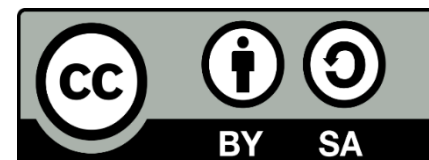


Webinar (Nederlandstalig): TRANSKRIBUS – geavanceerd.



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



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](https://doi.org/10.5281/zenodo.3555097)

Powerpoint: [10.5281/zenodo.3558864](https://doi.org/10.5281/zenodo.3558864)

Transkribus

INHOUD WEBINAR - geavanceerd

Ultra korte samenvatting basiswebinar

Geavanceerd

- Coördinatie 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>


BeNeLux – Coördinatie Transkribus

#Transkribus

#webinargeavanceerdNL

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.



Home Over KIA Agenda Groepen Forum Help Inloggen / Registreren

Kennisnetwerk
Informatie en Archief

#kennisdelen
zit in ons DNA.
Deel je ook mee?

Welkom bij hét Kennisnetwerk Informatie en Archief!

KIA is dé ontmoetingsplaats voor professionals in de archief- en informatiewereld. Word lid van de kennisplatforms, neem deel aan discussies, werk online samen en wissel kennis uit: het kan allemaal! KIA wordt mogelijk gemaakt door het Nationaal Archief, KVAN/BRAIN en Archief 2.0. kia.pleio.nl maakt deel uit van Pleio, hét samenwerkingsplatform voor de publieke zaak.

Heb je vragen? Lees de handleiding [snel aan de slag op \[kia.pleio.nl\]\(https://kia.pleio.nl\)](#), stel je vragen online in de [helpdeskgroep](#) of [mail](#) de beheerder.

Direct naar

- Alle activiteiten >
- Archiefwiki >
- Blogs >
- Discussie >
- Helpdesk >
- Twitterfeed >
- Vacaturebank >

Contactinfo

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

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

“Gadgets”: Scantent & DocScan

#Transkribus

#webinargeavanceerdNL

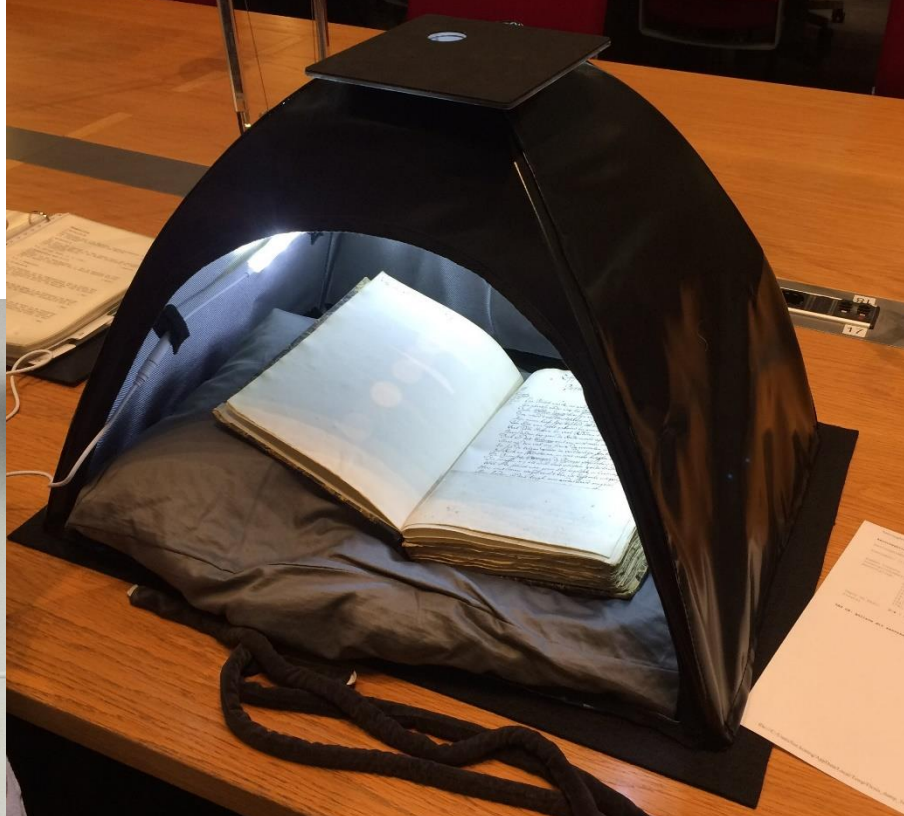


ScanTent And DocScan

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



Scantent



DocScan

The screenshot shows the Google Play Store interface for the DocScan app. At the top, the Google Play logo is on the left, and a search bar with the text 'Zoeken' is on the right. Below the search bar, there are navigation options: 'Categorieën', 'Homepage', 'Populairste items', and 'Nieuwe releases'. A green bar highlights the 'Apps' category. On the left side, a vertical menu lists options: 'Mijn apps', 'Winkelen', 'Games', 'Gezin', 'Keuze van de redactie', 'Account', 'Betaalmethoden', 'Mijn abonnementen', 'Tegoe inwisselen', 'Cadeaukaart kopen', 'Mijn verlanglijstje', 'Mijn Play-activiteit', and 'Gids voor ouders'. The main content area displays the app 'DocScan' by 'HofApps' in the 'Fotografie' category. It has a 4.5-star rating from 20 reviews and a PEGI 3 rating. A green button indicates the app is 'Geïnstalleerd'. Below the app title, there is a video player showing a document being scanned, with a play button overlay. To the right of the video, there are two promotional cards: 'LED LIGHT' and 'THE TENT', both describing features of the app's scanning process.

Google Play Zoeken

Apps Categorieën Homepage Populairste items Nieuwe releases

Mijn apps
Winkelen
Games
Gezin
Keuze van de redactie

Account
Betaalmethoden
Mijn abonnementen
Tegoe inwisselen
Cadeaukaart kopen
Mijn verlanglijstje
Mijn Play-activiteit
Gids voor ouders

DocScan
HofApps Fotografie
★★★★☆ 20
3 PEGI 3
Deze app is geschikt voor al je apparaten.

Geïnstalleerd

LED LIGHT
NON-ABSTRACTIVE LED LIGHT
UNIFORM ILLUMINATION IMPROVES PAGE QUALITY

THE TENT
SEMI-TRANSPARENT SURFACE PROVIDES ILLUMINATION
DIFFUSES AMBIENT LIGHT
EASY TO ASSEMBLE AND PORTABLE

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

#Transkribus

#webinargeavanceerdNL

Volgorde van TekstRegio's aanpassen

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)]

The screenshot shows the Transkribus interface. On the left, there is a table with columns: Type, Text, Structure, Rea..., ID, and Coords. The table lists 7 TextRegio entries. The main area shows a document page with text regions highlighted in green. A small 'Item visibility' dialog is open over the document, showing a list of options with checkboxes. The options are: Show regions (checked), Show lines (checked), Show baselines (checked), Show words (checked), Render blackenings (unchecked), Show regions reading order (checked), Show lines reading order (checked), and Show words reading order (unchecked).

Type	Text	Structure	Rea...	ID	Coords
TextRegio			1	region...	510,375 682,375 682,45
TextRegio			2	region...	503,473 1587,473 1587
TextRegio			3	region...	503,1303 1579,1303 15
TextRegio			4	region...	510,1684 1594,1684 15
TextRegio			5	region...	1318,353 1901,353 190
TextRegio			6	region...	2387,383 2738,383 273
TextRegio			7	region...	1632,480 2731,480 273

The 'Item visibility' dialog box is shown, listing the following options:

- R Show regions
- L Show lines
- B Show baselines
- W Show words
- Render blackenings
- R Show regions reading order
- L Show lines reading order
- W Show words reading order

Correcties van vormen etc.

Transkribus v1.10.0.3-SNAPSHOT (10_12_2019_12:54), Loaded doc: 168882_dupl, ID: 220644, Page 25, file: 00000025.png* [Image Meta Info: (Resolution:-1.0, w*h: 3050 * 4697)] [current lin

The screenshot shows the Transkribus software interface. On the left, there is a sidebar with navigation and management tools. The main area displays a document page with a highlighted text region. A context menu is open over the highlighted region, listing various actions. Below the page, a list of documents is visible, with the current document '168882_dupl' selected.

ID	Title	Pages	Uploader	Uploaded
288...	TRAINING_TESTSET_testNamur	2	info@caro...	Thu Dec 05...
288...	TRAINING_TESTSET_Romein_1...	10	info@caro...	Thu Dec 05...
288...	TRAINING_TESTSET_French_18...	6	info@caro...	Thu Dec 05...
284...	TRAINING_TESTSET_Latin_Arto...	2	info@caro...	Thu Dec 05...
284...	TRAINING_TESTSET_Frans	6	info@caro...	Wed Dec 0...
284...	TRAINING_TESTSET_Frans_Print	2	info@caro...	Wed Dec 0...
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 0...
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...	KBNLB410014860_png	434	sara.veldho...	Thu Jun 20 ...
175...	UBU000026528_png	356	sara.veldho...	Thu Jun 20 ...
174...	KBNLB390007382_png	608	sara.veldho...	Tue Jun 18 ...
174...	GENT900000222586_2_png	954	sara.veldho...	Tue Jun 18 ...
174...	GENT900000222480_png	1062	sara.veldho...	Tue Jun 18 ...

Create top level text region with size of image

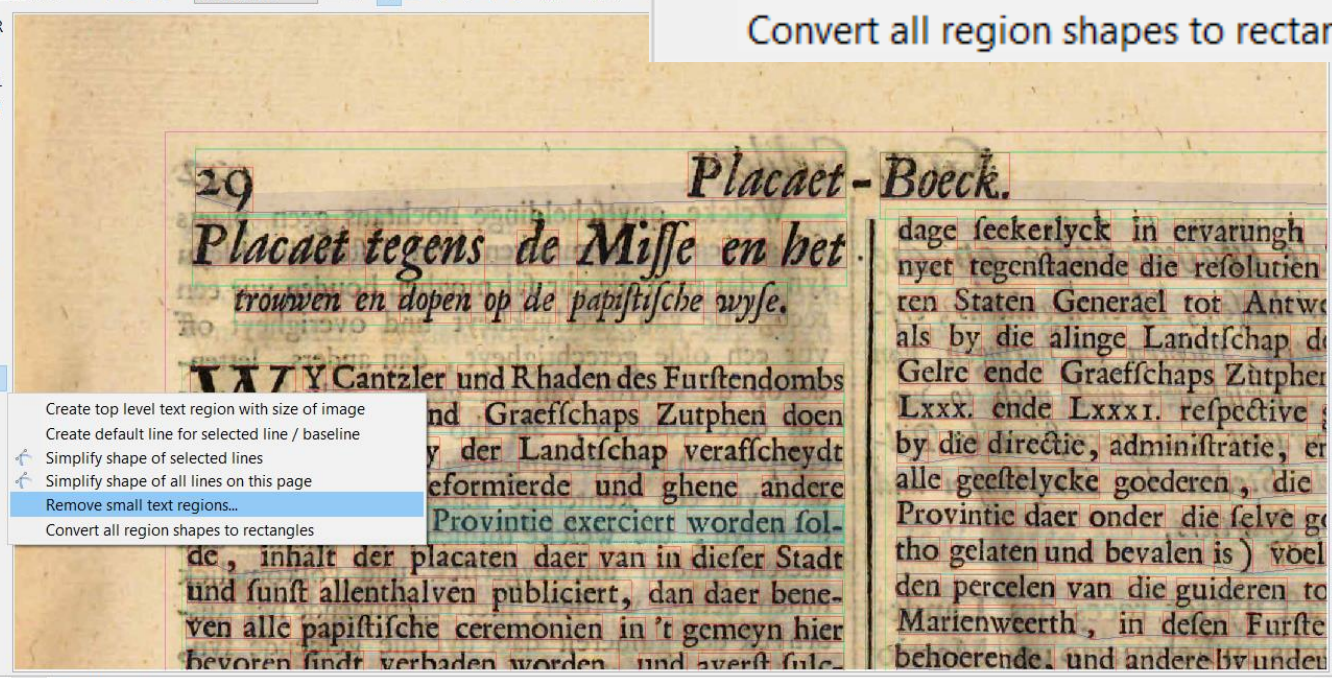
Create default line for selected line / baseline

Simplify shape of selected lines

Simplify shape of all lines on this page

Remove small text regions...

Convert all region shapes to rectangles



- 1-1 29 Placact
- 2-1 Placaet tegens de Misse en het
- 3-1 trouwen en dopen op de papistische wyse.
- 4-1 WY Cantzler und Rhaden des Furstendombs
- 4-2 Gelder und Graefschaps Zutphen doen
- 4-3 kondt, Alsoo by der Landschap verafscheydt
- 4-4 dat allein die Reformierde und ghene andere
- 4-5 religion in dieser Provincie exerciert worden sol-

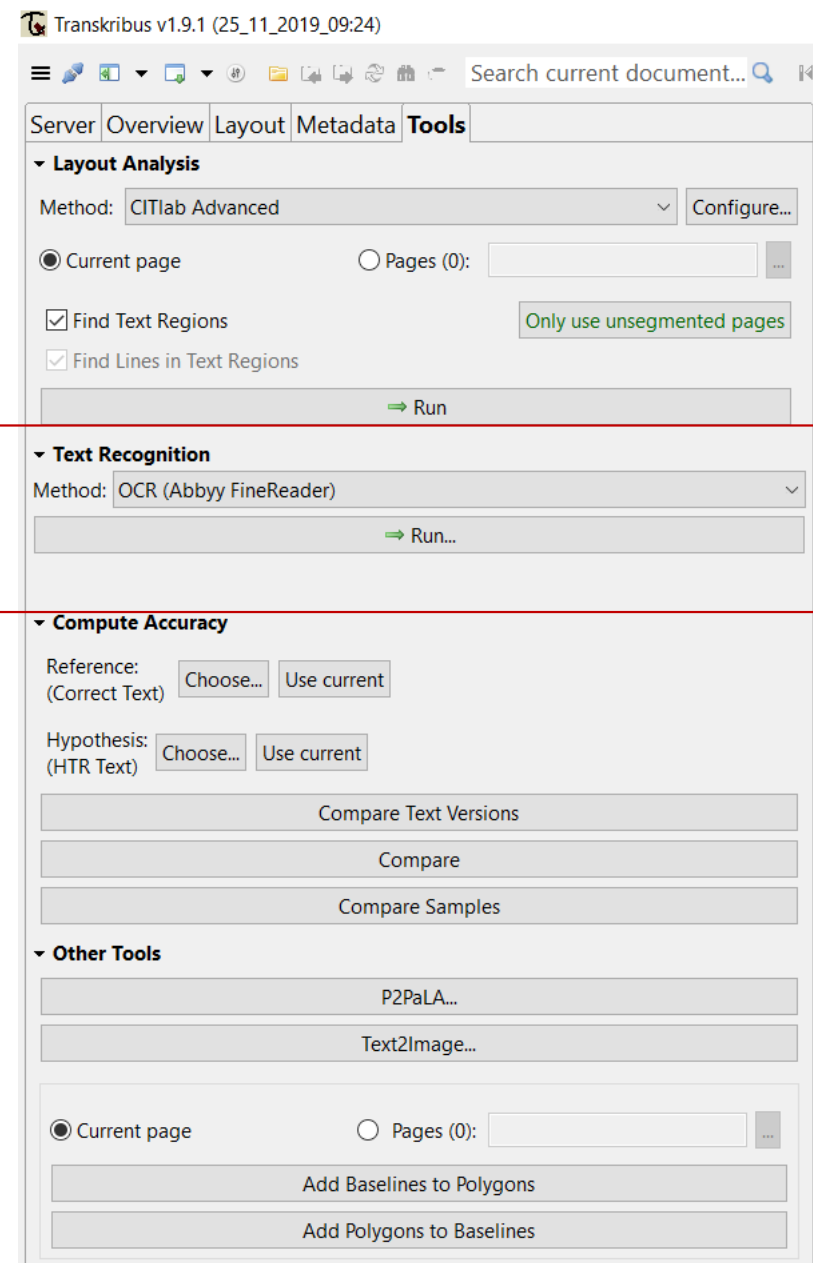
OCR in Transkribus (ABBYY Finereader v.11)

#Transkribus

#webinargeavanceerdNL

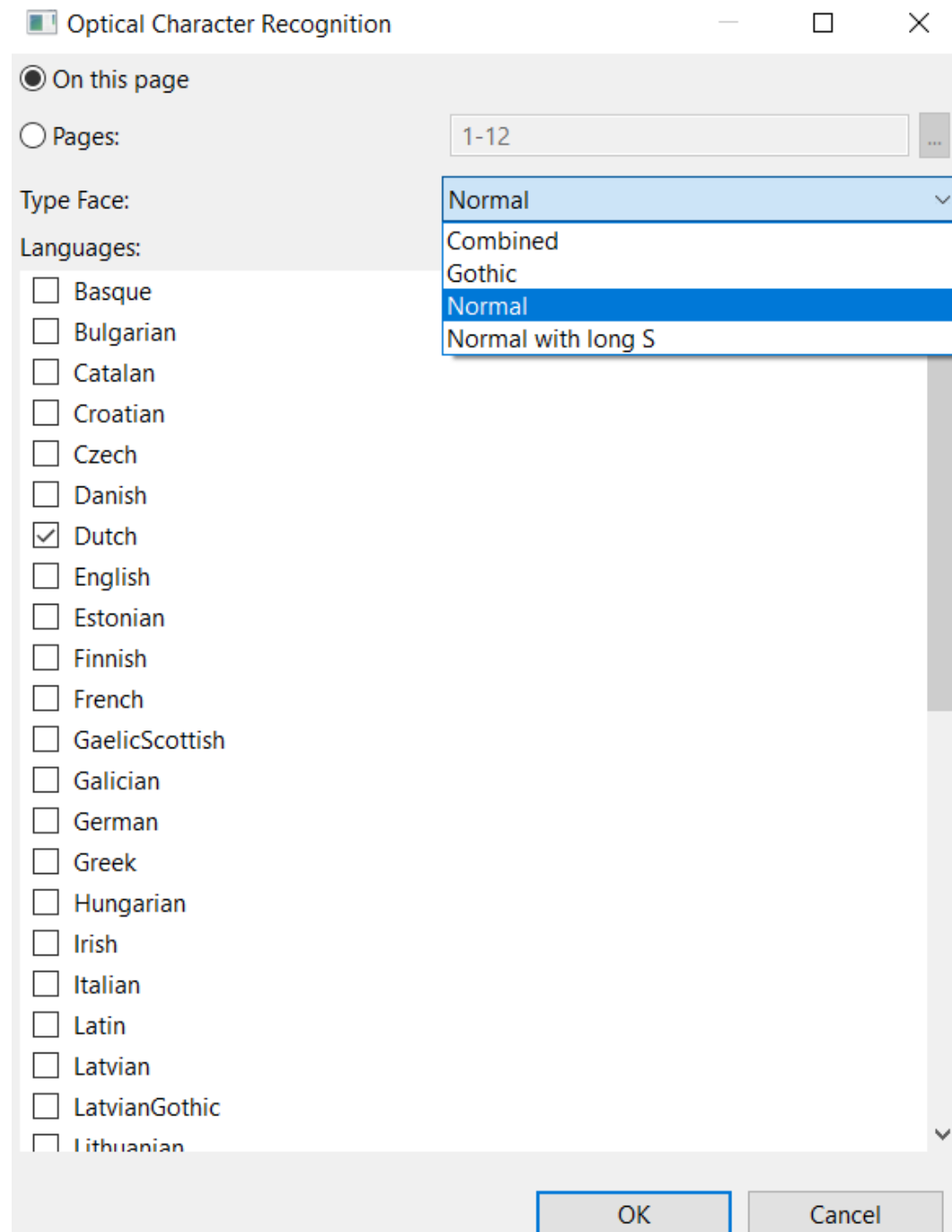
Optical Character Recognition

- 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).



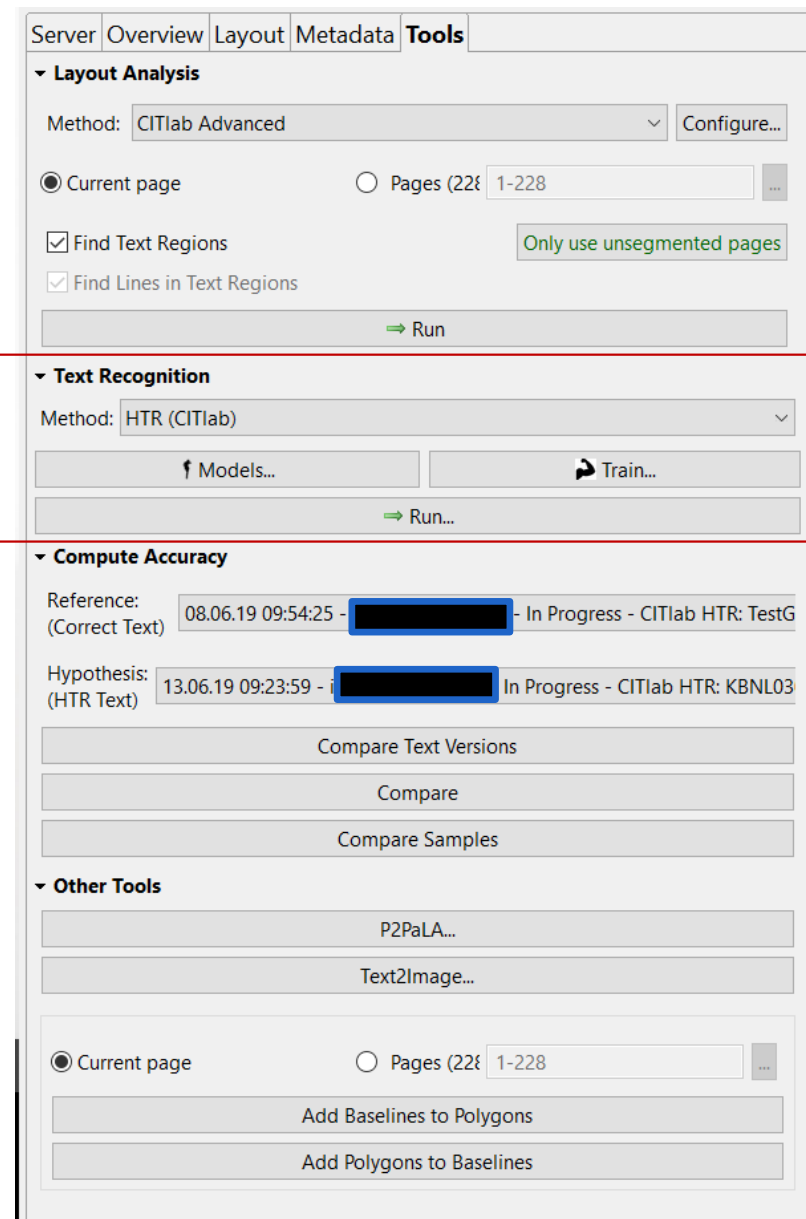
Modellen:
maken, beoordelen, toepassen

#Transkribus

#webinargeavanceerdNL

Train een eigen model

- Tab-menu, klik op Tools.
- Ga naar Tekst Recognition.
- **Kies bij “Method” voor: HTR.**
- 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 uitvoeren: dan pas selecteer je het model.



The screenshot shows the 'Tools' tab in the CITIab interface. The 'Text Recognition' section is highlighted with a red box. It includes a 'Method' dropdown set to 'HTR (CITlab)', a 'Models...' button, and a 'Train...' button. Below this, there are 'Run...' and 'Compute Accuracy' sections. The 'Compute Accuracy' section shows a 'Reference' and 'Hypothesis' comparison with buttons for 'Compare Text Versions', 'Compare', and 'Compare Samples'. The 'Other Tools' section includes 'P2PaLA...' and 'Text2Image...' buttons. At the bottom, there are radio buttons for 'Current page' and 'Pages (228 1-228)', and buttons for 'Add Baselines to Polygons' and 'Add Polygons to Baselines'.

HTR

HTR Training

Model Name: Language:

CITlab HTR CITlab HTR+

Nr. of Epochs: Learning Rate:

Noise: Train Size per Epoch:

Base Model:

Description:

Documents HTR Model Data

> 280830 - part1 (29 pages)

Search:

Overview

Transcript version

Training Set

ID	Title	Pages
----	-------	-------

Validation Set

ID	Title	Pages
----	-------	-------

Model Name:

Language:

Description:

CITlab HTR **CITlab HTR+**

Nr. of Epochs:

Base Model:

Documents | HTR Model Data

> 📁 280830 - part1 (29 pages)

Search:

Overview

Transcript version

Training Set


ID	Title	Pages
----	-------	-------













Validation Set

ID	Title	Pages
----	-------	-------

-
-

Publieke modellen (10-12-2019)

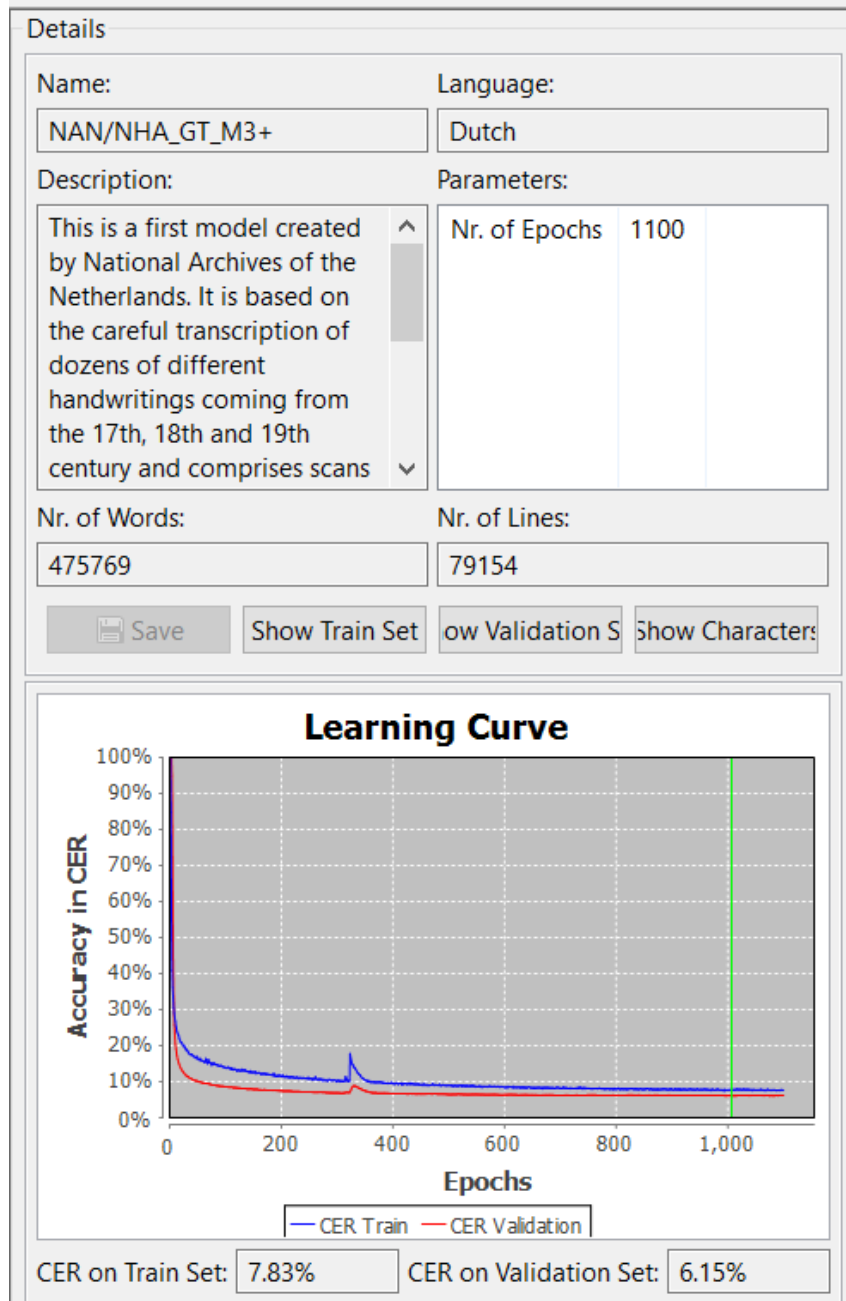
 Choose a model

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

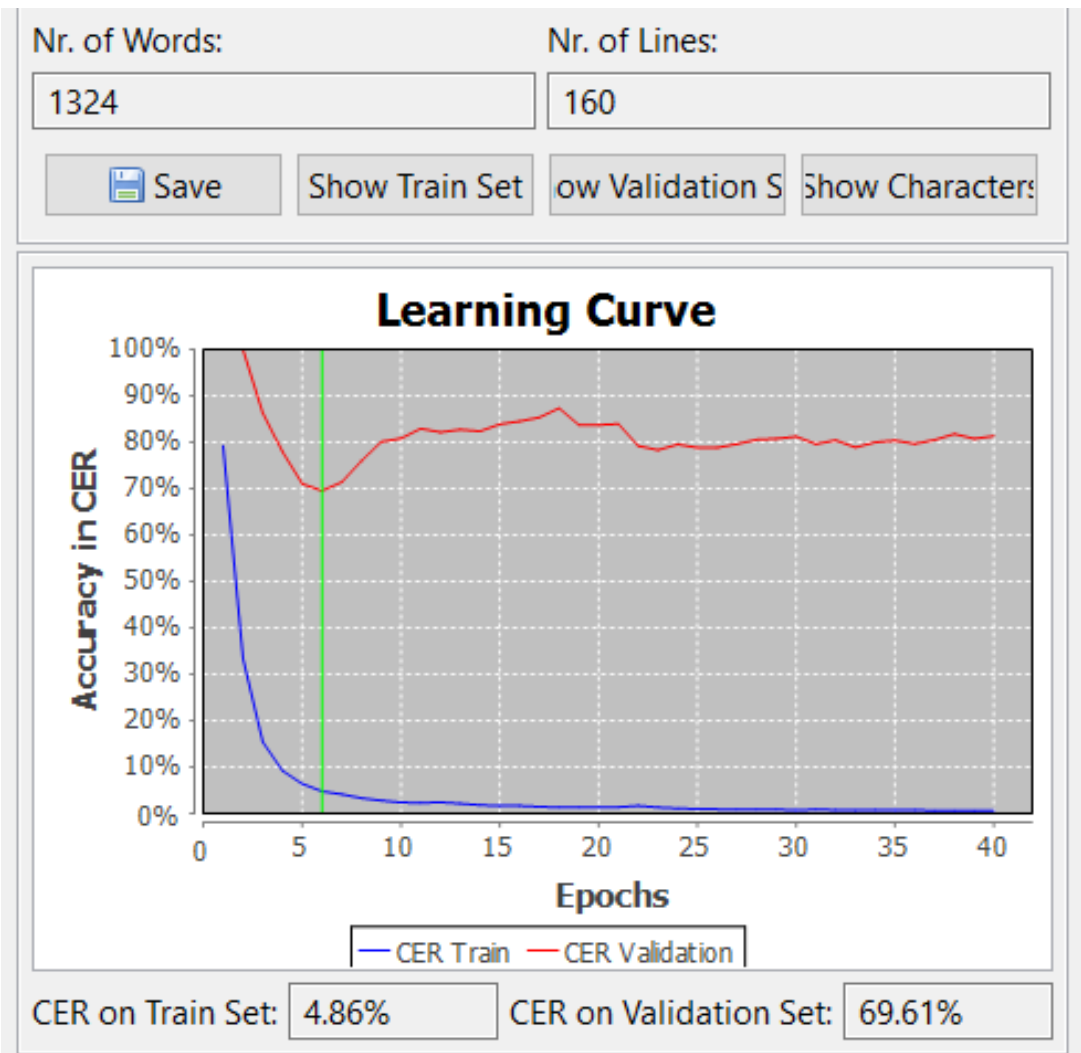
- Modellen voor politieke, economische, sociale, culturele, literaire teksten (etc.) kunnen enorm verschillen en minder goed werken op teksten van een ander type;
- Modellen kunnen regioafhankelijk zijn (dialect);
- Periodisering is belangrijk

Vermeldt deze informatie aub in het 'description' field!

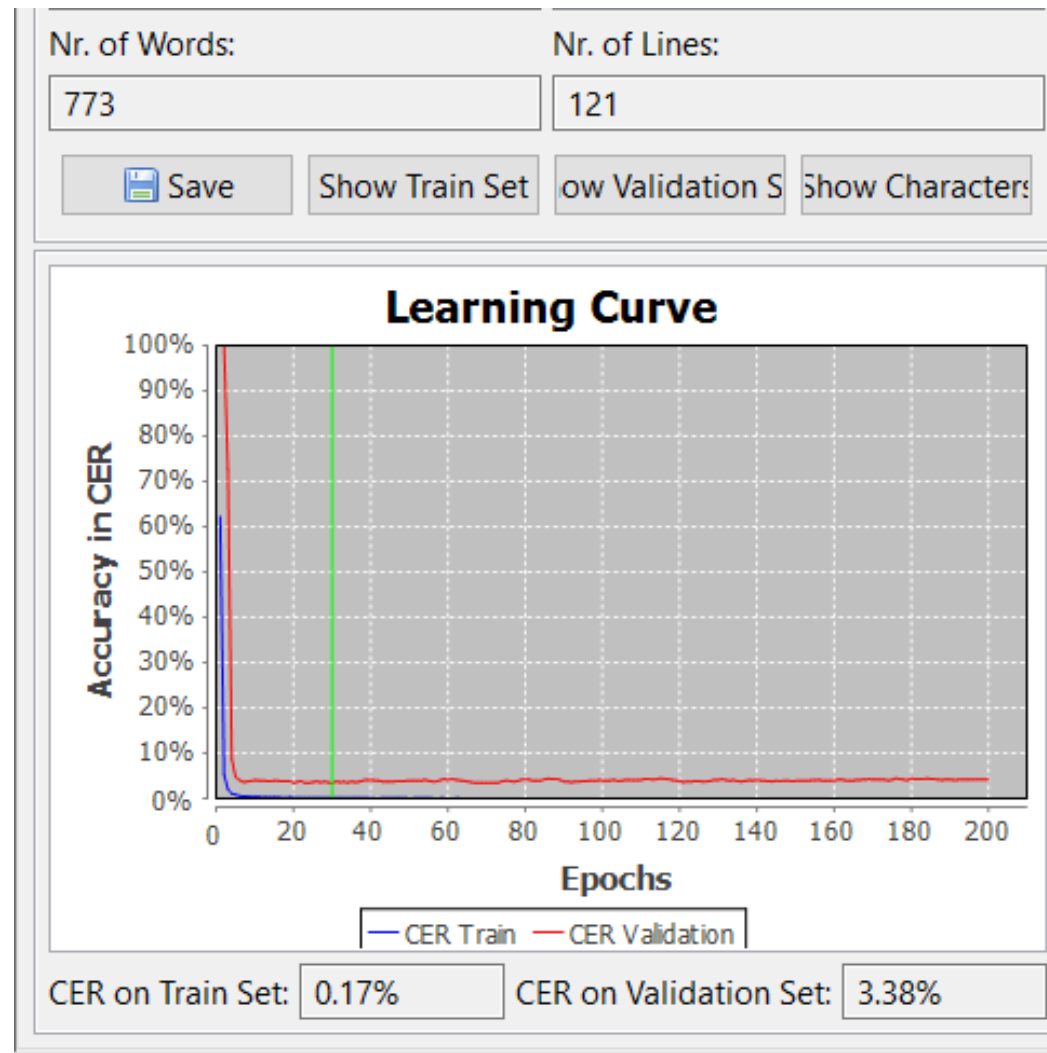


- CER= Character Error Rate (op de Trainingsset of Validatieset)
- Ook zichtbaar: hoeveel woorden en hoeveel regels (min >1000) er zijn gebruikt.
- Aantal Epochs.
- Show Characterset

Grafieken zijn niet altijd zo mooi...



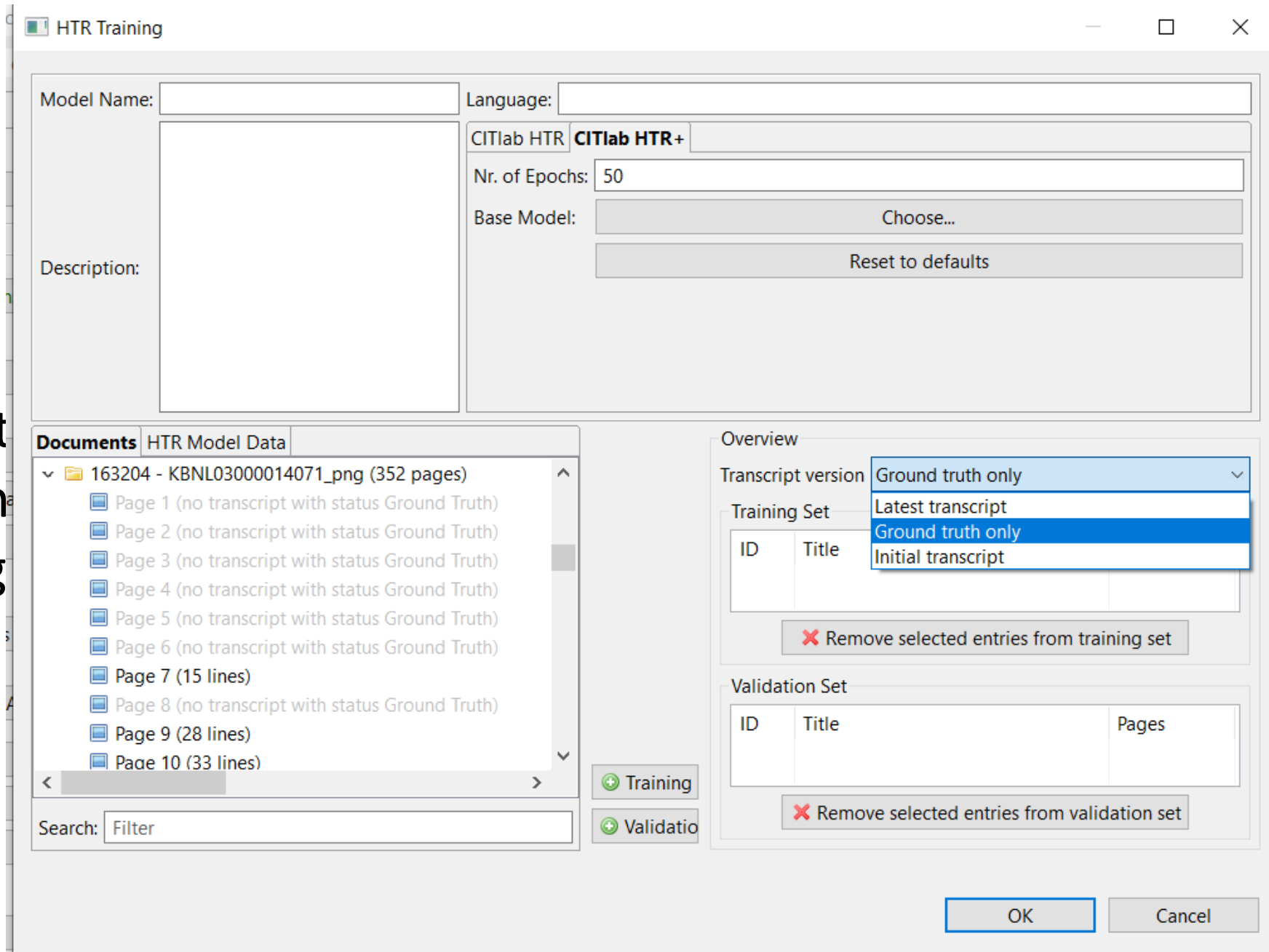
Leigh Hunt (o.a. te weinig training/ validatie)



Latijn drukwerk (overfitting)

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)

HTR Training

Model Name: Language:

Description:

CITlab HTR CITlab HTR+

Nr. of Epochs: 50

Base Model:

Documents HTR Model Data

- > 13743 - NederlandsGotisch
- > 13243 - NederlandsGotischPrint
- ▼ 12664 - NZZ Gold Standard M1+
 - ▼ Train Set (150 pages)
 - Page 1 (76 lines, 507 words)
 - Page 2 (78 lines, 522 words)
 - Page 3 (74 lines, 465 words)
 - Page 4 (74 lines, 510 words)
 - Page 5 (74 lines, 488 words)
 - Page 6 (74 lines, 539 words)
 - Page 7 (84 lines, 688 words)
 - Page 8 (73 lines, 467 words)

Search:

Overview

Transcript version

Training Set

ID	Title	Pages
----	-------	-------

Validation Set

ID	Title	Pages
----	-------	-------

Modellen:
Werkt het ok?

#Transkribus
#webinargeavanceerdNL

Server Overview Layout Metadata **Tools**

Layout Analysis
Method: CITIab Advanced Configure...
 Current page Pages (85): 1-852
 Find Text Regions Only use unsegmented pages
 Find Lines in Text Regions
Run

Text Recognition
Method: HTR (CITIab)
Models... Train...
Run...

Compute Accuracy
Reference: (Correct Text) 21.07.19 22:05:24 - info@caromein.nl - In Progress - CITIab HTR: Oxford
Hypothesis: (HTR Text) 09.10.19 13:12:46 - info@caromein.nl - In Progress - Abbyy Finereader 11
Compare Text Versions
Compare
Compare Samples

Other Tools
P2PaLA...
Text2Image...
 Current page Pages (85): 1-852
Add Baselines to Polygons
Add Polygons to Baselines

Text Recognition Close
 Current page Pages (852): 1-852
 Do polygon simplification
 Keep original line polygons
 Enable Keyword Spotting
Restrict on structure tags
CITIab RNN HTR
Net Name: Dutch_Gothic_Print
Language: Dutch (16th, 17th, 18th century)
Dictionary: No dictionary
Select HTR model...
OK Cancel

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: 27.11.19 10:20:13 - info@caromein.nl - Ground Truth [Use current]
 Hypothesis: 29.11.19 20:11:12 - info@caromein.nl - In Progress - Abbyy Finereader 1
 [Compare Text Versions] (circled in red)
 [Compare]
 [Compare Samples]

Other Tools
 [P2PaLA...]
 [Text2Image...]
 Current page Pages (352): 1-352
 [Add Baselines to Polygons]
 [Add Polygons to Baselines]

TR

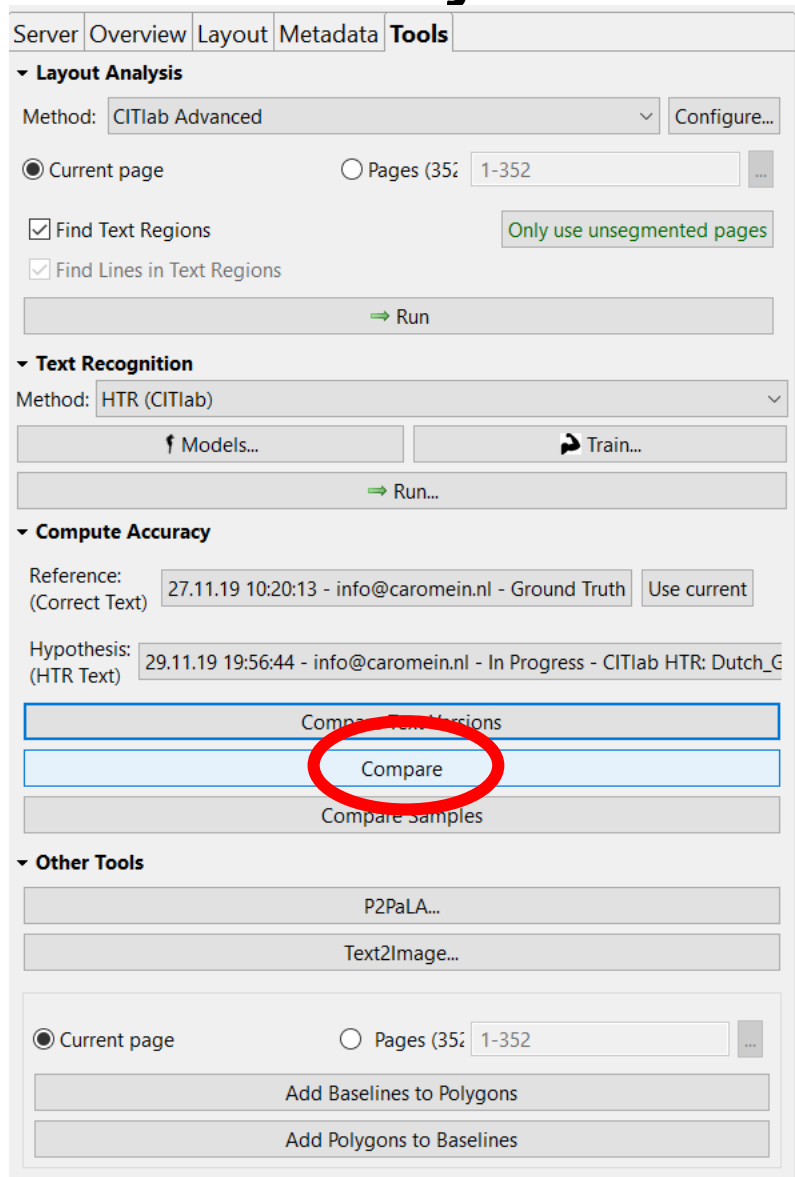
Version Comparator

Show line numbers

42 Generale
 fhappm schappen de volgende't volgende 't Placcaet ofte LM Lijste confiscatie
 subject zhn / zijn/ ende in vier vier voegen aengehaelt aengehaest worden/ dre die
 selve sullen in Sequestro blaven / blijven/ ende verwant worden verwaert worden
 vp Ve restiective by de respective Collecteurs ofte Gerievs dienaers Gerichts-dienaers in
 Stadt ende Lande / Lande/ ter thdt tijd 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 P?rovinciael 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 verkof/ ende die penningen daer van procederende procederende/ ge-
 distawleert wo?den naemholvan distribueert worden nae inholt van den Placcate Placcate/ in dier
 voegen / voegen/ dat des Pachters ende Breuck pachtrs Breuck- pachters anpar-
 ten ontfangen sullen worden by 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. daernvoven daerenboven
 te verbeuren.
 ¶ H 1 3 3 XXXIII.
 Die respective Pachters sullen vermogn vermogen soo dick dick-
 wyls alg wijls als 't haer gelieft aen die Lomtopren comparern/ Comtoyren compareren/
 Boecken / Boecken/ en Journalen concernerende haren pacht
 ende geen anderen examineren/ ende op de frauden in-
 gmrenen quireren/ daer toe de Collecteurs haer susten admitterm sullen admitteren
 ende haer behoMK anwyfinge doen / susten behoortlijck aenwijfinge doen/ sullen niet te
 Min min die Heerm 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-
 4-2 se/ sullen dieselde dieselve ter instantie van den Pachter V?euck

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



Server Overview Layout Metadata **Tools**

▼ **Layout Analysis**

Method: CITlab Advanced Configure...

Current page Pages (35): 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 19:56:44 - info@caromein.nl - In Progress - CITlab HTR: Dutch_G

Compare New Missions

Compare

Compare Samples

▼ **Other Tools**

P2PaLA...

Text2Image...

Current page Pages (35): 1-352 ...

Add Baselines to Polygons

Add Polygons to Baselines

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

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 19:56:44 - info@caromein.nl - In Progress - CITlab HTR: Dutch_Gothic_Print

Compare Text Versions

Compare

Compare Samples

Other Tools

P2PaLA...

Text2Image...

- TR
- L
- BL
- W



Compare ✕

Compare Advanced Compare

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

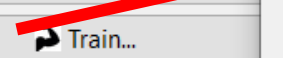
Hypothesis: (HTR Text) 29.11.19 19:56:44 - info@caromein.nl - In Progress - CITlab HTR: Dutch_Gothic_Print Use current

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



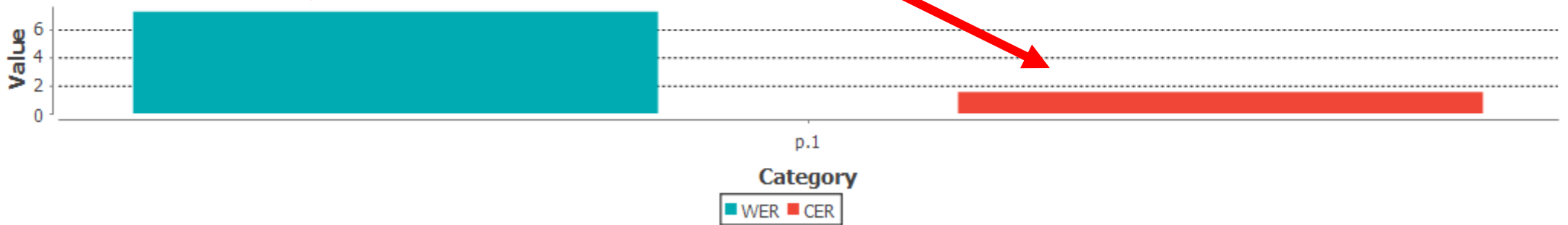
1-3 subject zijn/ ende in dier voegen aengehaest worden/ die
 1-4 selve sullen in Sequestro blijven/ ende verwaert worden
 1-5 by de respective Collecteurs ofte Gerichts-dienaers in
 1-6 Stadt ende Lande/ ter tijdt daer over verclaert ofte voor
 1-7 de waarde van dien geneoetsame cautie gestellet sal zijn/

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 ..

Error Rate Chart | Ref: GT | Hyp: CITIab HTR: Dutch_Gothic_Print



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

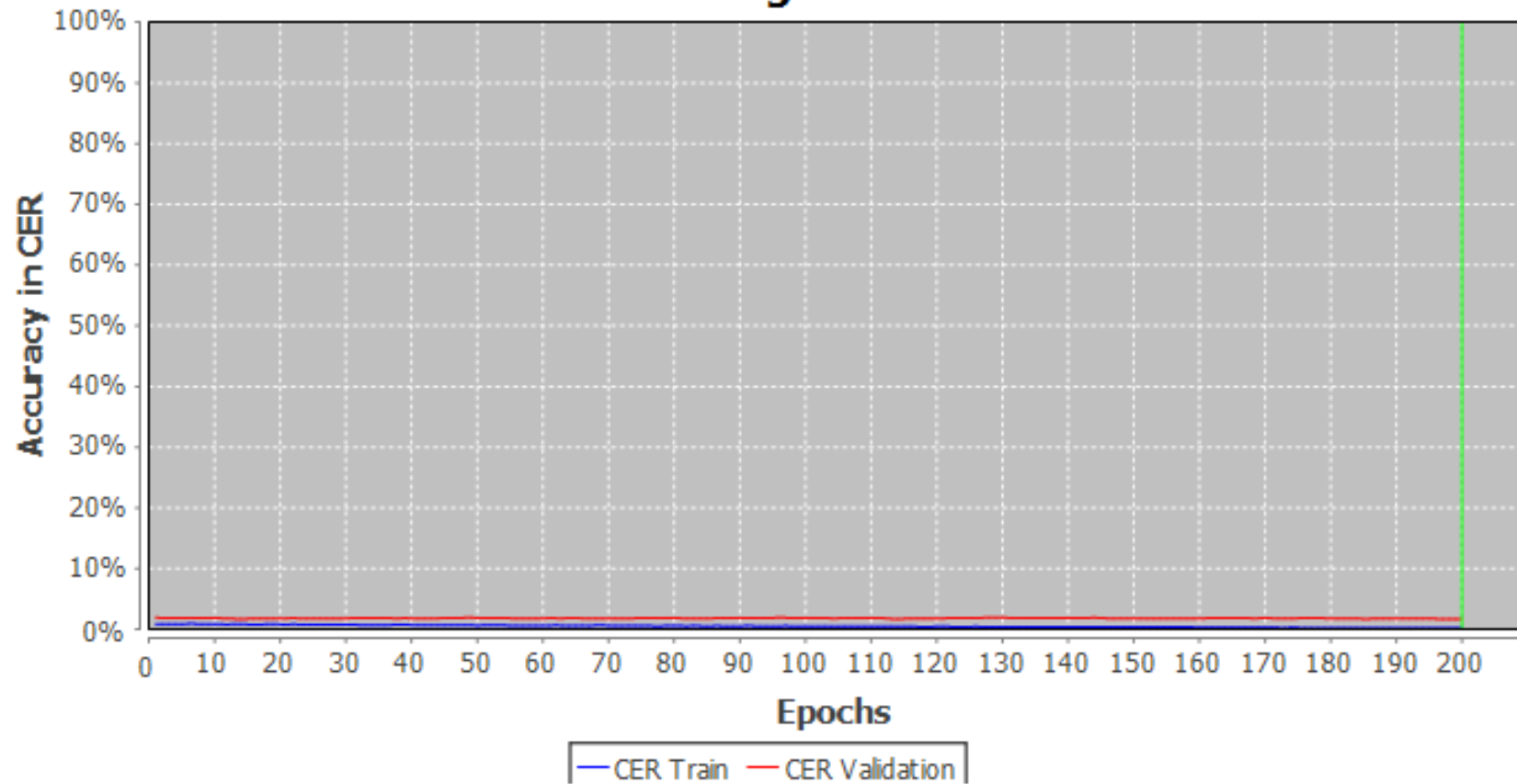
 Save

Show Train Set

Show Validation Set

Show Characters

Learning Curve



CER on Train Set: 0.22%

CER on Validation Set: 1.71%

Sample

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)

f 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 Use 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

Compare Samples

Documents Samples

Sample Title:
Sample_Sample_UBA000159632_png

Description:

Nr. of lines

Collection

- > 293324 - TRAINING_VALIDATION_SET
- > 293217 - Sample_UBA000159632_png
- > 288837 - TRAINING_TESTSET_testNar
- > 288816 - TRAINING_TESTSET_Romein
- > 288578 - TRAINING_TESTSET_French_
- > 284942 - TRAINING_TESTSET_Latin_Ar
- > 284693 - TRAINING_TESTSET_Frans (6
- > 284621 - TRAINING_TESTSET_Frans_Print (2 pages)
- > 272256 - TRAINING_TESTSET_DutchGr
- > 269147 - TRAINING_TESTSET_EarlyMc
- > 268671 - TRAINING_TESTSET_gothisch
- > 220687 - UBU000005064_png_duplica
- > 220644 - 168882_dupl (486 pages)
- > 178570 - UBL000045368_png (852 pa
- > 177692 - UU15755 (1514 pages)
- > 175406 - UBL000046178_png (1162 pa
- > 175064 - UBA000066806_png (220 pa
- > 175060 - GENT900000065568_png (26
- > 175059 - UBL000045369_png (748 pa
- > 175020 - KBNLB410014860_png (434
- > 175019 - UBU0000026528_png (356 pa

+ Add to Sample Set

Documents added to Sample Set

ID	Title	P
----	-------	---

✗ Remove selected entries from train set

📄 Create Sample

🔗 Help Cancel

Sample

The screenshot shows a software interface for document analysis. At the top, there is a toolbar with various icons for navigation and editing. Below the toolbar, the interface is divided into several sections:

- Layout Analysis:** This section includes a dropdown menu for the method (currently set to "CITlab Advanced"), a "Configure..." button, and radio buttons for "Current page" and "Pages (424): 1-424". There are also checkboxes for "Find Text Regions" and "Find Lines in Text Regions", with a "Only use unsegmented pages" option. A "Run" button is present at the bottom of this section.
- Text Recognition:** This section includes a dropdown menu for the method (currently set to "HTR (CITlab)"), a "Models..." button, and a "Train..." button. A "Run..." button is also present.
- Compute Accuracy:** This section includes a "Reference: (Correct Text)" field with a timestamp "11.12.19 21:24:47" and a "Use current" button. Below it is a "Hypothesis: (HTR Text)" field with a timestamp "11.12.19 21:24:47 - ir" and a "Use current" button. There are three buttons: "Compare Text Versions", "Compare", and "Compare Samples".
- Other Tools:** This section includes buttons for "P2PaLA...", "Text2Image...", "Add Baselines to Polygons", and "Add Polygons to Baselines".

The main workspace displays a scanned document page with a highlighted text region. The text in the highlighted region is:

gen Huwelijk 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

At the bottom left of the workspace, there is a small box containing the text "1-1".

KeyWord Spotting (KWS)

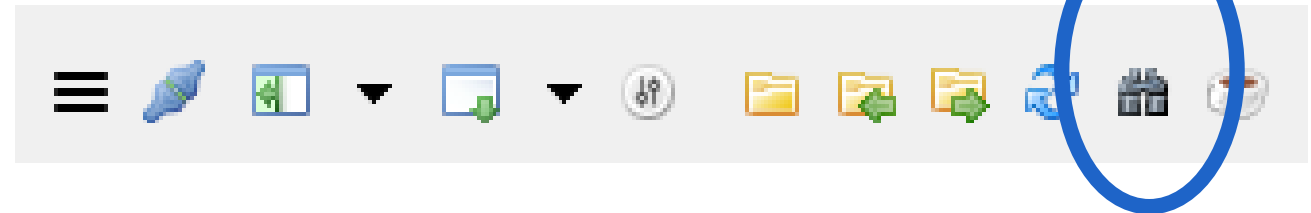
#Transkribus

#webinargeavanceerdNL



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 $\pm 30\%$ CER
- Toont op basis van:
 - Confidence Value (probability)
 - Kan altijd worden getoont



Search for...

Documents **Fulltext (Solr)** Tags KWS

Search transcribed text for words or phrases

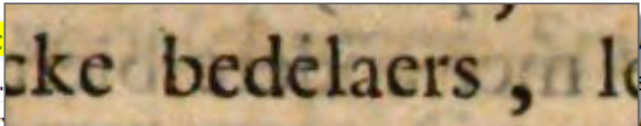
Search for:

Show word preview
 Current document
 Word-based text
 Line-based text

Case sensitive
 Fuzzy search

Search results

Showing Pagehits 0 to 6 of 6

Context	Document	Page
foodanighe heerloofc, flercke bedelac	168882_dupl	131
tntejlato fuccedeert. den 29 May 1630. 	168882_dupl	344
voor- gaande placcaten men dagelyc	168882_dupl	95
huyfluyden, om fodanige vagebonden, lant-loopers, en vreem- de bedelaers, foo haeft de felve	168882_dupl	95
-lopers, bedelacrs, heydens en der- geljicke Vagabonden. Copia. WY Raeden, in naeme van de H	168882_dupl	200
ordon- neren ende bevelen, alle foodane yagabonden, Vreemde bedelacrs ende lcdiggangcrs, c	168882_dupl	122
, vagebonden, vremde bedelacrs ende Jedichgan- gers binnen drie dagen na de publicatie van d	168882_dupl	105
bedelacrs ende ledighgan- gers na opiganek van de voorschreven drie da- gen, Too wanneer fy	168882_dupl	105

- Full tekst search (met fuzzy optie)

Search for...

Documents **Fulltext (Solr)** Tags KWS

Search transcribed text for words or phrases

Search for:

Show word preview
 Current document
 Word-based text
 Line-based text

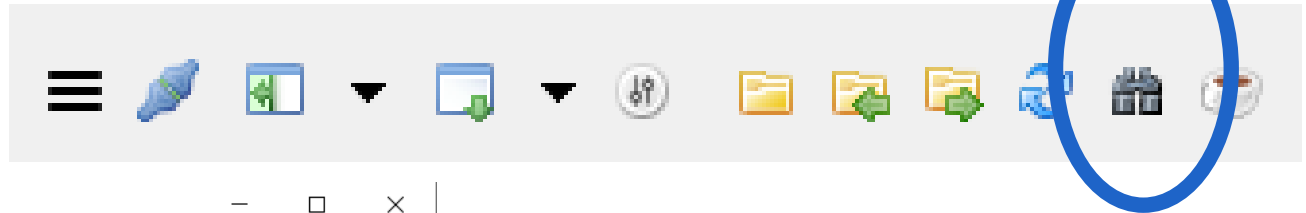
Case sensitive
 Fuzzy search

Search results

Showing Pagehits 0 to 6 of 6

Context	Document	Page
foodanighe heerloofc, flercke bedelac	168882_dupl	131
tntejlato fuccedeert. den 29 May 1630. cke bedelaers,	168882_dupl	344
voor- gaande placcaten men dagelyc	168882_dupl	95
huyfluyden, om fodanige vagebonden, lant-loopers, en vreem- de bedelaars, foo haeft de felve	168882_dupl	95
-lopers, bedelacrs, heydens en der- geljicke Vagabonden. Copia. WY Raeden, in naeme van de H	168882_dupl	200
ordon- neren ende bevelen, alle foodane yagabonden, Vreemde bedelacrs ende lcdiggangcrs, c	168882_dupl	122
, vagebonden, vremde bedelacrs ende Jedichgan- gers binnen drie dagen na de publicatie van d	168882_dupl	105
bedelacrs ende ledighgan- gers na opiganek van de voorschreven drie da- gen, Too wanneer fy	168882_dupl	105

KWS



Search for...

Documents | Fulltext (Solr) | Tags | **KWS**

Search in: Partial Matches Case-sensitivity Expert Syntax

Confidence Threshold: < >

Queries

Keyword 1

Keyword 2

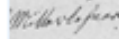
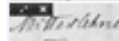
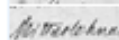
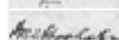



Search Results

Created ^	Status	Queries	Duration	Scope	

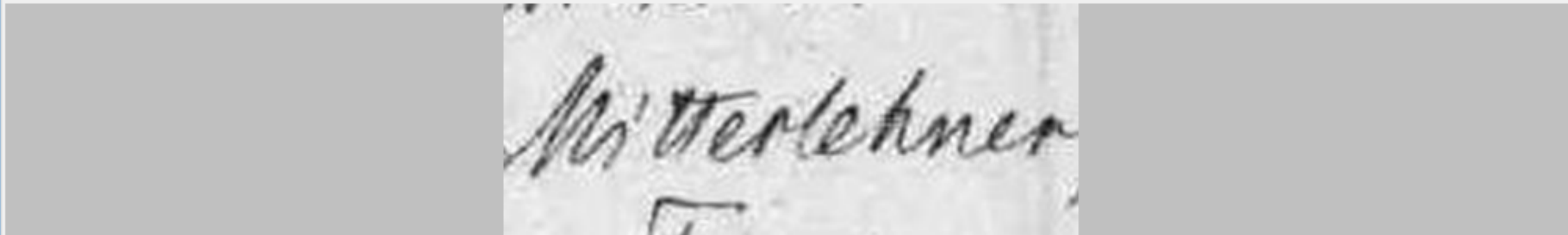
- Zoekt gedeeltelijk visueel maar deels ook transcripties (niet geheel duidelijk wat onderliggende techniek is)

Keyword Spotting Results

"Franz Mitterlehner" (3 hits) "Franz" (5 hits) "Mitterlehner" (7 hits) "Linimeir" (2 hits)

Confidence	Page Nr.	Line transcription	Previ...
0.4983	3	des Mitherlehaer	
0.4707	1	dranz Miterlekner, dienskaachtsein	
0.2080	1	Moiveshekner Maurrts ded siusbesitzas	
0.1445	4	Jofuch M strlofet	
0.1429	5	Josiuch Mberlehent	
0.0574	1	z e Mensrgaaeeghdnetuhes. Msmernensgriduach Mti	
0.0530	4	t keeste seht Mnet lesnnehen Eni Me	

Preview



Close



KWS Interface

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

Woordenboek

#Transkribus

#webinargeavanceerdNL

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)

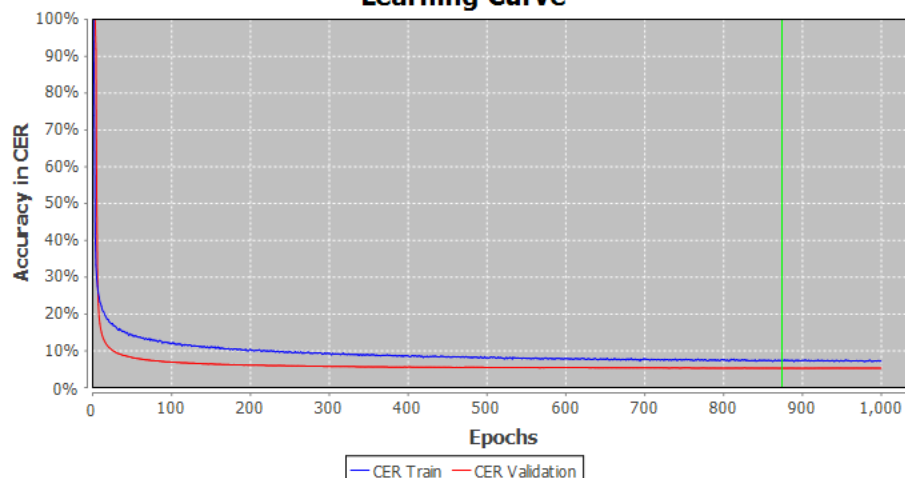
Search: All Technology:

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:	<p>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.</p> <p>All notaries (except Van Hoon and Van Esterwege) have 10 scans validation included (2671 scans training, 130 validation): Maten de Jonge (200) Van Homrigh (127)</p>				
Parameters:	<table border="1"> <tr> <td>Nr. of Epochs</td> <td>1000</td> </tr> </table>			Nr. of Epochs	1000
Nr. of Epochs	1000				
Nr. of Words:	622904	Nr. of Lines:	116698		
<input type="button" value="Save"/> <input type="button" value="Show Train Set"/> <input type="button" value="Show Validation Set"/> <input type="button" value="Show Characters"/>					

Learning Curve

CER on Train Set: CER on Validation Set:

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_Iav_ser.d...
- Church_Slavonic_VMC_Apostolos_F...
- Church_Slavonic_VMC_Apostolos_...
- Cochin_Court_Records_M1.dict
- Combined_Dutch_Model_M1.dict

OK

Cancel

Text2Image

#Transkribus

#webinargeavanceerdNL

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).

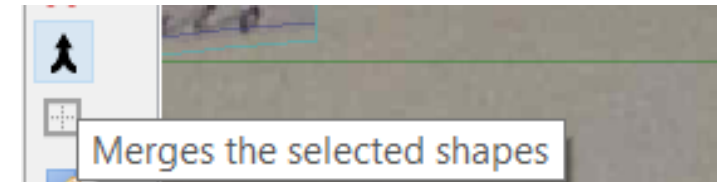
The screenshot shows a software interface with a 'Tools' tab selected. The interface is divided into several sections:

- Layout Analysis:** Method: CITIab Advanced. Options: Current page, Pages (228 1-228). Checkboxes: Find Text Regions, Find Lines in Text Regions. A button 'Run' is present.
- Text Recognition:** Method: HTR (CITIab). Buttons: 'Models...', 'Train...', and 'Run...'.
- Compute Accuracy:** Reference: (Correct Text) 08.06.19 09:54:25 - [redacted] - In Progress - CITIab HTR: TestG. Hypothesis: (HTR Text) 13.06.19 09:23:59 - [redacted] - In Progress - CITIab HTR: KBNL03. Buttons: 'Compare Text Versions', 'Compare', and 'Compare Samples'.
- Other Tools:** Buttons: 'P2PaLA...', 'Text2Image...', 'Add Baselines to Polygons', and 'Add Polygons to Baselines'.

At the bottom of the interface, there are radio buttons for 'Current page' (selected) and 'Pages (228 1-228)'.

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 afbeeldingspagina** 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:

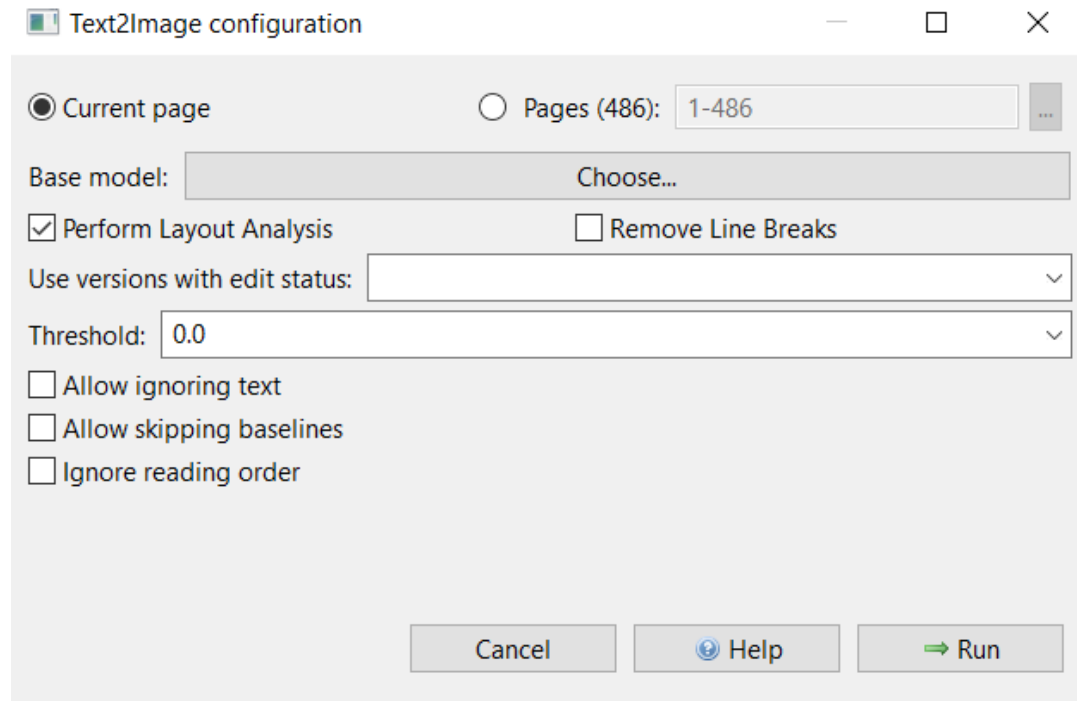
Perform LA

Allow skip words.

Ignore reading order.

N.B. Binnenkort komt er ook nog een knopje ‘assume Hyphenations’ (“-”, “=”, “:“).

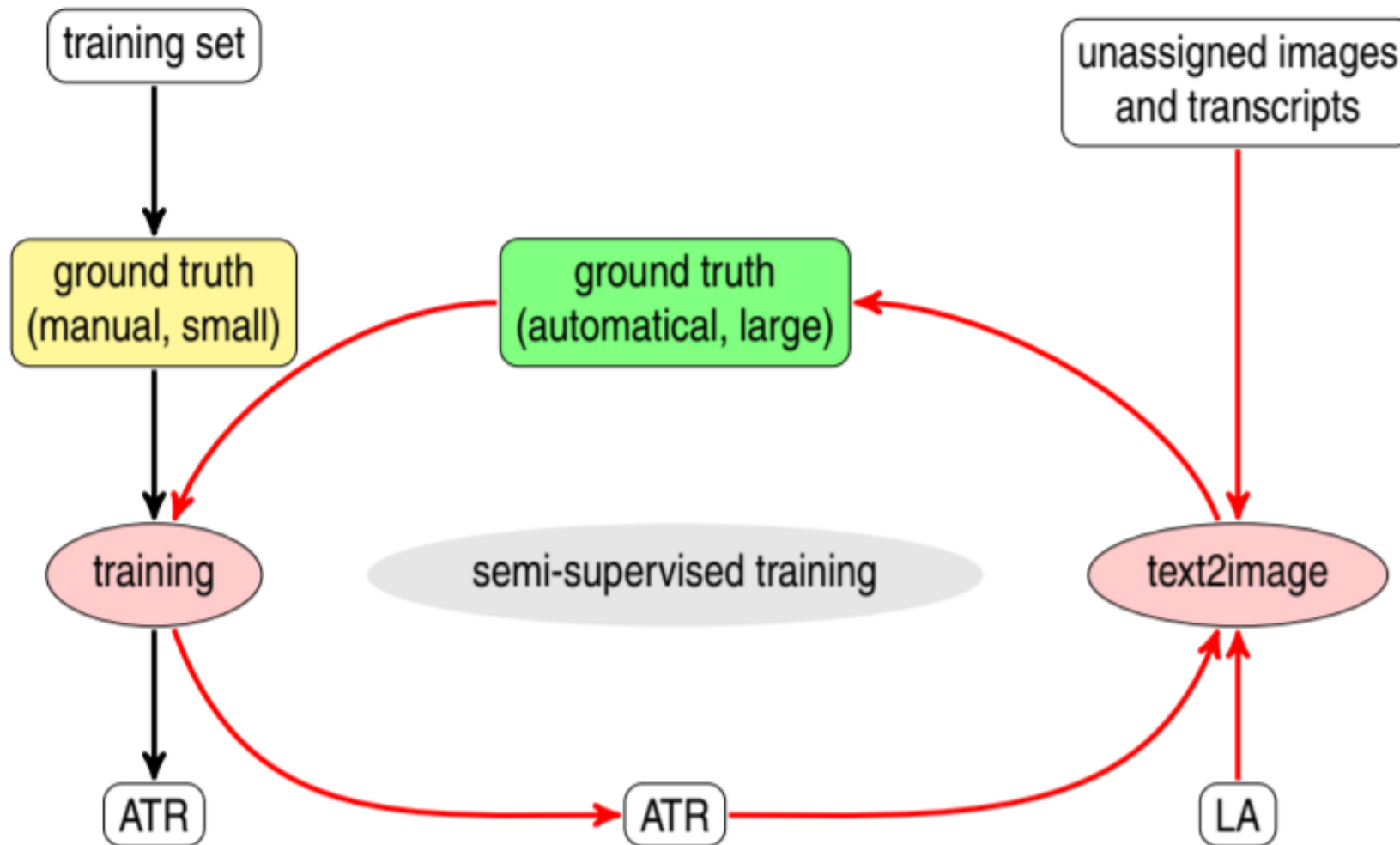
- Run



The image shows a software configuration window titled "Text2Image configuration". It contains several settings:

- Radio buttons for "Current page" (selected) and "Pages (486):" with a text input field containing "1-486".
- A "Base model:" label followed by a "Choose..." button.
- Checkboxes for "Perform Layout Analysis" (checked) and "Remove Line Breaks" (unchecked).
- A "Use versions with edit status:" dropdown menu.
- A "Threshold:" label followed by a dropdown menu showing "0.0".
- Checkboxes for "Allow ignoring text", "Allow skipping baselines", and "Ignore reading order", all of which are unchecked.
- Buttons for "Cancel", "Help" (with a question mark icon), and "Run" (with a right-pointing arrow icon).

semi-supervised training workflow



Labelen (structuur & woorden)

#Transkribus

#webinargeavanceerdNL

Toevoegen van Metadata aan je document

Server Overview Layout **Metadata** Tools

Document Structural Textual Comments

Title: KBNL03000092135_png

Authority: NA

Backlink: NA

External ID: NA

Hierarchy:

Author:

Uploaded: Thu May 16 14:45:56 CEST 2019

Genre:

Writer:

Dutch

Language: Arabic Basque Bulgarian Catalan

Script type: Printed Normal

Date of writing:

From: // To: //

Editorial Declaration...

Description:

Save

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

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

Transcription Options

ID	Text
70	Blackeing 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...	KRNI B410014860 png	434	sara.veldho...	Thu Jun 20 ...

25

Filter

Lay-out/ Structuur Metadata

The screenshot shows the 'Metadata' tab in a document editor, with the 'Structural' sub-tab selected. The 'Page type' and 'Links' fields are visible at the top. Below, the 'Selected element type' and 'Structure Type' sections are shown. A table lists various structure types with their corresponding colors and shortcuts.

Structure type	Color	Shortcut
paragraph	Cyan	
heading	Magenta	
caption	Green	
header	Blue	
footer	Purple	
page-number	Yellow	
drop-capital	Teal	
credit	Pink	
floating	Brown	
signature-mark	Dark Blue	
catch-word	Light Green	
marginalia	Tan	
footnote	Dark Green	

- 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).

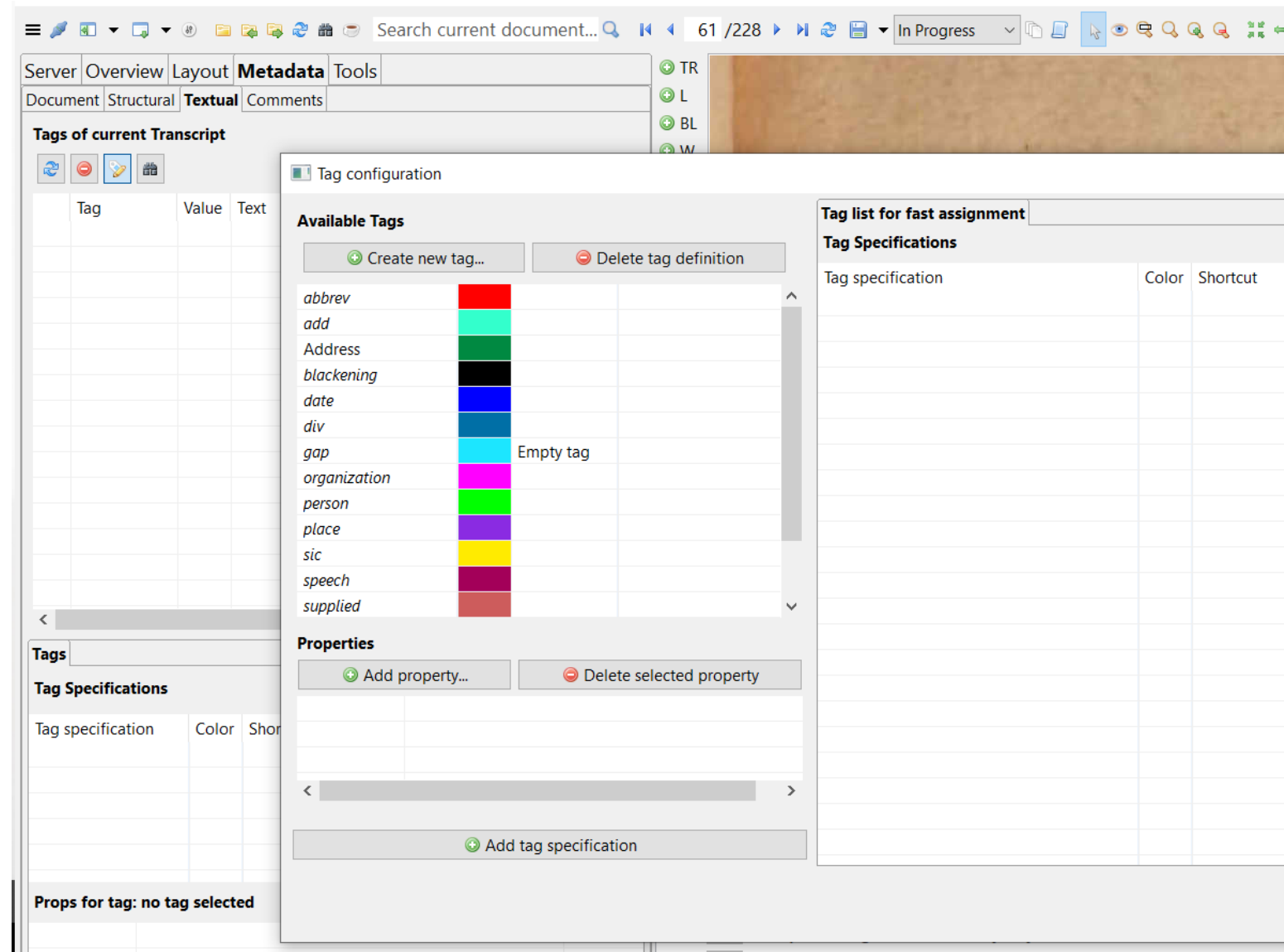
I. Tekstuele tags

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.

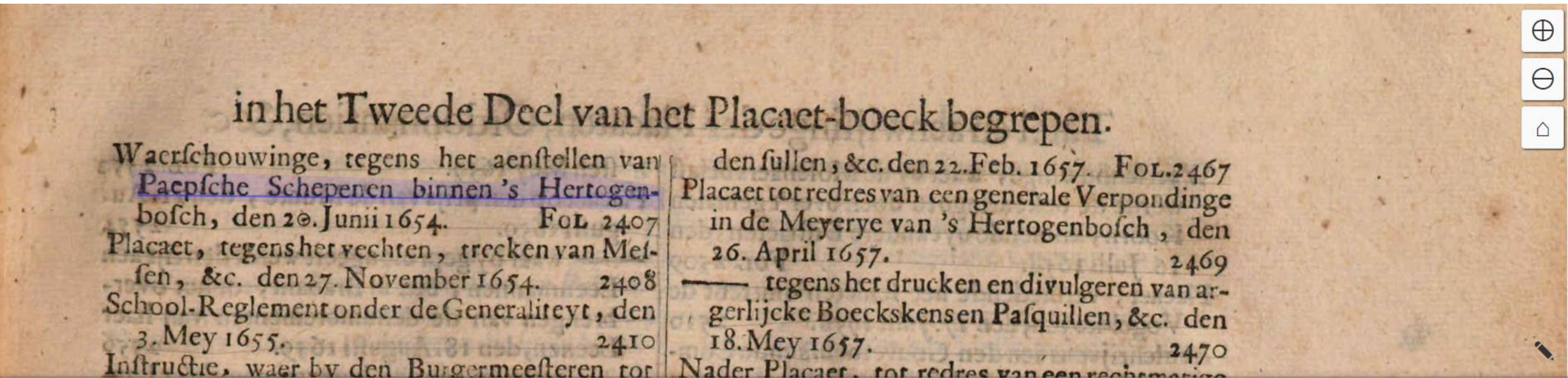
The screenshot shows a digital manuscript viewer interface. The main text is in Dutch and includes the following lines: "deren.", "Extract uittet Reces des Landdaghs binnen Zutphen in Martio, Aprili en Majo", "vervolgens in Martio", "dinarie gehouden.", "Mercurii den 29. Aprilis.", and "Op verzoeken van Adolph Gimbo". A right-click context menu is open over the text, displaying a list of tagging options: Address, abbrev, add, blackening, comment, date, div, gap, organization, person, place, sic, speech, supplied, unclear, and work. Below the main text, a secondary menu is visible with options: Copy text, Paste text, All tags (highlighted), Add a comment, and Delete all tags for current selection.

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!



Web interface Transkribus



Web interface Transkribus showing a document scan with a digital interface overlay. The interface includes a toolbar with options like x^2 , x_2 , **B**, *I*, U, **ab**, **?! Unclear**, **Special Characters**, and **Annotate ...**. A dropdown menu is open under **Annotate ...**, listing options: **Annotate ...**, **ABBREV**, **COMMENT**, **DATE**, **ORGANIZATION**, **PERSON**, **PLACE** (highlighted), and **UNCLEAR**. The interface also displays a table of text regions:

Text Region	Text	Annotation
Text Region 1	1 in het Tweede	#
Text Region 2	1 Waerschouwinge, tegens het aenstellen van	#
	2 Paepsche Schepenen binnen 's Hertogen-	#
	3 bosch, den 29. Junii 1654 Fol. 2407	#
Text Region 3	1 Placaet, tegens het vechten, trecken van Mes-	

Additional interface elements include a status bar with **In Progress** and **Save Changes** buttons.

Layout-analyse trainen P2PaLA

#Transkribus

#webinargeavanceerdNL

Geef je tekst structuur-tags (dan pas BL)

Server Overview Layout **Metadata** Tools

Document **Structural** Textual Comments

Page type:

Links:

Selected element type: TextRegion

Structure Type

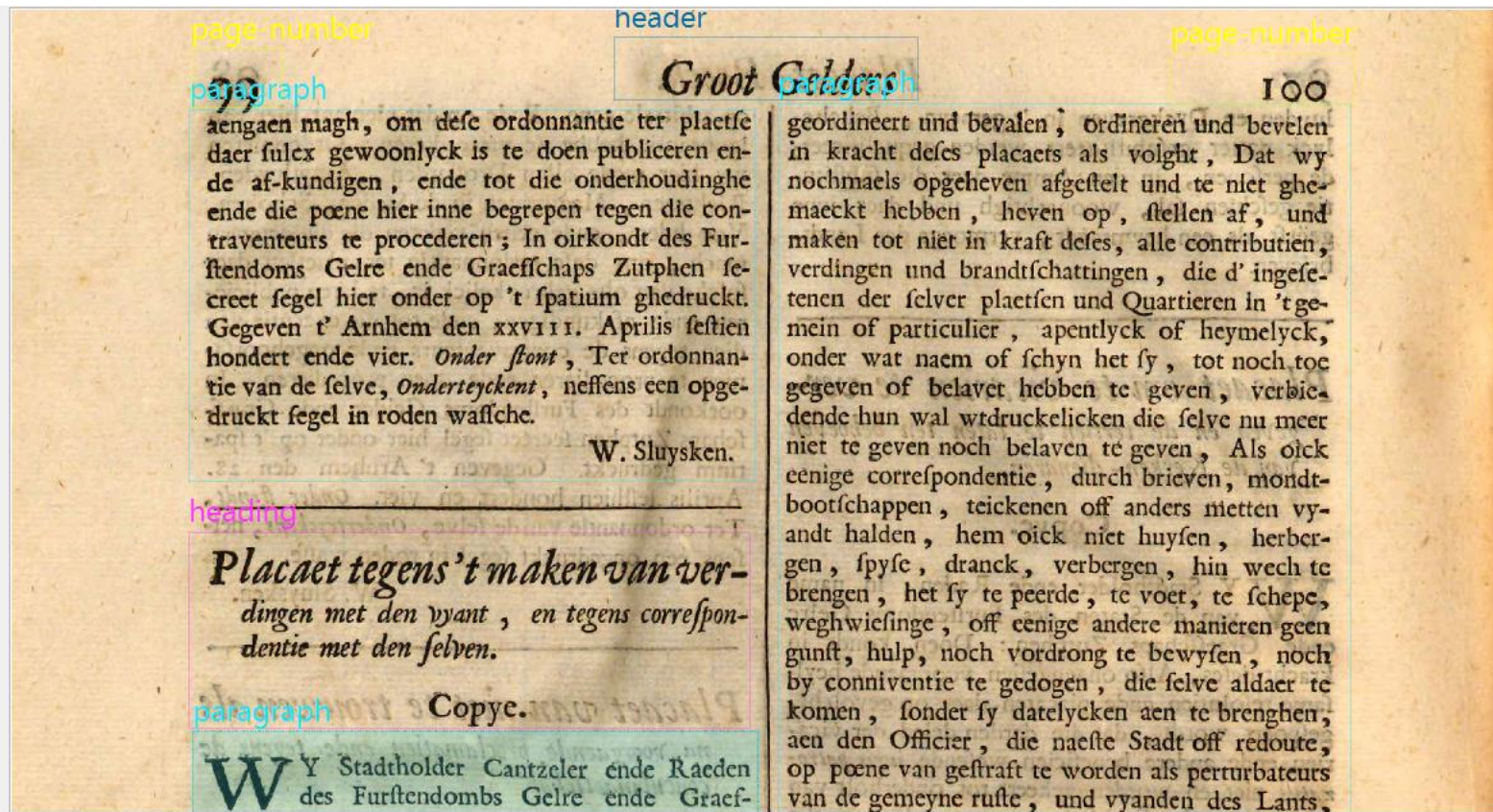
Type of selected: paragraph

Draw struct type Draw default colors

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
footer		
page-number		
drop-capital		
credit		
floating		
signature-mark		
catch-word		
marginalia		
footnote		

Layout

Type	Structure	Text	ID
TextRegion	page-number		region...
TextRegion	paragraph		region...
TextRegion	heading		region...
TextRegion	paragraph		region...



P2PaLA

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

The screenshot displays the 'Tools' tab of a software interface. It is organized into several sections:

- Layout Analysis:** Includes a 'Method' dropdown set to 'CITlab Advanced', a 'Configure...' button, radio buttons for 'Current page' (selected) and 'Pages (486): 1-486', checkboxes for 'Find Text Regions' (checked) and 'Find Lines in Text Regions' (checked), a 'Run' button, and a green button labeled 'Only use unsegmented pages'.
- Text Recognition:** Includes a 'Method' dropdown set to 'HTR (CITlab)', 'Models...' and 'Train...' buttons, and a 'Run...' button.
- Compute Accuracy:** Includes 'Reference: (Correct Text)' and 'Hypothesis: (HTR Text)' fields with corresponding text, and buttons for 'Compare Text Versions', 'Compare', and 'Compare Samples'.
- Other Tools:** Includes a highlighted 'P2PaLA...' button and a 'Text2...' button with a tooltip that reads 'Creates regions with structure tags and ba...'. Below this are radio buttons for 'Current page' (selected) and 'Pages (486): 1-486', and buttons for 'Add Baselines to Polygons' and 'Add Polygons to Baselines'.

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: 17.09.19 20:10:56 - info@caromein.nl - In Progress - CITlab HTR: Gelde
 Hypothesis: 01.10.19 12:03:07 - info@caromein.nl - In Progress - Abbyy Finereader 11
 [Compare Text Versions]
 [Compare]
 [Compare Samples]

Other Tools
 Current page Pages (486): 1-486
 [Add Baselines to Polygons]
 [Add Polygons to Baselines]

TR
L
BL
W
...
H

99
aengaen magh, om dese ordonnantie ter plaetse
daer sulcx gewoonlyck is te doen publiceren ende
af-kundigen, ende tot die onderhoudinghe
ende die poene hier inne begrepen tegen die con-
traventeurs te procederen; In oirkondt des Fur-
stendoms Gelre ende Graeffschaps Zutphen se-
creet segel hier onder op 't spatium ghedruckt.

Groot Gel

W Y Stadtholder Cantzeler ende Raeden
des Furstendoms Gelre ende Graef-

Copy.

P2PaLA structure analysis tool

Current page Pages (486): 1-486

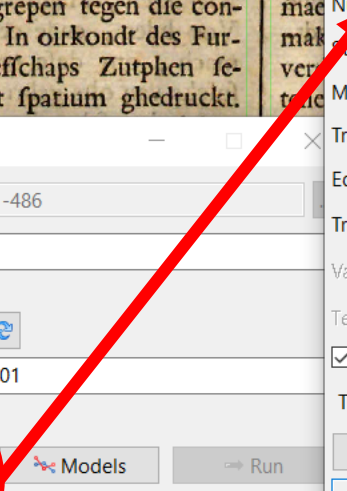
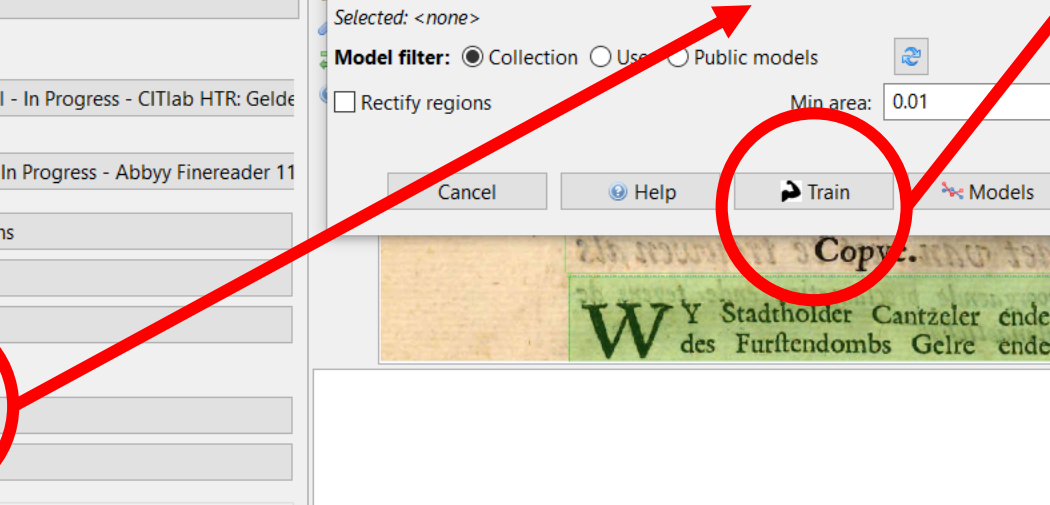
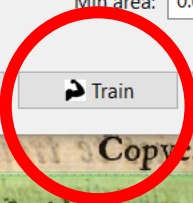
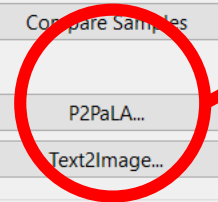
Select a model for recognition: []
 Selected: <none>

Model filter: Collection User 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]



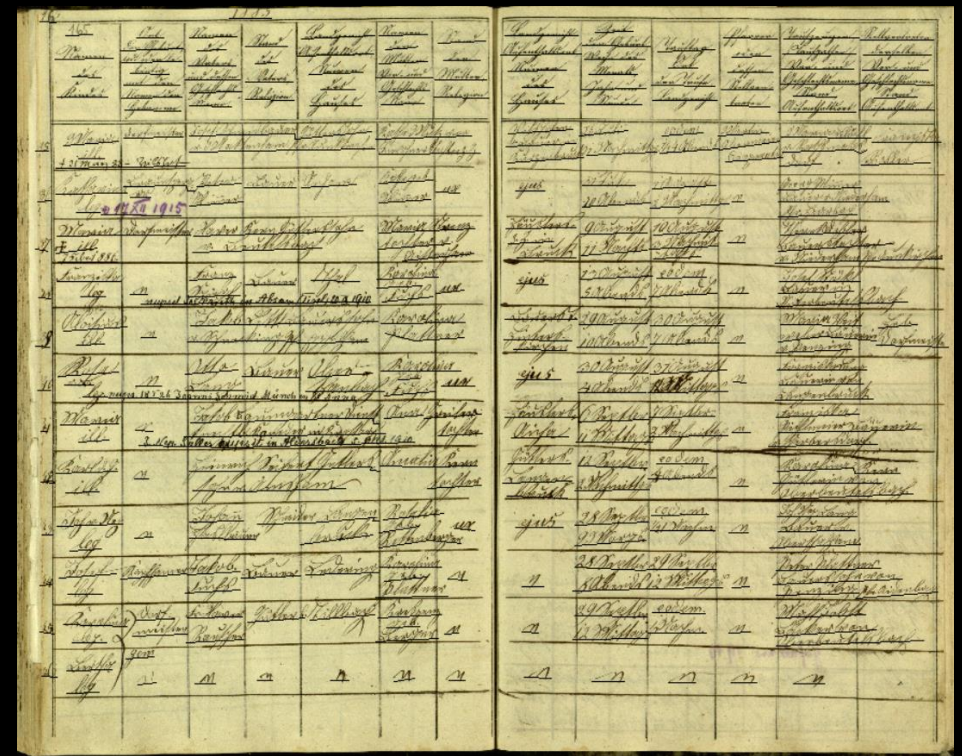
Layout-analyse trainen

- Tabellen (in ontwikkeling)
- NLE Document Understanding

#Transkribus

#webinargeavanceerdNL

@naverlabseurope



→ Information extraction

NLE Document Understanding – Naverlab Europe

Hervé Dejean

Jean-Luc Meunier



Stap 1. Tabellen markeren!

The screenshot shows the Transkribus v1.9.1 interface. The main window displays a document page with a table of handwritten text. The table has columns for 'NOMS.', 'PRÉNOMS.', 'LIEU DENAISSANCE', 'AGE', and 'ÉTUDES PR'. The table content is as follows:

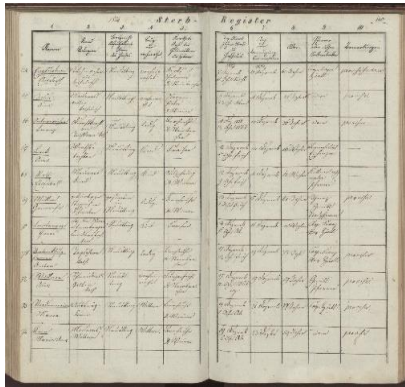
	NOMS.	PRÉNOMS.	LIEU DENAISSANCE	AGE	ÉTUDES PR
		Auguste	Francois	20	Elève
		Jean	Coulon pour de Louvain	19	Elève
		Angélique	Genève	17	Elève
		Wenceslas	Etalbrions pour de Louvain	20	Elève
		Joseph	Thourout	23	Candidat
		Adolphe	Genève	17	Elève
2302	Van Renswoude	Jean	Beersel	28	Candidat
2303	Van Affche	Louis Gerbma	Beersel	22	Elève
2304	Dequenne	Leon	Beersel	16	Elève

A red circle highlights the 'Table' option in the classification menu. A red arrow points from this circle to the 'Table' option in the expanded menu on the right. The menu includes options like Table, Printspace, Advert, Chart, Chem, Graphic, Image, LineDrawing, Maths, Music, Noise, Separator, UnknownRegion, Blackening, and Article (experimental).

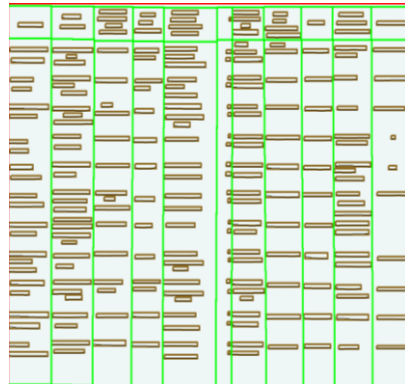
Below the table, a text box contains the following text:

Rijen en kolommen in een tabel zijn lastig: niet systematische scheidinglijn, 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)

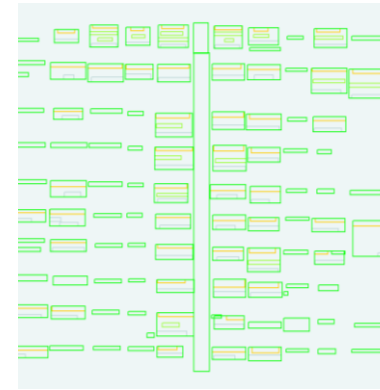


Table Understanding (NLE)

vor	Stand Lage	Lebenszeit Aufenthaltsort oder Geburtsort	Stand oder Verheiratet	Standort nach Geburtsort oder Geburtsort	Tag Monat Jahr Geburtsdatum	Tag Monat Jahr Todesdatum	Alter	Physische oder sonstige Anmerkungen	Beziehungen
1					18	18	0		
2					18	18	0		
3					18	18	0		
4					18	18	0		
5					18	18	0		
6					18	18	0		
7					18	18	0		
8					18	18	0		
9					18	18	0		
10					18	18	0		
11					18	18	0		
12					18	18	0		
13					18	18	0		
14					18	18	0		
15					18	18	0		
16					18	18	0		
17					18	18	0		
18					18	18	0		
19					18	18	0		
20					18	18	0		
21					18	18	0		
22					18	18	0		
23					18	18	0		
24					18	18	0		
25					18	18	0		
26					18	18	0		
27					18	18	0		
28					18	18	0		
29					18	18	0		
30					18	18	0		
31					18	18	0		
32					18	18	0		
33					18	18	0		
34					18	18	0		
35					18	18	0		
36					18	18	0		
37					18	18	0		
38					18	18	0		
39					18	18	0		
40					18	18	0		
41					18	18	0		
42					18	18	0		
43					18	18	0		
44					18	18	0		
45					18	18	0		
46					18	18	0		
47					18	18	0		
48					18	18	0		
49					18	18	0		
50					18	18	0		
51					18	18	0		
52					18	18	0		
53					18	18	0		
54					18	18	0		
55					18	18	0		
56					18	18	0		
57					18	18	0		
58					18	18	0		
59					18	18	0		
60					18	18	0		
61					18	18	0		
62					18	18	0		
63					18	18	0		
64					18	18	0		
65					18	18	0		
66					18	18	0		
67					18	18	0		
68					18	18	0		
69					18	18	0		
70					18	18	0		
71					18	18	0		
72					18	18	0		
73					18	18	0		
74					18	18	0		
75					18	18	0		
76					18	18	0		
77					18	18	0		
78					18	18	0		
79					18	18	0		
80					18	18	0		
81					18	18	0		
82					18	18	0		
83					18	18	0		
84					18	18	0		
85					18	18	0		
86					18	18	0		
87					18	18	0		
88					18	18	0		
89					18	18	0		
90					18	18	0		
91					18	18	0		
92					18	18	0		
93					18	18	0		
94					18	18	0		
95					18	18	0		
96					18	18	0		
97					18	18	0		
98					18	18	0		
99					18	18	0		
100					18	18	0		

Handwritten Text Recognition (URO)

```
<RECORD lastname="Doxleitner" firstname="Elisabeth" occupation="Wuchmacner" location="Neuwötting" situation="verheiratet" deathreason="Unierine de" age="64" />
<RECORD lastname="Kastl" firstname="Anna" location="Neuwötting" situation="verheiratet" deathreason="Fieber" doktor="der" deathDate="Dezember" age="15" />
<RECORD lastname="Seternicher" firstname="Lorenz" occupation="Dienstknecht" religion="kath" location="Neuwötting" situation="ledig" deathreason="Brechruhr"
doktor="Steinbae" age="26" />
<RECORD lastname="Denk" firstname="Anna" occupation="Wirth-" location="Neuwötting" situation="Kind" deathreason="Frasen" doktor="In" age="16" />
<RECORD firstname="Elisabeth" occupation="Maurer-" location="Neuwötting" situation="Kind" deathreason="Auszehrung" doktor="Wiiher" age="13" />
<RECORD lastname="Müller" firstname="Genovefa" location="Neuwötting" situation="ledig" doktor="Wiiher" deathDate="Dezember" age="61" />
<RECORD lastname="Lemberger" firstname="Theres" location="Neuwötting" situation="Kind" deathreason="Frasen" age="120" />
<RECORD lastname="Sammer" firstname="Anton" occupation="Tagelöhner" religion="kath" location="Neuwötting" situation="ledig" deathreason="Brechruhr"
doktor="Steinbne" />
<RECORD lastname="Wallner" firstname="Afa" occupation="Schneiders- Jattin" religion="kath" location="Lang" situation="verhei" deathreason="Lungensucht"
doktor="Scteimbre-" deathDate="Dezember" age="57" />
<RECORD lastname="Vordenmaier" firstname="Maria" occupation="Austrägler" location="Nerdötting" deathreason="Brechruhr" doktor="Wiiher" deathDate="Dezember"
age="17" />
<RECORD lastname="Kern" firstname="Maria Anna" occupation="Maurer Mittwe" location="Neuwötting" doktor="Wiiher" deathDate="Dezember" age="569" />
</PAGE>
```

Information Extraction (NLE)



Proces van Document Understanding

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

OCR/HTR



Raw text

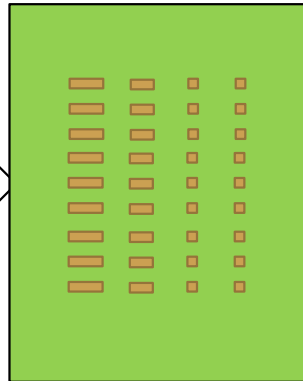
Final Grades - CSC 101

Student	Final Exam	Midterm	Homework	Lab	Project	Total
1001	85	70	90	80	95	85
1002	75	60	80	70	85	75
1003	90	80	95	85	90	90
1004	60	50	70	60	70	60
1005	80	70	85	75	85	80
1006	70	65	75	65	75	70
1007	85	75	90	80	85	85
1008	65	55	65	55	65	65
1009	95	85	95	85	90	90
1010	70	60	70	60	70	70

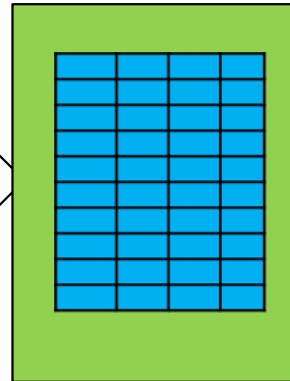


Our Input

Text, position



Layout Structure



Information (database)

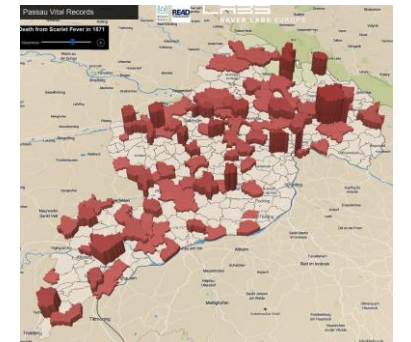
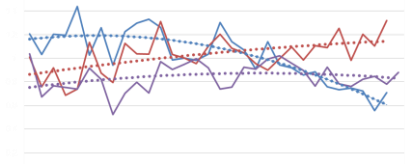
```

avdrental# select title, release_year, length, replacement_cost from film
avdrental# where length > 100 and replacement_cost > 20.00
avdrental# order by title desc;

```

title	release_year	length	replacement_cost
West Lion	2006	159	29.99
Virgin Daisy	2006	178	29.99
Unsub	2006	172	29.99
Tracy Cider	2006	142	29.99
Song Holiday	2006	106	29.99
Slacker Liaisons	2006	179	29.99
Sassy Packer	2006	154	29.99
River Outlaw	2006	149	29.99
Right Graves	2006	153	29.99
Quest Hexaflex	2006	177	29.99
Posedon Forever	2006	159	29.99
Loadin' Legally	2006	148	29.99
Lawless Vision	2006	181	29.99
Jingle Sagebrush	2006	124	29.99
Jetsmo Walk	2006	171	29.99
Japanese Run	2006	136	29.99
Gi'lore Boiled	2006	163	29.99
Flatts Garden	2006	146	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
Close Theory	2006	159	29.99
Clockwork Paradise	2006	143	29.99
Baltimore Mockingbird	2006	173	29.99

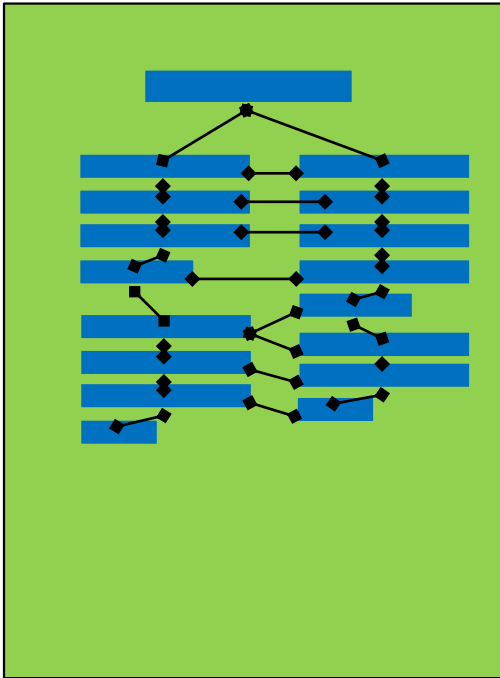
Analysis



Aanpak: één tool voor Document Layout/Understanding

Graph Creation

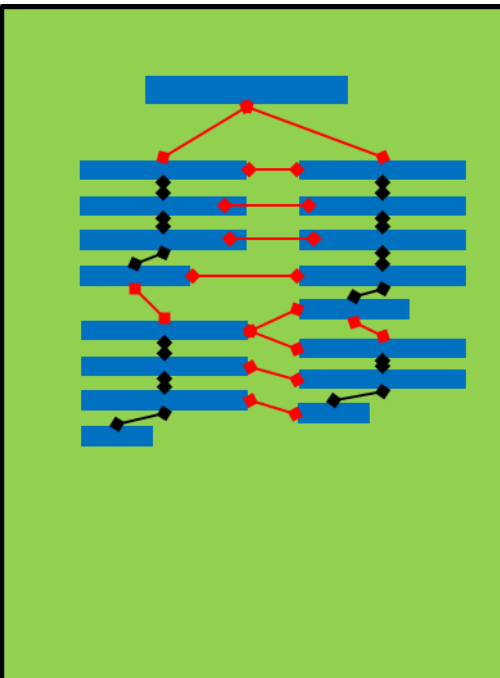
- Nodes: textlines
- Edges: positional relations



Page Layout

Segmentation

- Edge classification



Document Reconstruction

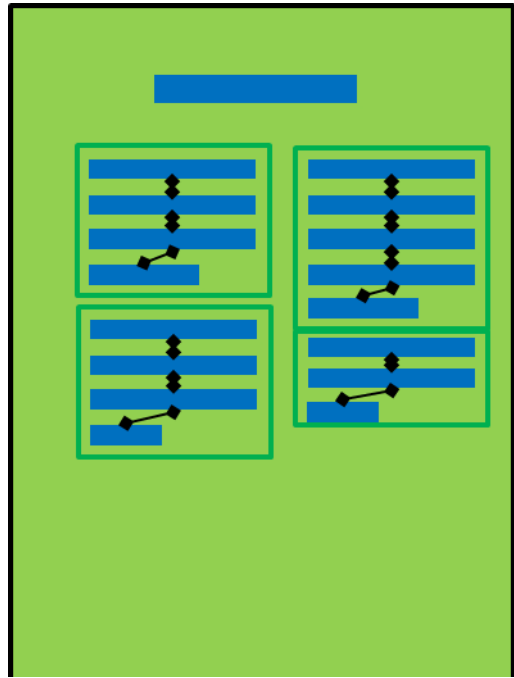
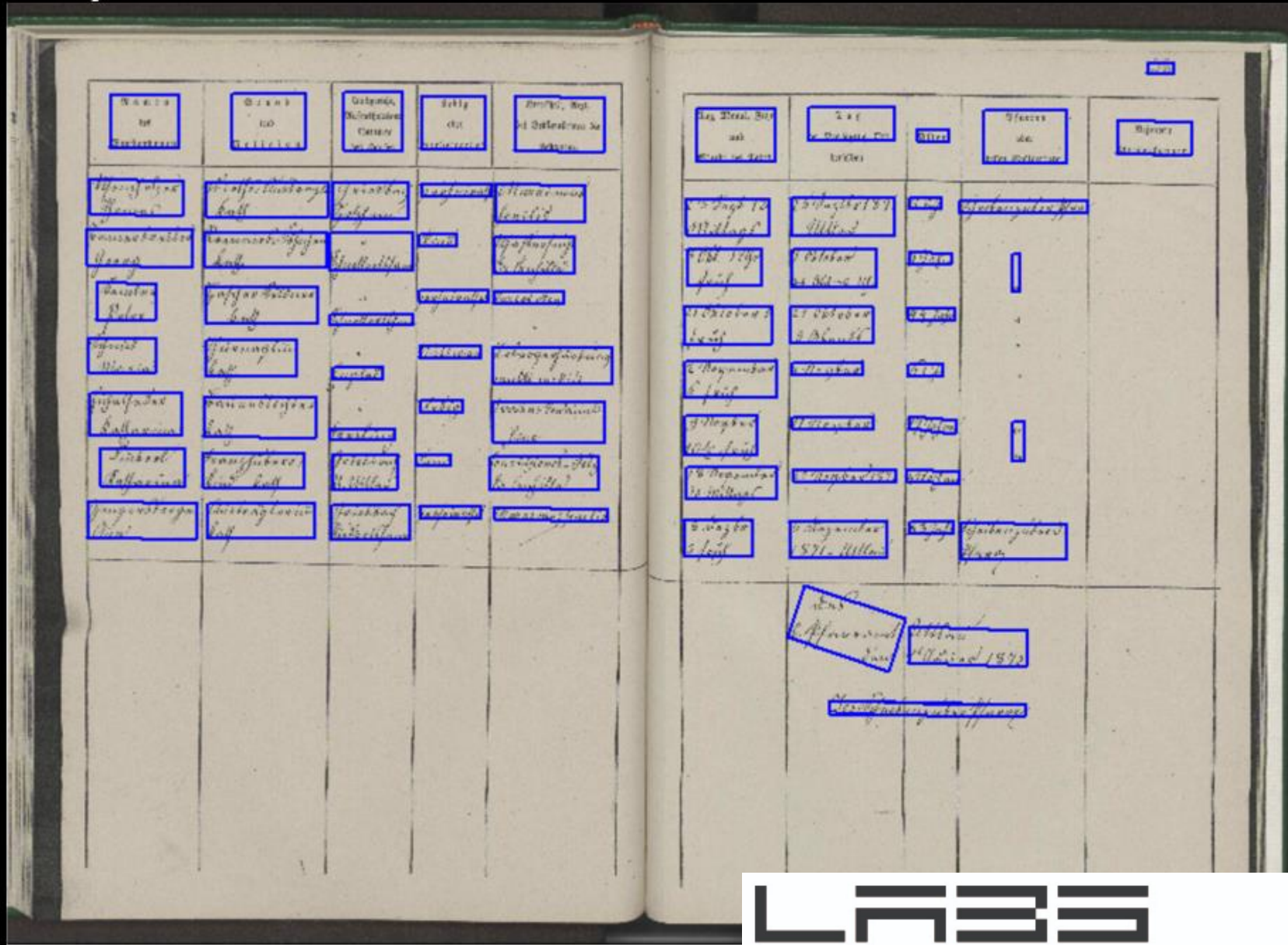


Table Understanding (Cells)

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

Table Understanding (Rows)

Training size : 990 pages

Test size : 48 pages

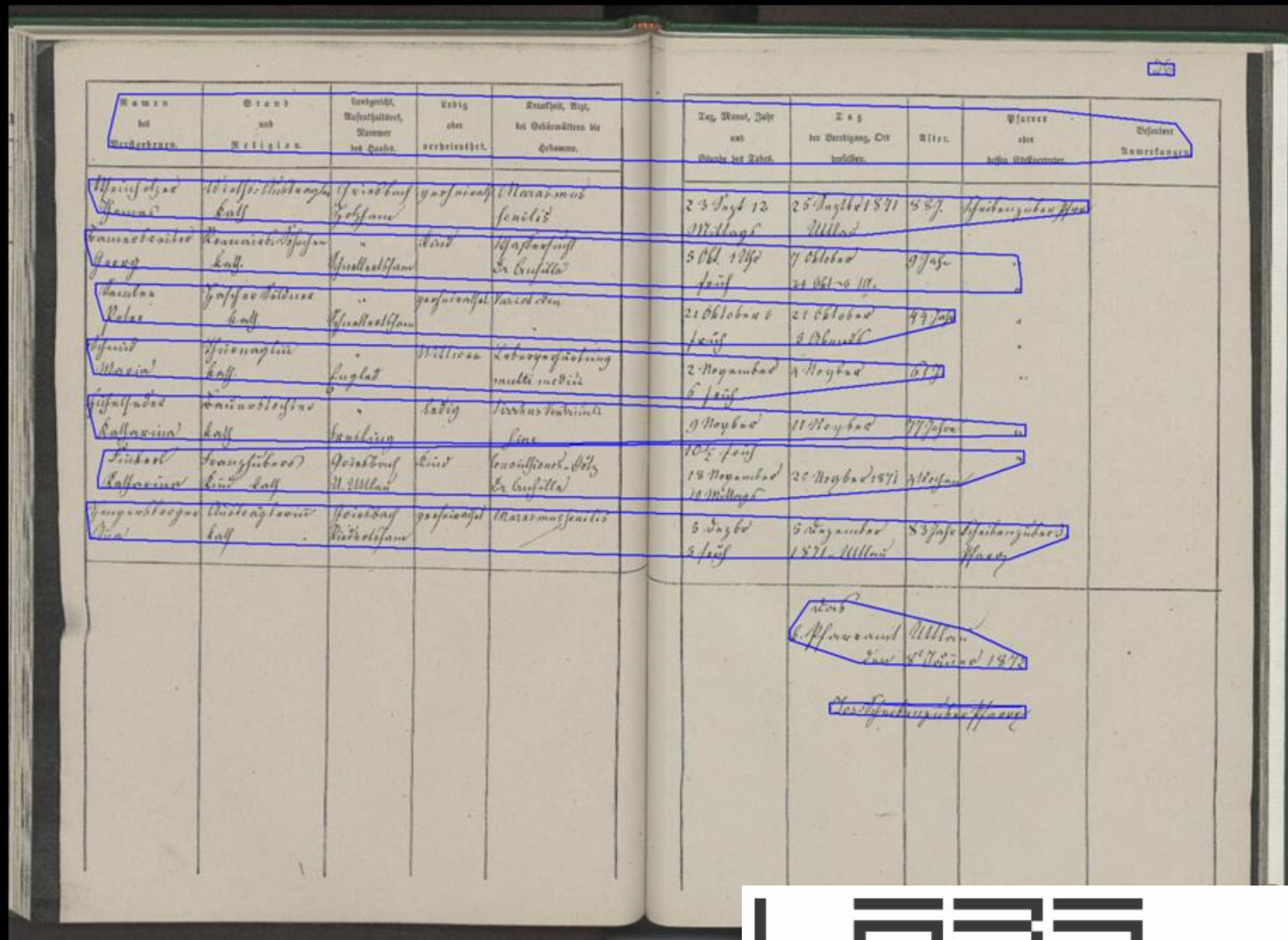
Training time : 3h

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



<https://data.remote.be/kijpe.nasveelab.com/>

#Transkribus
#webinargeavanceerdNL

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](https://transkribus.eu/wiki/index.php/How_to_Guides)

Youtube Channel en Tutorials:

<https://www.youtube.com/channel/UC-txVgM31rDTGIBnH-zpPjA>

Facebook:

@Transkribus Users (user based)

Web-Interface:

<https://transkribus.eu/r/read/projects/>

DocScan en ScanTent:

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

6-7 February 2020
transkribus user conference
in Innsbruck.

<https://bit.ly/2QhnpjX0>



Bedankt voor uw aandacht!