READ
Recognition and Enrichment of Archival Documents

How to Use the Structural Tagging Feature and how to Train it

Version v.1.9.1
Last update of this guide 12/12/2019

This guide will show you how to enrich your documents with structural tags like ”paragraph”, “heading”, “caption” or “footer”. This mark-up makes it possible to define the structure of your documents. By now it is also possible to train the structure of a document in Transkribus.

In case you are searching for information regarding the word and phrase based tags like persons, places etc. please have a look at the How to Enrich Transcribed Documents with Mark-up and the Transkribus Transcription Conventions guides.

Download the Transkribus Expert Client, or make sure you are using the latest version:

- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:

- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact:

- The Transkribus Team: email@transkribus.eu

Contents

The READ project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943.

Introduction

With the structural tagging feature, you can mark up the structure of your documents.

Moreover it is possible to train models to automatically recognise the structure of your documents. Adding structural tags creates training data for this process.

There is no need to tag every feature of your documents – focus on marking up the sections that are of interest to you.

The structural tagging interface in Transkribus enables you to

- divide up your documents into structural sections like paragraphs, headings or page numbers.
- add customized tag categories for your individual needs.
- in future use this structural information for the training of a model.

Structural tagging interface

- First, open your document in Transkribus
- The structural tagging interface can be found by clicking the “Metadata” tab, and then the “Structural” tab.

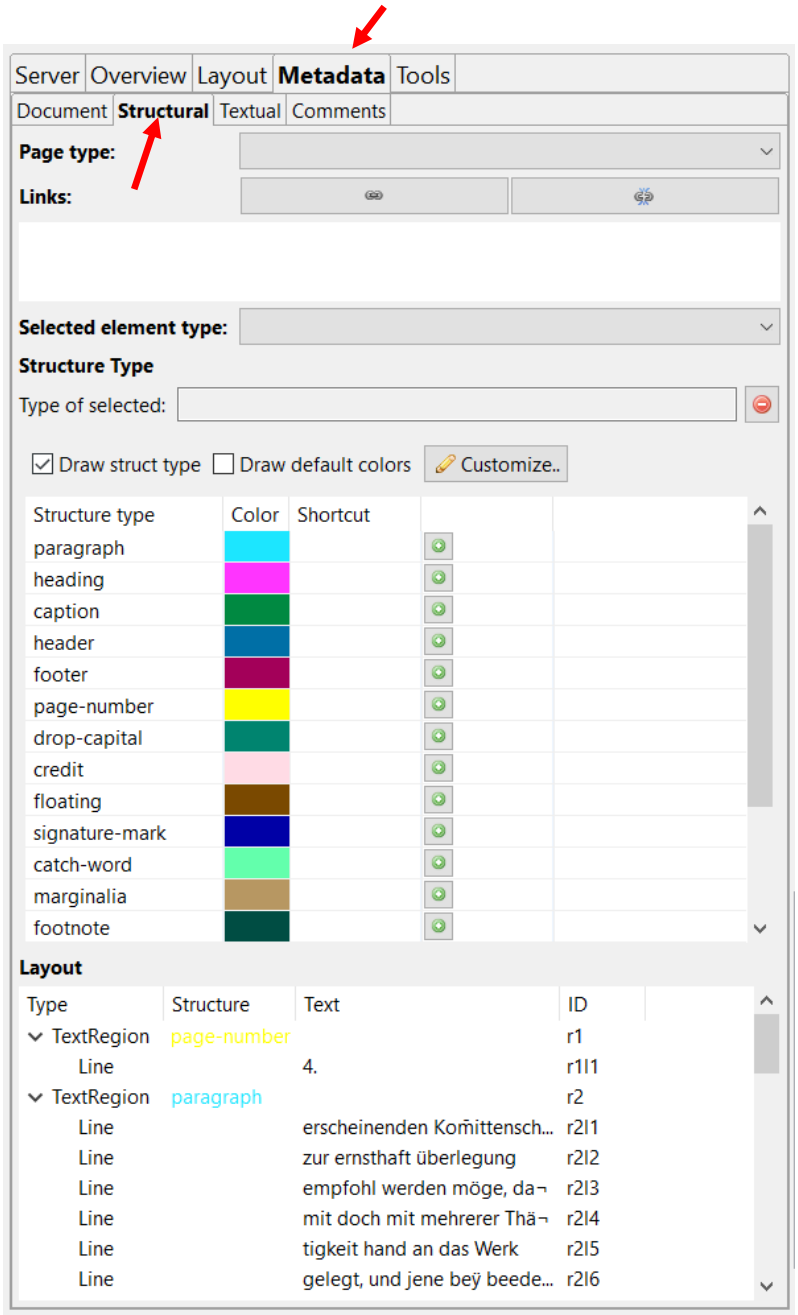


Figure 1 Where to find the structural tagging options

- In the centre of the tab you can see the different predefined structure types.

Create your own tag categories

- To create your own tag categories, click the “Customize” button. The “Tag configuration” window will open up.

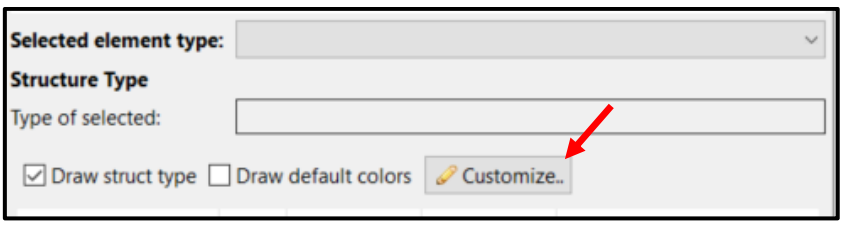


Figure 2 Customize button

- In order to create a new tag category simply type in the name in the blank box at the bottom of the window, then click the green plus button.

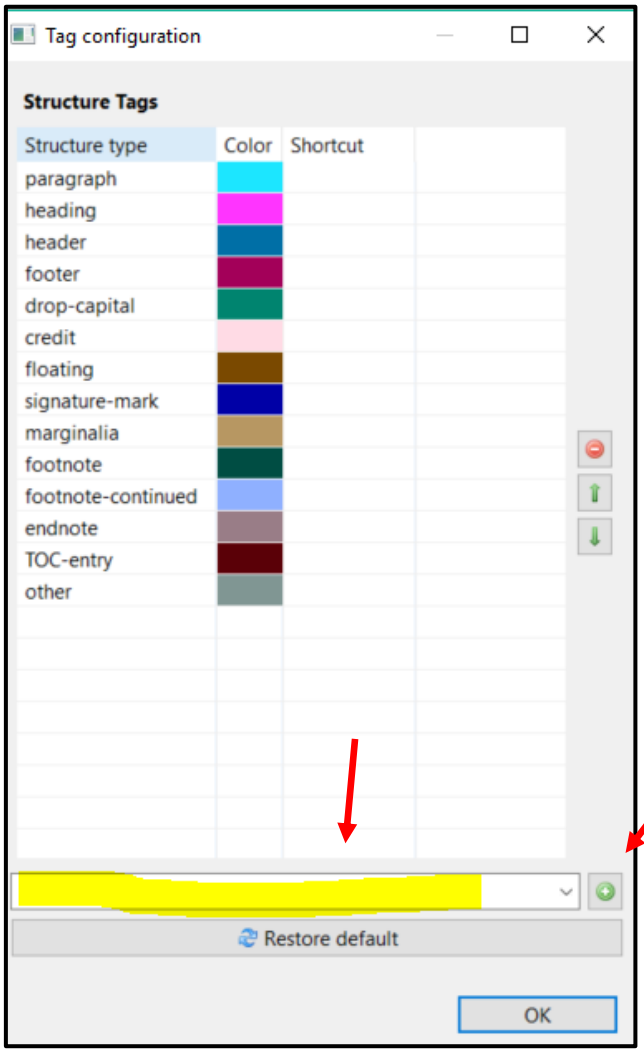


Figure 3 Create a new tag category

- In this window you can also customize the tag colours by clicking on the coloured section next to a tag and then choosing your desired colour.

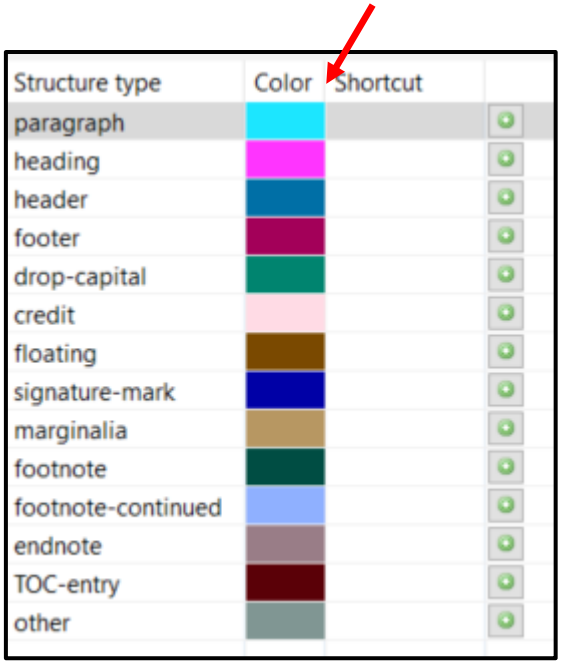


Figure 4 Customize colours

Figure 5 Choose colour

- The new tags you created will also be automatically available for all your documents in all your collections.

Assigning tags to elements of your document

- You can assign tags to text regions and line regions on each page in your document.
- The system of automated structural recognition is region-based, so it makes most sense to tag lines.
- Note: You do not need to tag every feature of your documents – the aim is to mark the sections that you are interested in.
- To place a tag first, click on the “Item visibility” button in the Main menu and make sure that text regions and line regions are visible on your document.

Figure 6 "Item visibility" button

- Click on a text or line region in your document. You can select several regions at once by holding down the "CTRL" key on your keyboard and then clicking on your document.
- You then have two options:
    - You can either add the tag by clicking the green plus button on the right of the desired tag category in the "Structural" tab.

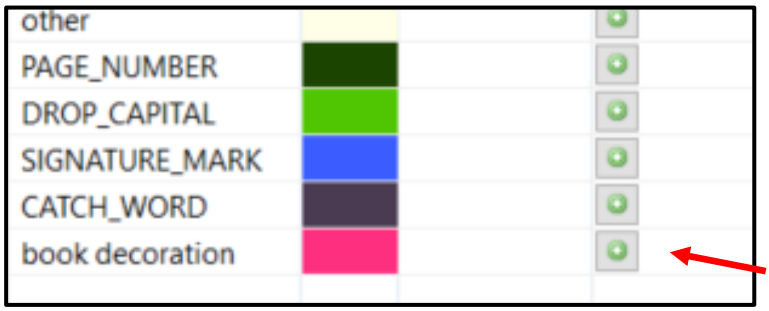


Figure 7 Assigning tags with the green white cross button

    - Or right click the marked section in your document and then choose the desired tag under "Assign structure type".

Figure 8 Assigning tags by right clicking

Link shapes

- You can link two structural tags together with the "Links" buttons in the "Structural" tab, e.g. a link between a line and the footnote connected with that line.

- The first button is to create such a link and the second one to remove it.
- Please note that for the training the linking of shapes is not relevant.

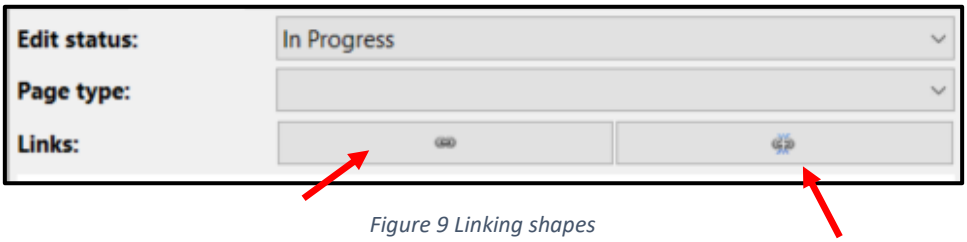


Figure 9 Linking shapes

Page type

- You can choose to assign a “Page type” to each page of your document.
- Possible options are:
  - Front cover
  - Back cover
  - Title
  - Table-of-contents
  - Index
  - Content
  - Blank
- When you have your page open choose the appropriate definition by clicking the arrow next to the “Page type” options and then choosing the desired type.
- Also the page type is not relevant for the structure-training.

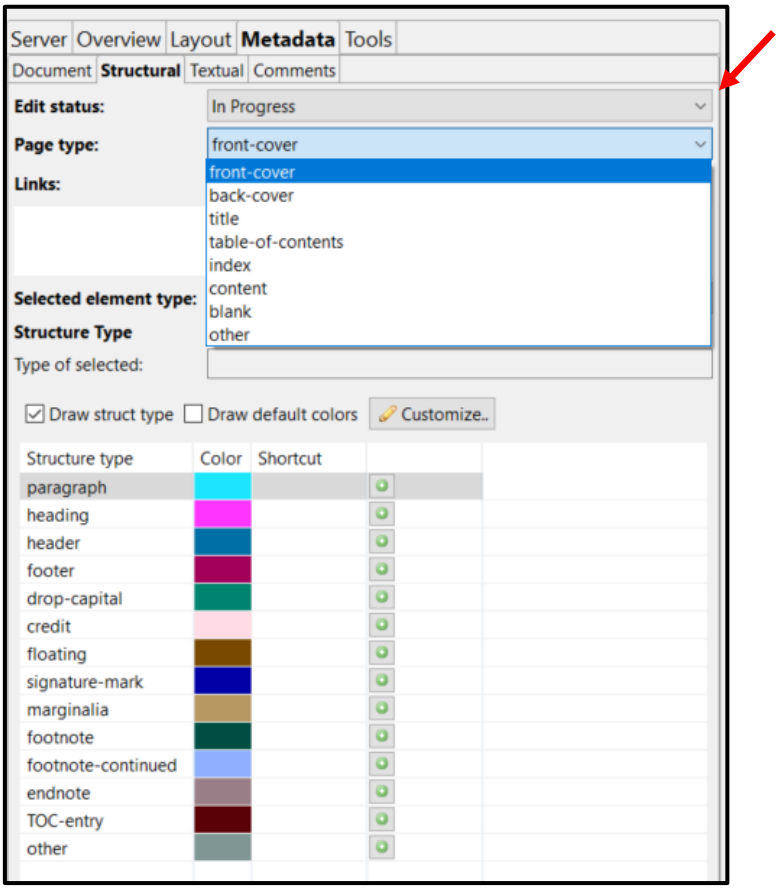


Figure 10 Choosing a page type

Layout section

- Within the “Layout” section of the “Structural” tab you can jump between the different structural types in your document.
- In this section, you will find an overview of the structural types in your document and snippets of any transcribed text. You may find it quicker to consult this list, rather than search for a particular line or text region in the image.
- To go to the desired text or line region, double-click the region in the “Layout” section. The image and the Text Editor will automatically jump to this line.

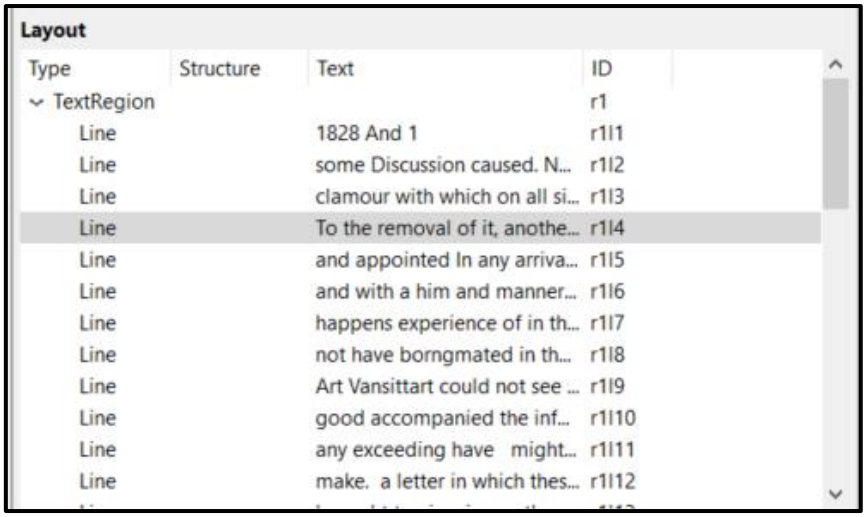


Figure 11 “Layout” section

- The tags you have added will be shown in the “Structure” column. Next to the structure type there is a small downward arrow. By clicking it, you can quickly change the structure type.

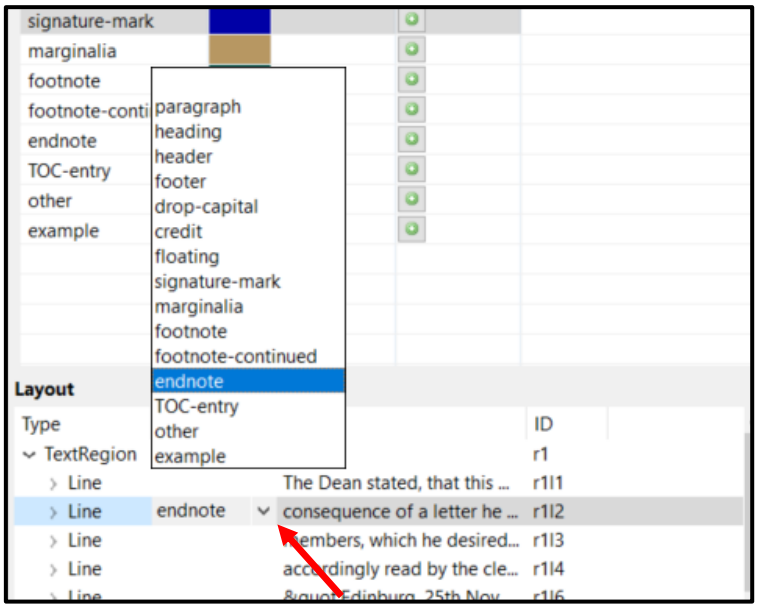


Figure 12 Changing the structure type via the “Layout” section

More Options

Deleting structural tags

- In order to delete a structural tag, click on it in the “Structure” column and in the menu selection choose the “—delete—” option at the top.

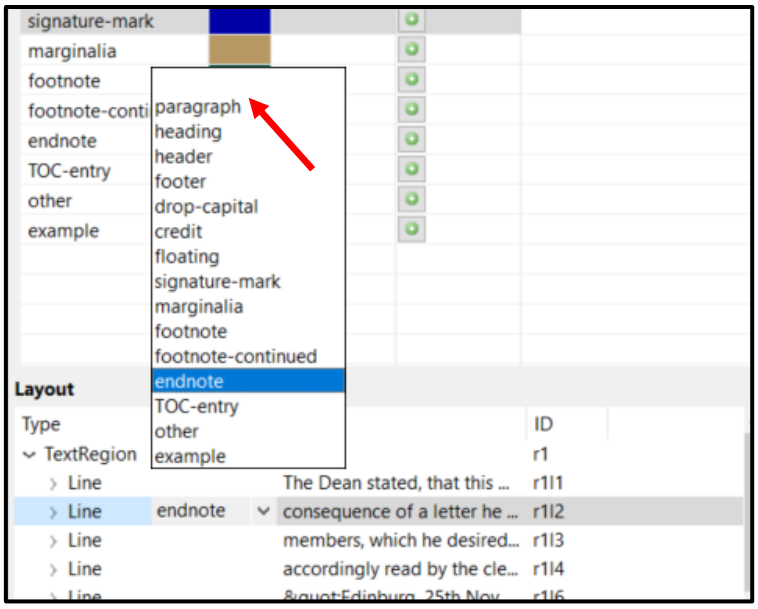


Figure 13 How to delete structural tags

Draw struct type

- If you choose this option, text labels will appear on your document describing each structural tag you have added.
- If you do not choose this option, the text labels will be hidden.

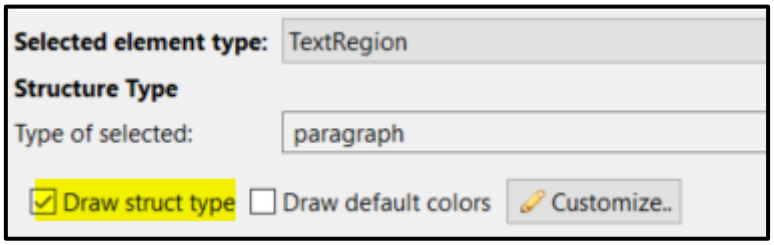


Figure 14 Show designations of the structural tags in the image

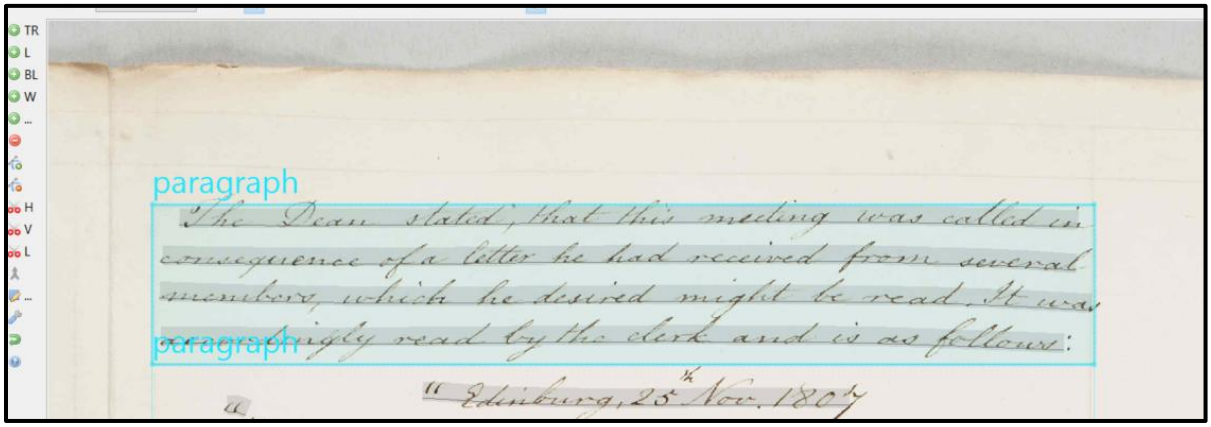


Figure 15 Draw struct type option

Draw default colours

- The structural tags are assigned different colours to the default colours for text and line regions.

- If you add structural tags to your document, the colours on your document image will change.
- If you would like to return to the default colour display in Transkribus, choose the “Draw default colours” option.
- Your tags will not be deleted – but the default colours will be displayed in your document, instead of the ones relating to the structural tags.

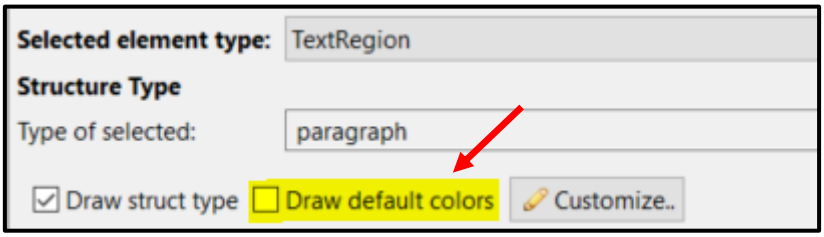


Figure 16 Show default colors

Type of selected

- When you click on a text or line region in your document, the “Type of selected” line in the “Structural” tab shows you which structural tag has been assigned to it.

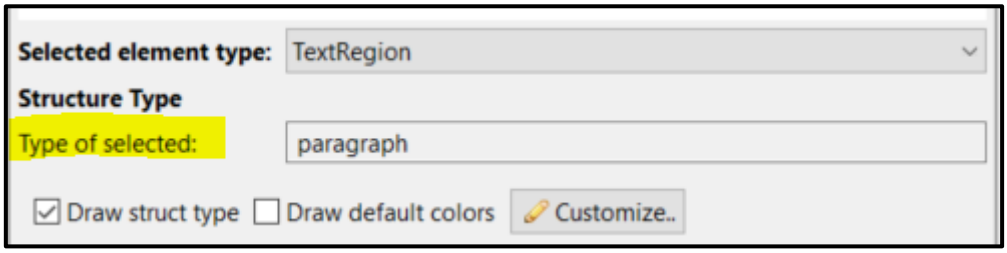
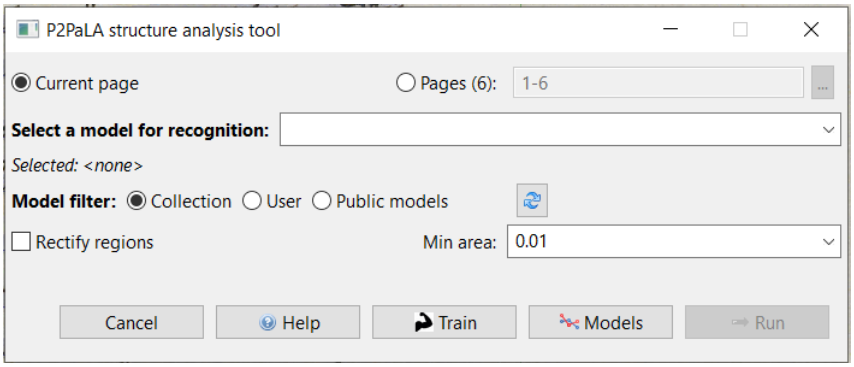


Figure 17 Currently marked structural tag

Structure-Training

With the structural training feature you will get a model, that can recognise the structure of your documents. The efficiency will depend, as with the Handwritten Text Recognition, on the quality of the training data. If you have tagged about 50 examples of every structure type, which should be trained this should be fair enough to start training, so 50-100 pages of training material should be suitable to create a useful model. Of course it is possible to start training earlier, with decrease in efficiency.

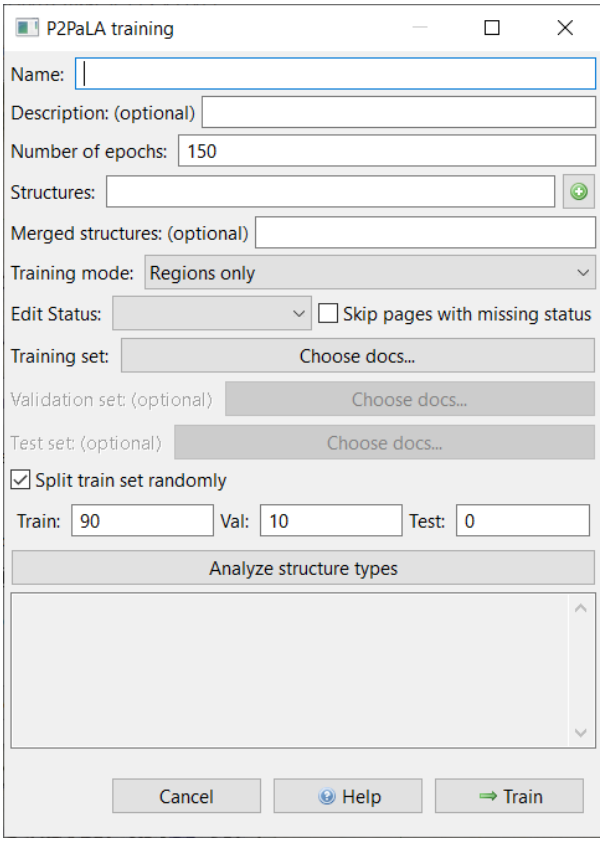
After finishing the tagging process you can start training. For this open the “Tools”-tab and click the “P2PaLA”-button in the “Other tools”-section. The following window will open:



Relevant settings for the training here:

- “Rectify regions”: all regions will be simplified to the bounding box of the actual recognized shape
- “Min area”: Shapes with an *area* smaller than this fraction of the image *width* will be removed after the recogniton. Use this parameter to remove small "garbage" regions. The default value is 0.01

If you click on “Train” the training parameters will open up:



In the upper sections some details about the model need to be added.

“Structures”: here you can add the structure types, which should be trained. When entering please pay attention to case-sensitivity and not to use the space bar. We recommend to use only lower case. Moreover we recommend to use hyphens (-) and underlines (_) as the only special characters.

- `Example: paragraph heading footnote page-number`

“Merged Structures”: are used to treat certain structure types the same as others during training (e.g. 'footnote-continued' or 'footer' like 'footnote'). Expected is a list of the structure types, separated by a colon with the structure types to merge.

- `Beispiel: footnote:footnote-continued,footer heading:header`

“Training mode”: here you can decide if you would like to train regions only, lines only or both. Please be aware that the baseline-training doesn’t mean, that structure types are trained on line basis. It is instead about the recognition of the baselines.

“Edit status”: if you would like to use the latest version, you don’t need to choose anything, otherwise you can choose, which status of the document should be trained.

“Training set”: this is the place to choose the training data.

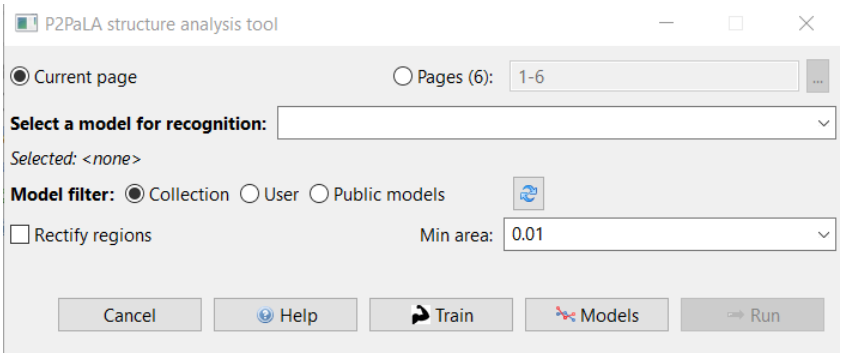
“Analyze structure types”: gives an overview about the number and types of structure tags within the chosen document.

To start the training, click “Train”.

After the training process is finished, the model is available for your collection and can be shared with other collections too.

Applying a Structure-Model

If you would like to apply a structure model to a document in order to let structure types be recognised, open the “P2PaLA”-feature within the “Tools”-tab.



Choose which pages should be recognised.

“Model filter”:

- “Collection”: when the desired model is in your collection
- “User”: when you have trained the model
- “Public models”: if you would like to use a public model.

After choosing one of the options, the available models will appear next to: “Select a model for recognition”. Choose the model you would like to use. An overview of all the models you get by clicking on “Models”.

“Rectify regions”: all regions will be simplified to the bounding box of the actual recognized shape

“Min area”: Shapes with an *area* smaller than this fraction of the image *width* will be removed after the recogniton. Use this parameter to remove small "garbage" regions. The default value is 0.01

To start the recognition, click “Run”.

Credits

We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission under grant agreement No 674943.