READ co·op Account

(https://readcoop.eu)

Transkribus (/transkribus/) > How-To Guides (https://readcoop.eu/transkribus/resources/how-to-guides/) >

How to Train Baseline Models in Transkribus

How to Train Baseline Models in Transkribus

Transkribus Tools Transkribus Expert Client

Last update 6 months ago

« How-To's overview (https://readcoop.eu/transkribus/resources/how-to-guides/)

About Transkribus ()

Table of Contents

Introduction

Layout Analysis (https://readcoop.eu/glossary/layout-analysis/) (LA) is a fundamental step before applying a HTR model to transcribe the documents automatically. It segments the image into text regions and baselines, and it is necessary to connect image and text for HTR to work.

Usually, Layout Analysis is performed automatically by clicking on the “Tools” tab and, under the section called “Layout Analysis”, selecting the pages on which to run the segmentation, as explained here (https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/#:~:text=be%20permanently%20deleted.-,2.%20Segmentation%20%E2%80%93%20Layout%20An
Once%20you%20have).

The default Layout Analysis tool works well for most document typologies but may not be as accurate with documents with complex layouts, such as newspapers, postcards, registers, annotated documents, etc.

If the default automatic Layout Analysis tool works well on your documents, you can continue using it, and you do not need to train a Baseline model.

On the contrary, if the default Layout Analysis is unsatisfactory for your documents, you can train a Baseline model specific to your document typology. After the training, you can apply your customised Baseline model to your documents, which will be segmented following the examples you provided for training.

Before starting training a Baseline model, remember the difference between it and P2Pala. P2Pala (https://readcoop.eu/transkribus/howto/how-to-use-the-structural-tagging-feature-and-how-to-train-it/) recognises the structure of your documents automatically, enriching them with structural tags. On the contrary, a Baseline model detects only baselines but has the advantage of being specifically trained on the layout of your documents. For this reason, it should be more accurate than the default Layout Analysis recognition tool.

Preparation

The first step is to prepare the pages on which to train the Baseline model. A good number to start with is 50 pages, but the model efficiency depends on the complexity of the layout. After the first training with 50 pages, you could decide if the Baseline model is good enough or if it needs more training material.

To prepare the pages, it is only necessary to segment, automatically or manually, the text regions and the baselines. To work more easily on the layout, you can activate the Segmentation view at the viewing profiles, as shown in the figure below. In this way, the text editor is hidden, and there is more space for the image to be shown.

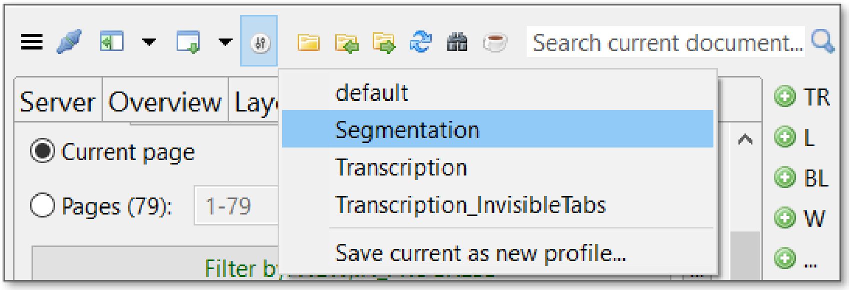


Figure 1. Segmentation view

Depending on the layout complexity, there are three options to segment the pages:

1. Run the default automatic Layout Analysis that you find under the “Tools” tab, as explained here (https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/#:~:text=be%20permanently%20deleted.-,2.%20Segmentation%20%E2%80%93%20Layout% Once%20you%20have), and then correct it manually using the Canvas menu to the left of the image.

2. Draw the Text Regions manually using the TR button in the Canvas menu. Then, under the “Tools” tab, run the automatic Layout Analysis to detect the baselines: before running it, remember to uncheck the “Find Text Regions” option. Finally, go through the pages and correct them manually using the Canvas menu.

3. Draw both the Text Regions and the Baselines manually, using respectively the TR button and the BL button in the Canvas menu to the left of the image.

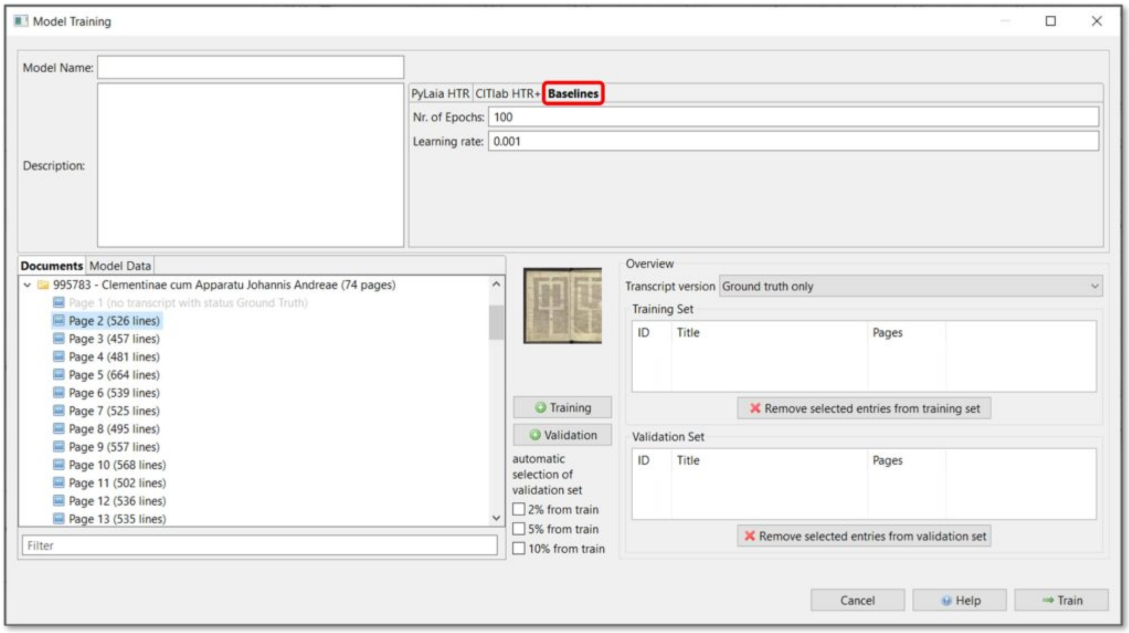
Which option to choose depends on the document type and how poorly the default automatic Layout Analysis recognition performs. We suggest trying the first option and then moving to the other ones if you realise that correcting the generated segmentation is more time-consuming than drawing it manually.

No transcription is required to be added to the pages before the Baseline model training since it focuses only on the baselines and the presence of transcribed text is irrelevant.

Training

Once the 50 or more pages are segmented, it is time to train the Baseline model. Click on the “Tools” tab. Under the “Model Training” section, click on “Train a new model”.

The Model Training window pops up, and on the right, you can choose which engine to train: for the Baseline model, please select “Baselines”, as shown in the figure below.



Model Training

Model Name:

Description:

PyLaia HTR | CITlab HTR+ | **Baselines**

Nr. of Epochs: 100

Learning rate: 0.001

Documents | Model Data

995783 - Clementinae cum Apparatu Johannis Andreae (74 pages)

- Page 1 (no transcript with status Ground Truth)
- Page 2 (526 lines)
- Page 3 (457 lines)
- Page 4 (481 lines)
- Page 5 (664 lines)
- Page 6 (539 lines)
- Page 7 (525 lines)
- Page 8 (495 lines)
- Page 9 (557 lines)
- Page 10 (568 lines)
- Page 11 (502 lines)
- Page 12 (536 lines)
- Page 13 (535 lines)

Filter

automatic selection of validation set

□ 2% from train
□ 5% from train
□ 10% from train

Overview

Transcript version: Ground truth only

Training Set

ID	Title	Pages

Validation Set

ID	Title	Pages

Training
Validation
Remove selected entries from training set
Remove selected entries from validation set
Cancel
Help
Train

Figure 2. Model Training window

Before starting training:

- On the top left, enter the name and the description of your model.
- On the top right, under the “Baselines” tab you just selected, there are the training parameters, i.e. the number of epochs and the learning rate. For the first training and if you are not familiar with machine learning, please do not change these parameters.
- At the bottom, you need to select the pages you want to use to train the model, i.e. the pages you previously segmented into text regions and baselines.
  On the left, select the whole collection or the relevant pages. Click the Training button in the centre to add the selected pages to the Training Set. If you want to consider only the pages with Ground Truth status, select “Ground Truth only” in the drop-down menu on the right, under “Overview”.
  Do the same for the Validation Set. Remember that a good Validation Set should comprise all the different examples you would like the trained Baseline model to be able to segment. The Validation Set should be around 10% of the Training Set, so we suggest, for the first training, including 45 pages in the Training Set and 5 pages in the Validation Set. If you want to automatically assign a percentage of the Training Set to the Validation Set, tick a percentage in the “automatic selection of validation set” option, before clicking the “Training” button.

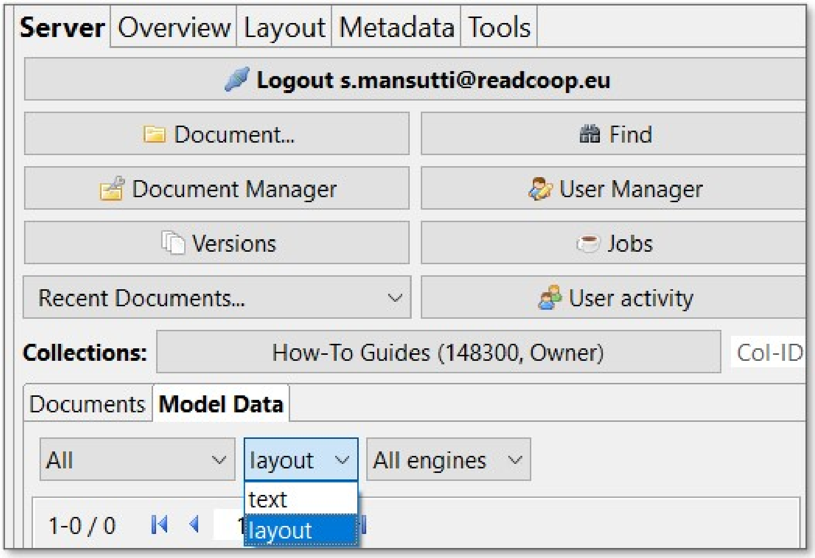
- On the right, under “Overview”, you can see all the pages assigned to the Validation Set and the Training Set.

After completing this phase, you can start training the Baseline model by clicking on the “Train” button in the bottom right-hand corner of the window.

Your Output

The training of the Baseline model could take from several hours to a couple of days, depending on the number of pages and the learning machine parameters. You can check the training progress by clicking on the “Jobs” button under the “Server” tab.

When the training is finished, the Baseline model will appear in the “Server” tab, under “Model Data”. To see it, please select “layout” instead of “text” as model output type in the second drop-down menu, as shown below.



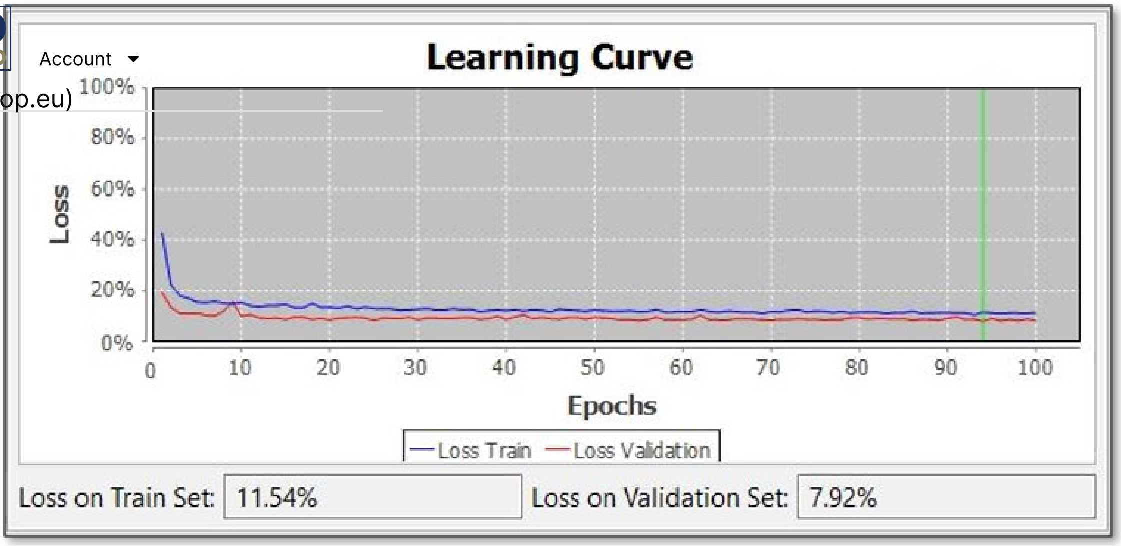
Collections: How-To Guides (148300, Owner)

layout

Figure 3. Layout as model output type

Double-clicking on the Baseline model name, you will see all the details and its learning curve. The “Learning Curve” graph shows the Baseline model’s accuracy. The x-axis indicates the number of Epochs, i.e. the number of times the training data is evaluated. The y-axis measures the Loss, i.e. the percentage amount of pixels classified incorrectly.

The program trains itself first on the Training Set; then, it tests itself on the pages of the Validation Set. For this reason, there are two lines in the graph. The blue line indicates the progress of the training; the red line indicates the progress of the evaluation on the Validation Set. Note that it is important that the two curves do not differ too much. If the two curves diverge, it is most likely that the Training Set differs too much from the Validation Set and the resulting model is not effective.



Loss on Train Set: 11.54%

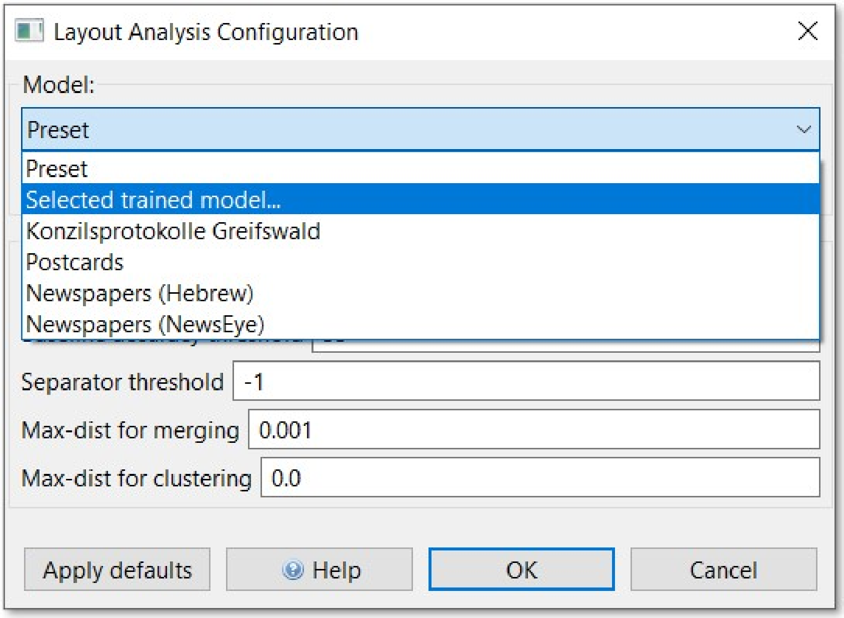
Loss on Validation Set: 7.92%

Figure 4. Learning Curve

Underneath the graph, the two percentages indicate how the Baseline model performs on the Training Set and the Validation Set in terms of Loss. The Loss on the Validation Set is the most significant value because it indicates how the Baseline model performs on new pages that it has not been trained on. Results with a Loss of 10% or below mean that the Baseline model is effective.

Applying your Baseline model

To apply the trained Baseline model to your documents, go to the “Tools” tab. Under the “Layout Analysis” top section, leave the “Transkribus LA” Method selected as it already is and click the “Configure” button. The “Layout Analysis Configuration” window pops up, and under “Neural Net” you can choose the trained Baseline model you want to apply.


Layout Analysis Configuration

Model: Preset

Separator threshold: -1

Max-dist for merging: 0.001

Max-dist for clustering: 0.0

Figure 5. Layout Analysis Configuration

By default, the Neural Net is set to “Preset”. To choose another model, click on the drop-down menu and select the trained model that best suits the layout of your documents.

The settings below enable you to configure the Layout Analysis when the default settings are not satisfactory on the layout of your documents. In particular, the setting you can configure are:

- Model: leave the "Preset" model if you haven't trained a specific baselines model on the layout of your documents.
  The Preset Transkribus LA model works well for most document typologies. Only if your documents have a complex layout and the preset model is unsatisfactory, you can train a Baselines model specific to your document typology, as explained here (https://readcoop.eu/transkribus/howto/how-to-train-baseline-models-in-transkribus/).

- Minimal baseline length: it indicates the minimum length of baselines in pixels. Baselines shorter than this length will not be detected.

- Baseline accuracy threshold: in the first stage of the layout analysis, each pixel is labelled as baseline, separator or other. The baseline accuracy threshold applies to the baseline labelling at this stage. It ranges between 0 and 255, and higher values enforce higher accuracy in the detected baselines.
  If you have low-resolution images and no or only a few baselines are detected, try to reduce the value. Bear in mind, however, that the results can get noisy for lower thresholds.

- Separator threshold: separators are small vertical lines drawn beside each baseline; they mark the beginning and end of each baseline (do not confuse them with actual separators in printed document images). As for the baseline accuracy threshold, the separator threshold refers to the first stage, when pixels are labelled.
  The separator threshold ranges between 0 and 255: 0 means that separators are not used at all; with a higher value, separators are used, thus, nearby baselines tend not to be merged.
  Usually, low values are sufficient to prevent a connection between nearby baselines. Use, for instance, 1 to use separator information "sometimes" and larger values to use them pretty much all the time, for instance, when text lines are close together but have to be separated because belonging to different columns.

- Max-dist for merging: in the second stage, the algorithm tries to merge nearby baselines but only when their distance is smaller than the set value. The value is not measured in pixels but is a fraction of the image's width. By default, it is set to 0.01: when two baselines are closer than the 0.01 fraction of the image width, they will be merged; if they are more distant than this value, they will not be merged. According to your layout and image's width, you can increase the fraction value to merge more distant lines or reduce it to prevent nearby baselines from being merged together.

- Max-dist for clustering: this value refers to the text region creation: after the baselines are detected, they are clustered in text regions based on their distance. The Max-distance for clustering is a fraction of the image's width: baselines that are closer than this fraction are clustered together in a text region.
  If too many text regions are created with the default settings, you can try to increase the value so that more baselines are clustered together. If it is set to -1, no region clustering will be performed, and only one text region will be produced as the bounding box of all lines.

For more information about the Transkribus LA algorithm and setting, consult this page (https://readcoop.eu/transkribus/docu/layout-analysis-help/).

Finally, click the “OK” button at the bottom of the “Layout Analysis Configuration” window. Your trained model has now been selected.

Under the “Tools” tab, choose the pages on which to apply the Layout Analysis and click the “Run” button: the Layout Analysis job will now start. You can check its progress by clicking on the “Jobs” button under the “Server” tab. Once the job is finished, reload the page/pages and the text regions, and baselines will appear in the images. No credit will be used to apply the Baseline model to your documents.

The COOP

About us (https://readcoop.eu/about/)

Join us! (https://readcoop.eu/join/)

Our Members (https://readcoop.eu/members/)

Success Stories (https://readcoop.eu/success-stories/)

Work with us (https://readcoop.eu/work-with-us/)

Products & Services

Transkribus (https://readcoop.eu/transkribus/)

Transkribus lite (https://transkribus.eu/lite)

Read&search (https://readcoop.eu/readsearch/)

ScanTent (https://readcoop.eu/scantent/)

Useful information

News (https://readcoop.eu/news/)

Download Transkribus (https://readcoop.eu/transkribus/download/)

Public Models (https://readcoop.eu/transkribus/public-models/)

Payment and shipping (https://readcoop.eu/payment-and-shipping/)

Helpful resources

Resource center (https://readcoop.eu/transkribus/resources/)

How-to Guides (https://readcoop.eu/transkribus/resources/how-to-guides/)

Getting started with Transkribus (https://readcoop.eu/transkribus/start/)

FAQs (https://readcoop.eu/transkribus/questions/)

Videos (https://readcoop.eu/transkribus/resources/video/)

Community

(https://twitter

(https://www.linkedin.com/

(https://www.youtube.com/cha

(https://www.facebo

Account

(https://readcoop.eu)

com/Transkribus)
om/company/read-coop-sce)
nnel/UC-txVgM31rDTGIBnH-zpPjA)
ebook.com/transkribus)

Terms & conditions (https://readcoop.eu/terms-and-conditions/)
Privacy Policy (https://readcoop.eu/privacy-policy/)
Contact (https://readcoop.eu/contact/)
Imprint (https://readcoop.eu/imprint/)

EN