Digi-workshop #3
Transkribus: updates.

By Dr. C.A. (Annemieke) Romein
(Huygens ING)

Digital workshop, 9 October 2020

Noon CEST

Pylaia

- What is Pylaia and does it differ from HTR(+).
- How to use it.

Tables and LA

- Developments in the field of tables/ how to add seperation lines in your LA.

Small regions

- How to remove small regions (noise).

Sample sets

- How to create a sample set (to function as a sturdy basis for your models).

Other updates

- Small update on P2PaLA
- Read and Search
- Webversion
- masking text

Pylaia

- In this section you will learn more about Pylaia.

HTR → HTR(+) and Pylaia

- Keep in mind **HTR stands for Handwriting Text Recognition**.

  From that, there are different ‘techniques’ that are being used for the recognition. These have (confusingly enough) similar names:

  1. HTR: original tool with RNN to recognise the handwriting
  2. HTR+: faster tool. Created by Planet (Germany) and licensed to Transkribus.
     This will be phased out next year, replaced by IDA (which will be faster than HTR+).
  3. Pylaia: open source tool by the University of Valencia (no license, hence the lower costs).

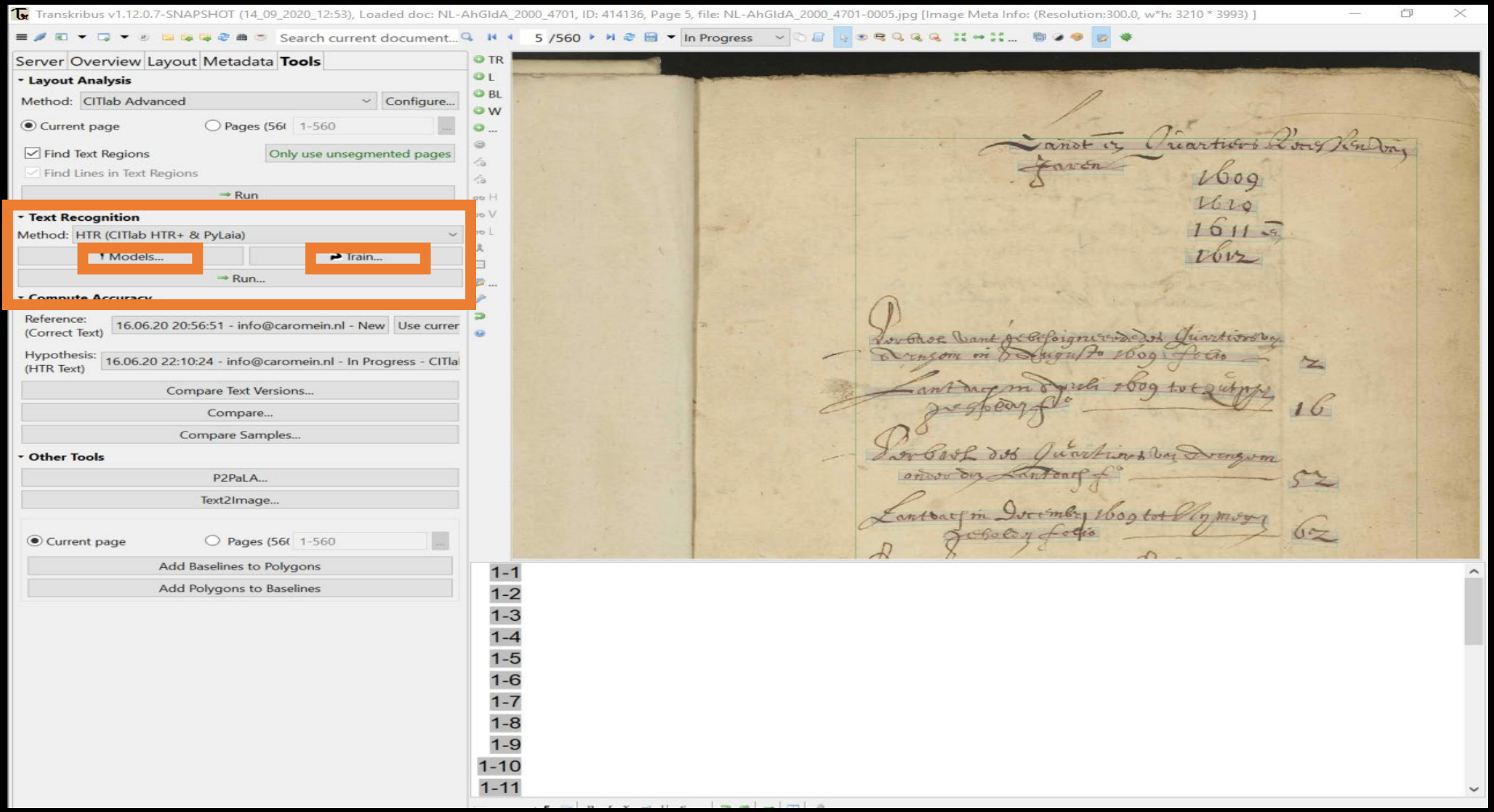
Pylaia

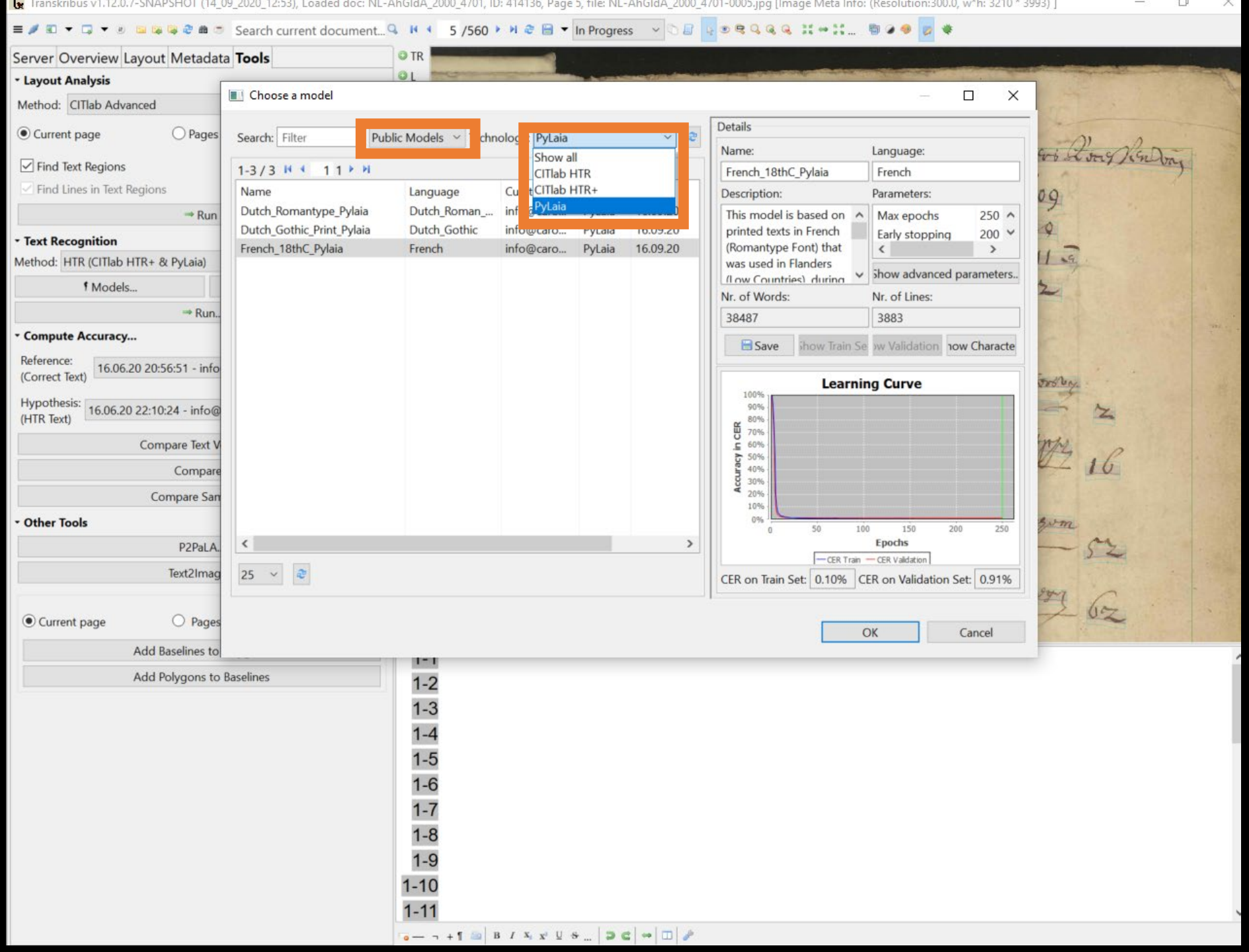
- If you have **curved lines (or pages) HTR+ generally works better**; but this will improve in pylaia.
- **Pylaia models cannot be used for Text2Image**, you'll need HTR+.
- Pylaia is PyTorch based (aka Machinelearning Library).
- Benefits Pylaia: you can adjust more Parameters and the Netstructure can be changed (if you are familiar with machine learning). More can be found on: https://github.com/jpuigcerver/pylaia
- Same amount of pages necessary; but preprocessing is important.
- You can train Pylaia to fill out abbreviations by using $ as a sign between the abbreviated form and the filled out form. *new sign will come!*
- Pylaia can export word-coordinates, which is interesting if you want to be able to show the exact word location in an image/text (e.g. IIIF).

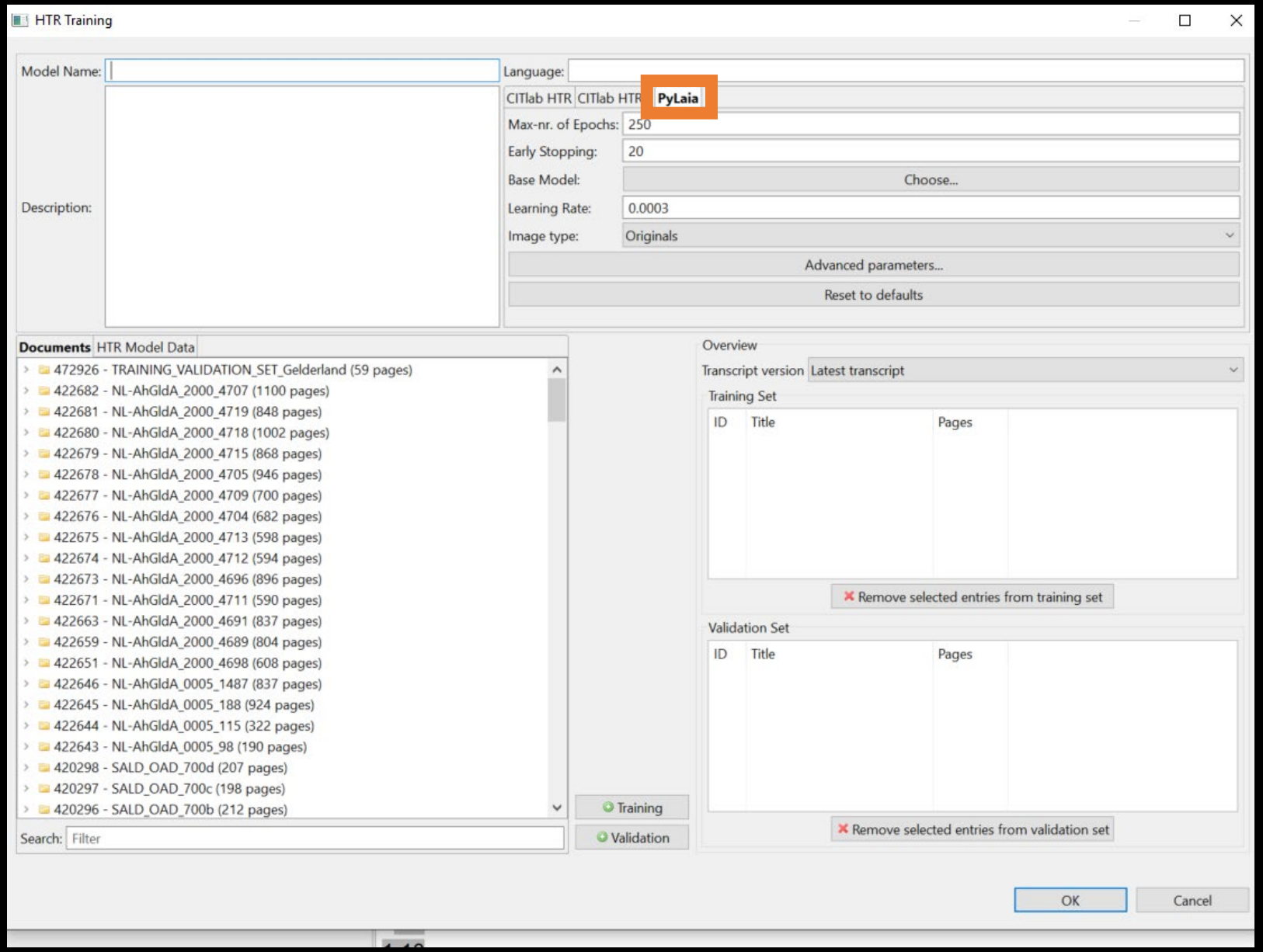
Trainingsbutton

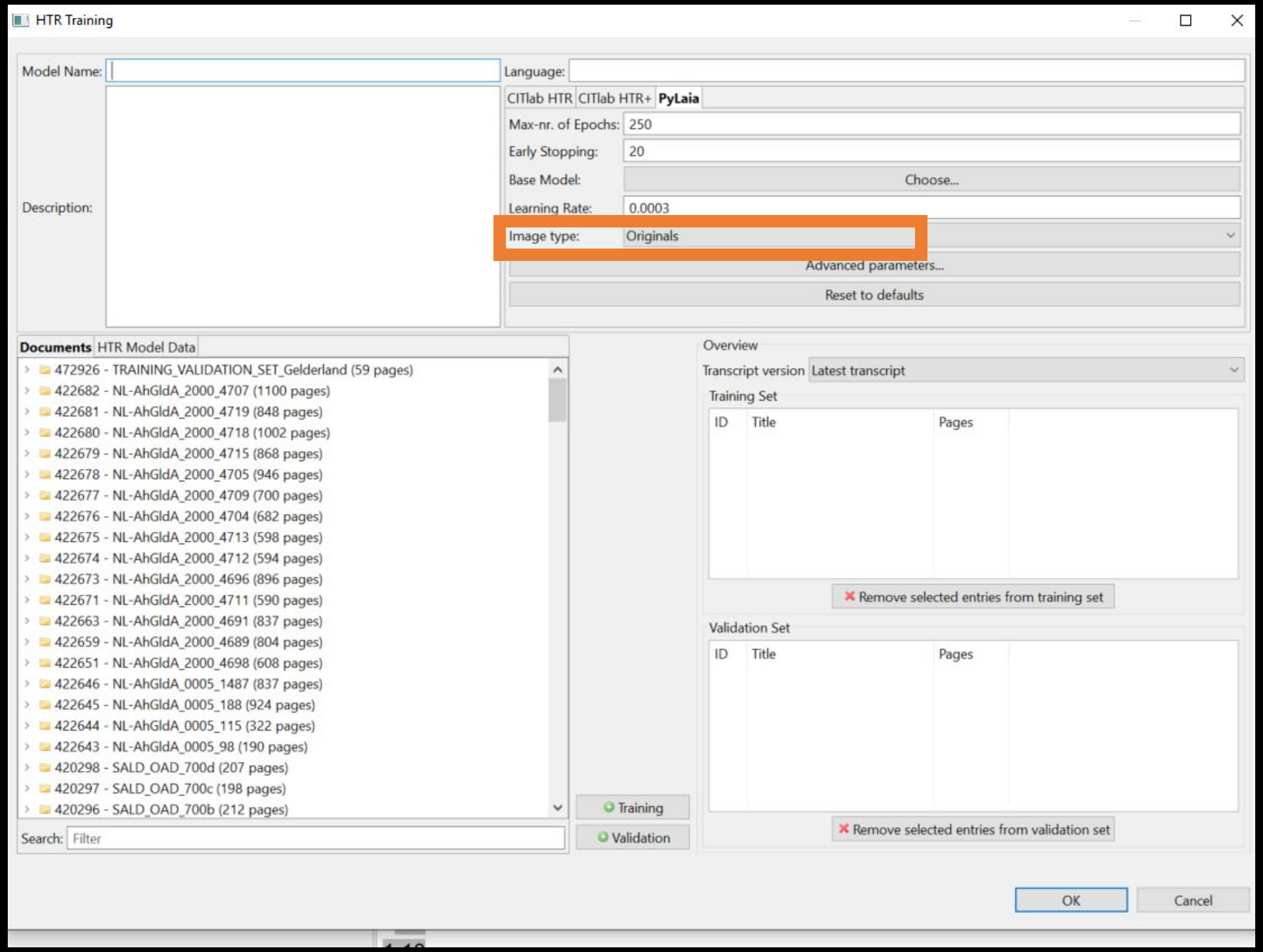
- In order to train Pylaia-models you will need – like with HTR+ - ask the Read-Coop team to provide you with the button.
  - Info@readcoop.eu

  This is because not everyone is going to be using this.


Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701-0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)]
Search current document...
5 /560
In Progress
Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method: CITlab Advanced
Configure...
Current page
Pages (56
1-560
Find Text Regions
Only use unsegmented pages
Find Lines in Text Regions
Run
Text Recognition
Method: HTR (CITlab HTR+ & PyLaia)
Models...
Train...
Run...
Compute Accuracy
Reference: (Correct Text)
16.06.20 20:56:51 - info@caromein.nl - New
Use currer
Hypothesis: (HTR Text)
16.06.20 22:10:24 - info@caromein.nl - In Progress - CITla
Compare Text Versions...
Compare...
Compare Samples...
Other Tools
P2PaLA...
Text2Image...
Current page
Pages (56
1-560
Add Baselines to Polygons
Add Polygons to Baselines
TR
L
BL
W
1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11

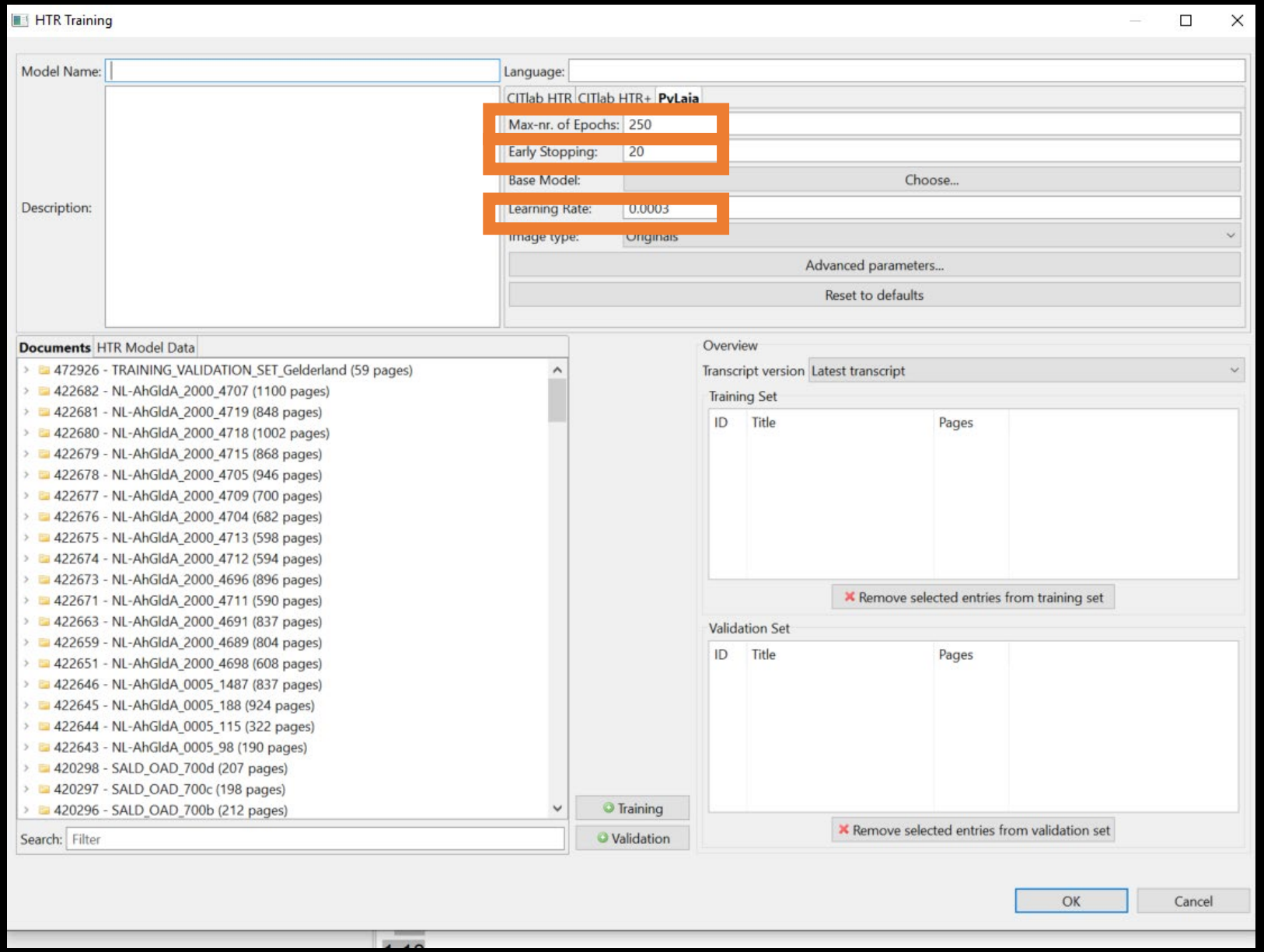

Choose a model
Search:
Filter
Public Models
PyLaia
Show all
CITlab HTR
CITlab HTR+
PyLaia
1-3 / 3
Name
Language
Dutch_Romantype_Pylaia
Dutch_Roman_...
Dutch_Gothic_Print_Pylaia
Dutch_Gothic
French_18thC_Pylaia
French
info@caro...
PyLaia
16.09.20
Details
Name:
French_18thC_Pylaia
Language:
French
Description:
This model is based on printed texts in French (Romantype Font) that was used in Flanders
Parameters:
Max epochs 250
Early stopping 200
Show advanced parameters...
Nr. of Words:
38487
Nr. of Lines:
3883
Save
Learning Curve
Accuracy in CER
Epochs
CER Train
CER Validation
CER on Train Set: 0.10%
CER on Validation Set: 0.91%
OK
Cancel
Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method: CITlab Advanced
Current page
Pages
Find Text Regions
Find Lines in Text Regions
Run
Text Recognition
Method: HTR (CITlab HTR+ & PyLaia)
Models...
Run...
Compute Accuracy...
Reference: (Correct Text)
Hypothesis: (HTR Text)
Compare Text V
Compare
Compare San
Other Tools
P2PaLA..
Text2Imag
Add Baselines to
Add Polygons to Baselines
1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11


HTR Training
Model Name:
Language:
CITlab HTR
CITlab HTR
PyLaia
Max-nr. of Epochs:
250
Early Stopping:
20
Base Model:
Choose...
Description:
Learning Rate:
0.0003
Image type:
Originals
Advanced parameters...
Reset to defaults
Documents
HTR Model Data
472926 - TRAINING_VALIDATION_SET_Gelderland (59 pages)
422682 - NL-AhGldA_2000_4707 (1100 pages)
422681 - NL-AhGldA_2000_4719 (848 pages)
422680 - NL-AhGldA_2000_4718 (1002 pages)
422679 - NL-AhGldA_2000_4715 (868 pages)
422678 - NL-AhGldA_2000_4705 (946 pages)
422677 - NL-AhGldA_2000_4709 (700 pages)
422676 - NL-AhGldA_2000_4704 (682 pages)
422675 - NL-AhGldA_2000_4713 (598 pages)
422674 - NL-AhGldA_2000_4712 (594 pages)
422673 - NL-AhGldA_2000_4696 (896 pages)
422671 - NL-AhGldA_2000_4711 (590 pages)
422663 - NL-AhGldA_2000_4691 (837 pages)
422659 - NL-AhGldA_2000_4689 (804 pages)
422651 - NL-AhGldA_2000_4698 (608 pages)
422646 - NL-AhGldA_0005_1487 (837 pages)
422645 - NL-AhGldA_0005_188 (924 pages)
422644 - NL-AhGldA_0005_115 (322 pages)
422643 - NL-AhGldA_0005_98 (190 pages)
420298 - SALD_OAD_700d (207 pages)
420297 - SALD_OAD_700c (198 pages)
420296 - SALD_OAD_700b (212 pages)
Search:
Filter
Training
Validation
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

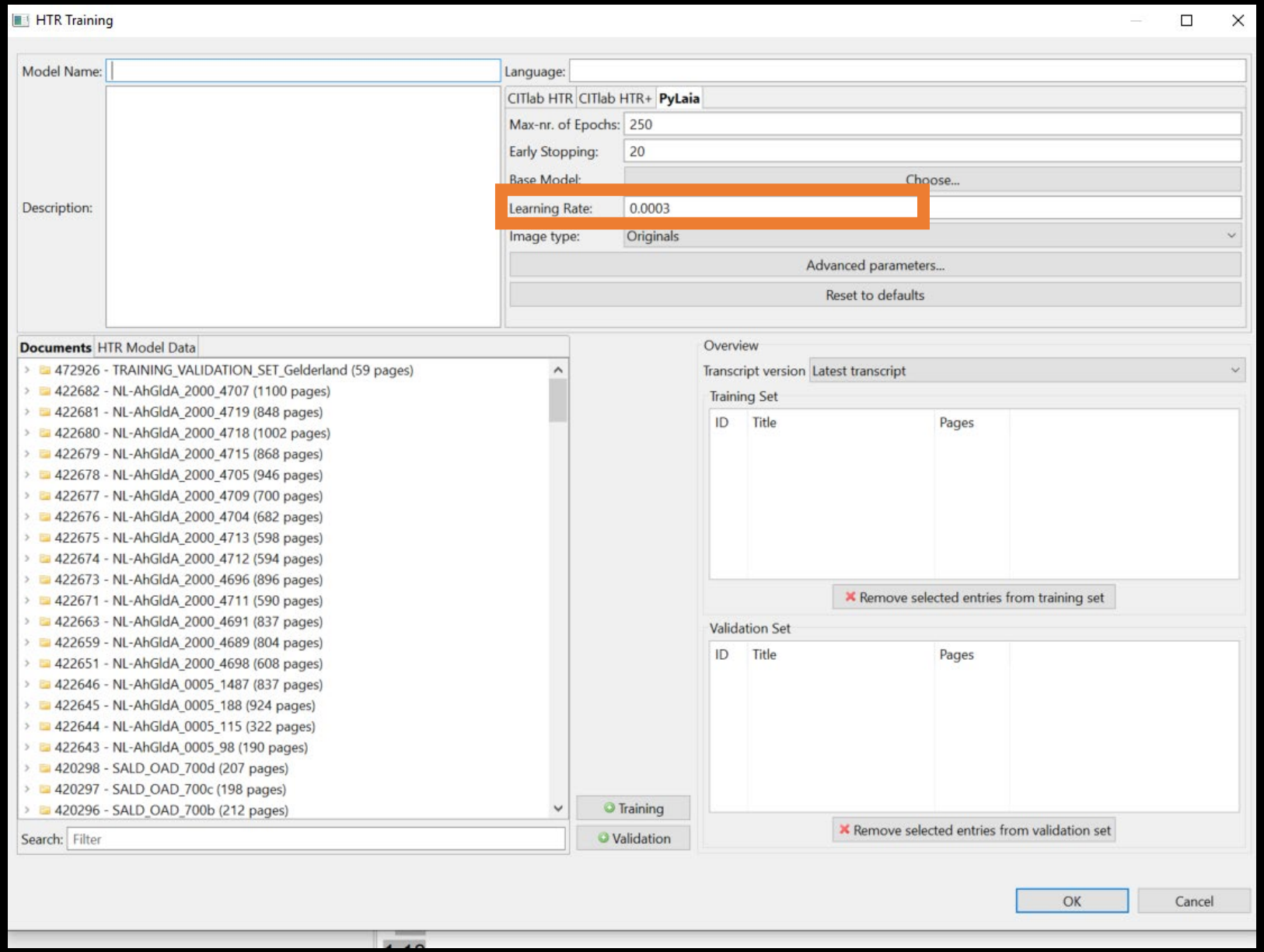


- Image type – do you want to have the model work on original or compressed?
  - If it takes to long to create, stop the training and use the compressed! (Not that much influence on the result.)

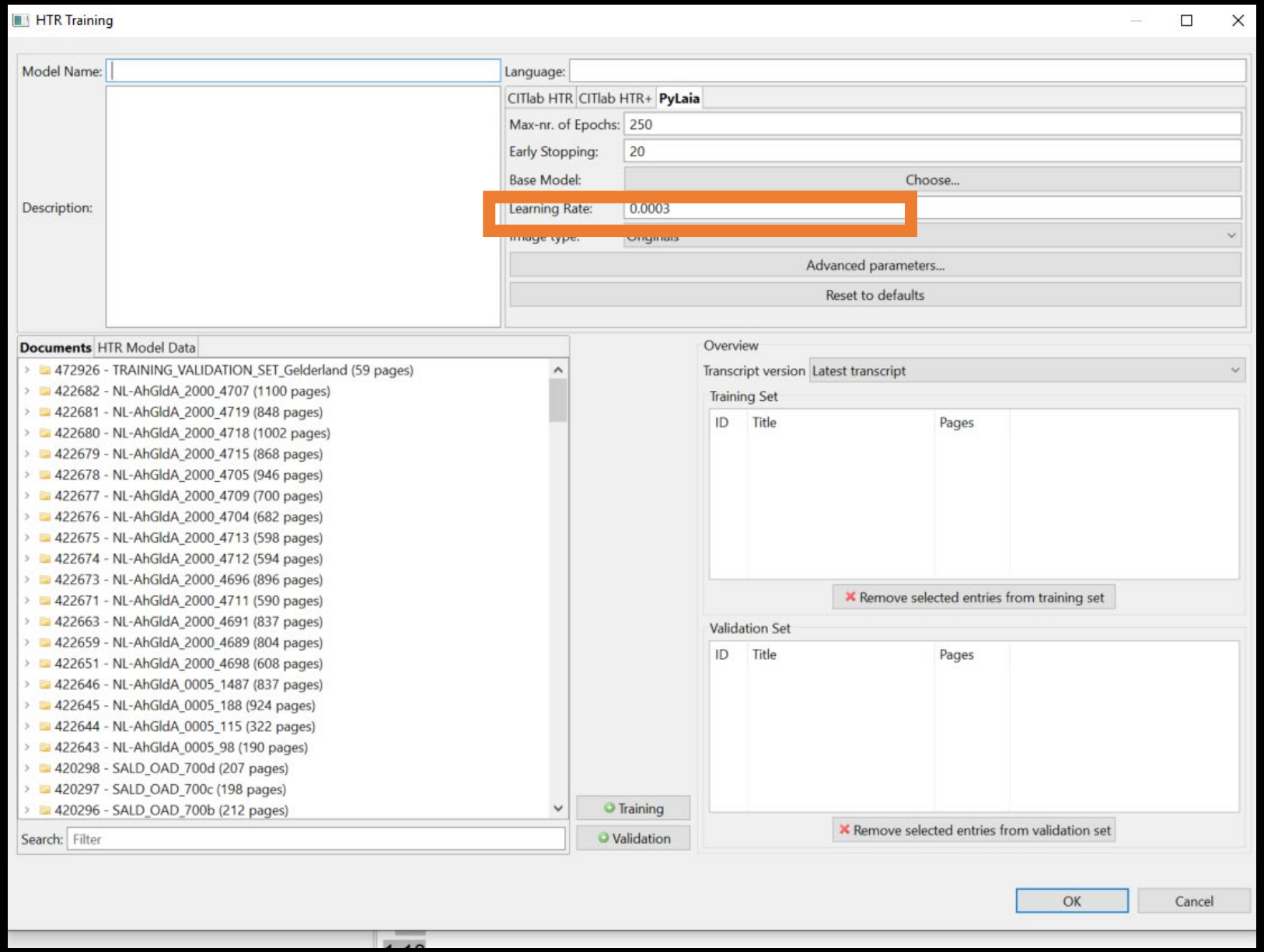
- Number of epochs (same as in HTR+)



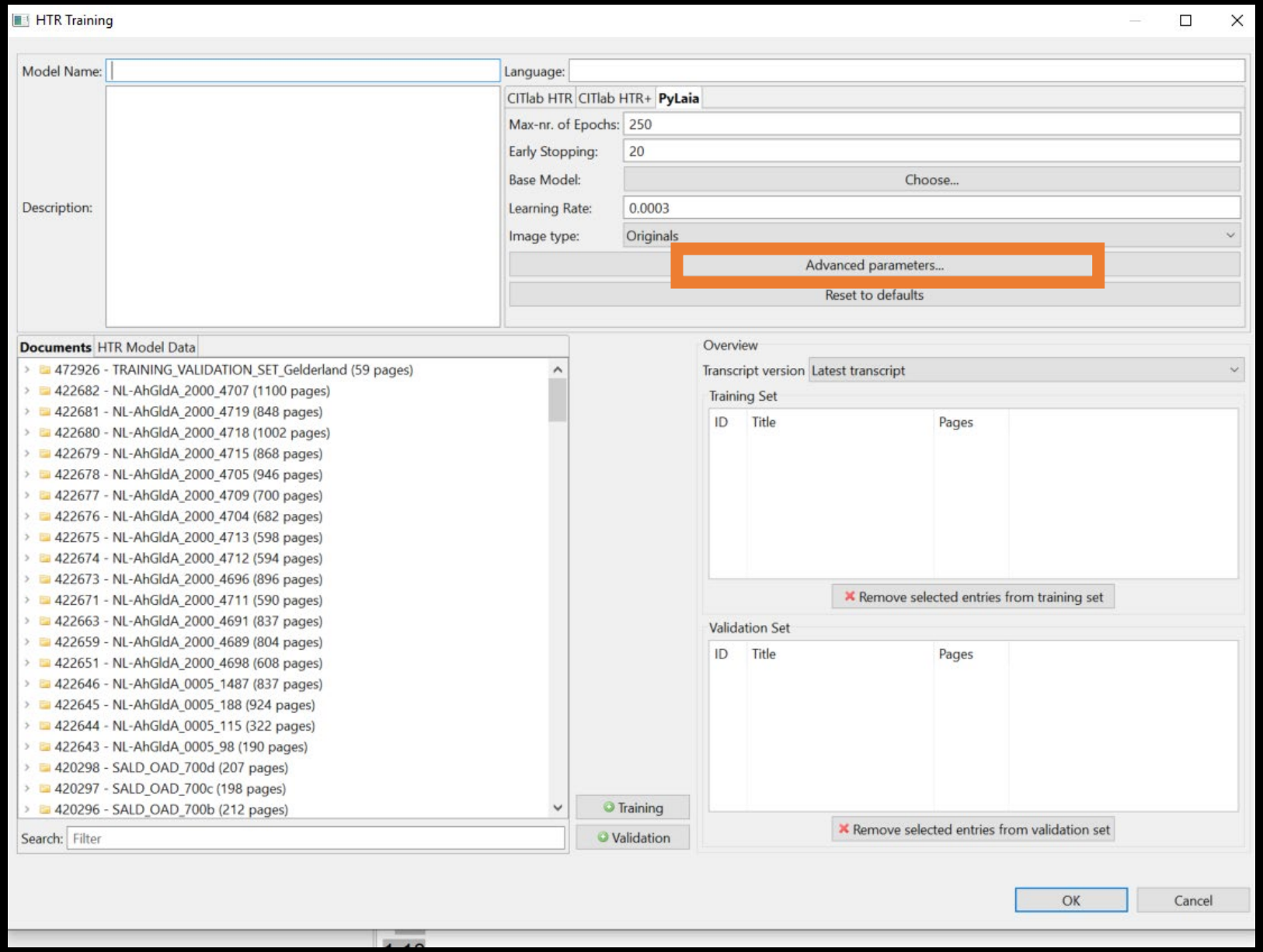
- Early stopping & learning rate combined
- → when the improvement on the validationset is too small, the system stops. (with very small validation sets: make this number a bit higher).

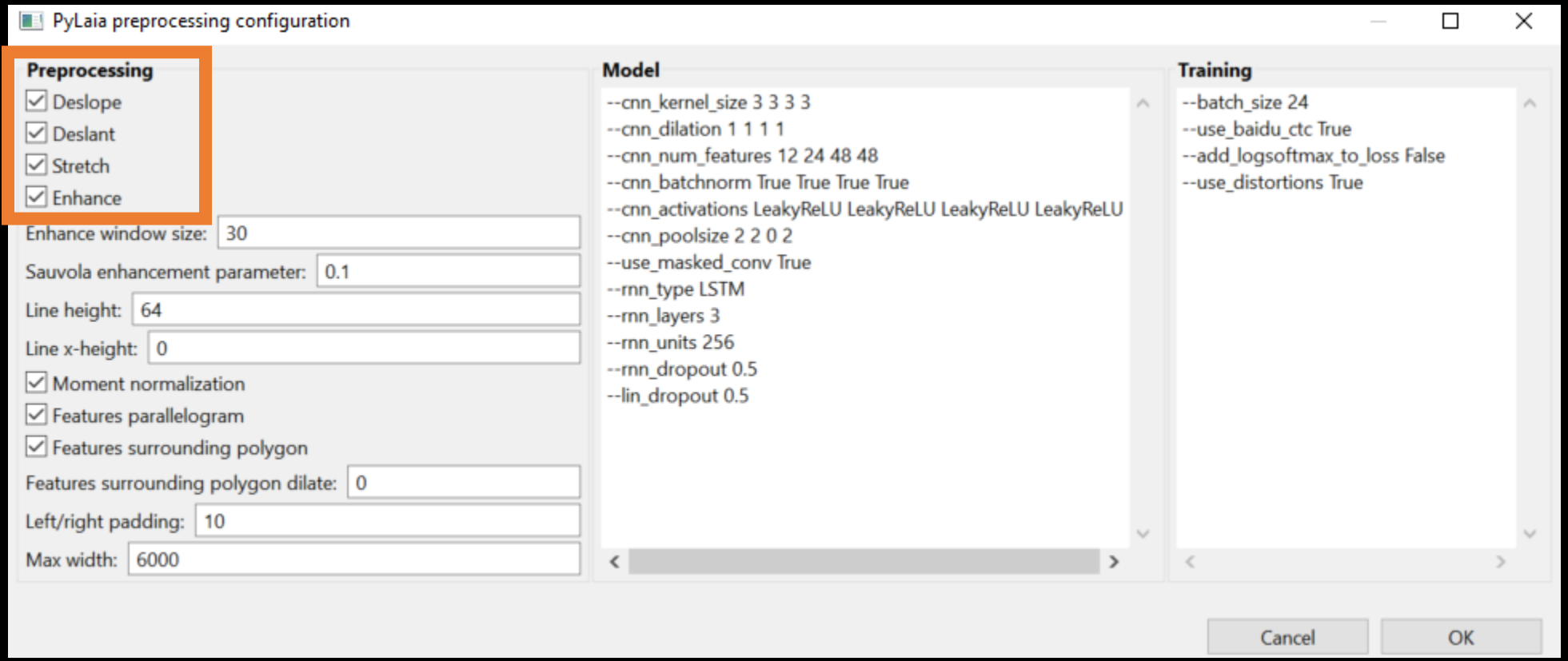


- Learning rate: the worth between each step in the training. If the steps are large, the CER will drop quicker, but is less accurate. **Leave it at default if you are not experienced.**

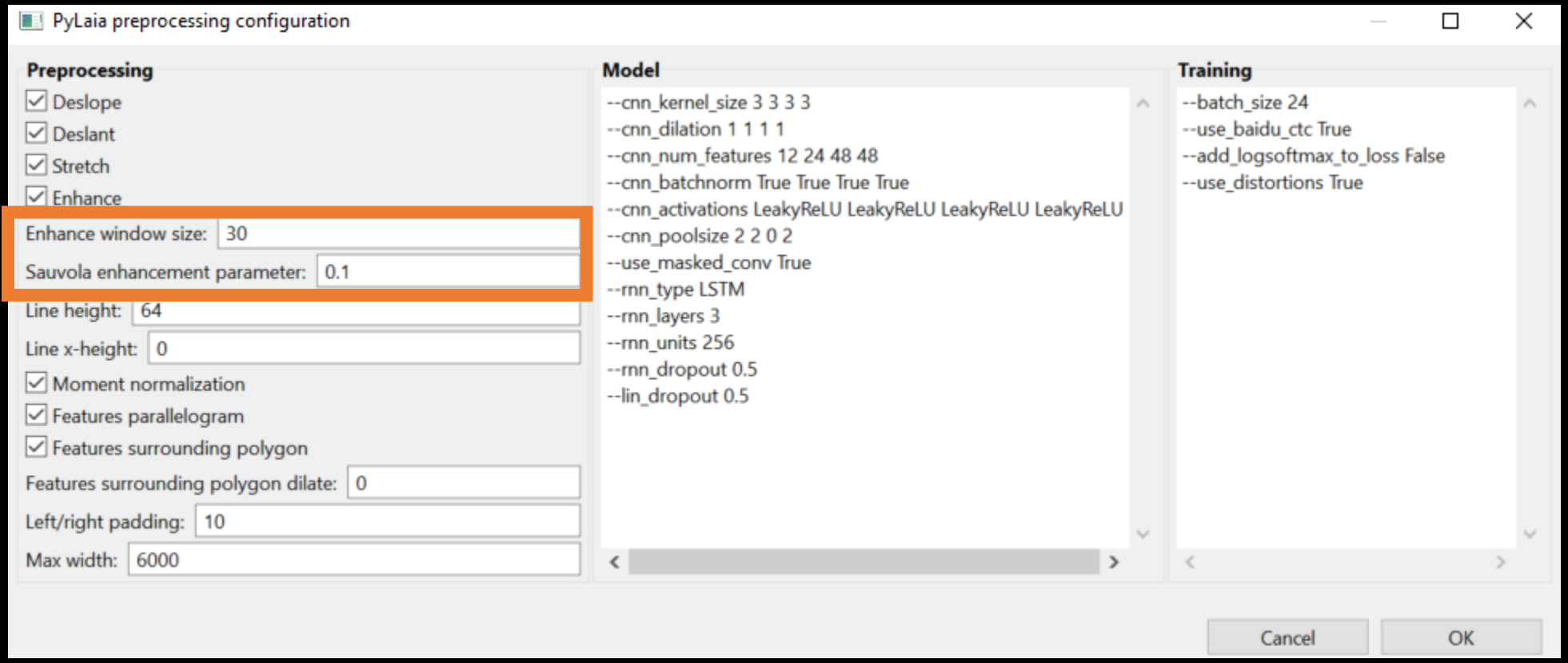


- Learning rate: the worth between each step in the training. If the steps are large, the CER will drop quicker, but is less accurate. **Leave it at default if you are not experienced.**

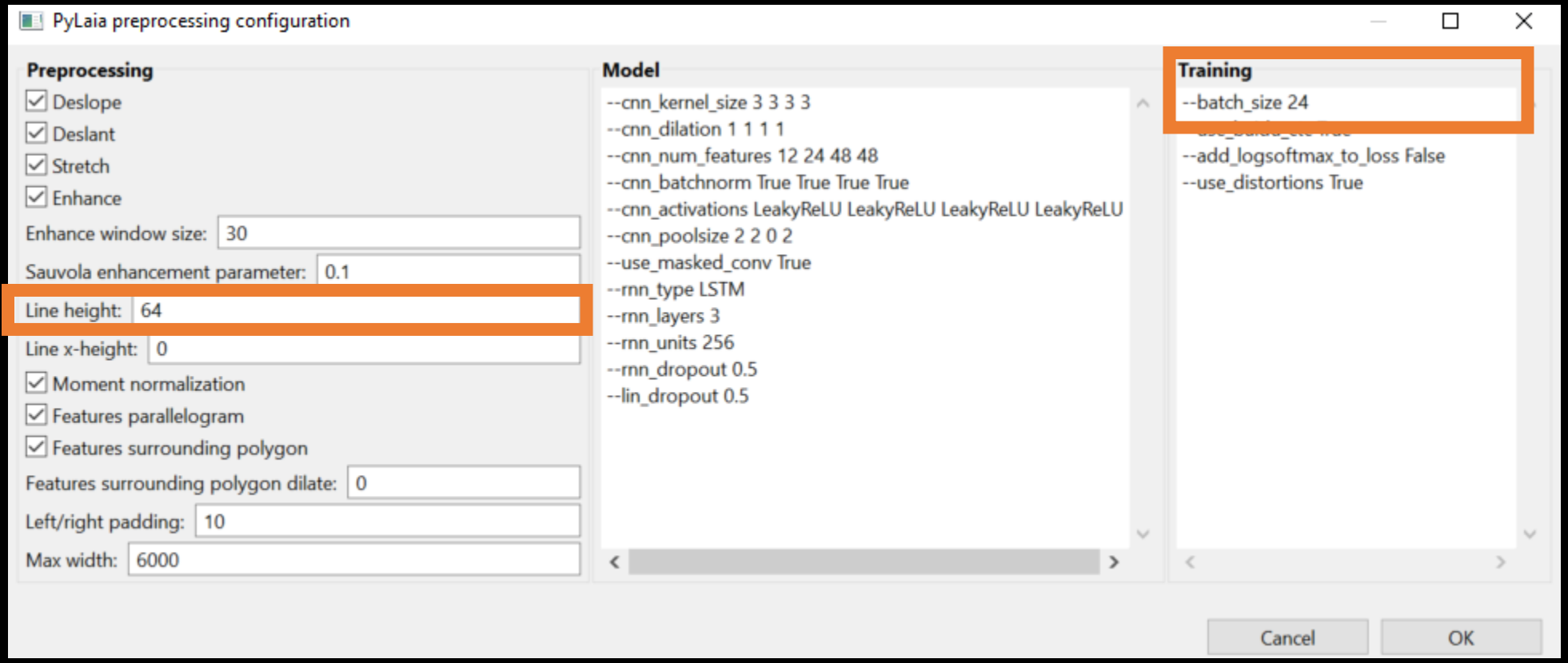

HTR Training
Model Name:
Language:
Description:
CITlab HTR
CITlab HTR+
PyLaia
Max-nr. of Epochs:
250
Early Stopping:
20
Base Model:
Choose...
Learning Rate:
0.0003
Image type:
Originals
Advanced parameters...
Reset to defaults
Documents
HTR Model Data
472926 - TRAINING_VALIDATION_SET_Gelderland (59 pages)
422682 - NL-AhGldA_2000_4707 (1100 pages)
422681 - NL-AhGldA_2000_4719 (848 pages)
422680 - NL-AhGldA_2000_4718 (1002 pages)
422679 - NL-AhGldA_2000_4715 (868 pages)
422678 - NL-AhGldA_2000_4705 (946 pages)
422677 - NL-AhGldA_2000_4709 (700 pages)
422676 - NL-AhGldA_2000_4704 (682 pages)
422675 - NL-AhGldA_2000_4713 (598 pages)
422674 - NL-AhGldA_2000_4712 (594 pages)
422673 - NL-AhGldA_2000_4696 (896 pages)
422671 - NL-AhGldA_2000_4711 (590 pages)
422663 - NL-AhGldA_2000_4691 (837 pages)
422659 - NL-AhGldA_2000_4689 (804 pages)
422651 - NL-AhGldA_2000_4698 (608 pages)
422646 - NL-AhGldA_0005_1487 (837 pages)
422645 - NL-AhGldA_0005_188 (924 pages)
422644 - NL-AhGldA_0005_115 (322 pages)
422643 - NL-AhGldA_0005_98 (190 pages)
420298 - SALD_OAD_700d (207 pages)
420297 - SALD_OAD_700c (198 pages)
420296 - SALD_OAD_700b (212 pages)
Search:
Filter
Training
Validation
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



- Deslope: ✓ accept a higher variety with the baselines .
- Deslant: ✓ necessary for written texts; NOT for printed texts.
- Stretch: ✓ if texts are written extremely tight to one another
- Enhance: ✓ if there is a lot of ‘noise’ on the page (not with very white pages e.g.)

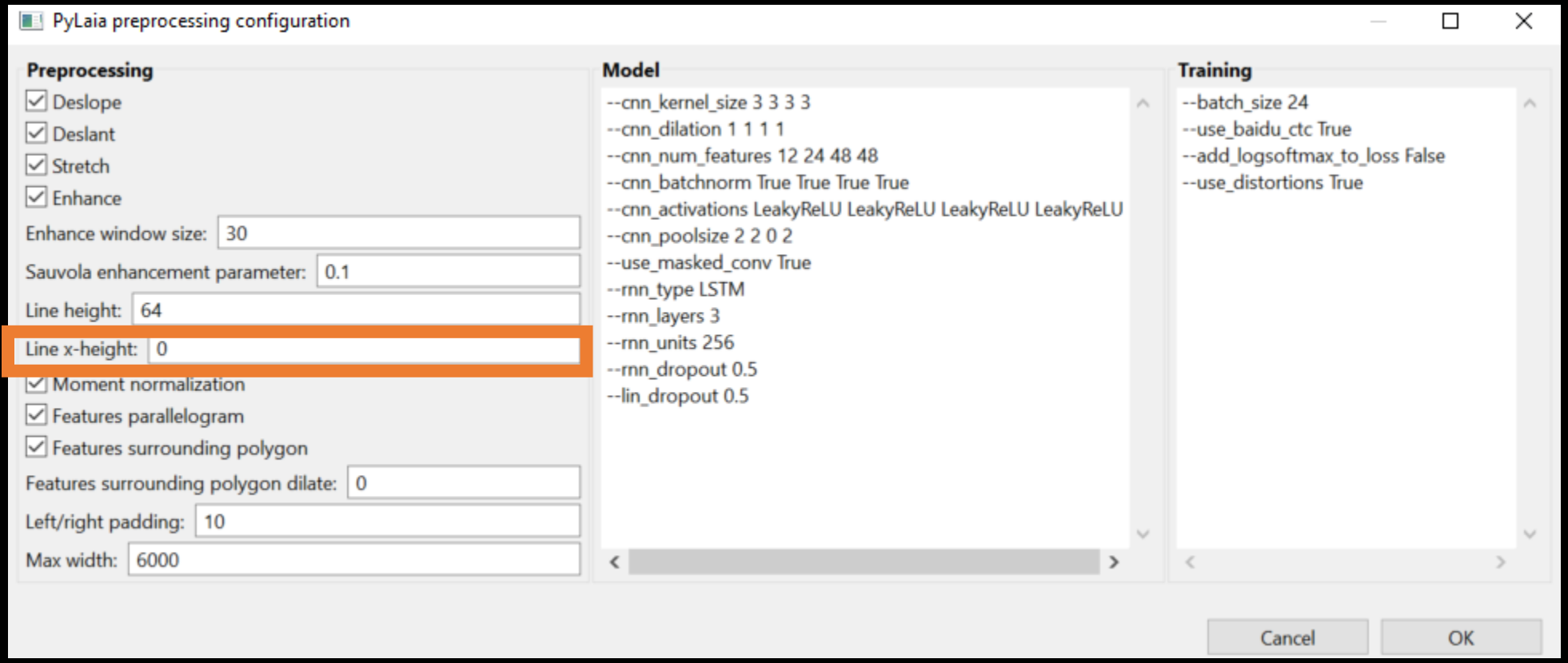


- Enhance window size: the size of the “optimizing tool” that goes over the pages.
- Sauvola enhancement parameter: ✘ don’t touch, has to do with special algoritms.

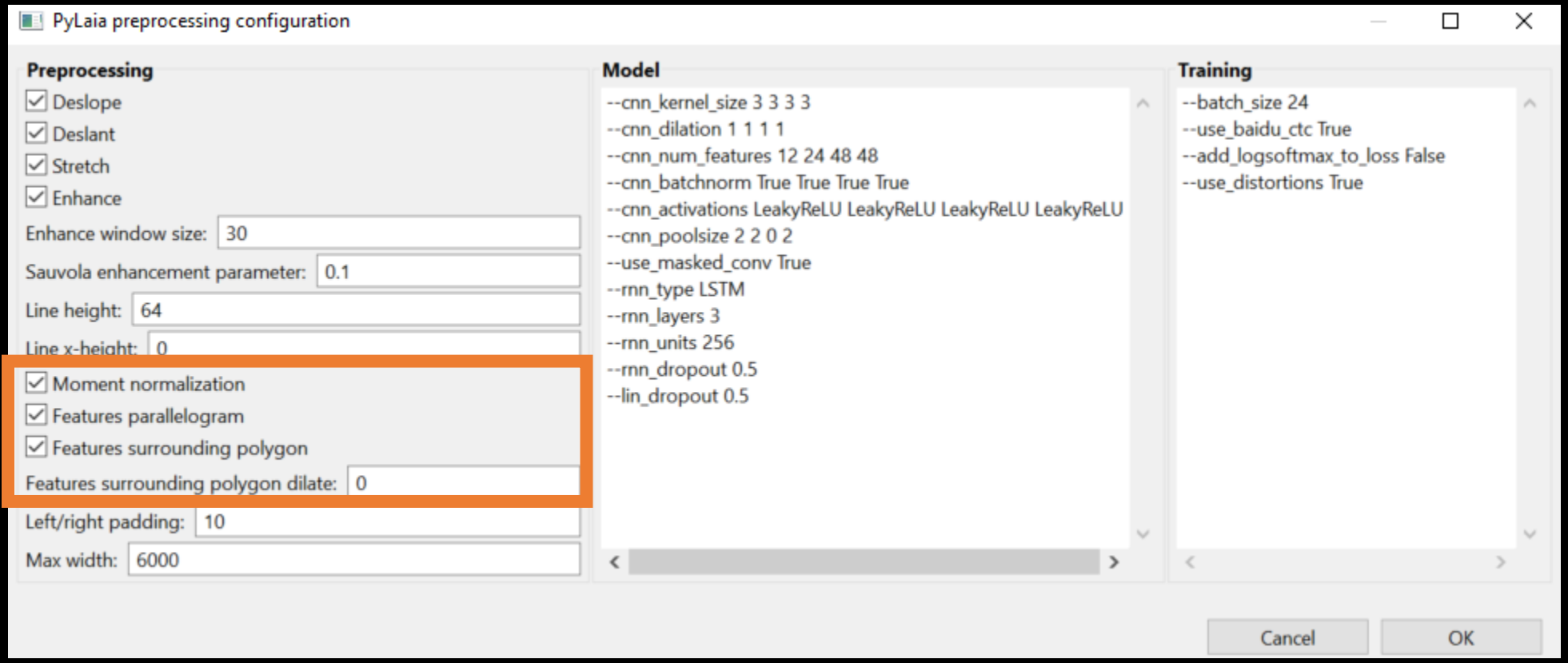


- Line height: the height of the text measured in pixels. If it is not high enough, change it to 100.
- If you get an ‘out of memory order’ you may change the “batch size” to a lower number. The smaller this number is, the slower the training becomes.

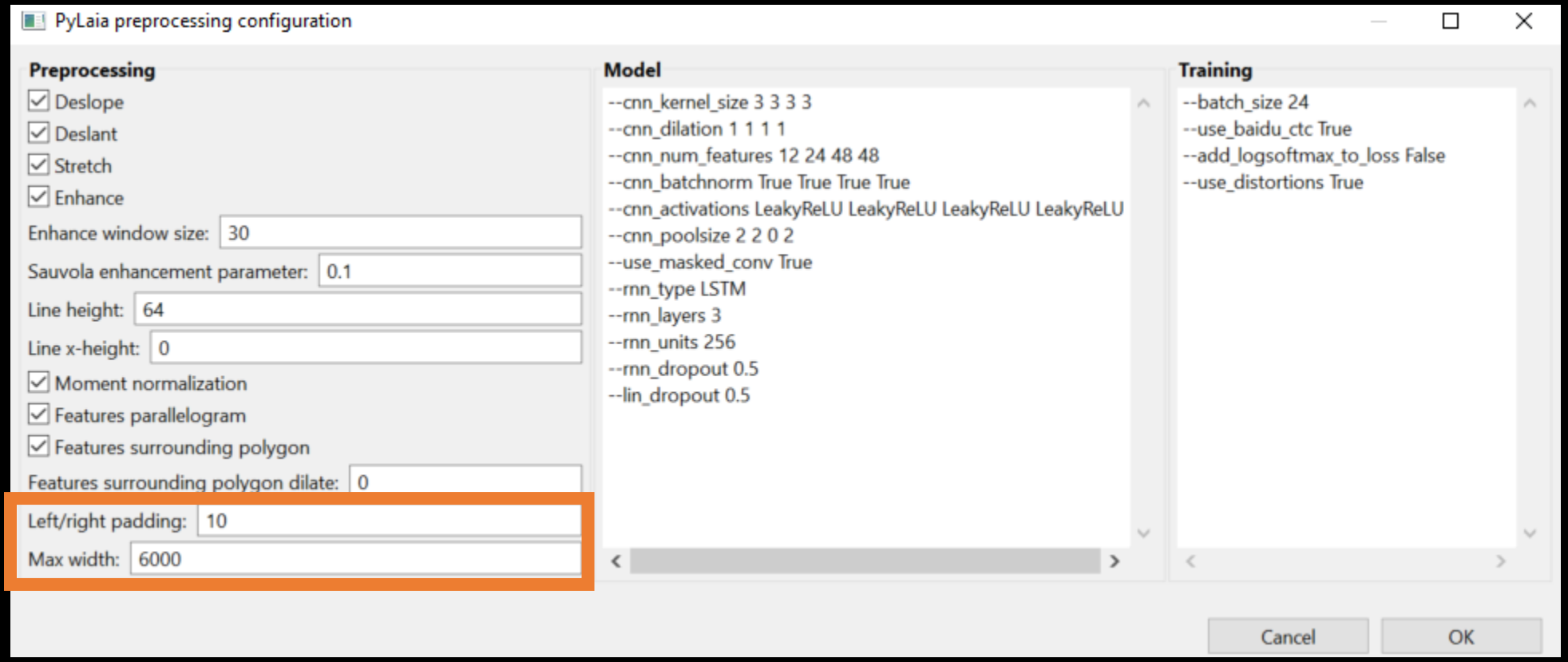
  In a next version this will be changed into an automatic batch-size.



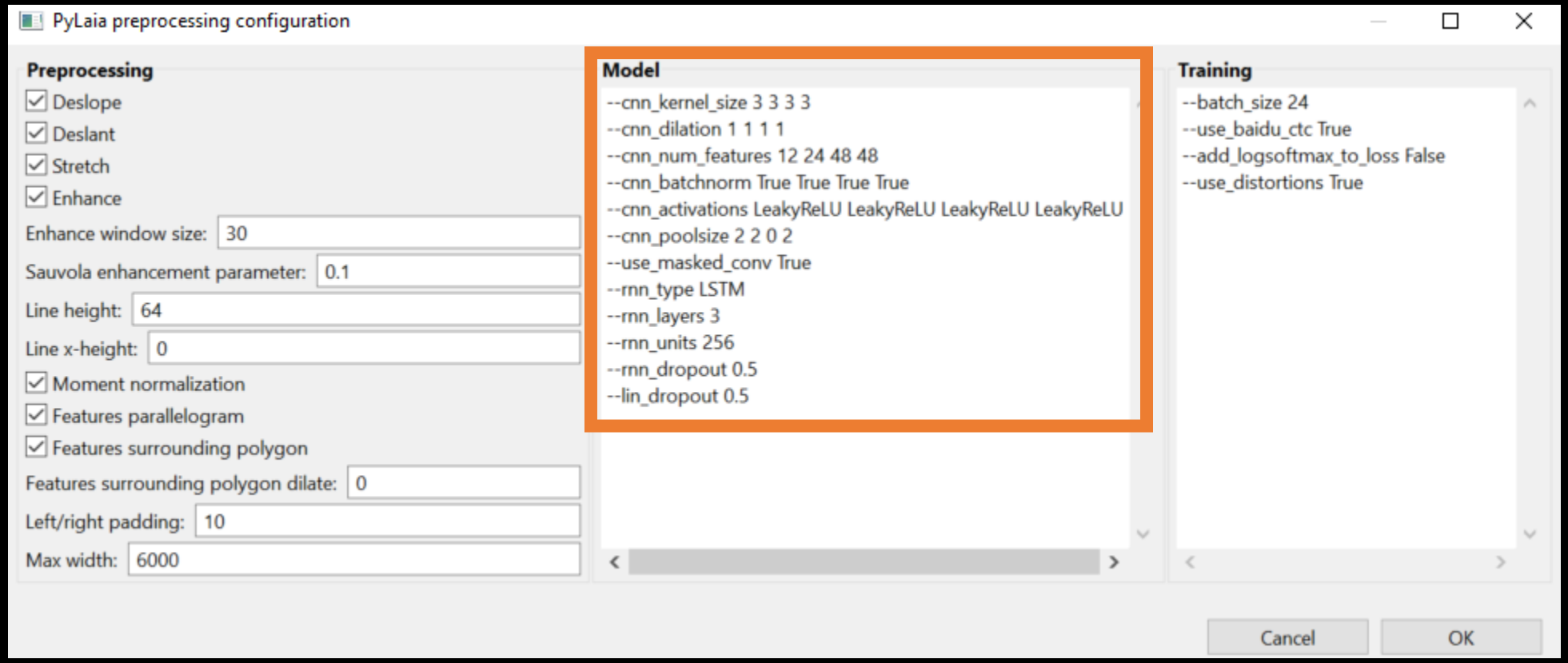
- Line x-height: this concerns descenders and ascenders. If you chose to fill this one in, the Line-height parameter is ignored.



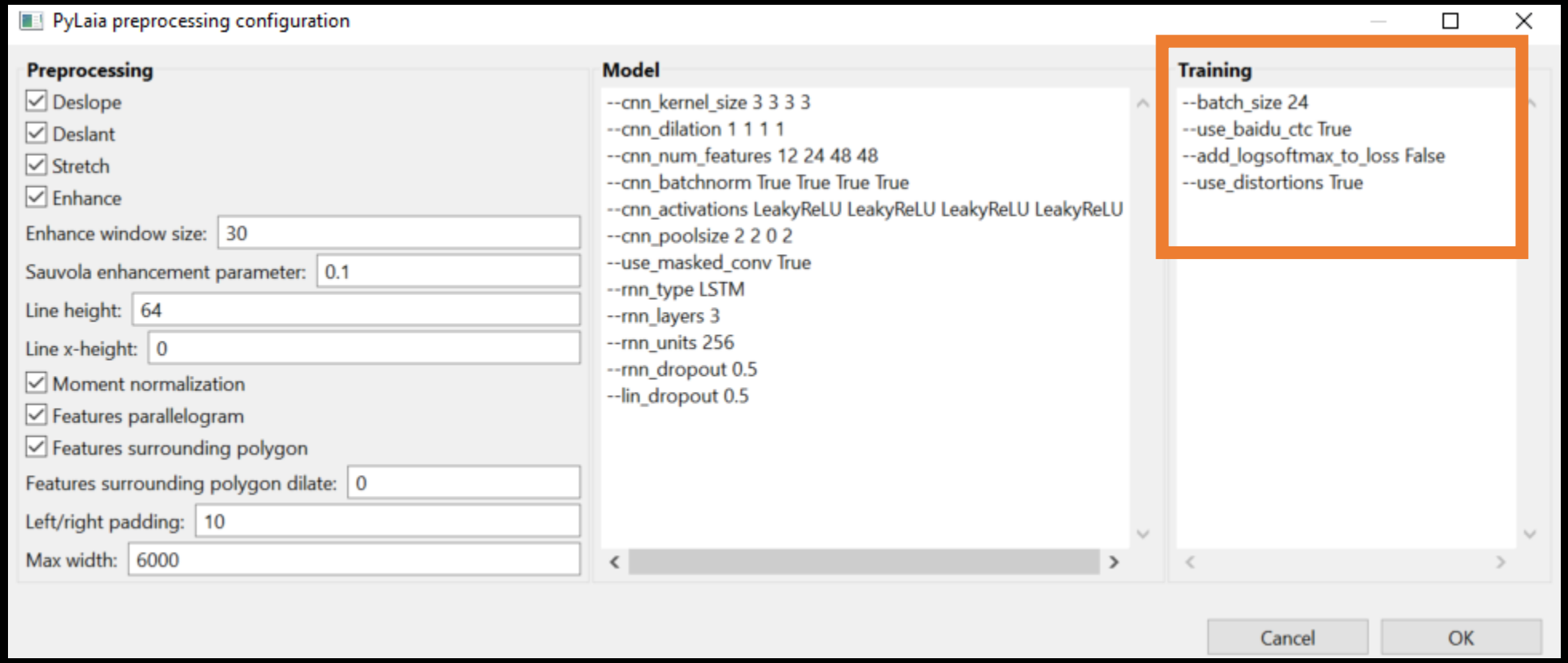
- Moment normalization: ✗ don’t touch (leave the ✓ there)
- Features parallelogram: ✗ don’t touch (leave the ✓ there)
- Features surrounding polygon: ✗ don’t touch (leave the ✓ there)
- Features surrounding polygon dilate: ✗ don’t touch (leave the ✓ there)



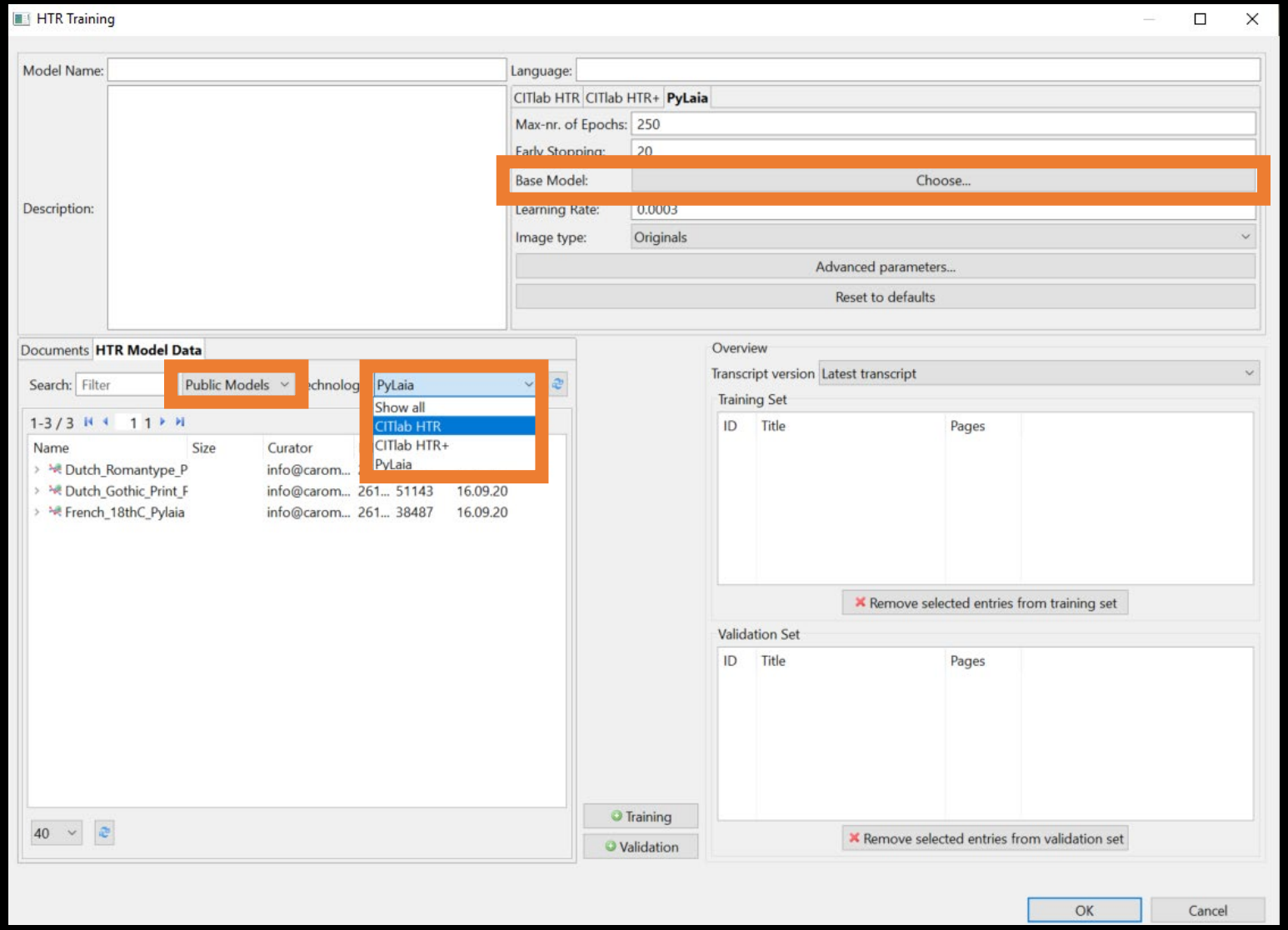
- Left/ right padding: the number 10 here means ‘10 pixels’ to prevent the risk of cutting a part of the row.
- Max width: 6000 is default (pixels). If you have huge pages, you will need to adjust this to a higher number, otherwise you don’t need to touch this.



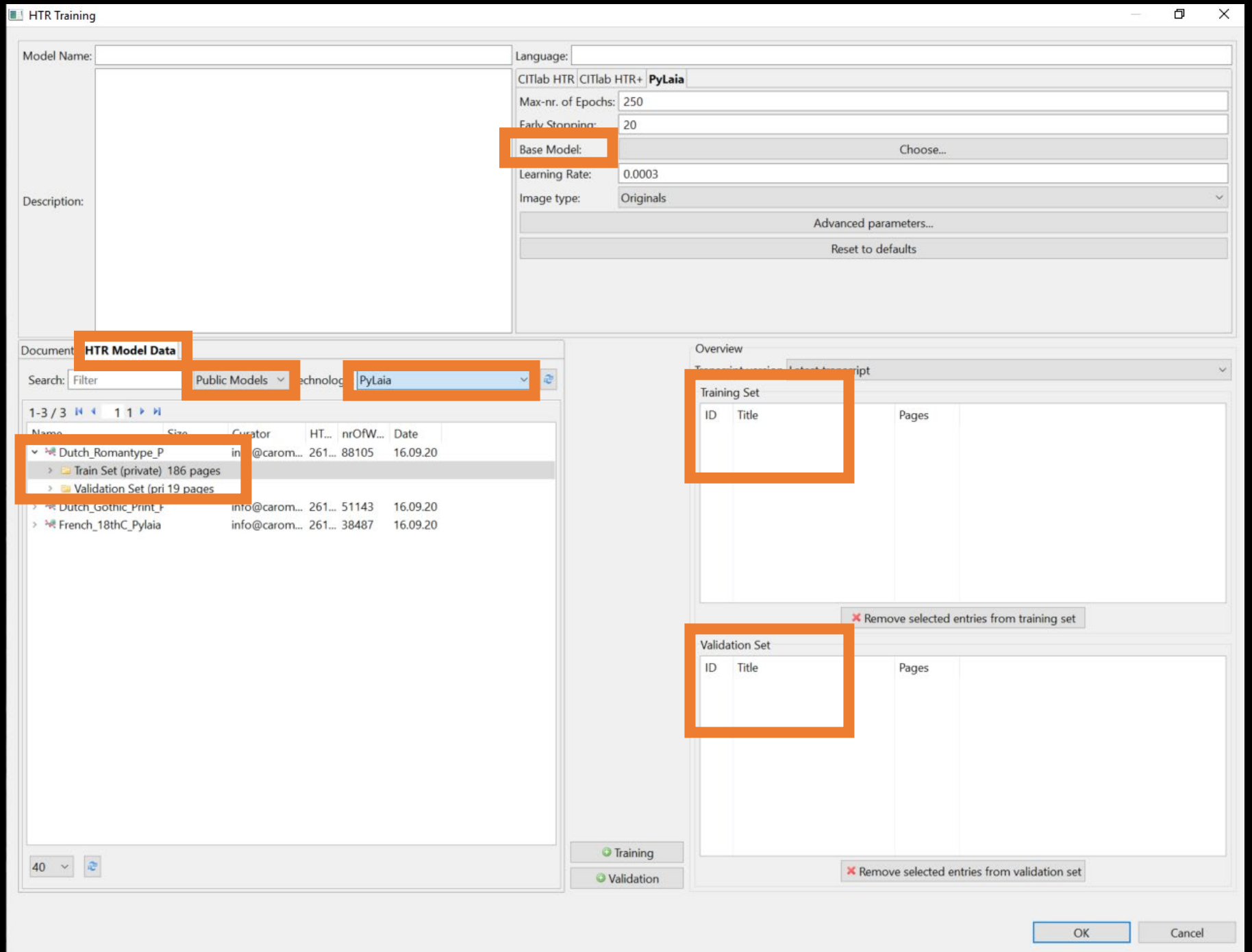
- Model: all the parameters here are meant for people who know their way around with machine learning. It has to do with modifications to the Neural Networks.



- Batch size: number of pages that are being processed at once at the GPU (Graphic Processing Unit aka: the core in the server).
- Use_distortions: “True” means that the computer will rotate or change the shape and position of the trainingsset to make the model more stable.
  - Not necessary for printed and very regularly written texts and really good quality scans. You can then change this statement from ‘True’ by typing ‘False’ there.

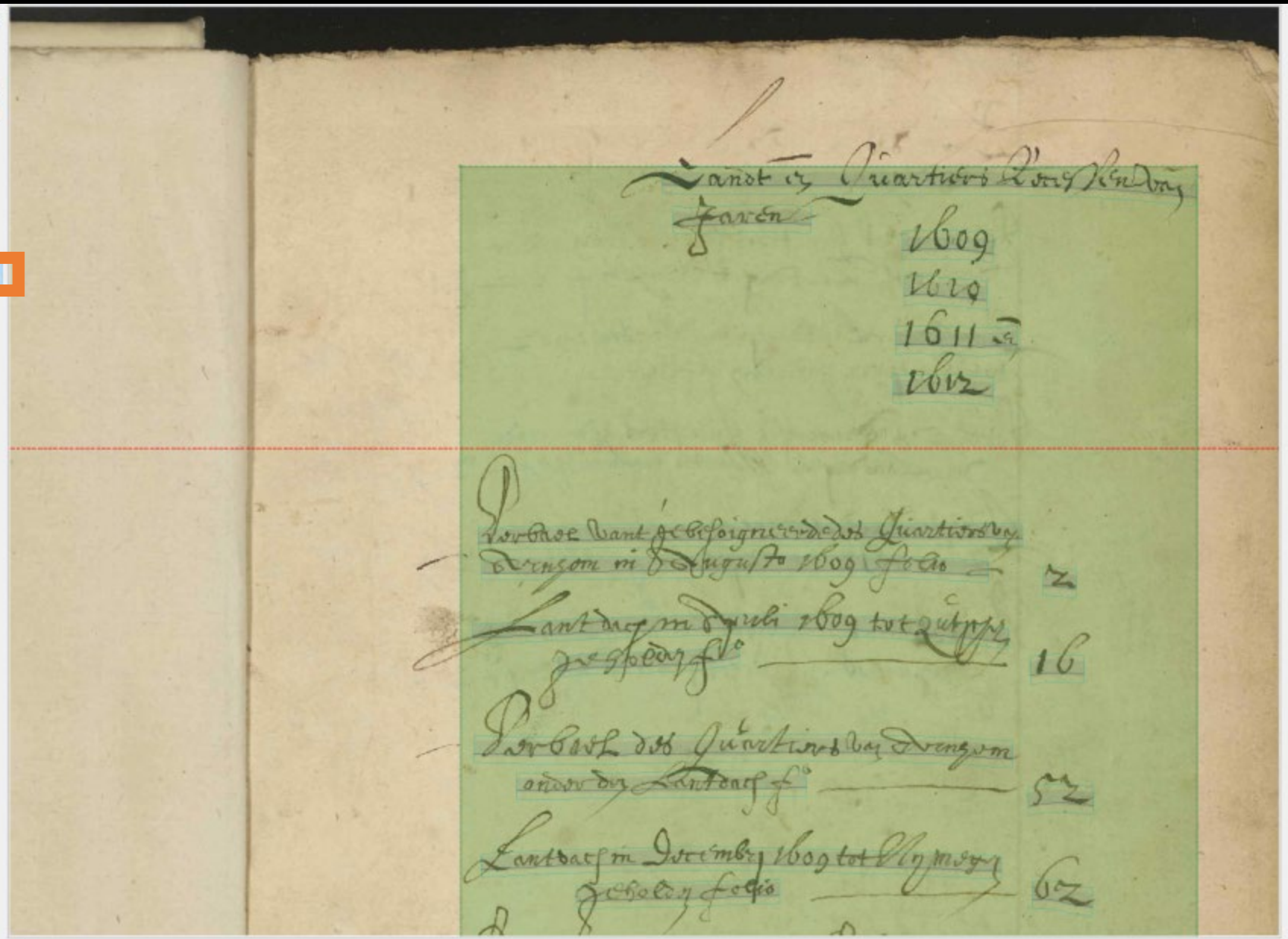


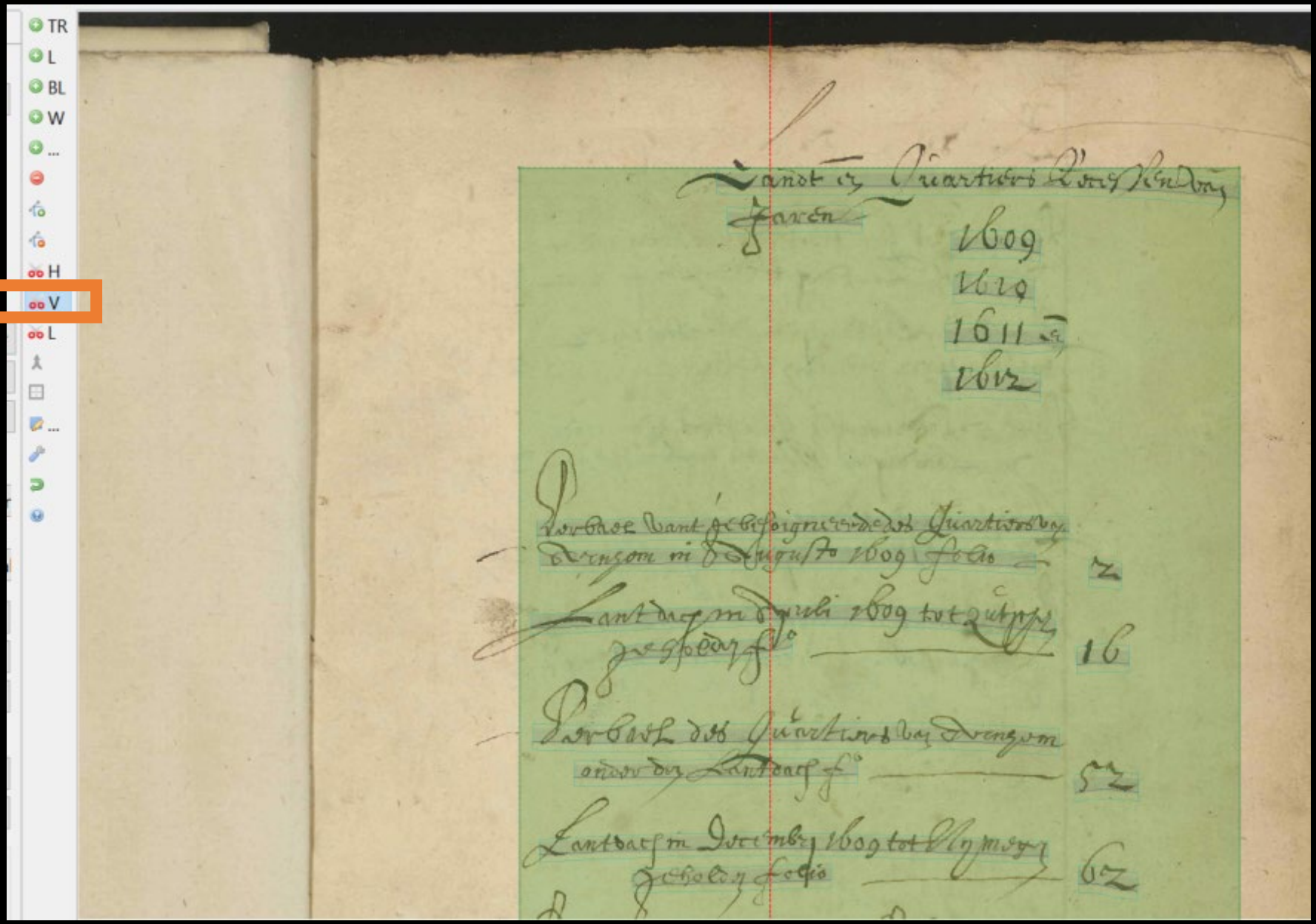
- Pylaia only works with Pylaia models as a base (as does HTR+ with HTR+ models)

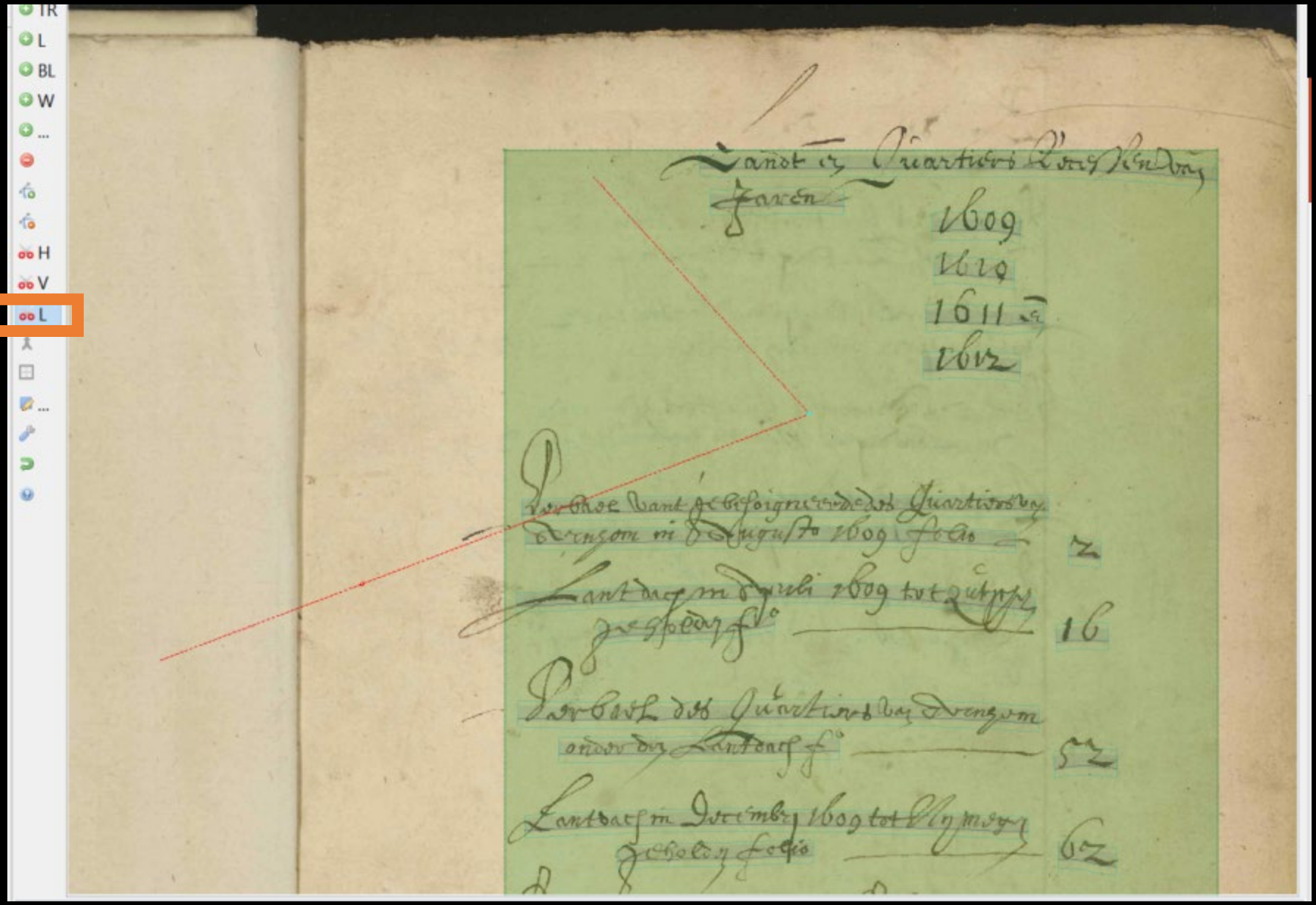

HTR Training
Model Name:
Language:
CITlab HTR
CITlab HTR+
PyLaia
Max-nr. of Epochs:
250
Early Stopping:
20
Base Model:
Choose...
Description:
Learning Rate:
0.0003
Image type:
Originals
Advanced parameters...
Reset to defaults
Document
HTR Model Data
Search:
Filter
Public Models
PyLaia
1-3 / 3
Name
Size
Curator
HT...
nrOfW...
Date
Dutch_Romantype_P
261...
88105
16.09.20
Train Set (private) 186 pages
Validation Set (pri 19 pages
Dutch_Gothic_Print_F
261...
51143
16.09.20
French_18thC_Pylaia
261...
38487
16.09.20
40
Training
Validation
Overview
Training Set
ID
Title
Pages
Remove selected entries from training set
Validation Set
ID
Title
Pages
Remove selected entries from validation set
OK
Cancel

Tables or adjustments of LA with additional lines

- There is a lot of improvement regarding tables. It’s easier to draw the lines.
- Cells are not yet recognised in P2PaLA but you could opt for columns.
- Naverlab Document Understanding will be integrated sometime in the coming months (this focusses on the AI-understanding of the content).







Remove Small regions

- If there is a lot of noise on your pages (because of stains or discolourations).

Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701-0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)] [current ...

Search current document... 5 /560 In Progress

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

Layout Analysis

Method: CITlab Advanced Configure...

Current page Pages (56(1-560

Find Text Regions Only use unsegmented pages

Find Lines in Text Regions

Run

Text Recognition

Method: HTR (CITlab HTR+ & PyLaia)

Models... Train...

Run...

Compute Accuracy...

Reference: (Correct Text) 16.06.20 20:56:51 - info@caromein.nl - New Use currer

Hypothesis: (HTR Text) 16.06.20 22:10:24 - info@caromein.nl - In Progress - CITla

Compare Text Versions...

Compare...

Compare Samples...

Other Tools

P2PaLA...

Text2Image...

Current page Pages (56(1-560

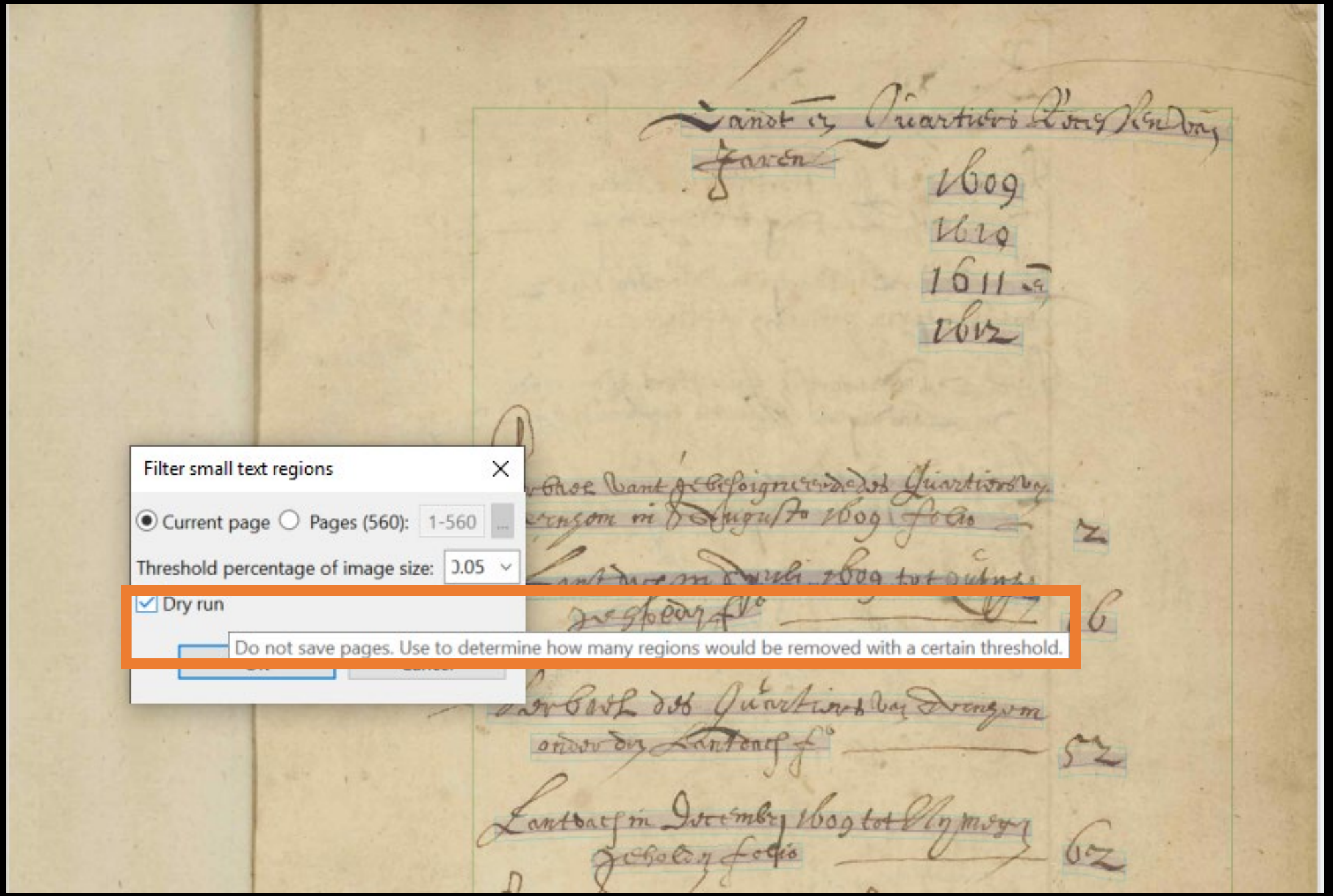
Add Baselines to Polygons

Add Polygons to Baselines

TR L BL W ... H V L ...

Create top level text region with size of image
Create default line for selected line / baseline
Simplify shape of selected lines
...is page
Remove small text regions...
Convert all region shapes to rectangles

1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11


Filter small text regions
Current page
Pages (560):
1-560
Threshold percentage of image size:
0.05
Dry run
Do not save pages. Use to determine how many regions would be removed with a certain threshold.

Create a sampleset

- Why?
  - Because you can create a representative set from a whole collection to transcribe and make that into a useful model for your collection.

Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701-0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)] [current ...

Search current document... 5 /560 In Progress

Server | Overview | Layout | Metadata | Tools

Logout info@caromein.nl

Document... | Find

Document Manager | User Manager

Versions | Add pages, add transcripts, choose symbolic images, ...

Recent Documents... | User activity

Collections: Gelderland (62362, Owner) Col-ID

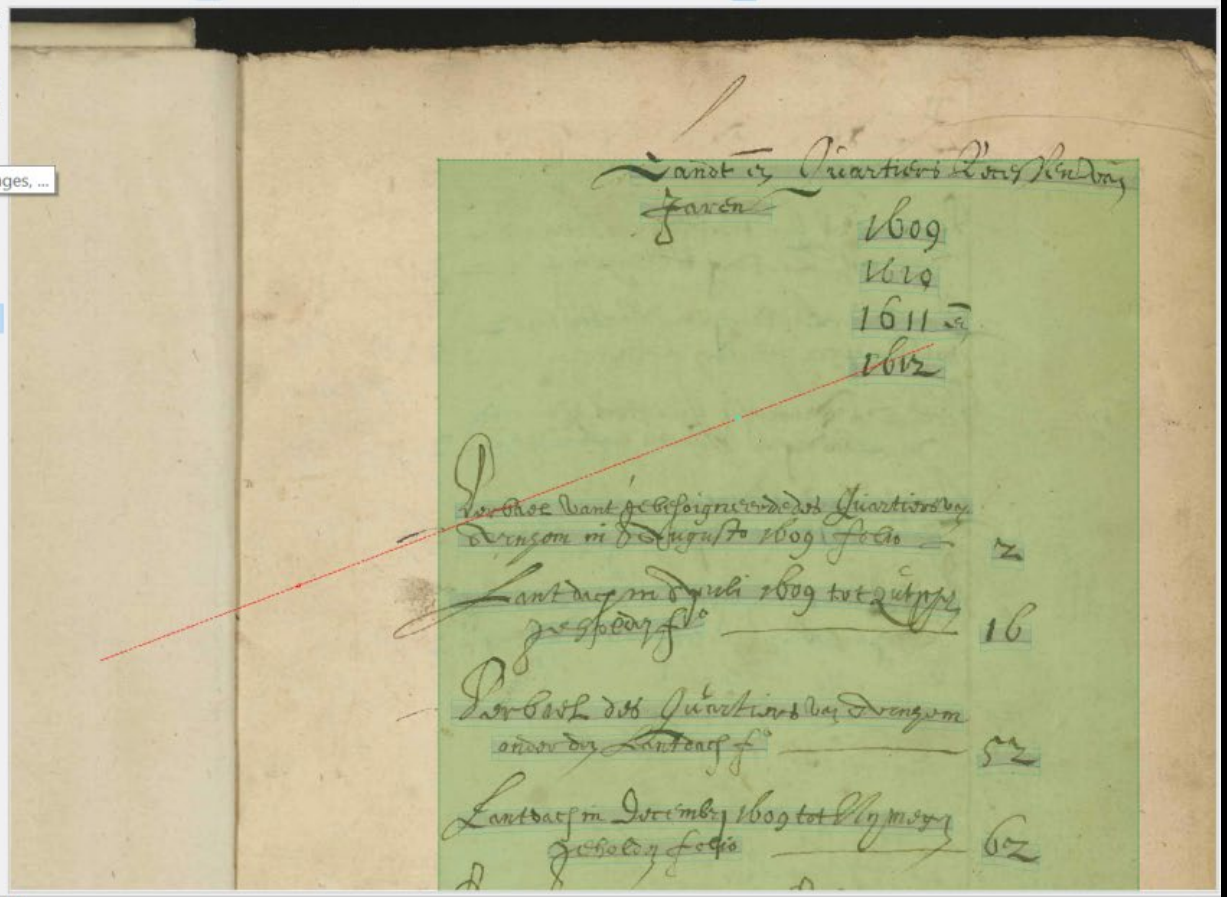
Documents | HTR Model Data

1-101 / 204 1 3 Doc-ID

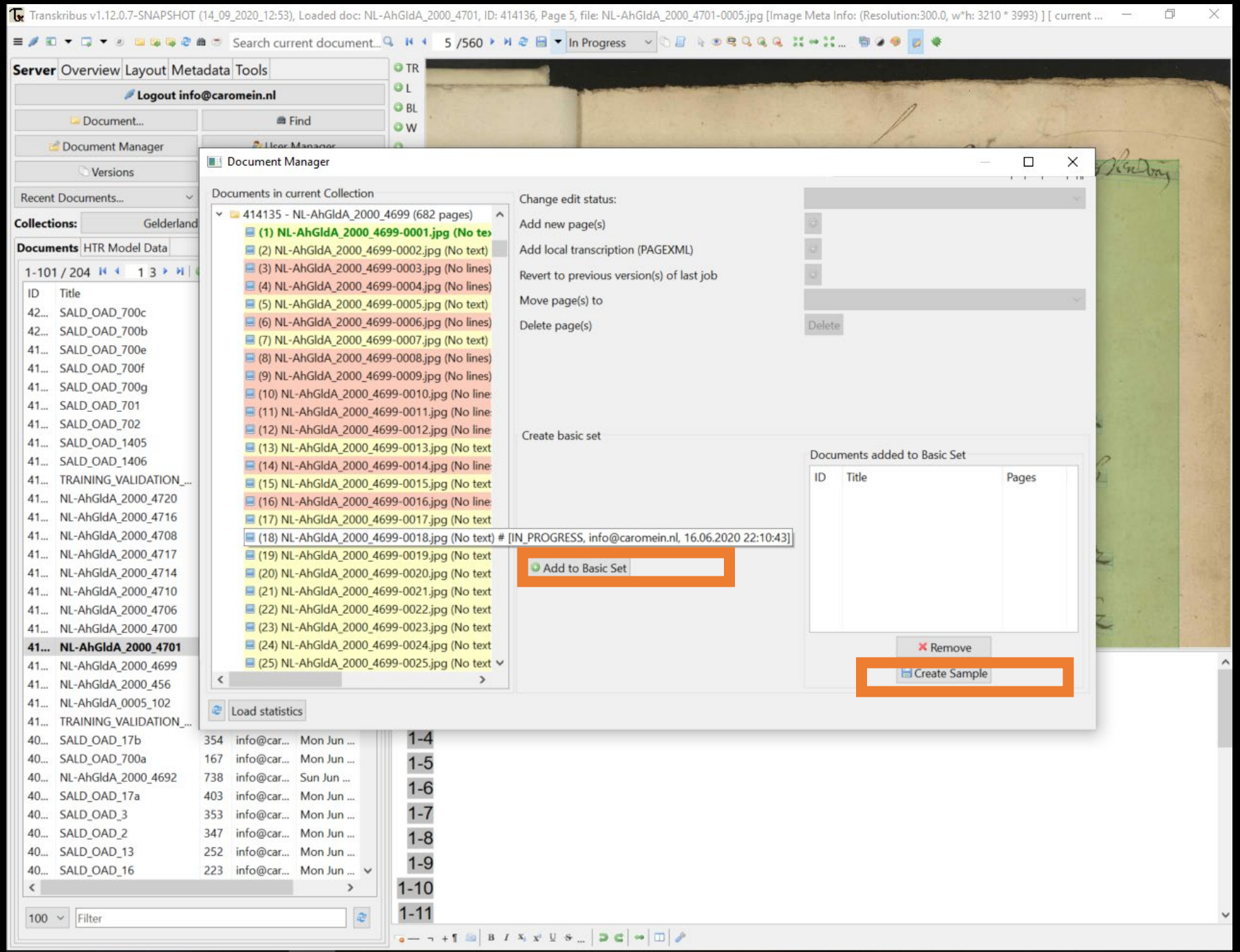
ID	Title	Pa...	Uploader	Uploaded
42...	SALD_OAD_700c	198	info@car...	Thu Jun ...
42...	SALD_OAD_700b	212	info@car...	Thu Jun ...
41...	SALD_OAD_700e	167	info@car...	Thu Jun ...
41...	SALD_OAD_700f	155	info@car...	Thu Jun ...
41...	SALD_OAD_700g	230	info@car...	Thu Jun ...
41...	SALD_OAD_701	129	info@car...	Thu Jun ...
41...	SALD_OAD_702	181	info@car...	Thu Jun ...
41...	SALD_OAD_1405	201	info@car...	Thu Jun ...
41...	SALD_OAD_1406	52	info@car...	Thu Jun ...
41...	TRAINING_VALIDATION_...	1	info@car...	Mon Jun ...
41...	NL-AhGldA_2000_4720	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4716	996	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4708	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4717	956	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4714	686	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4710	618	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4706	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4700	798	info@car...	Tue Jun 1...
41...	**NL-AhGldA_2000_4701**	**560**	**info@ca...**	**Tue Jun ...**
41...	NL-AhGldA_2000_4699	682	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_456	308	info@car...	Tue Jun 1...
41...	NL-AhGldA_0005_102	172	info@car...	Tue Jun 1...
41...	TRAINING_VALIDATION_...	6	info@car...	Sun Jun ...
40...	SALD_OAD_17b	354	info@car...	Mon Jun ...
40...	SALD_OAD_700a	167	info@car...	Mon Jun ...
40...	NL-AhGldA_2000_4692	738	info@car...	Sun Jun ...
40...	SALD_OAD_17a	403	info@car...	Mon Jun ...
40...	SALD_OAD_3	353	info@car...	Mon Jun ...
40...	SALD_OAD_2	347	info@car...	Mon Jun ...
40...	SALD_OAD_13	252	info@car...	Mon Jun ...
40...	SALD_OAD_16	223	info@car...	Mon Jun ...

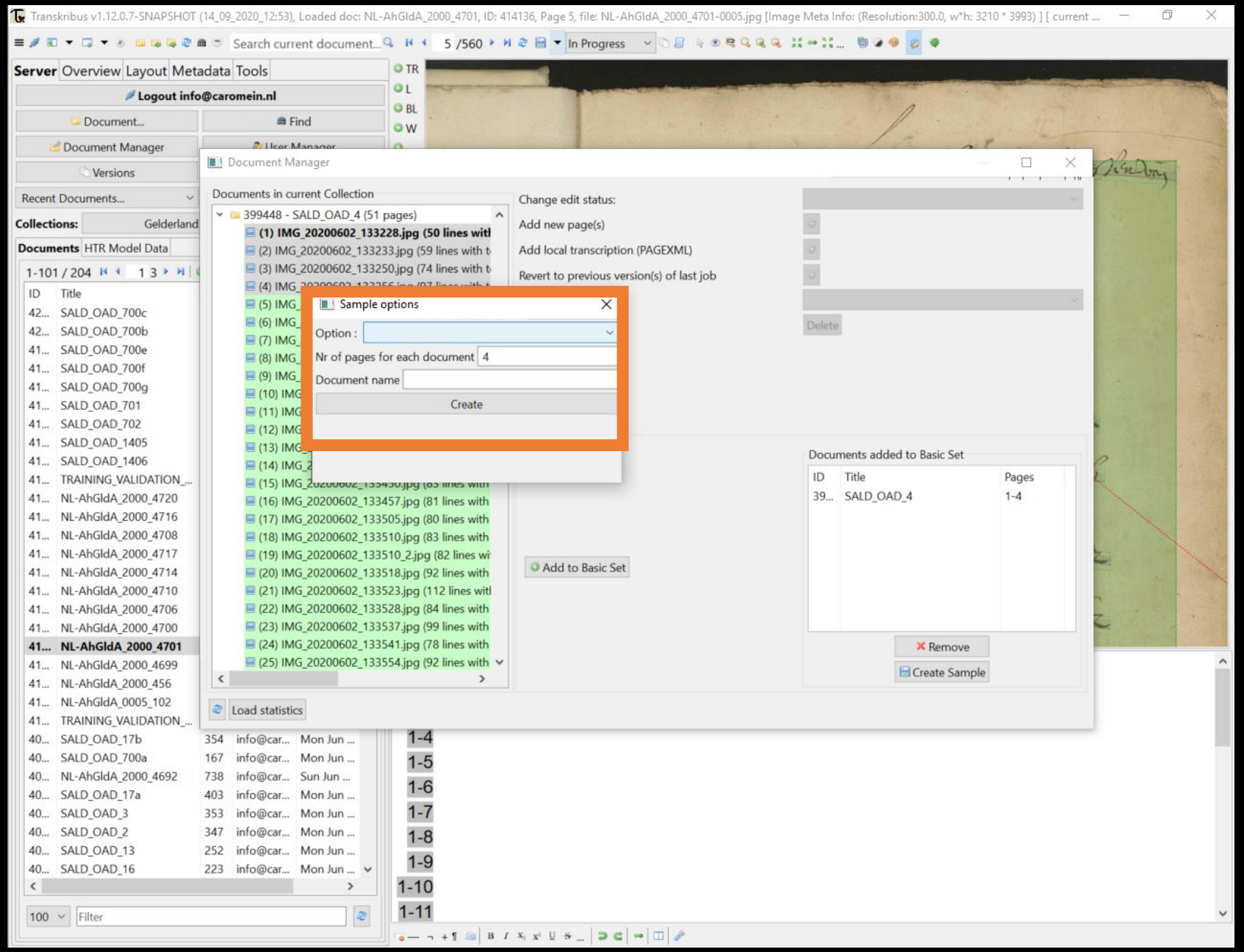
100 Filter

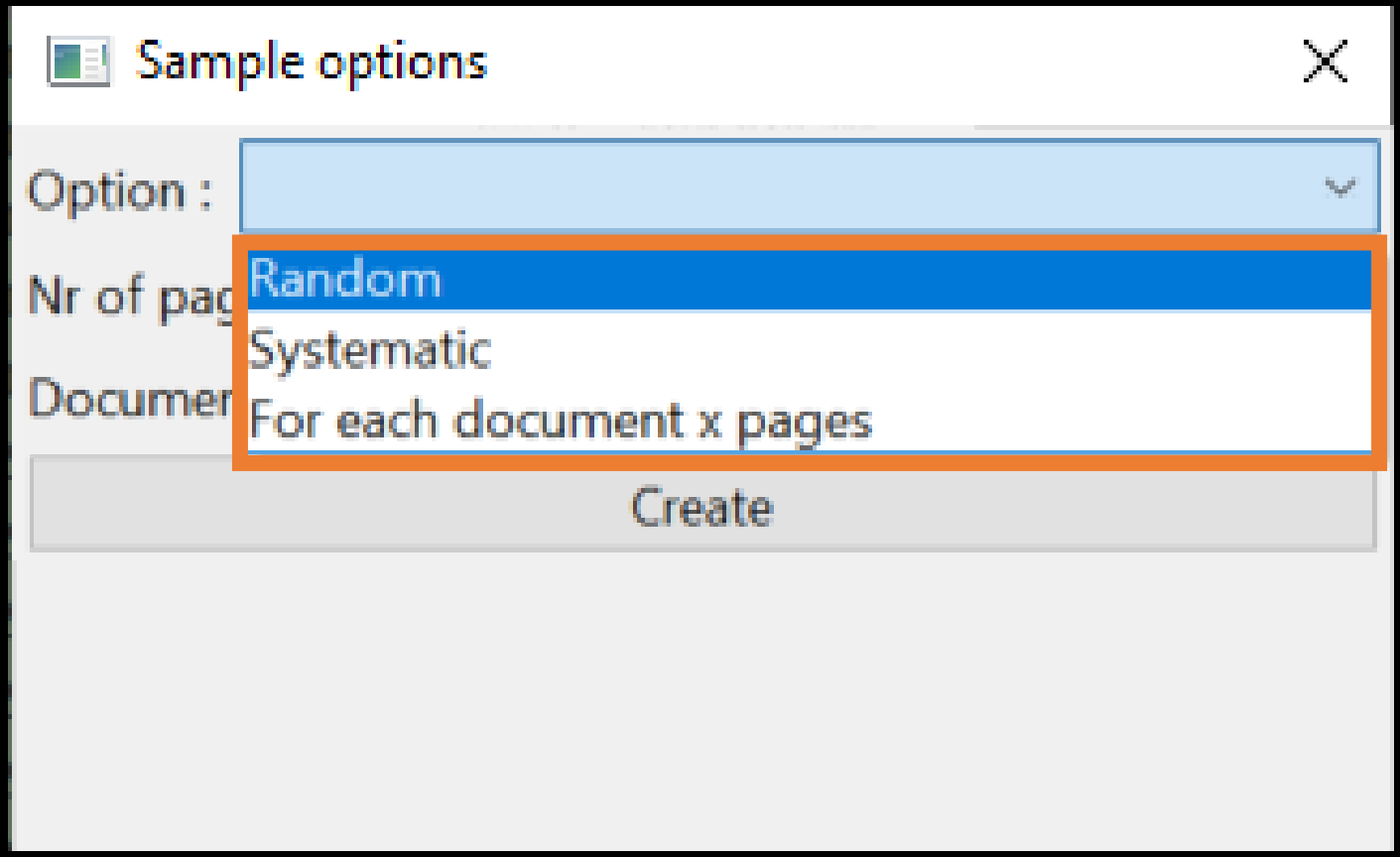
TR L BL W ... H V L ...



1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11


Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701-0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)] [current ...
Search current document...
5 /560
In Progress
Server
Overview
Layout
Metadata
Tools
Logout info@caromein.nl
Document...
Find
Document Manager
Versions
Recent Documents...
Collections:
Gelderland
Documents
HTR Model Data
1-101 / 204
ID
Title
42... SALD_OAD_700c
42... SALD_OAD_700b
41... SALD_OAD_700e
41... SALD_OAD_700f
41... SALD_OAD_700g
41... SALD_OAD_701
41... SALD_OAD_702
41... SALD_OAD_1405
41... SALD_OAD_1406
41... TRAINING_VALIDATION_...
41... NL-AhGldA_2000_4720
41... NL-AhGldA_2000_4716
41... NL-AhGldA_2000_4708
41... NL-AhGldA_2000_4717
41... NL-AhGldA_2000_4714
41... NL-AhGldA_2000_4710
41... NL-AhGldA_2000_4706
41... NL-AhGldA_2000_4700
41... NL-AhGldA_2000_4701
41... NL-AhGldA_2000_4699
41... NL-AhGldA_2000_456
41... NL-AhGldA_0005_102
41... TRAINING_VALIDATION_...
40... SALD_OAD_17b 354 info@car... Mon Jun ...
40... SALD_OAD_700a 167 info@car... Mon Jun ...
40... NL-AhGldA_2000_4692 738 info@car... Sun Jun ...
40... SALD_OAD_17a 403 info@car... Mon Jun ...
40... SALD_OAD_3 353 info@car... Mon Jun ...
40... SALD_OAD_2 347 info@car... Mon Jun ...
40... SALD_OAD_13 252 info@car... Mon Jun ...
40... SALD_OAD_16 223 info@car... Mon Jun ...
100
Filter
TR
L
BL
W
Document Manager
Documents in current Collection
414135 - NL-AhGldA_2000_4699 (682 pages)
(1) NL-AhGldA_2000_4699-0001.jpg (No te
(2) NL-AhGldA_2000_4699-0002.jpg (No text)
(3) NL-AhGldA_2000_4699-0003.jpg (No lines)
(4) NL-AhGldA_2000_4699-0004.jpg (No lines)
(5) NL-AhGldA_2000_4699-0005.jpg (No text)
(6) NL-AhGldA_2000_4699-0006.jpg (No lines)
(7) NL-AhGldA_2000_4699-0007.jpg (No text)
(8) NL-AhGldA_2000_4699-0008.jpg (No lines)
(9) NL-AhGldA_2000_4699-0009.jpg (No lines)
(10) NL-AhGldA_2000_4699-0010.jpg (No line
(11) NL-AhGldA_2000_4699-0011.jpg (No line
(12) NL-AhGldA_2000_4699-0012.jpg (No line
(13) NL-AhGldA_2000_4699-0013.jpg (No text
(14) NL-AhGldA_2000_4699-0014.jpg (No line
(15) NL-AhGldA_2000_4699-0015.jpg (No text
(16) NL-AhGldA_2000_4699-0016.jpg (No line
(17) NL-AhGldA_2000_4699-0017.jpg (No text
(18) NL-AhGldA_2000_4699-0018.jpg (No text) # [IN_PROGRESS, info@caromein.nl, 16.06.2020 22:10:43]
(19) NL-AhGldA_2000_4699-0019.jpg (No text
(20) NL-AhGldA_2000_4699-0020.jpg (No text
(21) NL-AhGldA_2000_4699-0021.jpg (No text
(22) NL-AhGldA_2000_4699-0022.jpg (No text
(23) NL-AhGldA_2000_4699-0023.jpg (No text
(24) NL-AhGldA_2000_4699-0024.jpg (No text
(25) NL-AhGldA_2000_4699-0025.jpg (No text
Load statistics
Change edit status:
Add new page(s)
Add local transcription (PAGEXML)
Revert to previous version(s) of last job
Move page(s) to
Delete page(s)
Delete
Create basic set
Add to Basic Set
Documents added to Basic Set
ID
Title
Pages
Remove
Create Sample
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11


Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701-0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)] [current ...
Search current document...
5 /560
In Progress
Server
Overview
Layout
Metadata
Tools
Logout info@caromein.nl
Document...
Find
Document Manager
Versions
Recent Documents...
Collections:
Gelderland
Documents
HTR Model Data
1-101 / 204
ID
Title
42... SALD_OAD_700c
42... SALD_OAD_700b
41... SALD_OAD_700e
41... SALD_OAD_700f
41... SALD_OAD_700g
41... SALD_OAD_701
41... SALD_OAD_702
41... SALD_OAD_1405
41... SALD_OAD_1406
41... TRAINING_VALIDATION_...
41... NL-AhGldA_2000_4720
41... NL-AhGldA_2000_4716
41... NL-AhGldA_2000_4708
41... NL-AhGldA_2000_4717
41... NL-AhGldA_2000_4714
41... NL-AhGldA_2000_4710
41... NL-AhGldA_2000_4706
41... NL-AhGldA_2000_4700
41... NL-AhGldA_2000_4701
41... NL-AhGldA_2000_4699
41... NL-AhGldA_2000_456
41... NL-AhGldA_0005_102
41... TRAINING_VALIDATION_...
40... SALD_OAD_17b 354 info@car... Mon Jun ...
40... SALD_OAD_700a 167 info@car... Mon Jun ...
40... NL-AhGldA_2000_4692 738 info@car... Sun Jun ...
40... SALD_OAD_17a 403 info@car... Mon Jun ...
40... SALD_OAD_3 353 info@car... Mon Jun ...
40... SALD_OAD_2 347 info@car... Mon Jun ...
40... SALD_OAD_13 252 info@car... Mon Jun ...
40... SALD_OAD_16 223 info@car... Mon Jun ...
100
Filter
TR
L
BL
W
Document Manager
Documents in current Collection
399448 - SALD_OAD_4 (51 pages)
(1) IMG_20200602_133228.jpg (50 lines with
(2) IMG_20200602_133233.jpg (59 lines with t
(3) IMG_20200602_133250.jpg (74 lines with t
(16) IMG_20200602_133457.jpg (81 lines with
(17) IMG_20200602_133505.jpg (80 lines with
(18) IMG_20200602_133510.jpg (83 lines with
(19) IMG_20200602_133510_2.jpg (82 lines wi
(20) IMG_20200602_133518.jpg (92 lines with
(21) IMG_20200602_133523.jpg (112 lines with
(22) IMG_20200602_133528.jpg (84 lines with
(23) IMG_20200602_133537.jpg (99 lines with
(24) IMG_20200602_133541.jpg (78 lines with
(25) IMG_20200602_133554.jpg (92 lines with
Load statistics
Change edit status:
Add new page(s)
Add local transcription (PAGEXML)
Revert to previous version(s) of last job
Delete
Add to Basic Set
Documents added to Basic Set
ID
Title
Pages
39... SALD_OAD_4 1-4
Remove
Create Sample
Sample options
Option :
Nr of pages for each document 4
Document name
Create
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11

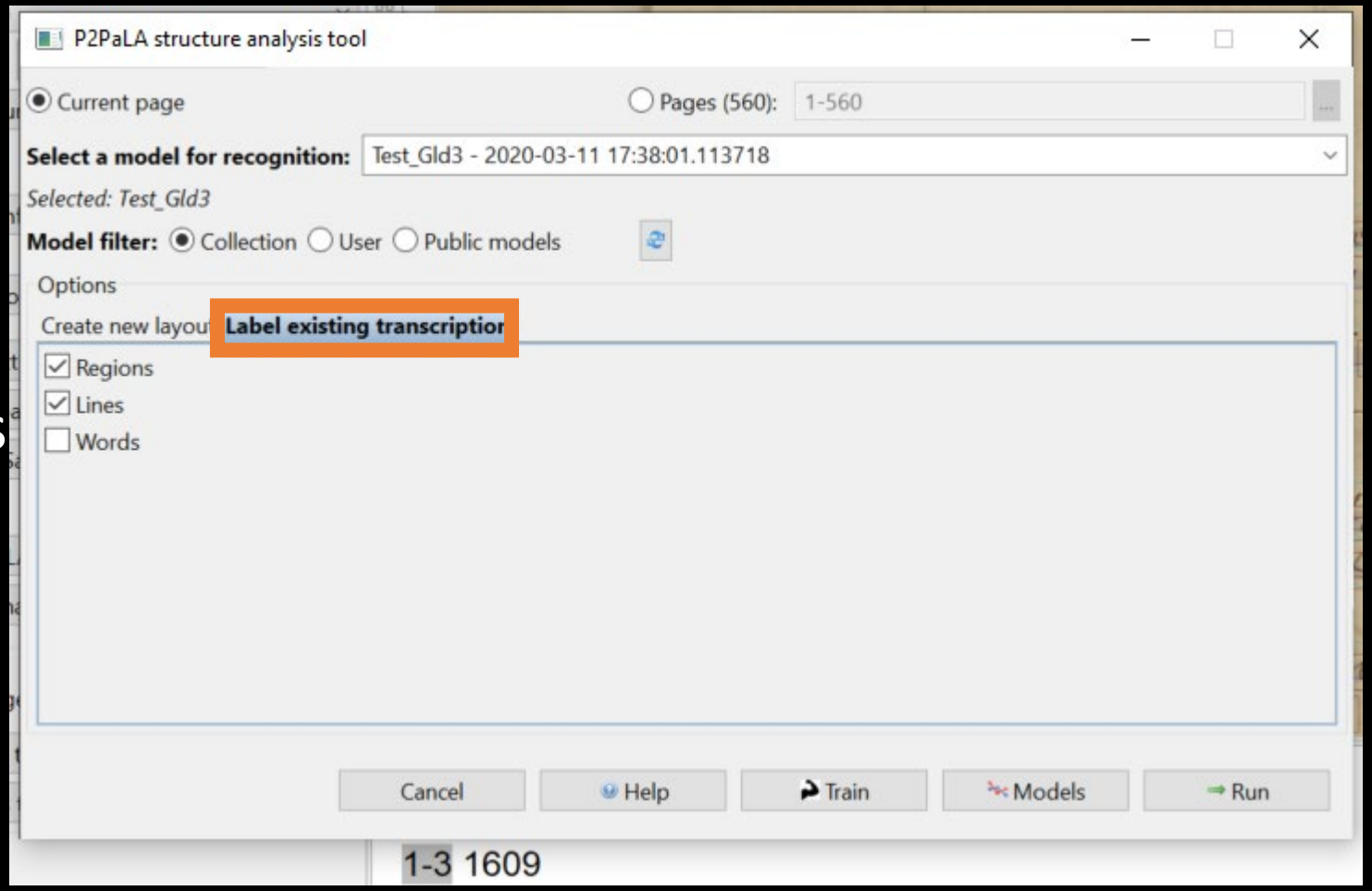

Sample options
Option :
Nr of pag
Documer
Random
Systematic
For each document x pages
Create

Some other small updates

- P2PaLA
- Read and Search (kws)
- Little bit on the webversion
- Masking text

P2PaLA

- “Label existing transcription”
  aka: rename the regions
- P2PaLA/ text regions that are labelled can (in the very near future) be used to apply a model just to certain fields and not on others if you want. (E.g. useful with multiple languages/translations.)

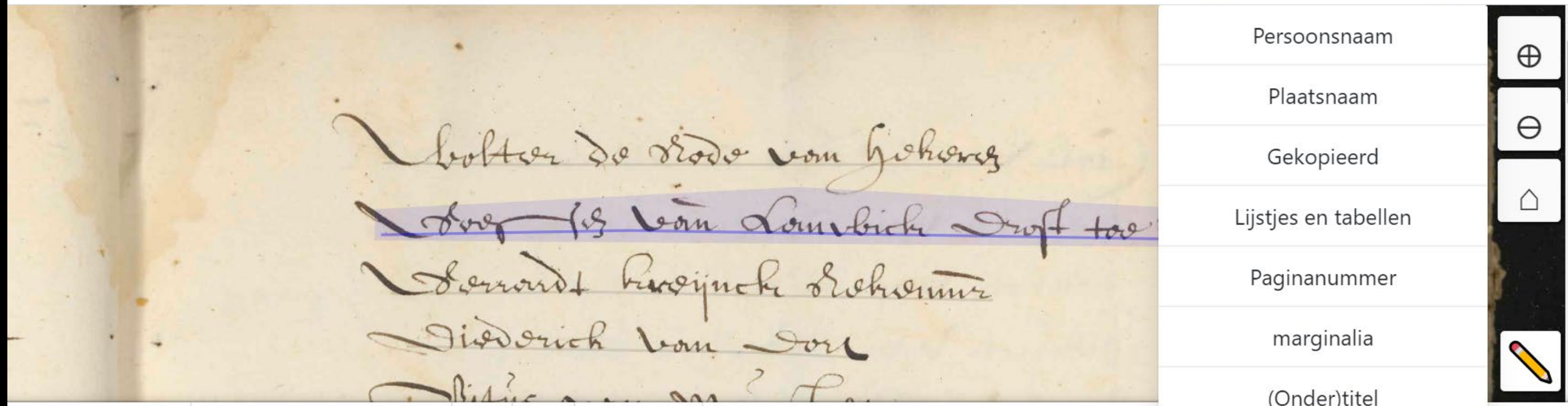


Read & Search service

- Keyword spotting and searching options are available through the Transkribus server (with log in through your own account).
  - See KWS in webinar 1 https://youtu.be/5YCfaFNMoI4?t=7266
- If you have a (small) project and you want to provide the results to a wider audience but you do not want to have to figure out how to host/ create a website and such for all this searchable data… Transkribus is going to offer this as a service. In other words: they can host your data and show it as a searchable website so others (without Transkribus accounts/ access to your collection) would be able to publicly view it.
- Works based on IIIF.

Webversion

- The rights of someone you shared the folder with will be made the same within the webversion as in the desktop version (expert client). (So a co-owner or editor will be the same in both versions, meaning that they can sign a page off as GT.)
- If you need to use special signs: use Unicode. (e.g. Alt133 = é)
- Italics and bold are not (yet) possible in the web interface. (Mind you, these are not going to be integrated into a model anyway for the coming time; so ask yourself if you need these to be tagged/ analysed).



x² x₂ U ab ?! Unclear Annotation **NOTE:** right-click to add an

Persoonsnaam
Plaatsnaam
Gekopieerd
Lijstjes en tabellen
Paginanummer
marginalia
(Onder)titel
Koptekst
Voettekst
Afbeelding
Gedrukte Tekst

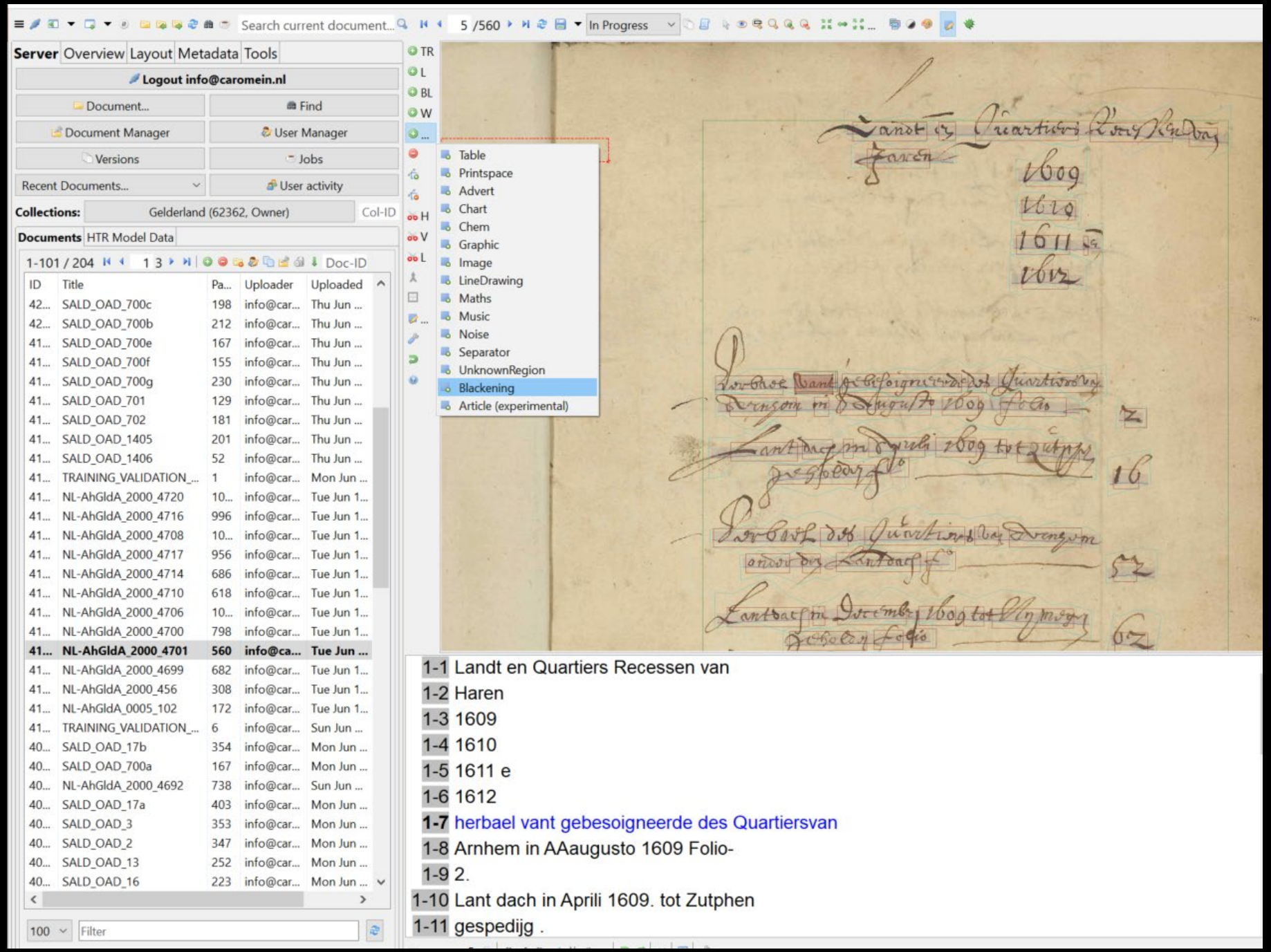
Text Region 3

1	3
2	Woster de Node van hekere
3	Coessen van Lenwick Drost toe Bre
4	Gerrardt kreijnck Rekenmr
5	Diederick van Dort

Save Changes

Masking

- “blackening”
- You then draw a box over the text that you want to hide.
- N.B. it’s only a “layer”, not ***in*** the image.
- It will be in pdf but textual export contains the original text



Transkribus v1.12.0.7-SNAPSHOT (14_09_2020_12:53), Loaded doc: NL-AhGldA_2000_4701, ID: 414136, Page 5, file: NL-AhGldA_2000_4701_0005.jpg [Image Meta Info: (Resolution:300.0, w*h: 3210 * 3993)] [current ...

Search current document... 5 /560 In Progress

Server | Overview | Layout | Metadata | Tools

Logout info@caromein.nl

Document... | Find | Document Manager | User Manager | Versions | Jobs | Recent Documents... | User activity

Collections: Gelderland (62362, Owner) Col-ID

Documents | HTR Model Data

1-101 / 204 1 3 Doc-ID

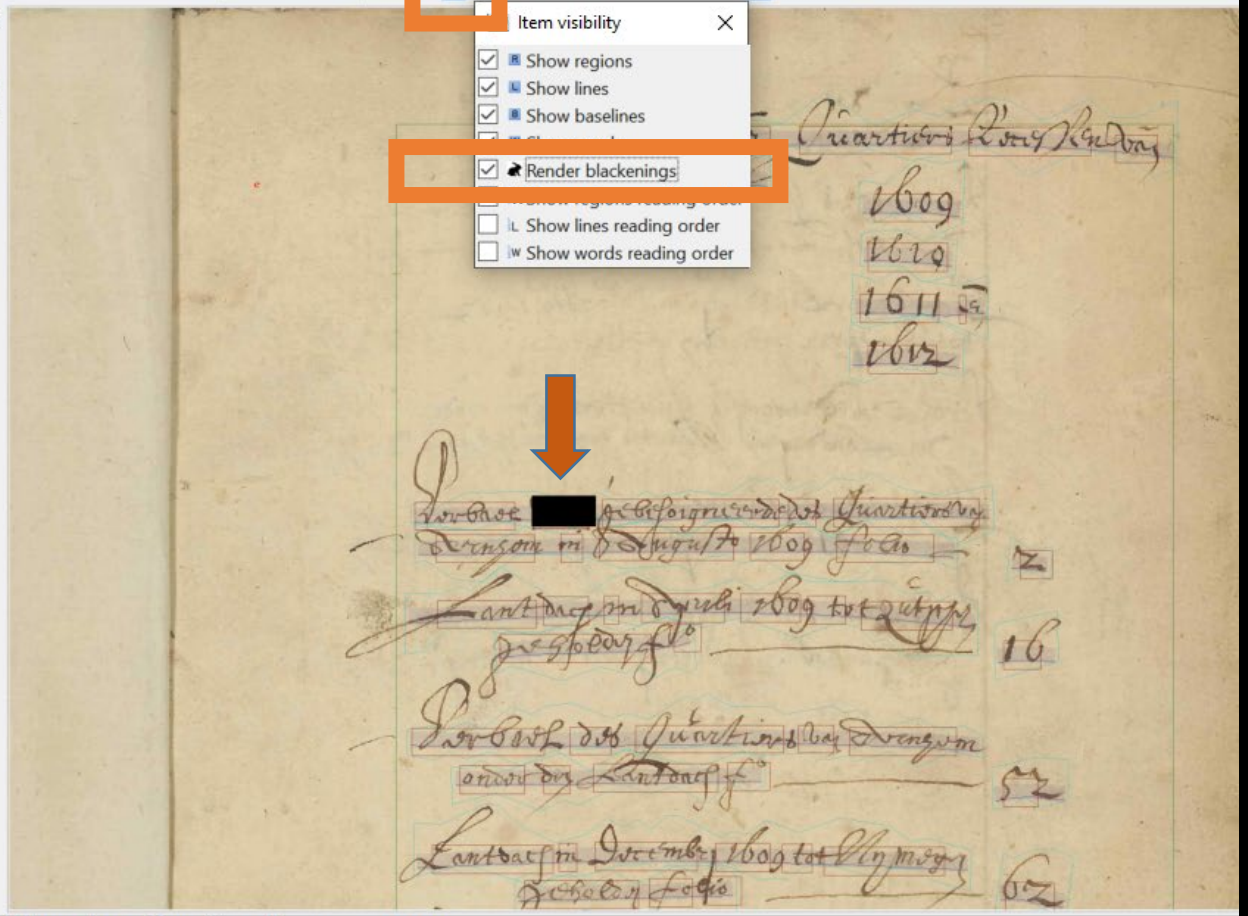
ID	Title	Pa...	Uploader	Uploaded
42...	SALD_OAD_700c	198	info@car...	Thu Jun ...
42...	SALD_OAD_700b	212	info@car...	Thu Jun ...
41...	SALD_OAD_700e	167	info@car...	Thu Jun ...
41...	SALD_OAD_700f	155	info@car...	Thu Jun ...
41...	SALD_OAD_700g	230	info@car...	Thu Jun ...
41...	SALD_OAD_701	129	info@car...	Thu Jun ...
41...	SALD_OAD_702	181	info@car...	Thu Jun ...
41...	SALD_OAD_1405	201	info@car...	Thu Jun ...
41...	SALD_OAD_1406	52	info@car...	Thu Jun ...
41...	TRAINING_VALIDATION_...	1	info@car...	Mon Jun ...
41...	NL-AhGldA_2000_4720	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4716	996	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4708	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4717	956	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4714	686	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4710	618	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4706	10...	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_4700	798	info@car...	Tue Jun 1...
41...	**NL-AhGldA_2000_4701**	**560**	**info@ca...**	**Tue Jun ...**
41...	NL-AhGldA_2000_4699	682	info@car...	Tue Jun 1...
41...	NL-AhGldA_2000_456	308	info@car...	Tue Jun 1...
41...	NL-AhGldA_0005_102	172	info@car...	Tue Jun 1...
41...	TRAINING_VALIDATION_...	6	info@car...	Sun Jun ...
40...	SALD_OAD_17b	354	info@car...	Mon Jun ...
40...	SALD_OAD_700a	167	info@car...	Mon Jun ...
40...	NL-AhGldA_2000_4692	738	info@car...	Sun Jun ...
40...	SALD_OAD_17a	403	info@car...	Mon Jun ...
40...	SALD_OAD_3	353	info@car...	Mon Jun ...
40...	SALD_OAD_2	347	info@car...	Mon Jun ...
40...	SALD_OAD_13	252	info@car...	Mon Jun ...
40...	SALD_OAD_16	223	info@car...	Mon Jun ...

100 Filter

TR L BL W ... H V L ...

Item visibility

- Show regions
- Show lines
- Show baselines
- Render blackenings
- Show lines reading order
- Show words reading order



1-1 Landt en Quartiers Recessen van
1-2 Haren
1-3 1609
1-4 1610
1-5 1611 e
1-6 1612
1-7 herbael vant gebesoigneerde des Quartiersvan
1-8 Arnhem in AAaugusto 1609 Folio-
1-9 2.
1-10 Lant dach in Aprili 1609. tot Zutphen
1-11 gespedijg .

- Thank you for your attention!