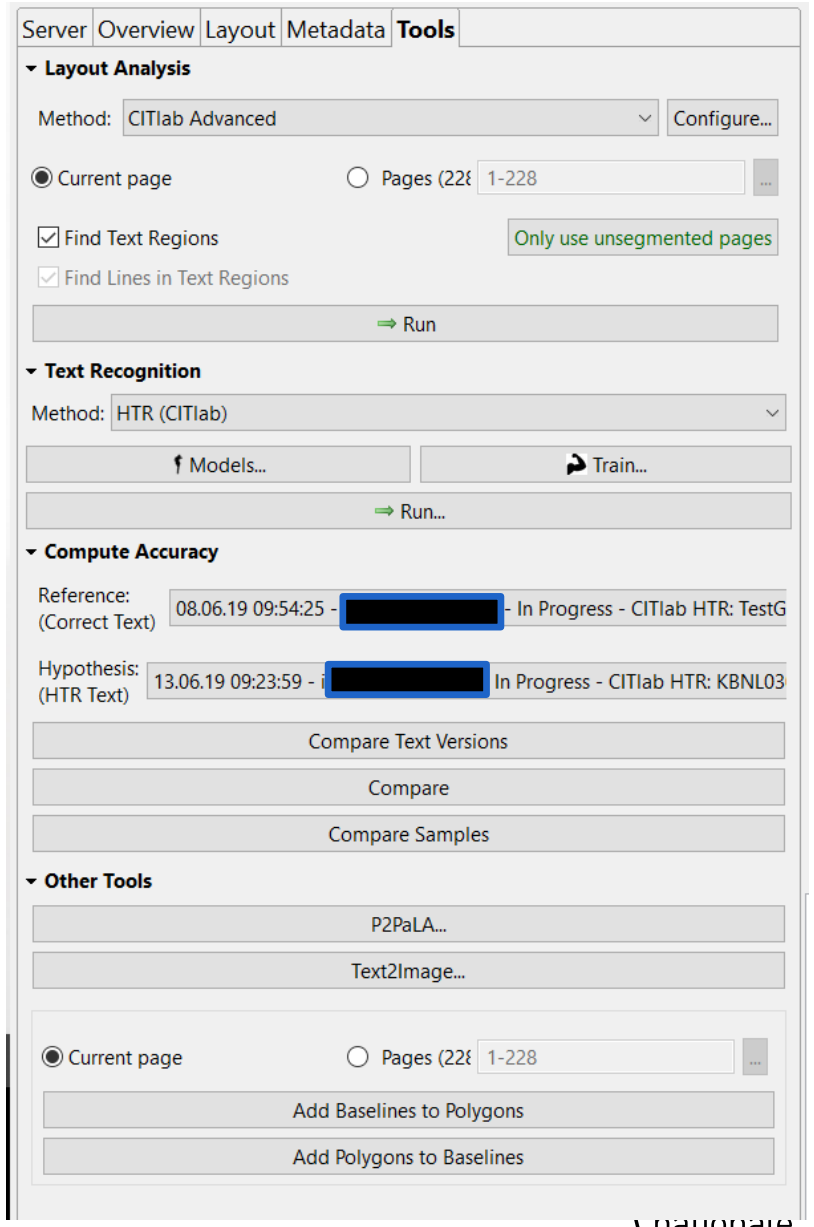
#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Text2Image

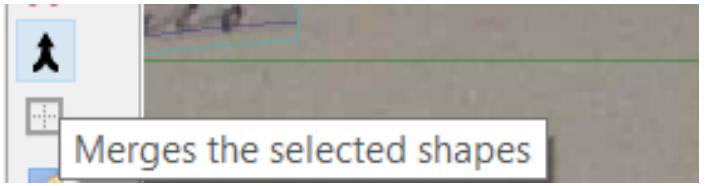
- Dit kan je gebruiken wanneer je al transcripties beschikbaar hebt.
- Er zijn drie manieren om dit aan te pakken.

- **Let op:** verwijder uit je transcriptie alle tekens waaruit blijkt dat je niet zeker was van een karakter (dus de vraagtekens etc).



Stap voor stap T2I (hoge accuratesse)

- Creëer een Layout Analyse voor je pagina (handmatig/automatisch)
- Controleer of de LA goed is, loop de regels even door.
  - Indien een regel gesplitst is (dus baseline is in twee delen)
    selecteer beide en klik op de “Merge” knop op de *Edit Toolbar*.
- Sla je LA op.



Methode 1. Transcripties met regeleindes (veel werk, hoge accuratesse)

- Kopieer je transcriptie van de hele pagina naar de eerste regel in het transcriptieveld.
- **Druk “ctrl+enter”** na de eerste transcriptieregel (zoals in de afbeelding); alles verplaatst automatisch naar de volgende regel (alleen de laatste regel moet je handmatig kopiëren en plakken…).

- NB. Als je iets te vaak enter hebt gedaan (2x ‘ctrl+enter’) dan kan je op **delete klikken MITS de bovenliggende regel leeg is** (anders functioneert ‘delete’ niet).

Methode 2.

Filmpje van methode 2,
met dank aan Steve Jackson
Rijksarchief van Oslo

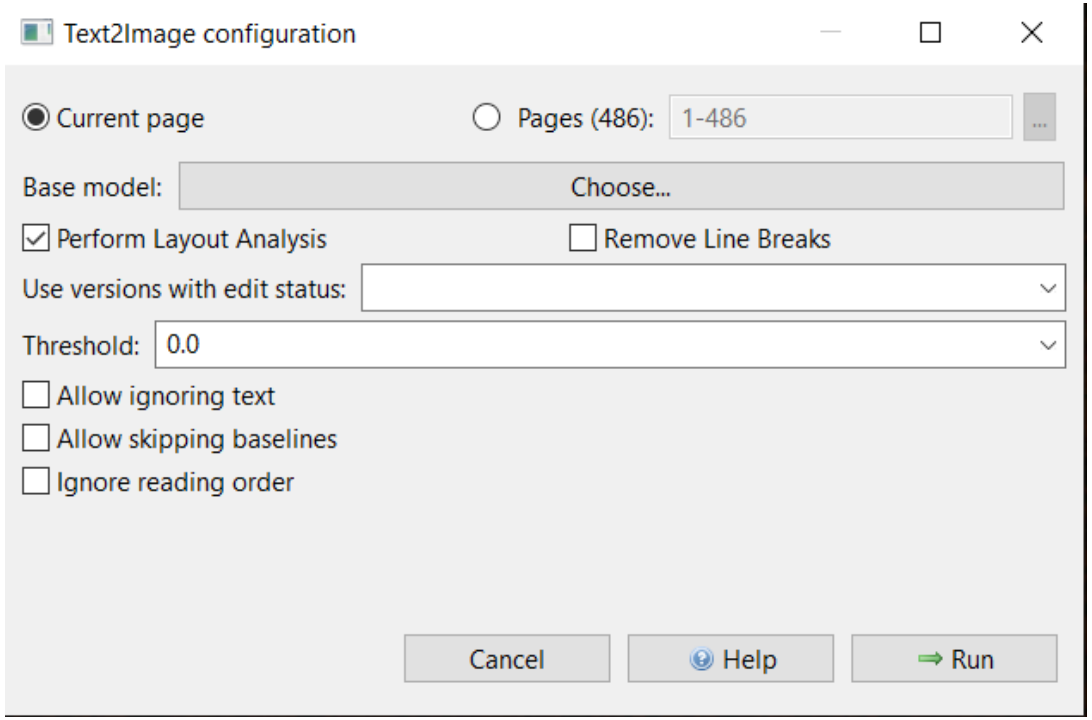
Methode 3: bulk (lage accuratesse, minder werk)

- Maak een map aan genaamd ‘txt’ in dezelfde map waar alle afbeeldingen staan.
- Maak **per** afbeeldings**pagina** een txt aan waar de tekst in staat.
- Importeer de hele map (foto’s én txt bestanden).
- Je gaat de tekst al direct in het transcriptieveld zien staan, per pagina… maar hélaas er is nog geen LA toegepast (dus tekst is niet verbonden aan regels!) <hier wordt een update over verwacht>

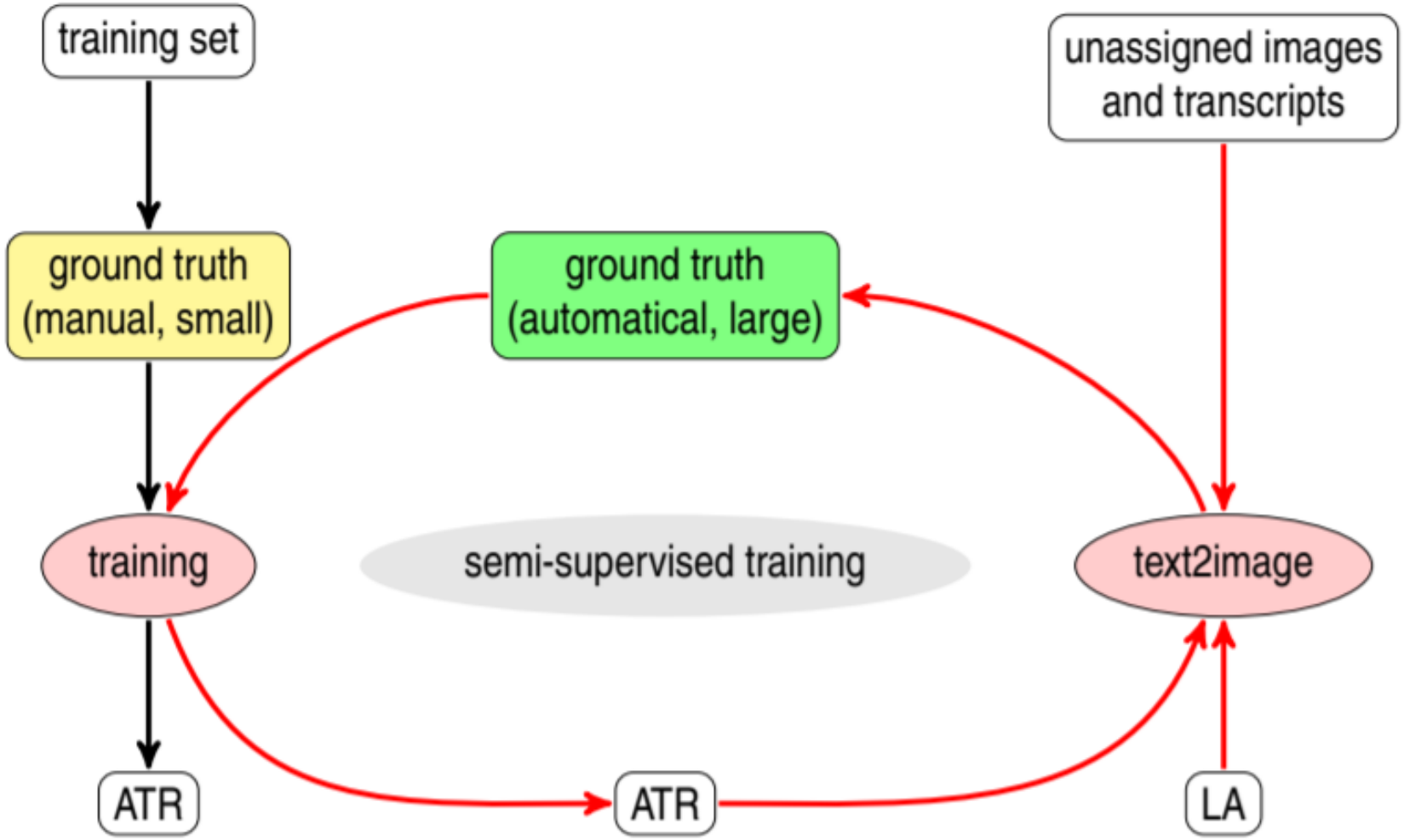
Methode 3.

- Klik op de “Text2Image” knop:
- Kies een model én, klik de vakjes aan:

  Perform LA
  Allow skip words.
  Ignore reading order.
  N.B. Binnenkort komt er ook nog een knopje ‘assume Hyphenations’ ("-", "=", ":“).
- Run



semi-supervised training workflow

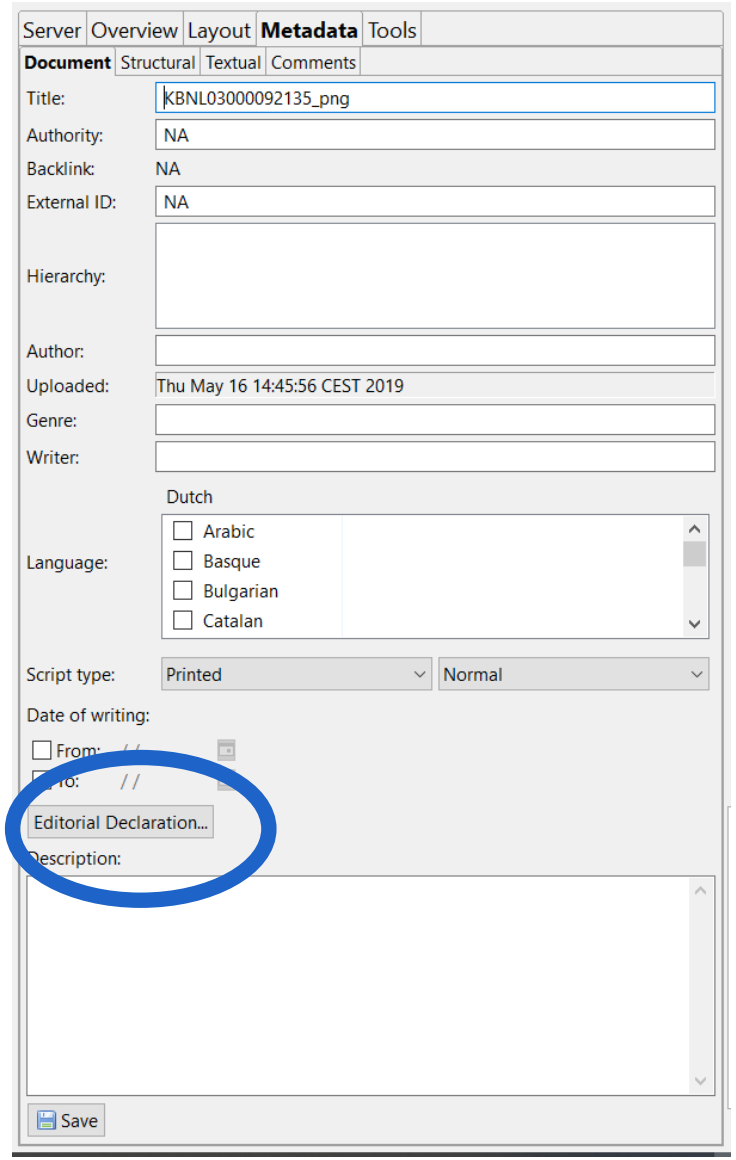


Labelen (structuur & woorden)

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Toevoegen van Metadata aan je document



- Algemene informatie over het bestand waar je mee werkt. <Dit kan je mee-exporteren, in het voorblad>

- Metadata tab → Document.

KB nationale bibliotheek

Editorial Declaration

Transcription Features

ID	Title	Description	Collection
1	Long S	Source uses long "s"	preset
2	u and v	Source uses v for u	preset
3	i and j	Source uses "i" and "j" differently ...	preset
5	Printspace	The printspace indicates the overa...	preset
6	Ligature "sz"	"sz" is set as ligature	preset
28	Text regions	Regions which contain handwritte...	preset
29	Line Regions	Contain the text of line	preset
30	Baselines	The baseline is defined as in Wiki...	preset
47	Omitted text	Even in diplomatic transcriptions t...	preset
48	Person names	Tagging of person names	preset
49	Geo-Names	Tagging of geo-names	preset
50	Abbreviations - ...	Common abbreviations are usuall...	preset
51	Abbreviations	Especially in medieval texts and e...	preset
52	Blackening	Sensible text can be marked as "b...	preset

Transription Options

ID	Text
70	Blackening was not applied
69	Blackening was applied to names of perso...

Selected Features

ID	Title	Description	S...
2	u and v	Source uses...	Tr...
3	i and j	Source uses...	Tr...
28	Text regions	Regions wh...	A...
29	Line Regions	Contain the...	A...
30	Baselines	The baselin...	A...
47	Omitted text	Even in dipl...	N...
48	Person nam...	Tagging of ...	P...
49	Geo-Names	Tagging of ...	G...
50	Abbreviati...	Common a...	C...
51	Abbreviati...	Especially i...	A...
52	Blackening	Sensible te...	B...

Copy to document:

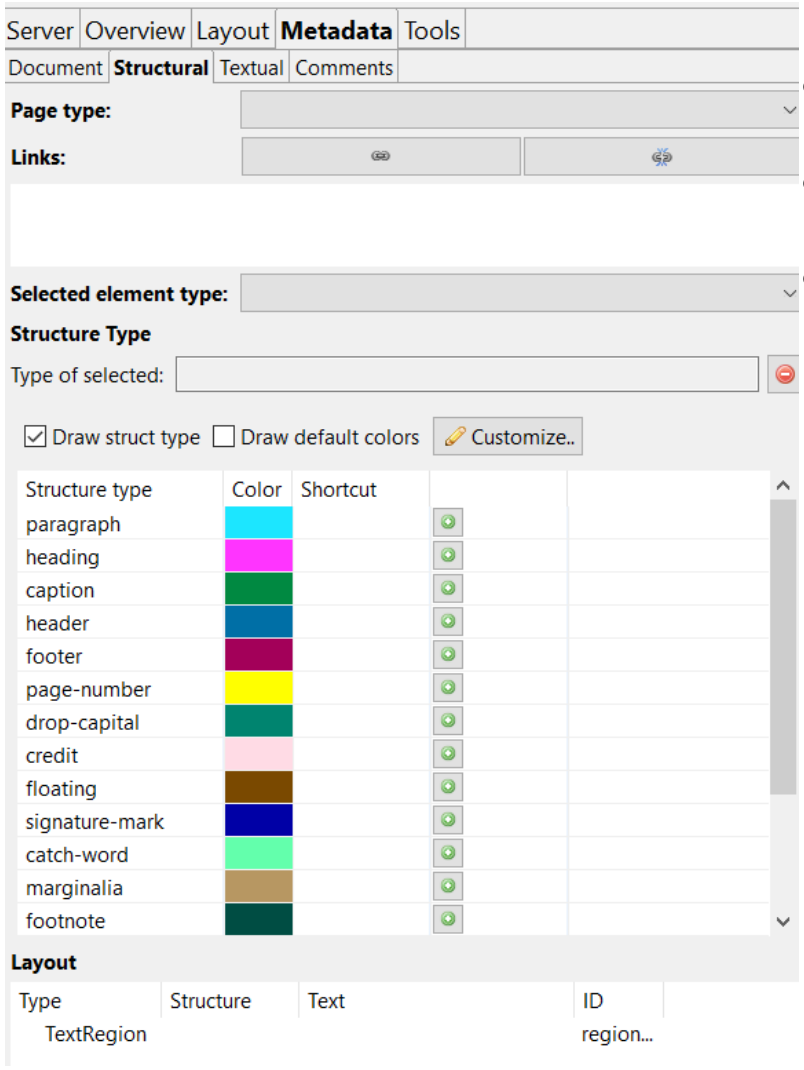
1-26 / 112 1 5

ID	Title	Pages	Uploader	Uploaded
272...	TRAINING_TESTSET_DutchGot...	12	info@caro...	Thu Nov 28...
269...	TRAINING_TESTSET_EarlyMod...	12	info@caro...	Tue Nov 26...
268...	TRAINING_TESTSET_gothisch1...	1	info@caro...	Mon Nov 2...
220...	UBU000005064_png_duplicaat	486	info@caro...	Tue Oct 01 ...
220...	168882_dupl	486	info@caro...	Tue Oct 01 ...
178...	UBL000045368_png	852	sara.veldho...	Tue Jul 02 1...
177...	UU15755	1514	sara.veldho...	Fri Jun 28 1...
175...	UBL000046178_png	1162	sara.veldho...	Fri Jun 21 1...
175...	UBA000066806_png	220	sara.veldho...	Thu Jun 20 ...
175...	GENT900000065568_png	264	sara.veldho...	Thu Jun 20 ...
175...	UBL000045369_png	748	sara.veldho...	Thu Jun 20 ...
175...	KBNL B410014860_png	434	sara.veldho...	Thu Jun 20 ...

25 Filter

KB nationale bibliotheek

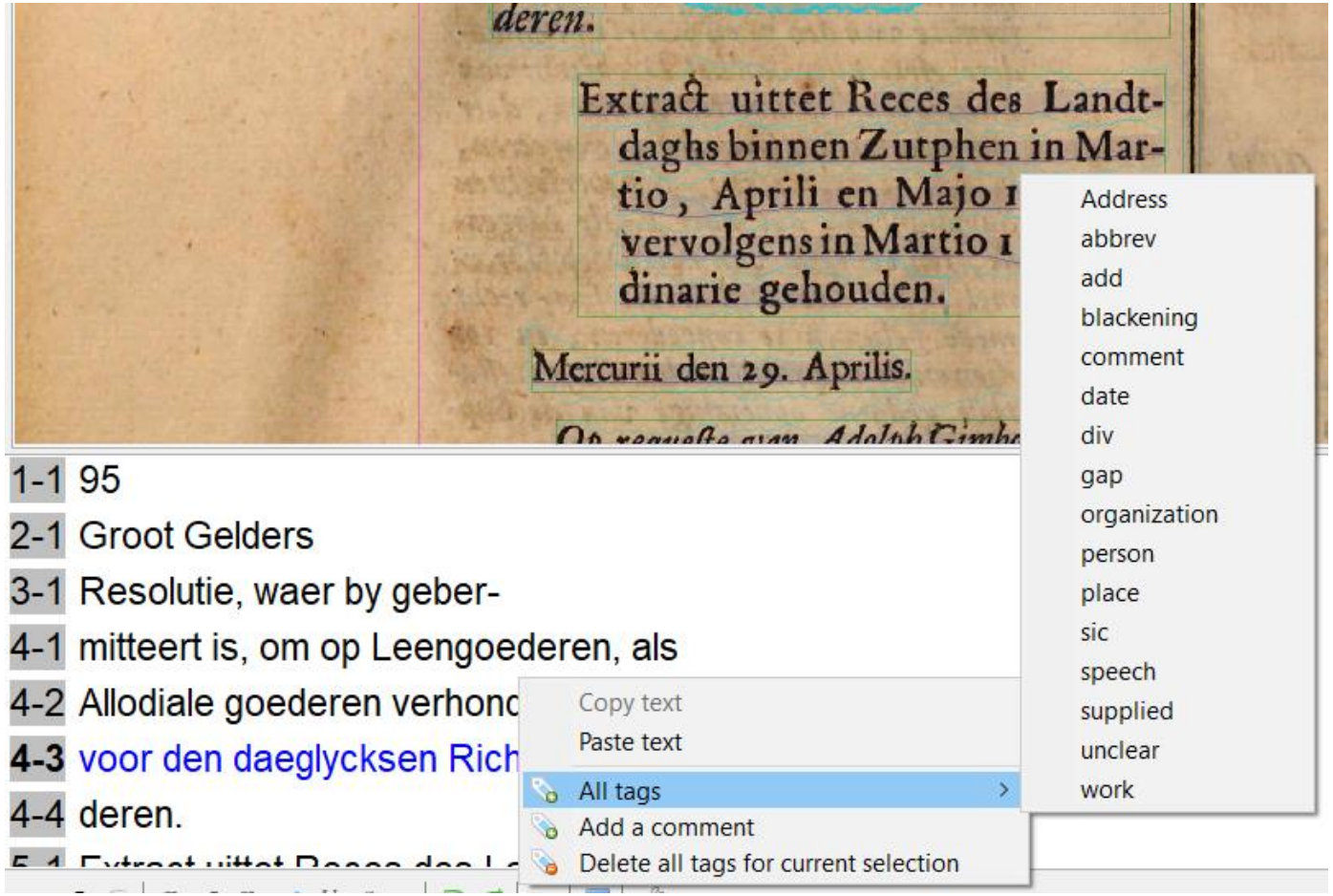
Lay-out/ Structuur Metadata



- T.b.v. het benoemen van structuur in je tekst.
- Te vinden bij: Metadata tab → Structural.
- Waarom:
  - Dit kan nuttig/ handig zijn om je XML-output te herkennen. (Zoeken op titels, paragraaf, paginanummers, etc.)
  - Wanneer je structuur wilt trainen, bijv. als voorbereiding op P2PaLA of NLE Document Understanding (bèta).

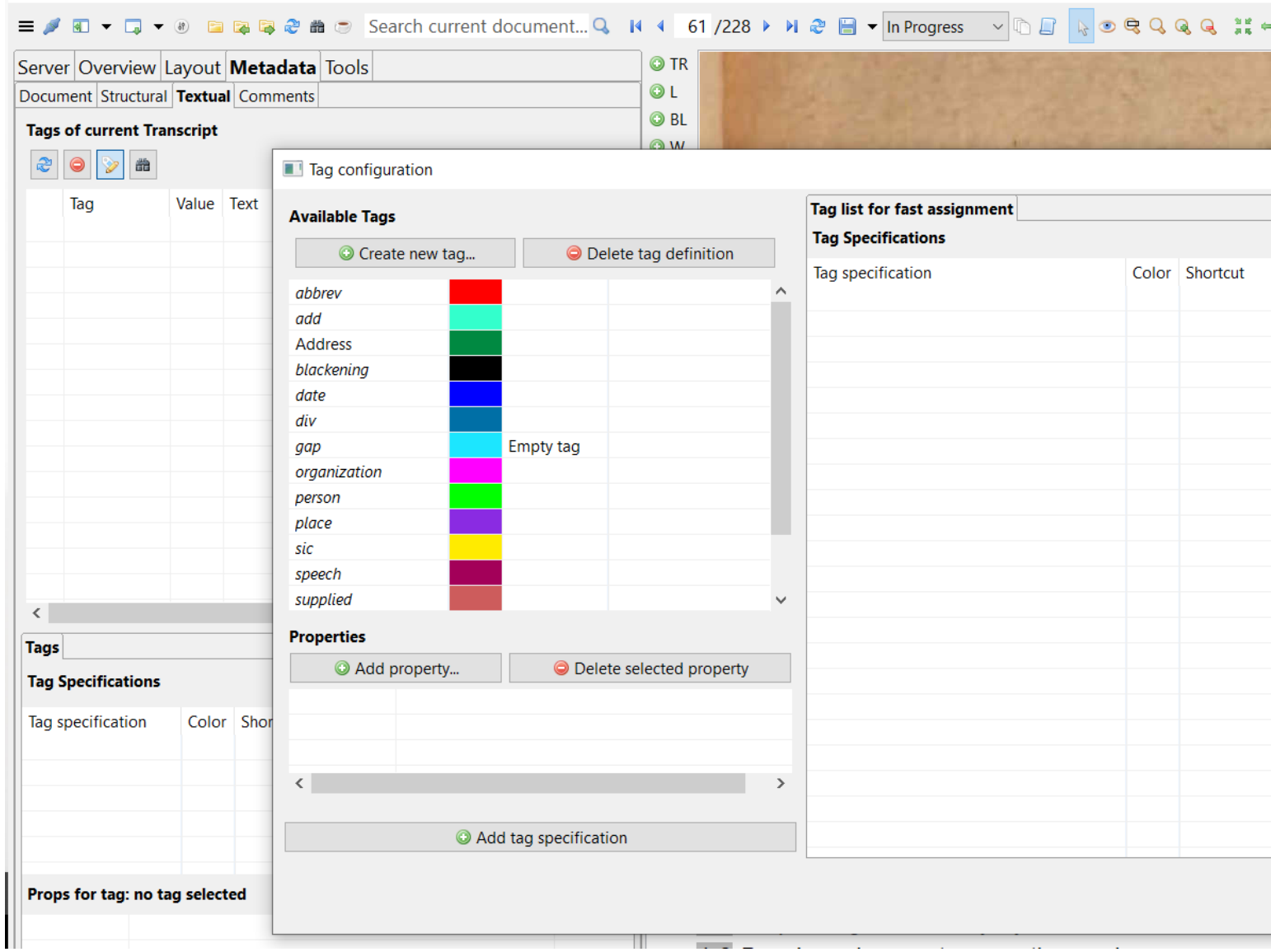
I. Tekstuele tags

Wanneer je in het transcriptieveld met de rechtermuisknop klikt, krijg je een uitschuifmenu. Hier kan je informatie toevoegen t.b.v. de tekst. Bijv. afkortingen.



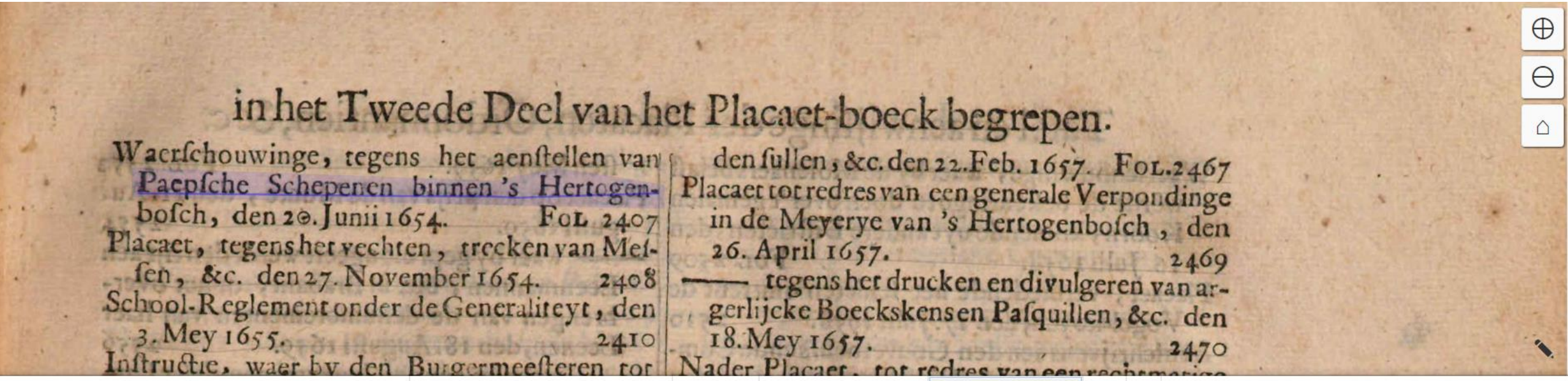
II. Tekstuele tags

- Specifieke woorden (namen, adressen, data, plaatsen, etc.)
- Metadata tab → Textual
- Dit kan handig zijn voor het zoeken van bepaalde woorden.
- XML-export.
- Het is handmatig – wordt nog niet herkend in een model!



KB nationale bibliotheek

Web interface Transkribus

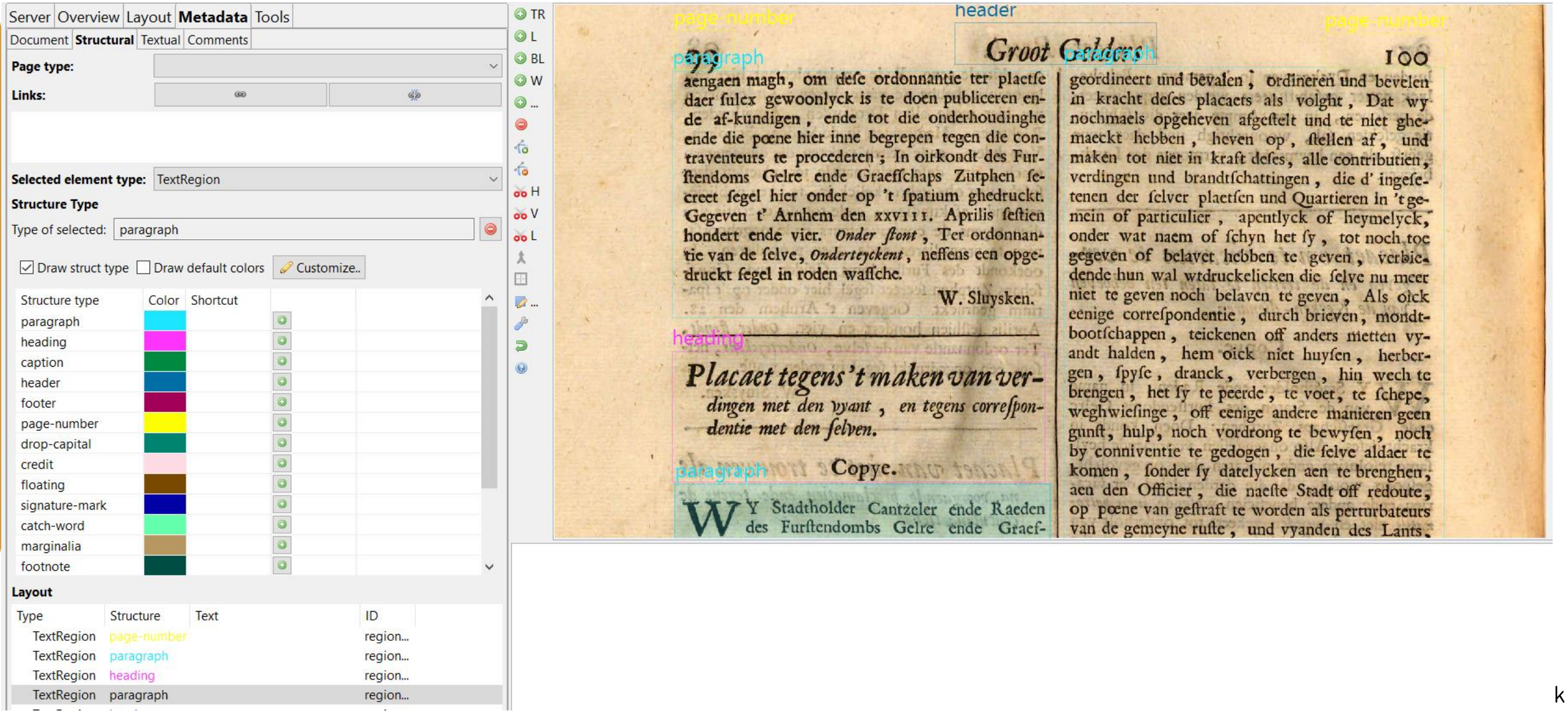


Layout-analyse trainen P2PaLA

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

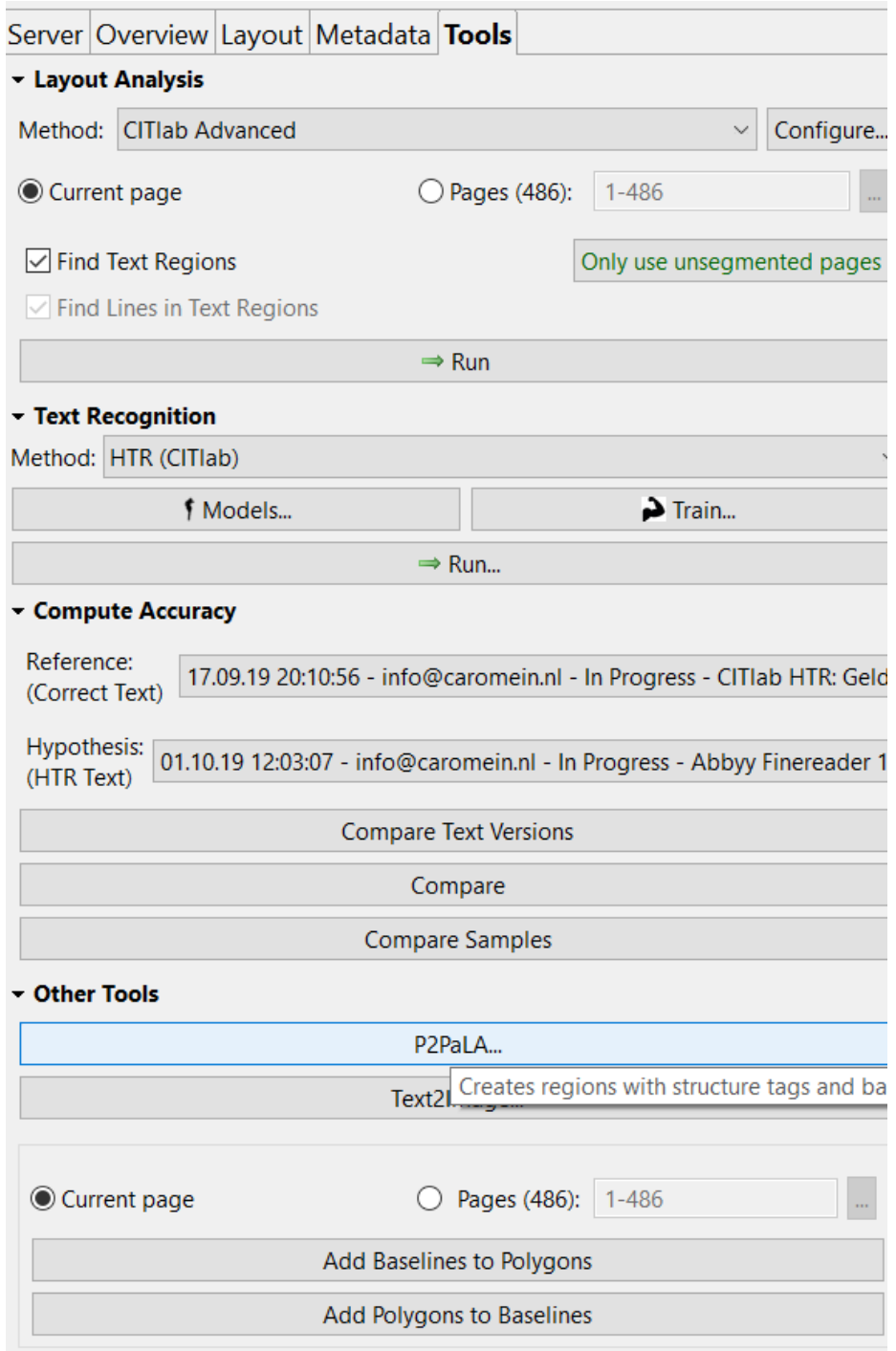
Geef je tekst structuur-tags (dan pas BL)

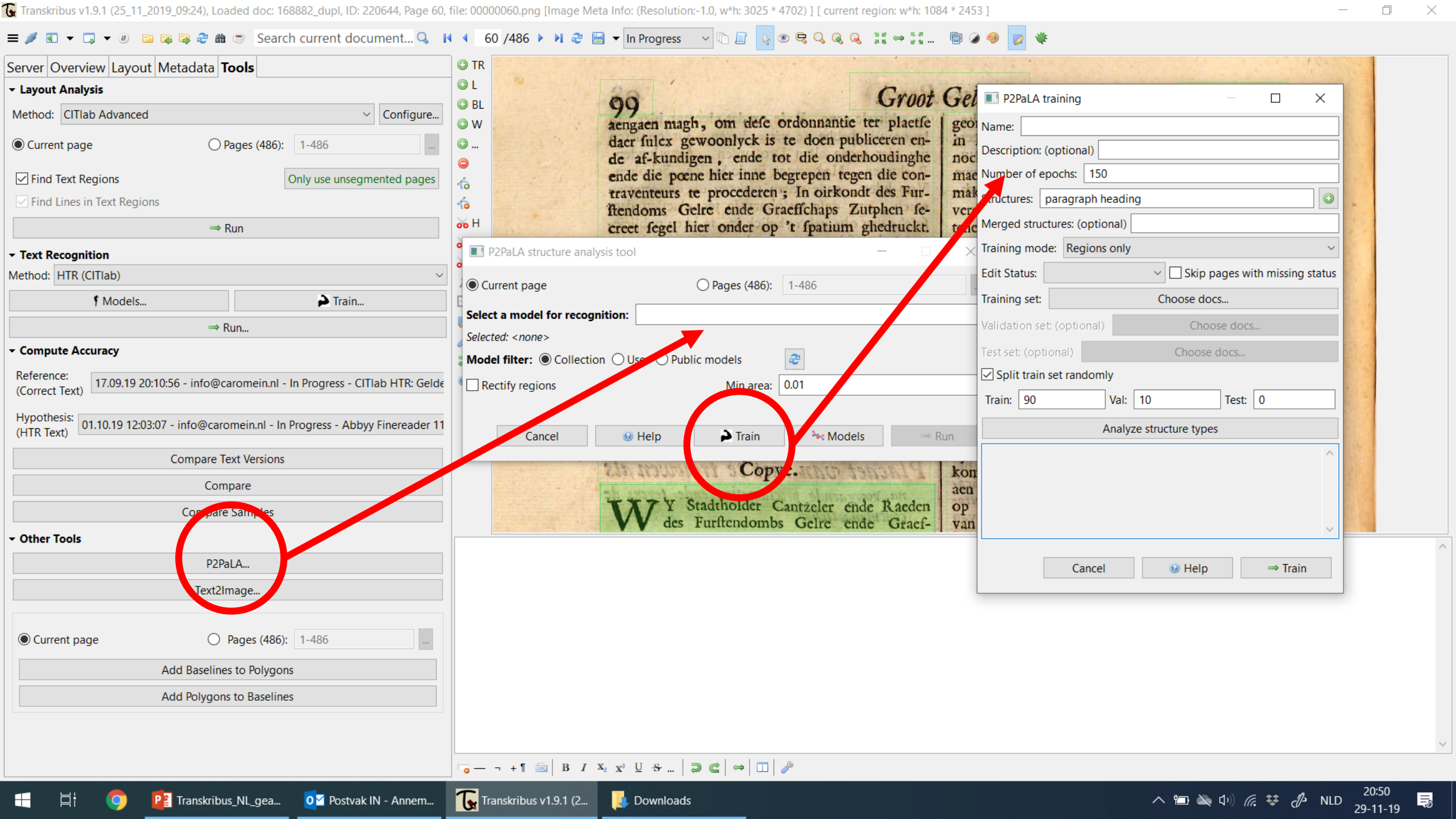


k

P2PaLA

- Je hebt dus een groot aantal pagina’s nodig die al structuurtags hebben.



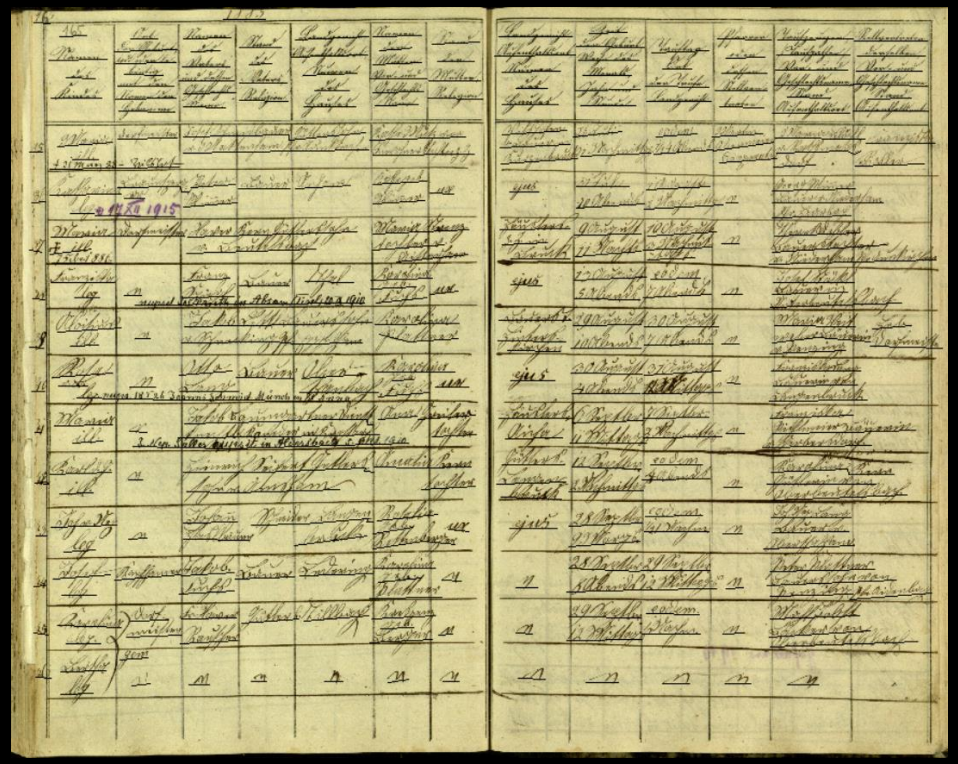

Transkribus v1.9.1 (25_11_2019_09:24), Loaded doc: 168882_dupl, ID: 220644, Page 60, file: 00000060.png [Image Meta Info: (Resolution:-1.0, w*h: 3025 * 4702)] [current region: w*h: 1084 * 2453]
Search current document...
60 /486
In Progress
Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method: CITlab Advanced
Configure...
Current page
Pages (486): 1-486
Find Text Regions
Only use unsegmented pages
Find Lines in Text Regions
Run
Text Recognition
Method: HTR (CITlab)
Models...
Train...
Run...
Compute Accuracy
Reference: (Correct Text)
17.09.19 20:10:56 - info@caromein.nl - In Progress - CITlab HTR: Gelde
Hypothesis: (HTR Text)
01.10.19 12:03:07 - info@caromein.nl - In Progress - Abbyy Finereader 11
Compare Text Versions
Compare
Other Tools
P2PaLA...
Text2Image...
Current page
Pages (486): 1-486
Add Baselines to Polygons
Add Polygons to Baselines
TR
L
BL
W
H
99
Groot Gel
aengaen magh, om dese ordonnantie ter plaetse
daer sulex gewoonlyck is te doen publiceren en-
de af-kundigen, ende tot die onderhoudinghe
ende die pœne hier inne begrepen tegen die con-
traventeurs te procederen; In oirkondt des Fur-
stendoms Gelre ende Graefschaps Zutphen se-
creet segel hier onder op 't spatium ghedruckt.
P2PaLA structure analysis tool
Current page
Pages (486): 1-486
Select a model for recognition:
Selected: <none>
Model filter: Collection Use Public models
Rectify regions
Min area: 0.01
Cancel
Help
Train
Models
Run
P2PaLA training
Name:
Description: (optional)
Number of epochs: 150
Structures: paragraph heading
Merged structures: (optional)
Training mode: Regions only
Edit Status:
Skip pages with missing status
Training set:
Choose docs...
Validation set: (optional)
Choose docs...
Test set: (optional)
Choose docs...
Split train set randomly
Train: 90
Val: 10
Test: 0
Analyze structure types
Cancel
Help
Train
WY Stadtholder Cantzeler ende Raeden
des Furstendombs Gelre ende Graef-
Transkribus_NL_gea...
Postvak IN - Annem...
Transkribus v1.9.1 (2...
Downloads
NLD
20:50
29-11-19

Layout-analyse trainen

- Tabellen (in ontwikkeling)
- NLE Document Understanding

#Transkribus
#webinargeavanceerdNL
@naverlabseurope

KB nationale bibliotheek



→ Information extraction

NLE Document Understanding – Naverlab Europe

Hervé Dejean
Jean-Luc Meunier

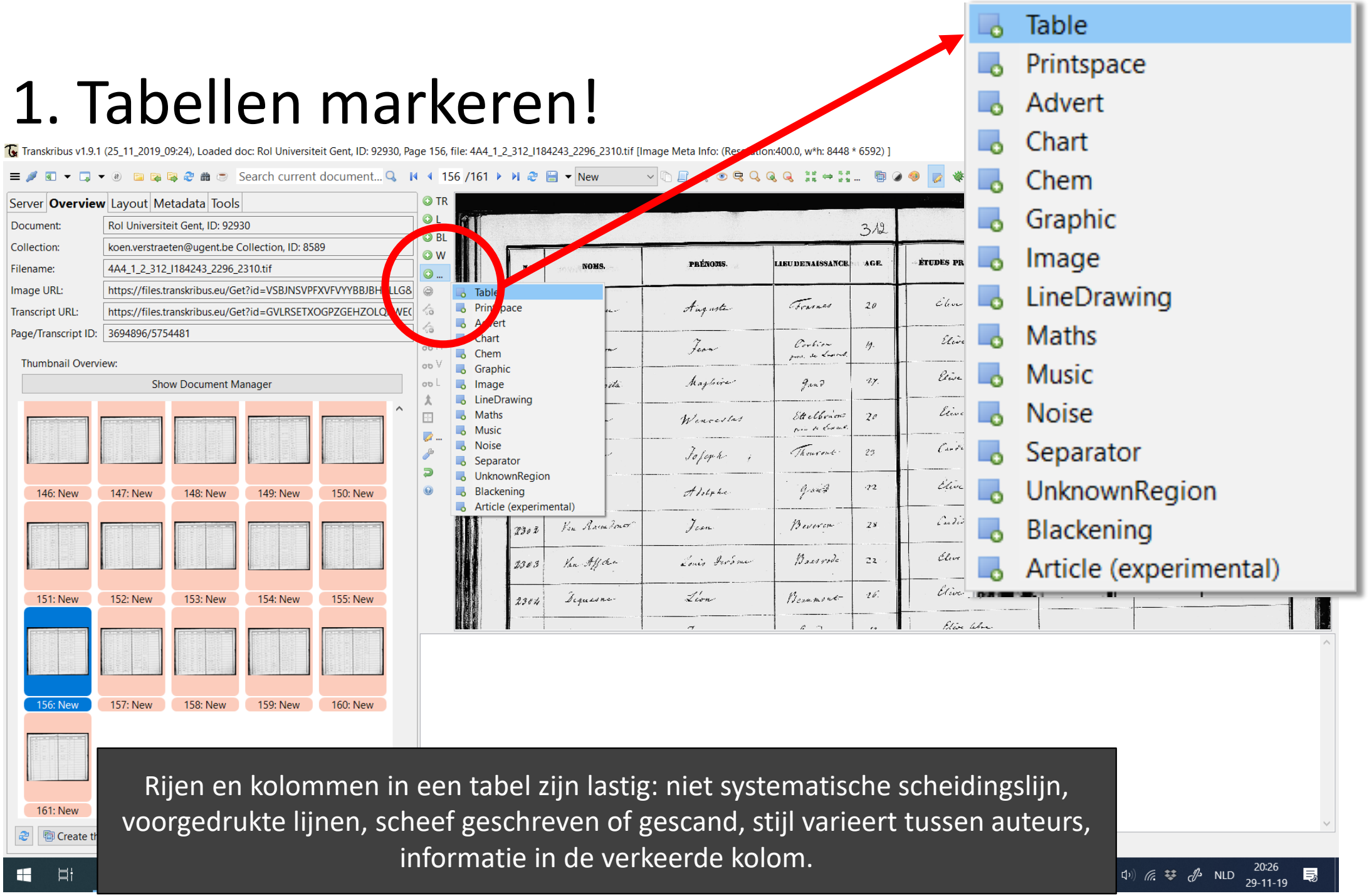
BISTUM PASSAU ARCHIV

CVL Computer Vision Lab

LABS NAVER LABS EUROPE

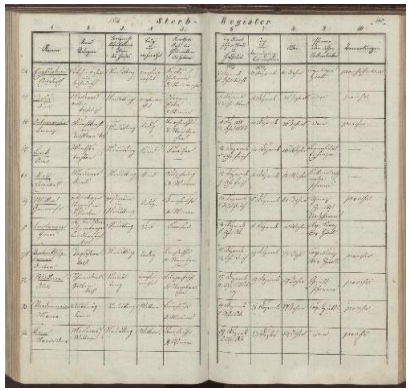
KB nationale bibliotheek

Stap 1. Tabellen markeren!



Rijen en kolommen in een tabel zijn lastig: niet systematische scheidingslijn, voorgedrukte lijnen, scheef geschreven of gescand, stijl varieert tussen auteurs, informatie in de verkeerde kolom.

Globale workflow Naverlab's Document Understanding



Data collection / Ground truth

Text line detection (URO)

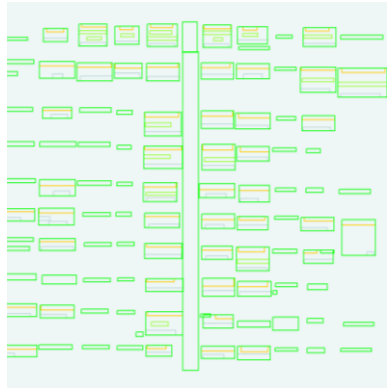
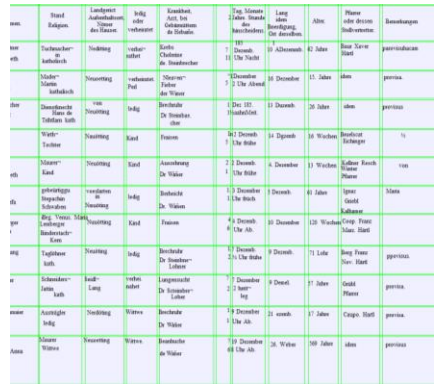


Table Understanding (NLE)



Handwritten Text Recognition (URO)

<RECORD lastname="Boxleitner" firstname="Elisabeth" occupation="Tuchmacher¬" location="Neuötting" situation="verhei¬" deathreason="Lungenlähme de" age="67" />
<RECORD lastname="Kastl" firstname="Anna" location="Neuoetting" situation="verheiratet" deathreason="Fieber" doktor="der" deathDate="Dezember" age="15"/>
<RECORD lastname="Seternicher" firstname="Lorenz" occupation="Dienstknecht" religion="kath" location="Neuötting" situation="ledig" deathreason="Brechruhr" doktor="Steimbae" age="26"/>
<RECORD lastname="Denk" firstname="Anna" occupation="Wirth¬" location="Nauötting" situation="Kind" deathreason="Fraisen" doktor="In" age="16"/>
<RECORD firstname="Elisabeth" occupation="Maurer¬" location="Neuötting" situation="Kind" deathreason="Auszehrung" doktor="Wimer" age="13"/>
<RECORD lastname="Müller" firstname="Genovefa" location="Nmuötting" situation="ledig" doktor="Wimen" deathDate="Dezember" age="61"/>
<RECORD lastname="Lemberger" firstname="Theres" location="Nuuötting" situation="Kind" deathreason="Fraisen" age="120"/>
<RECORD lastname="Sammer" firstname="Anton" occupation="Taglöhner" religion="kath" location="Neuöting" situation="ledig" deathreason="Brechruhr" doktor="Steinbne¬"/>
<RECORD lastname="Wallner" firstname="Aña" occupation="Schneiders¬ Jattin" religion="kath" location="Lang" situation="verhei" deathreason="Lungensucht" doktor="Scteimbre¬" deathDate="Dezember" age="57"/>
<RECORD lastname="Vordemmaier" firstname="Maria" occupation="Austrägler" location="Nerdöting" deathreason="Brechruhr" doktor="Wimer" deathDate="Dezember" age="17"/>
<RECORD lastname="Kern" firstname="Maria Anna" occupation="Maurer Wittwe" location="Neuoetting" doktor="Wimer" deathDate="Dezember" age="569"/>
</PAGE>

Information Extraction (NLE)

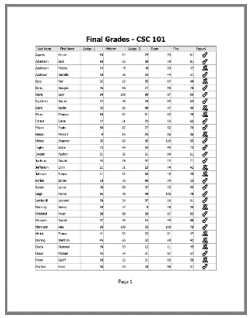
NAVER LABS EUROPE

Proces van Document Understanding

Begrip voor de document layout (organisatie) zodat Information Extraction kan plaatsvinden

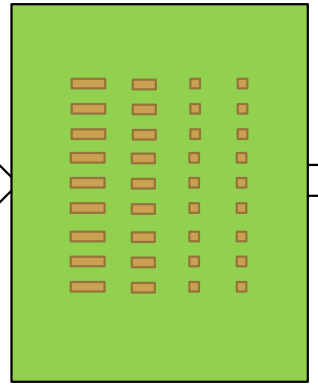
OCR/HTR

Raw text

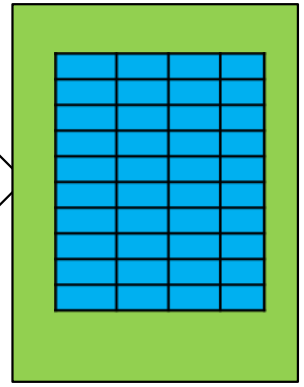


Our Input

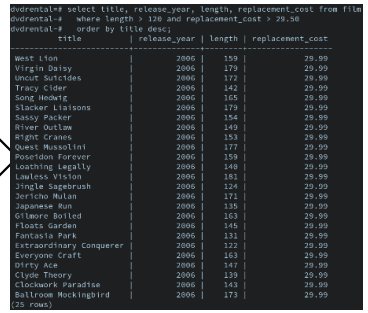
Text, position



Layout Structure

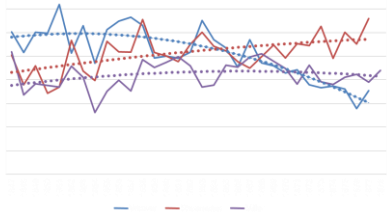


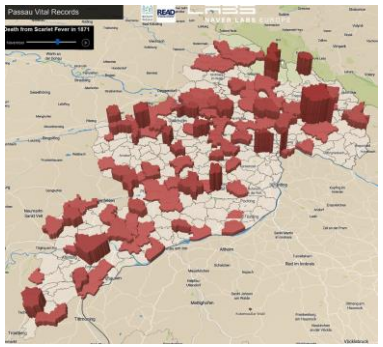
Information (database)



```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental-#   where length > 120 and replacement_cost > 29.50
dvdrental-#   order by title desc;
          title          | release_year | length | replacement_cost
-------------------------+--------------+--------+------------------
 West Lion               |         2006 |    159 |            29.99
 Virgin Daisy            |         2006 |    179 |            29.99
 Uncut Suicides          |         2006 |    172 |            29.99
 Tracy Cider             |         2006 |    142 |            29.99
 Song Hedwig             |         2006 |    165 |            29.99
 Slacker Liaisons        |         2006 |    179 |            29.99
 Sassy Packer            |         2006 |    154 |            29.99
 River Outlaw            |         2006 |    149 |            29.99
 Right Cranes            |         2006 |    153 |            29.99
 Quest Mussolini         |         2006 |    177 |            29.99
 Poseidon Forever        |         2006 |    159 |            29.99
 Loathing Legally        |         2006 |    148 |            29.99
 Lawless Vision          |         2006 |    181 |            29.99
 Jingle Sagebrush        |         2006 |    124 |            29.99
 Jericho Mulan           |         2006 |    171 |            29.99
 Japanese Run            |         2006 |    135 |            29.99
 Gilmore Boiled          |         2006 |    163 |            29.99
 Floats Garden           |         2006 |    145 |            29.99
 Fantasia Park           |         2006 |    131 |            29.99
 Extraordinary Conquerer |         2006 |    122 |            29.99
 Everyone Craft          |         2006 |    163 |            29.99
 Dirty Ace               |         2006 |    147 |            29.99
 Clyde Theory            |         2006 |    139 |            29.99
 Clockwork Paradise      |         2006 |    143 |            29.99
 Ballroom Mockingbird    |         2006 |    173 |            29.99
(25 rows)
```

Analysis

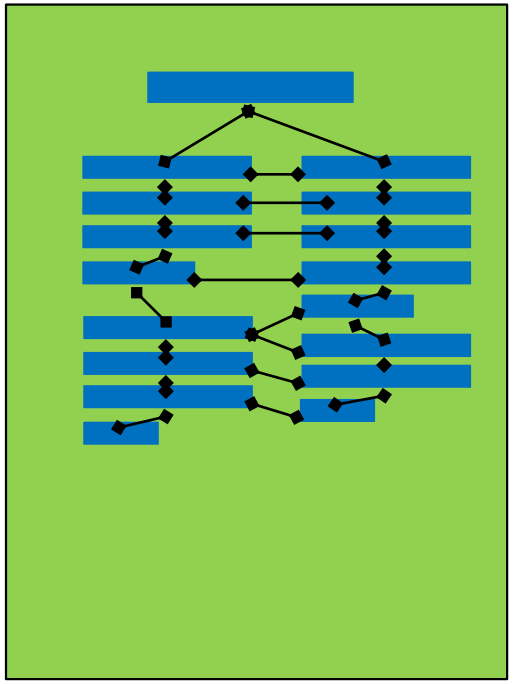




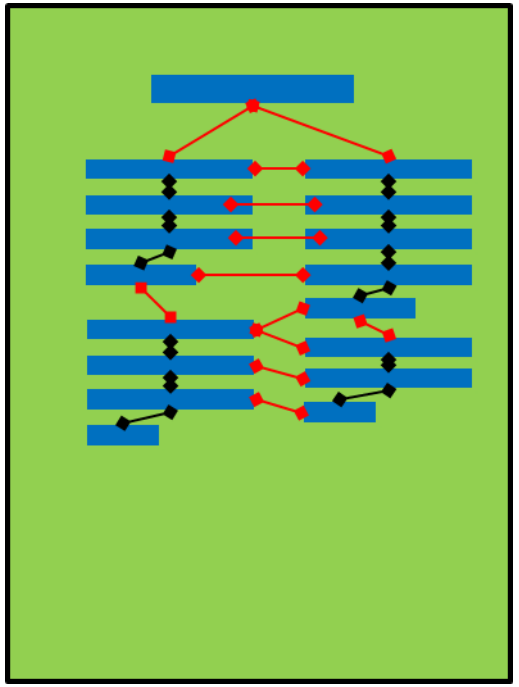
KB nationale bibliotheek

Aanpak: één tool voor Document Layout/Understanding

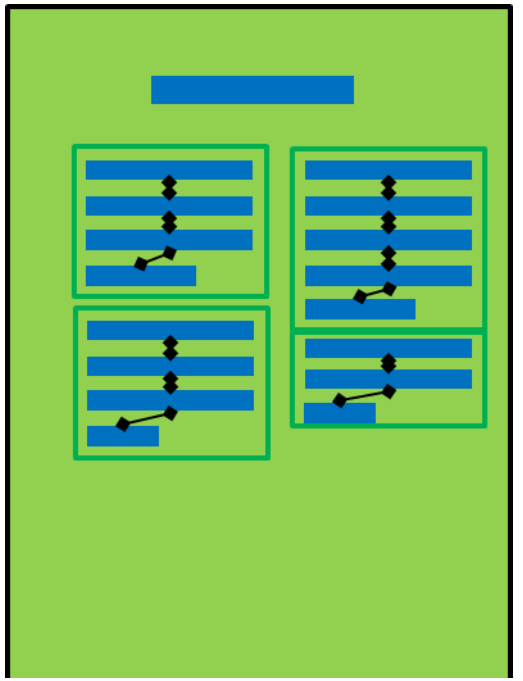
Graph Creation
- Nodes: textlines
- Edges: positional relations



Page Layout Segmentation
- Edge classification



Document Reconstruction



NAVER LABS EUROPE

Table Understanding (Cells)

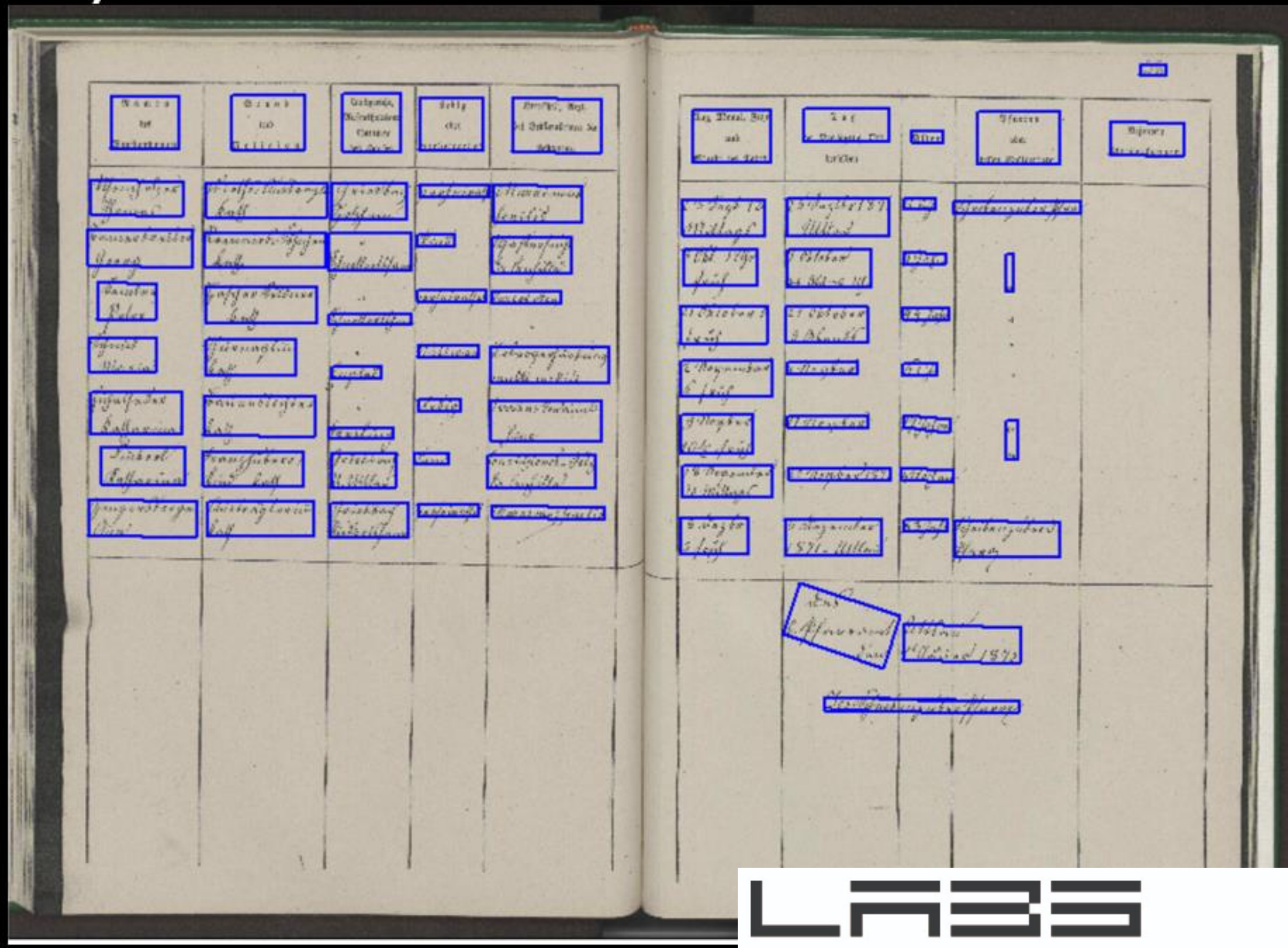
Training size : 990 pages
Test size : 48 pages
Training time : 3h

EVALUATION AT **EDGE** LEVEL

	Precision	Recall	F-1
keep	0.970	0.985	0.978
remove	0.996	0.991	0.994

EVALUATION AT **CLUSTER** LEVEL

	Precision	Recall	F-1
@80	0.957	0.953	0.955
@100	0.957	0.953	0.955



NAVER LABS EUROPE

Table Understanding (Rows)

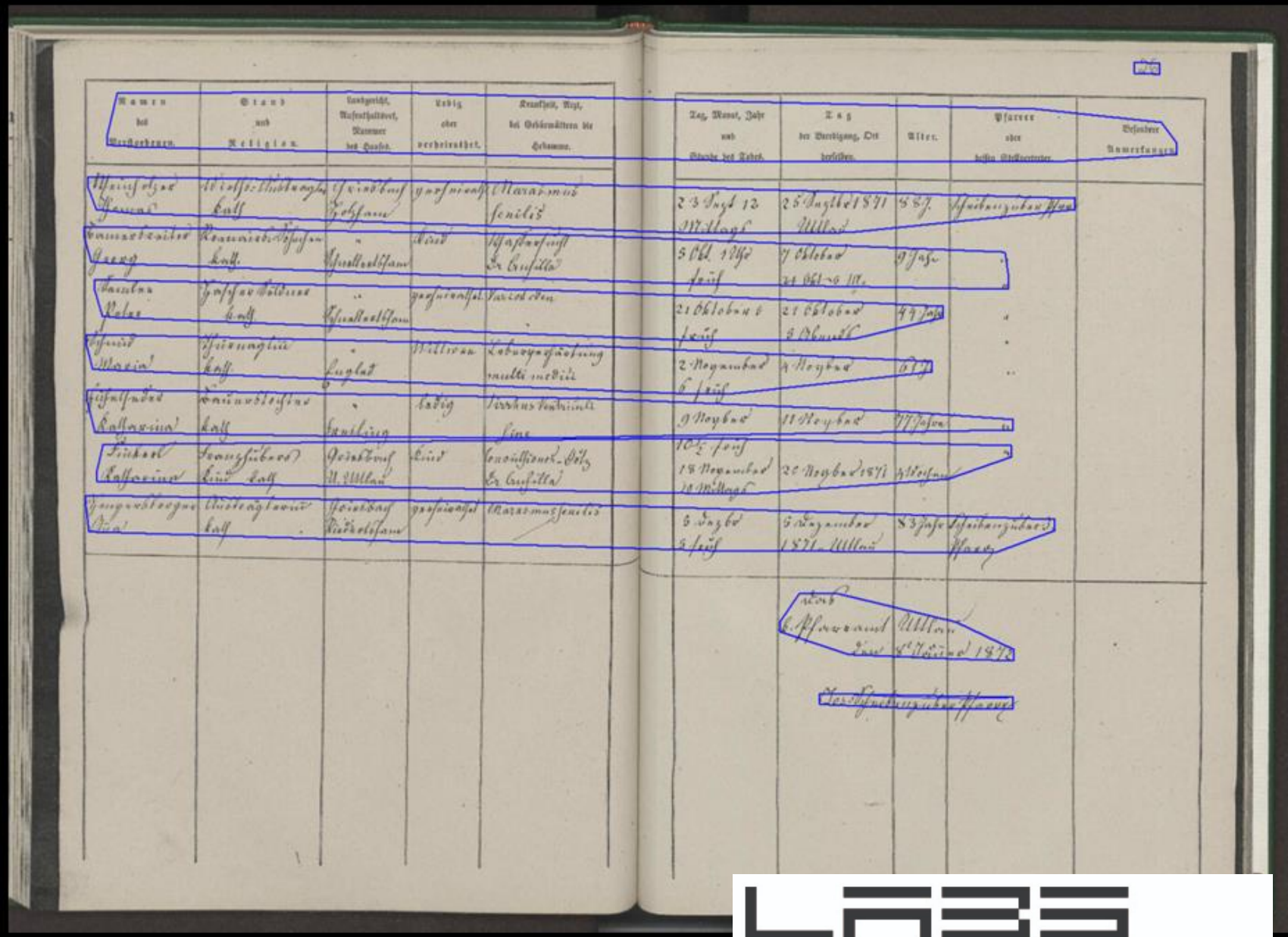
Training size : 990 pages
Test size : 48 pages
Training time : 3h

EVALUATION AT **EDGE** LEVEL

	Precision	Recall	F-1
Keep	0.986	0.992	0.989
Remove	0.982	0.969	0.976

EVALUATION AT **CLUSTER** LEVEL

	Precision	Recall	F-1
@80	0.863	0.931	0.896
@100	0.773	0.833	0.802



NAVER LABS EUROPE


NAVER LABS EUROPE


NAVER LABS EUROPE

Pauze / data records bekijken (viewer)

https://datarecords.bekijken.nisvlabs.com/

#Transkribus
#webinargeavanceerdNL

KB nationale bibliotheek

Extra information

READ en Transkribus:
https://read.transkribus.eu/

https://transkribus.eu/

E-Learning:
https://learn.transkribus.eu/

Contact:
email@transkribus.eu

Handleidingen:
https://transkribus.eu/wiki/index.php/How_to_Guides

Youtube Channel en Tutorials:
https://www.youtube.com/channel/UC-txVgM31rDTGlBnH-zpPjA

Facebook:
@Transkribus Users (user based)

Web-Interface:
https://transkribus.eu/r/read/projects/

DocScan en ScanTent:
https://scantent.cvl.tuwien.ac.at/en/

KB nationale bibliotheek

6-7 February 2020
transkribus user conference
in Innsbruck.
https://bit.ly/2QhnjXO

KB nationale bibliotheek

Bedankt voor uw aandacht!

KB nationale bibliotheek