> **Attention**
>
> The site is currently being revised. If you encounter any problems , please **contact us.**

4. Workflow

OCR4all basically offers two different variants of an OCR workflow, which can differ greatly in terms of the amount of work involved, but almost inevitably also in the verifiability of partial results and thus the quality of the data created. Both variants are presented and classified below.

4.1 Process Flow

The variant of the so-called " **Process Flow** " (main menu ☰ → Process Flow) offers the possibility of an almost fully automated workflow. Here, only the scans intended for processing are selected in the right-hand sidebar and then all those work steps that are to be carried out on the available data material are selected with a tick.

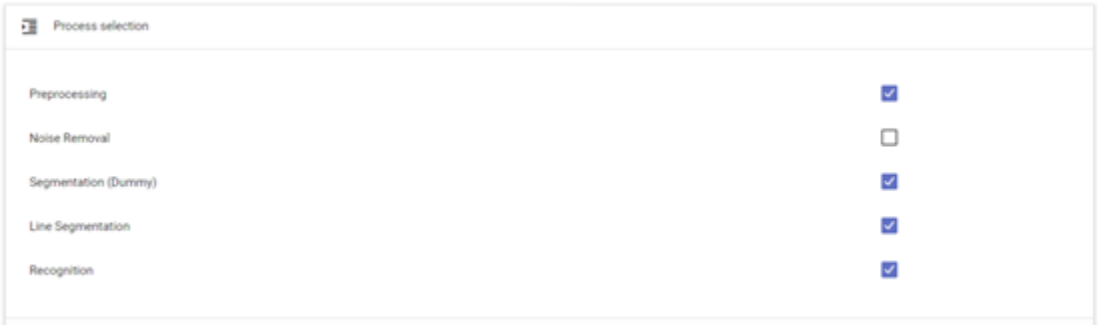

Fig. 5: Sub-components of the "Process Flow".

Only for the sub-module "Recognition" does a suitable OCR model or model ensemble (five individual models acting simultaneously and with one another, see also Chapter 4.7) have to be selected for recognition (this is done under "Settings" → "Recognition" → " General"), as shown in the figure below, from the list of all available OCR models ("Line recognition models – Available").

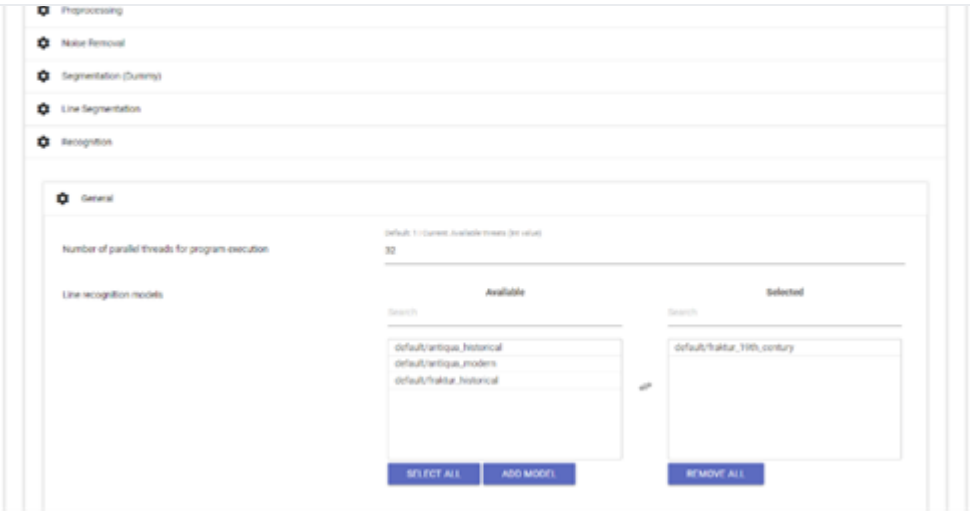

Fig. 6: Selection of a suitable OCR model.

In general, it is possible to select more than one model for recognition. However, this is only recommended if different types occur within the printed text to be recognized.

"EXECUTE" starts the "Process Flow". The current status of automated processing can be tracked via progress bars for the individual sub-modules. After the workflow has been completely run through, the results can be checked in the "Ground Truth Production" (☰) menu item.

Fig. 7: Line images with the corresponding OCR result.

If the OCR texts created correspond to the desired or required recognition accuracy on a line basis, final OCR results (TXT and/or PageXML) can already be generated under the menu item "Result Generation" (☰). If the results do not correspond to the desired accuracy, they can be corrected again before the results are output (see Chapter 4.8).

and their associated work steps in order to guarantee the correctness and quality of the data produced . Since the separate sub-modules build on each other, this approach appears to be particularly useful in the case of processing early modern prints with an elaborate and complex layout.

V. a. At this point, first-time users are advised to carry out the following step-by-step OCR workflow at least once in order to understand how the respective sub-modules work.

4.2 Preprocessing

Input: Original image (color, grayscale or binary)
Output: Corrected binary (and grayscale) image

- This processing step serves to create binary and normalized grayscale images, which are the basis for successful segmentation and OCR.
- All scans to be edited are selected in the right sidebar; All settings (“Settings (General)” and “Settings (Advanced)”) remain, ie the angle of the images to be processed remains unchanged, as does the automatically generated number of CPUs used by the sub-module (the latter affects all subsequent sub-modules of OCR4all!).

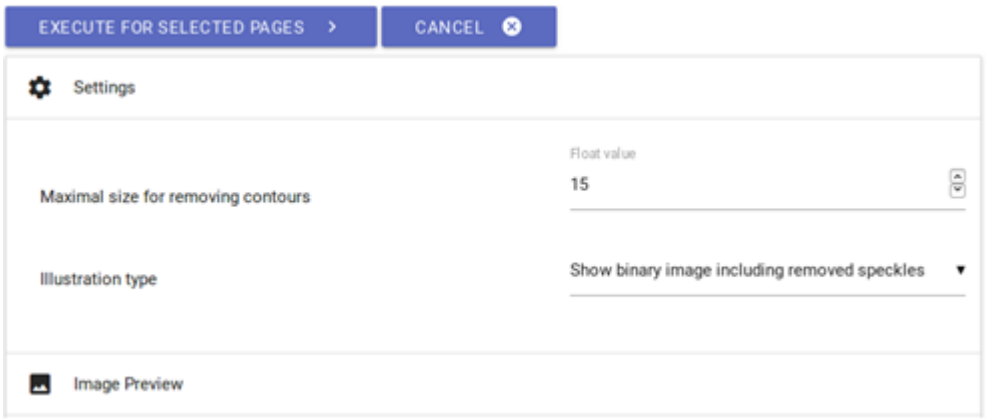

Fig. 8: Preprocessing settings.

- The binarization process can be started by clicking on "EXECUTE". The course of the work step can be followed in the console, more precisely the "Console Output". Warnings may appear in "Console Error" during the binarization process. However, these have no effect on the result of the binarization.

should appear in the “Preprocessing” column of the project overview for all processed image files.

4.3 Noise Removal

Input: contaminated binary images
Output: binary images with no or only a few impurities

- With the help of the Noise Removal option, for example, smaller impurities such as spots and dots on the scans can be removed.
- To use it, click on the “Noise Removal” step in the main menu and select which scans this process should be applied to on the right-hand side of the screen. Leave all the defaults as they are and, after pressing "EXECUTE", look at the result as a test by clicking on the lettering of the respective scan in the right-hand sidebar that you want to look at. The result is now displayed in a comparison with the unprocessed scan under “Image Preview”. Picture elements colored red were removed by the work step.

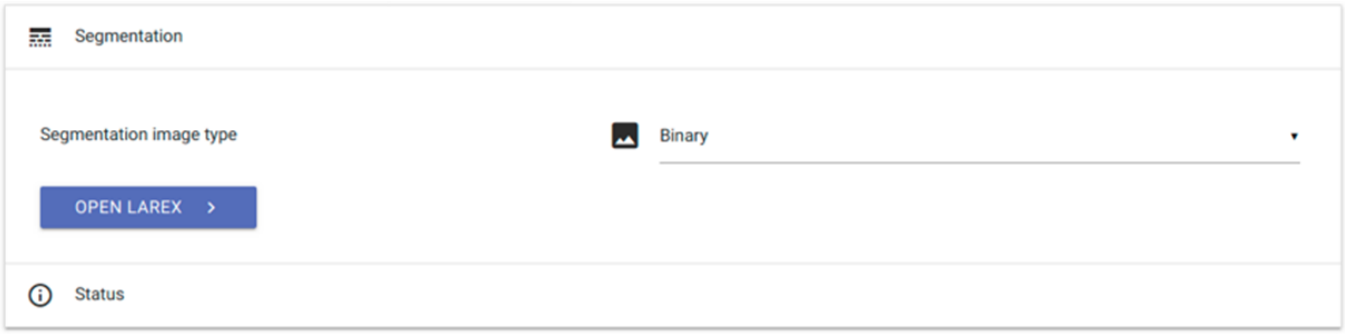

Fig. 9: Settings for the Noise Removal sub-module.

- If too many disturbing elements can still be seen on the scan, increase the value of the "Maximum size for removing contours" slightly, carry out the work step again by clicking on "EXECUTE" and check the result again.
- If too many image elements were removed, adjust the “Maximum size for removing contours” value downwards.
- Continue in this way until you are satisfied with the result.

4.4 Segmentation - LAREX

LAREX is used as a segmentation tool, ie for structuring and classifying the layout of printed pages with a view to further processing steps. The fundamental assumption here is that the pages of particularly early print products are composed of a recurring pool of different layout elements, and that their structure is also intrinsic to the work, i.e. more uniform to a certain degree. For this reason, the user has various tools and aids at his disposal to structure a print page, ie to segment it, so that all information on a page that is necessary for the subsequent components of the workflow and that relates to the page layout is adequately recorded. In addition to the basic distinction between text and non-text (e.g. text vs.

4.4.1 Preferences

- Menu: "Segmentation" → "LAREX"
- "Segmentation image type": "Binary" if you want to continue working with the binarized image files; "Despeckled" if the "Noise Removal" step was completed beforehand
- "OPEN LAREX" → LAREX opens in a new tab.

Fig. 10: LAREX settings.

The first of the selected scan pages is now displayed in the middle. The first segmentation results can already be seen. These arise due to an automatic segmentation of each scan page as soon as it is called up for the first time. These results are not saved. The task of the user is then to make settings in order to adapt the displayed automatic segmentation results to the layout of the present work or to make manual corrections to these results in order to obtain a correct segmentation result.

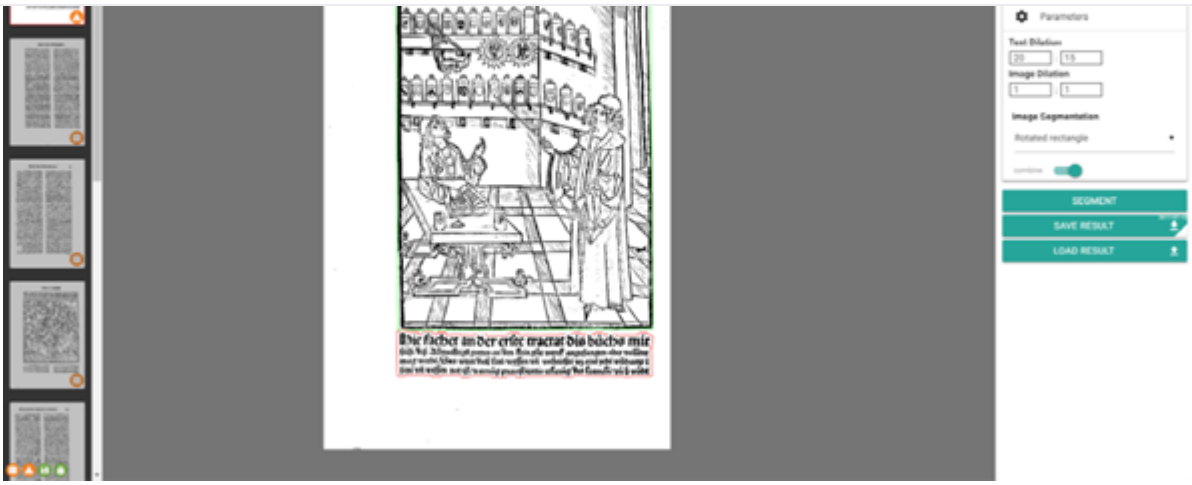

Fig. 11: Start screen and automatic segmentation results.

4.4.2 Overview and toolbar

All scans to be segmented and previously selected are displayed in the left sidebar. Depending on the current processing status, different colored markings appear in the lower right corner:

- Exclamation mark, orange: "There is no segmentation for this page." - There are currently no segmentation results for this scan page.
- Warning triangle, orange: "Current segmentation may be unsaved." - The current segmentation results have not yet been saved (see below).
- Disk, green: "Segmentation was saved in this session." - The segmentation results, saved as XML files, are available for the scan side.
- Lock, green: "There is a segmentation for this page on the server." - The individual, saved segmentation results were confirmed as correct after the segmentation of the entire work was complete (see below).

Fig. 12: Different display modes

- Using the '0' and '1' buttons, it is possible to switch between the binarized (black and white) and the normalized (grayscale) image display. The corresponding selection is remembered for all remaining pages of the work. It is possible to change the display mode again at any time.
- The header contains various tools and tool groups for navigation and editing:

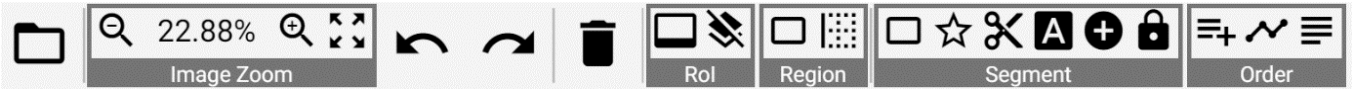

Fig. 13: Different menu items of the toolbar.

- *BILDCHEN* **Open a different book** : No settings or changes are necessary for the version of LAREX integrated in OCR4all!
- **Image Zoom** : The general display of scanned pages and image files in LAREX is regulated via the settings that are possible here, ie eg zoom settings. However, these settings and display options can also be controlled using the mouse and/or the touchpad (simply move

- *BILDCHEN* **Undo** and *BILDCHEN* **Redo** : Undo or redo the last user action. Common key combinations are also possible here (e.g. CTRL + Z = undo last action).
- *BILDCHEN* Delete selected items: Removes the currently selected regions.
- **RoI** , **Region** , **Segment** , **Order** : The various options for scan processing and segmentation are shown here, supplemented by the sidebar on the right. While the options listed in the toolbar are generally used for specific editing of the currently available scan page (see below), the sidebar on the right primarily shows scan-wide and work-related options.

Fig. 14: Right sidebar settings.

(“Parameters”) and the layout elements (“Regions”) that are present in a work and specified by the user under “Settings” at any time and the next time you use the tool to reuse. This enables working with factory-specific settings.

4.4.3 Factory Settings: Regions, Parameters, Reading Order, Settings

- " **Regions**“: According to the conception and idea of LAREX, each scanned and thus work and text page consists of different layout elements. This includes e.g. B. the main text, headings, marginalia, page numbers, etc. Each of these layout elements must be assigned a specific, defined "region" or layout region in LAREX. With a view to further processing steps and the actual recognition of the displayed content, this assignment will be carried out consistently across the entire work to be segmented! In addition to some predefined and fixed layout regions such as "image" (e.g. graphic representations such as woodcuts, decorative initials, etc.), "paragraph" (main text) or "page_number" (page number), the user can create further, work-specific layout regions under "Create". be added and defined, ie In addition to a display color, the minimum size of a text or image region on the scanned page to be recognized as a corresponding layout region can also be specified under "minSize". The layout region defined in this way is added to the work-specific list using the "SAVE" button.

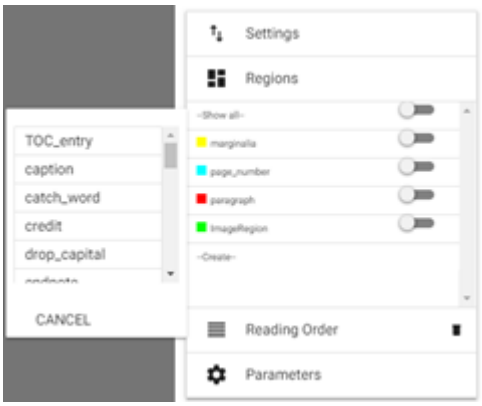

Fig. 15: Setting options under Regions.

- In addition, " **Regions** " offers the possibility of assigning certain layout regions a fixed and predefined place on a scanned page, which is taken over during the automatic segmentation of the following pages (when opening them for the first time), ie: The layout of a page is always repeated throughout a work again, a type of layout template can be created here, with the help of which the automatic segmentation can be improved and the number of

position of the layout regions can be displayed and then changed by simply selecting the regions on the scanned page.

Fig. 16: Display of layout regions and layout template.

- If the user defines a new "region", the position of this can be specified via the toolbar and the following option "Region" → "Create a region rectangle (Shortcut: 1)" and changed at any time afterwards. For "images" no layout region can be located on the scan page.

Fig. 17: Setting up new layout regions.

- At the same time, it does not always make sense to specify fixed locations for all layout regions across the entire work on scanned pages. V. a. If the position of certain "regions" such as headings, mottos, but also page numbers or sheet signatures varies again and again, the definition of defined places can lead to incorrect identifications. In this case, it makes more sense to manually correct the corresponding layout elements after the automatic segmentation. If the position of layout regions is to be completely deleted, they are simply selected with a click and deleted with "Del".

- " **Parameters**“: General parameters for text and image recognition are specified here. The need to set work-specific parameters can be explained by the very inconsistent layout and print image, especially of early modern prints. Words and entire lines can be printed at different distances from one another. In order to avoid, for example, that these are

Dilation". In this way, line and word spacing can be overcome and lengthy sections of text can be merged with one another. It is advisable to test different factory-specific settings in order to optimize them.

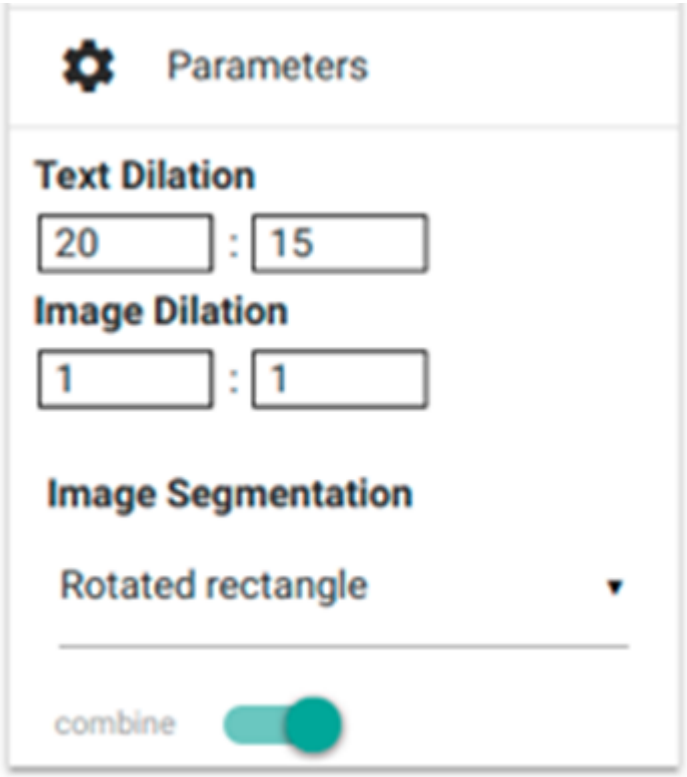

Fig. 18: Settings in Parameters.

- " **Settings**“: Under the “Settings” menu item, the segmentation and display options specified under “Regions” and “Parameters” can be saved and, if necessary, e.g. B. when resuming the segmentation of a work after an interruption. The “SAVE SETTINGS” and “LOAD SETTINGS” buttons are used for this. If you save, an XML file is generated that must be selected again when loading (click on "Load Settings", select and open the appropriate file in the window that opens). In addition, there is also the option here of reloading the segmentation results of pages that have already been saved and thus displaying them. To do this, click on "LOAD NOW" under "Advanced Settings". If an XML file with segmentation results was once saved for the present scan page, will now be loaded. At the same time, this last option can be implemented automatically from the start of LAREX, provided that the corresponding segmentation results are already available.

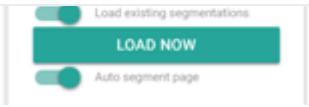

Fig. 19: Settings.

- " **Reading order** ": If the text on a page is to be reproduced in the correct order in the recognition results that follow segmentation and can be generated later, it is essential to define a reading order for those layout elements that contain text. This definition can be automated, for example with a clear and simple print image. In the case of more complex layout structures, however, it is advisable to set the reading order manually in order to avoid errors in the order. To do this, select between the tools “Auto generate a reading order” and “Set a reading order” in the “Order” group in the toolbar.

Fig. 20: Right: Reading Order in the toolbar.

- If you click on the automated creation of the reading order, a naive listing of all text-containing layout elements from top to bottom appears in the right sidebar under "Reading Order". If the order is set manually, the user must click on the individual elements on the scan page in the correct order in order to appear in the list mentioned (see below). The order of the individual elements of the reading order can be changed using drag and drop, and individual elements can be removed using the associated trash can icon. Like all other interventions in LAREX, the reading order can also be changed again and again before the segmentation results are finally saved.

4.4.4 Sample segmentation of a scan page

LAREX automatically creates the first segmentation results when loading a scan page. These must be corrected below.

The following segmentation pass refers to the fourth page of the standard work "Cirurgia", which can be downloaded when downloading the OCR4all folder structure LINKhierLINK.

recognized but affect the segmentation result?

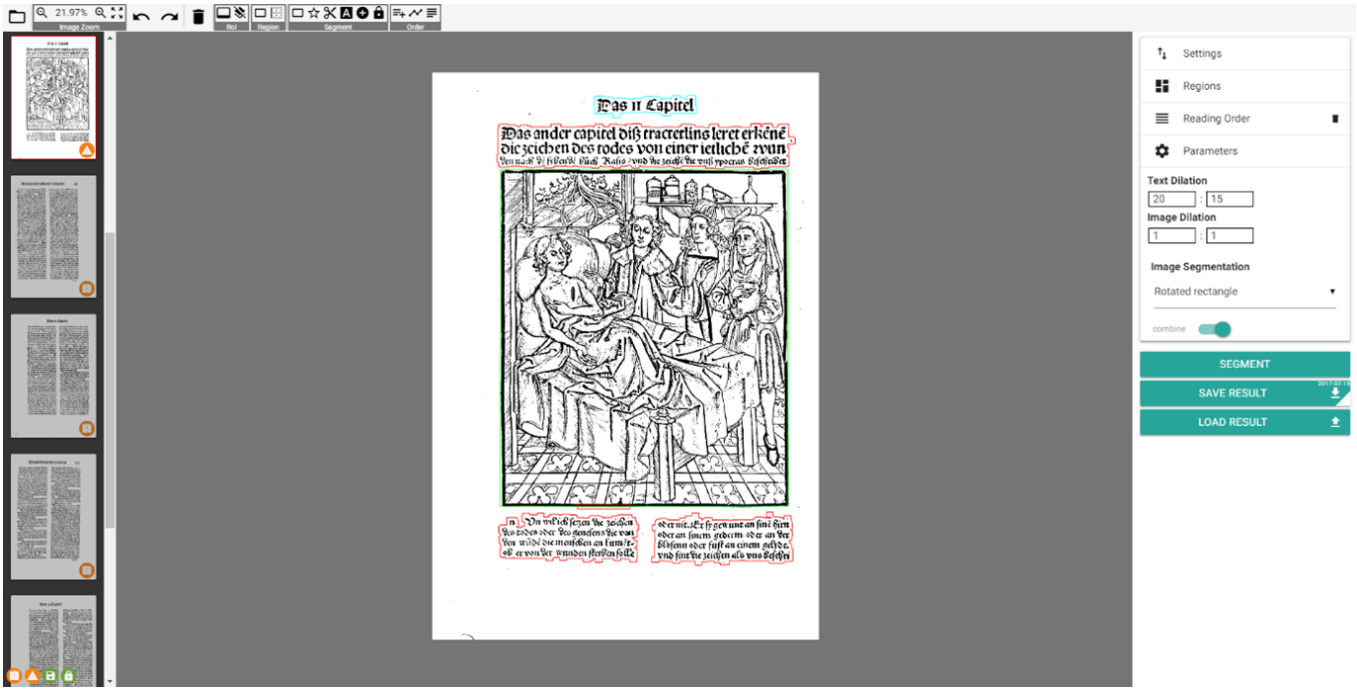

Fig. 21: Automatic segmentation result for the fourth page from "Cirurgia".

"Region of Interest" (RoI) : If there are elements outside of the sections of a scan page that are relevant for the recognition that have a negative impact on the segmentation result (e.g. user traces, dirt, library stamps, etc.), then a RoI can be defined to exclude these areas from the automatic segmentation from the outset. To do this, select the "Set the Region of Interest" option under "RoI" in the toolbar and place a rectangle around the content of the scan page to be segmented using the left mouse button.

Once the RoI has been defined, click on the "**SEGMENT**" field on the right-hand side - elements that are outside the RoI are no longer taken into account. Important: If an RoI is set, this is also transferred to all scan pages that are called up in the further course of work on a work. Since the sections relevant to segmentation on a page can be shifted again and again due to various factors, it is likely that the RoI will also have to be adjusted to the page conditions from time to time. To do this, you can simply click on individual areas of the ROI and move them with the mouse.

Independent of the RoI, a so-called ignore region can be created using the "Create an ignore rectangle" option, with the help of which smaller scan components can be ignored and thus excluded from the segmentation.

Correction of incorrectly recognized layout elements : Incorrectly recognized layout elements can be changed in their typing. To do this, right-click on the corresponding element - the correct region can be selected in the selection window that opens.

Fig. 23: Correction of an incorrect typing.

If the heading is to be separated from the text following it due to its conglomeration, this can be done in three ways:

in the toolbar, then use the mouse to draw a window around the corresponding region and select the correct name in the selection menu that opens. Second, the region to be classified can be selected using a polygon. This is particularly useful for complex, confusing or nested layouts in which sloping edges, curves in images and woodcuts or decorative initials or similar placed in the text block. occurrence. For this purpose, the option “Create a fixed segment polygon” (shortcut: 4) is selected and the layout region to be classified is enclosed in a dotted line, the end of which is linked to the beginning and thus combined into a polygon. Here, too, after connecting the start and end point, a selection menu appears in which the correct designation can be selected.

The third option involves dividing the text block recognized as a paragraph from the heading and main text using a cutting line. This is selected in the toolbar under "Segment" with the option "Create a cut line" (Shortcut: 5).

Fig. 24: Selection of the cutting line in the toolbar.

With the help of the left mouse button, the line is drawn polygon-like through several clicks across the layout element to be split. A double-click on the left mouse button sets the end point of the line.

Fig. 25: Definition of the cutting line between two areas of a layout element to be separated.

clicking and making the appropriate selection (see above).

Fig. 26: Correct typing of the separated areas.

If layout elements, incorrectly drawn cutting lines, distorted polygons, etc. are to be deleted, they can simply be marked with a left click of the mouse and then deleted with "Del" or in the toolbar with "Delete selected items".

Determination of the "Reading Order" (see above):

Fig. 27: Definition of the reading order.

Saving the segmentation result of the current scan : The results are saved by clicking on the "SAVE RESULT" button or by Ctrl + S. At this moment, an XML file with the segmentation results is stored in the OCR4all folder structure.

Fig. 28: Saving segmentation results.

The next scan can then be selected in the left sidebar. If the segmentation of a scan is to be changed again later, the new segmentation simply has to be saved - in this way the outdated XML file is then overwritten by the current and new one.

4.4.5 More editing options

In addition, there are generally **other processing** options for scans, which are shown below:

- For deletions or the merging of several layout elements into a contiguous region, it is useful to be able to **select them at the same time** . To do this, hold down the Shift key and use the mouse to draw a rectangle around the corresponding layout regions. The regions must be completely inside the rectangle. All layout regions selected in this way are now outlined in blue.
- **"Select contours to combine (with "C") to segments (see function combine)"** (Shortcut: 6): This tool can be used to achieve an optimal segmentation result even on very narrow and detailed printed pages. The basic idea is that layout elements are limited by the contours of the individual types of text they contain, or exactly by the edges of images and decorative initials - without excess margins created by manual segmentation, which repeatedly lead to element overlaps and thus to inaccuracies Consequences for the OCR can lead.
  - To execute the function, first click on the corresponding button in the toolbar or on the shortcut 6. Then all components recognized as layout elements of the page are colored blue.

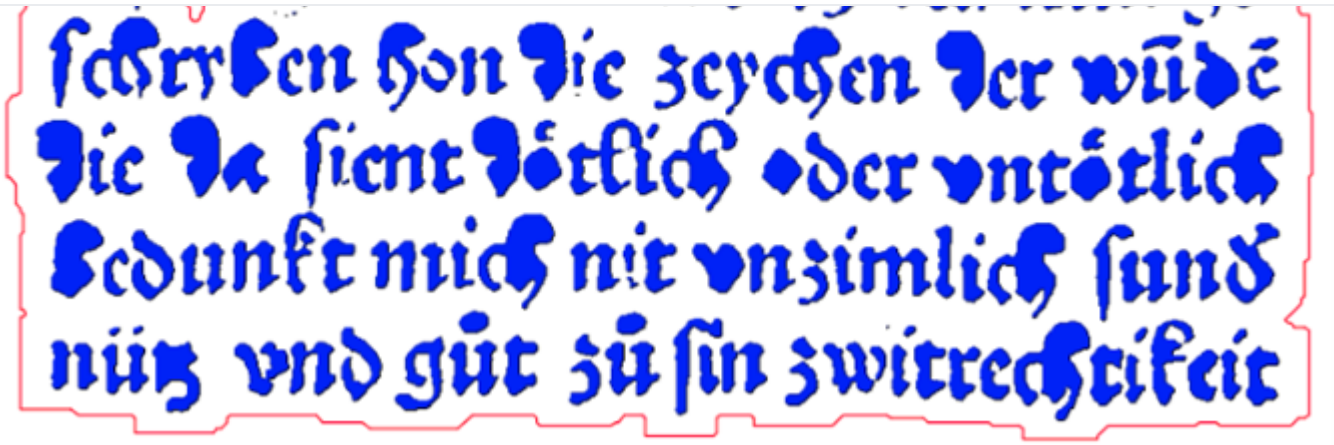

Fig. 29: Contour display.

- If you now click on individual types or even type components, they turn purple - they are now selected.

Fig. 30: Contour selection.

- Several types, entire words and lines or parts of entire layout elements can also be selected (like this: Shift + selection by drawing a rectangle). If the shortcut C is used after selecting certain types, words, lines, etc., all selected elements of the scanned page are combined into a separate layout element, regardless of their previous element affiliation. The delimitation of the resulting new layout element is much finer in comparison to the automatically recognized elements because, as discussed, it is based directly on the edges of individual

Fig. 31: Combination of selected contours into a new layout element.

- The subsequent click on "SEGMENT" fixes the intervention. Finally, the resulting, independent layout element can be renamed as desired according to the procedure above.

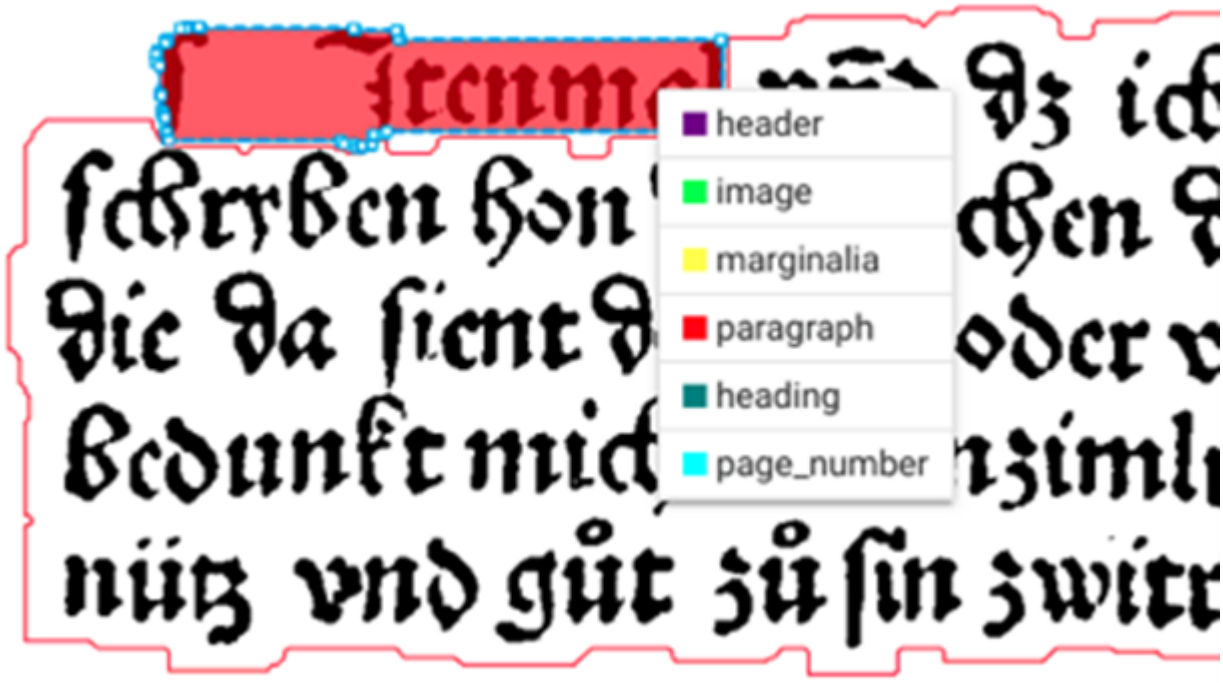

Fig. 32: Typing of the segmented layout element.

- **"Combine selected segments or contours"** (shortcut: C): To combine several individually recognized layout elements into one, select the desired regions completely (see above) and click "C" or the corresponding button in the toolbar.

page. To do this, the corresponding layout element is marked by clicking, followed by a click on "F" or the corresponding button. Fixed items appear with a dashed border. To discard the fixation, the process is simply repeated.

- **Zoom : The mouse wheel can be used to zoom** in on the scan when the text is very small or the layout is complicated. Pressing the spacebar resets the display to its original state.
- In the case of a particularly small and therefore complex layout, segmentation results can be further optimized by special **detailed interventions .** The outlines of the areas of a scanned page recognized as layout elements are displayed as a dotted line on closer inspection.

Fig. 33: Dotted line as an outline of layout elements.

- These points can be moved individually or in groups, for example to avoid overlaps with other adjacent layout elements in the case of a very narrow print image. Individual points can

- **“LOAD RESULTS”** : This function can be used to load existing segmentation results for a specific scan page directly from the OCR4all folder structure into LAREX.

4.4.6 Completion of segmentation with LAREX

- Once all segmentation work for a work in LAREX has been completed, ie results have been saved for each page of a work, these are now available in the well-known folder structure of OCR4all.
- Whether the segmentation and saving of the results was successful can then be checked in the "Post Correction" menu item in the "SEGMENTS" column (see below).

4.5 Line segmentation

Input : preprocessed images and segmentation information in the form of PageXML files
Output : extracted text lines in the PageXML files

- In direct preparation for the following OCR, in this step all layout elements that contain text and are defined and classified using LAREX are cut into lines (the OCR works line-based) and stored in the associated PageXML.

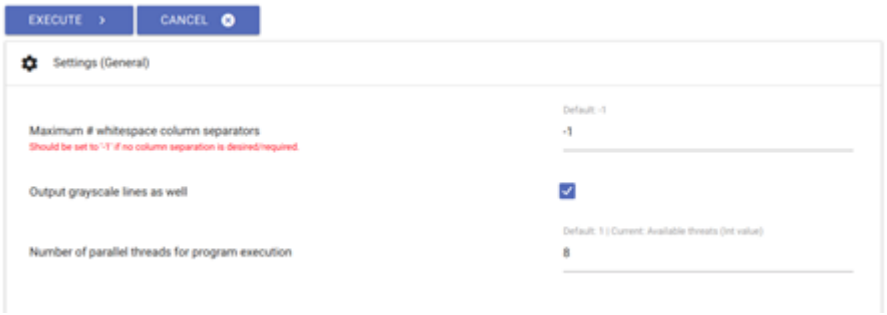

Fig. 34: Line segmentation settings.

- In general, the existing settings can also be retained here. **Important restriction with regard to the existing page layout** : If there is a two- or multi-column page layout and the corresponding text columns in LAREX were each segmented as independent main texts, the default value of -1 (confirmation that no multi-column layout available and a column separation is therefore not desired) can be changed as follows:

- In a **two-column layout** , the text of which is continuous in content, i.e. the first lines of the two text blocks do not form a unit in terms of content, the value for "Maximum # of whitespace column separators" must be set to 3: This specification results from the left whitespace of the left column of text, the right whitespace of the right column of text and the common whitespace between both columns of text.
- With a three- **column layout** , the value would have to be changed accordingly to 4, etc.

- Once all the settings are the way you want them, click EXECUTE and check the results again under Project Overview. Here, the individual lines are given as subitems of the individual layout elements (see above).
- Using the advanced settings (“Settings (Advanced)”) is always helpful, especially for line segmentation – especially when error messages are displayed in the console and the line segmentation could not be carried out correctly. For example, if the letters are too small, the minimum width of whole lines specified in the defaults is often undercut. However, this minimum width can be changed, for example, by reducing the value "Minimum scale permitted" under the menu item "Limits". The repeated execution of the line segmentation for the selected scan pages is then carried out correctly without an error message.
- The correct line segmentation can also be checked under the menu item “Post Correction” in the “Lines” tab (see below).

4.6 Recognition

Input : Lines of text and one or more OCR models
Output : OCR output in text form for each line present in the PageXML files

- The Recognition step represents the recognition of text based on the line images of all layout elements with text created during line segmentation (see above).
- To do this, select the menu item “Recognition”. In the sidebar on the right you will now only find scans or print pages of the edited work listed for which all the preconditions of the OCR have already been met, ie all the work steps described so far (with the exception of "Noise Removal") have been carried out. Select those for which you want to have text produced.
- Now under "Line recognition models" in the "Available" column, select all those models or model ensembles that are required to recognize your text according to the existing fonts and types (e.g. early modern or historical Fraktur, italics, historical Antiqua, etc.) are suitable. The use of model ensembles (five individual models acting simultaneously and together) instead

are a particularly large number of models to choose from.

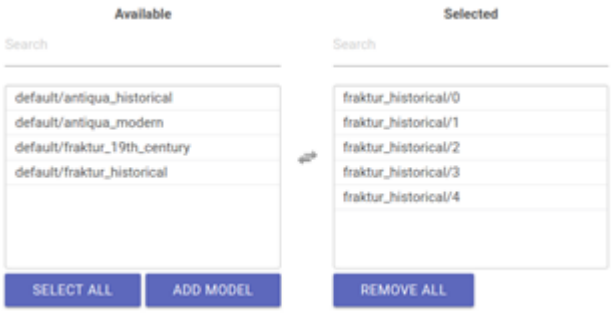

Fig. 35: Selection of a mixed model ensemble for text recognition.

- It is usually not necessary to adjust the advanced setting.
- Now click on "EXECUTE" and wait for the text recognition via the progress bar and the console.
- When recognition is complete, you can view the results for each line image under the Ground Truth Production menu item.

4.7 Ground Truth Production

Input : line image and corresponding OCR output if available
Output : line based ground truth

- Under the menu item "Ground Truth Production" the texts generated in the sub-module Recognition can be viewed, corrected and saved as a basis for training in the form of so-called Ground Truth.
- The correction tool is based on a three-column structure: on the left-hand side you will find the selectable pages, one below the other. The line images generated by the workflow from the text pages (see above) and the lines of OCR text generated from them are displayed in the middle. This display, which is displayed by default, is referred to as "Text View".

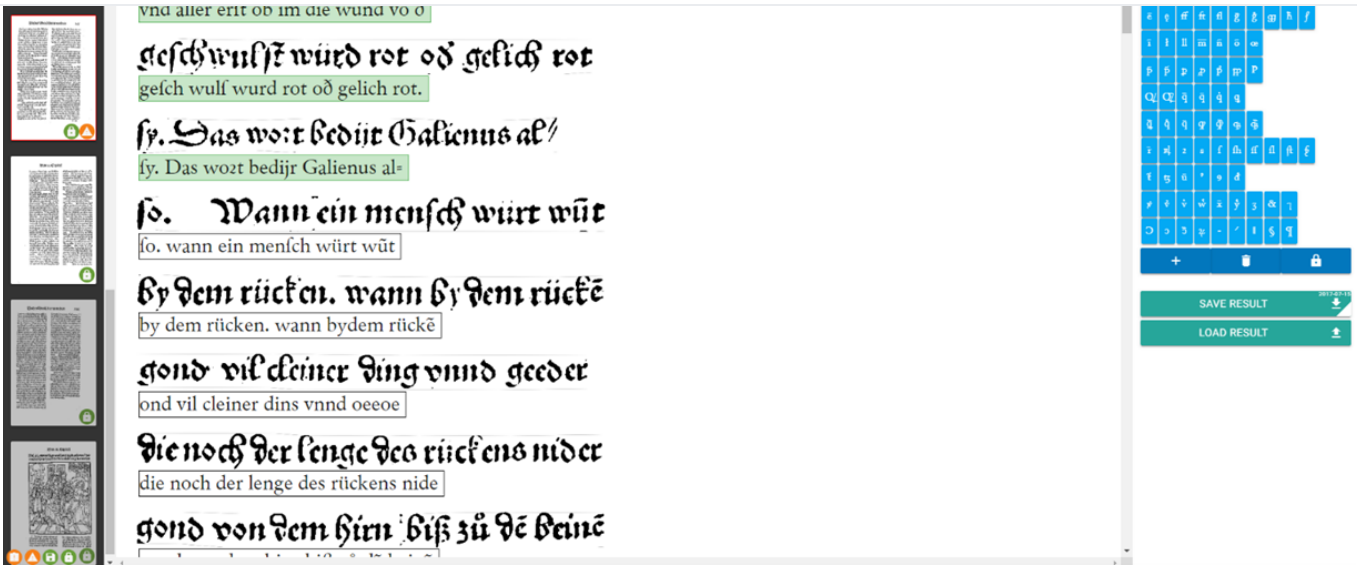

Fig. 36: Ground Truth Production with "Text View".

- You can use the "Switch to Page View" option in the toolbar to switch from the "Text View" to the "Page View". In this view, the individual lines of text can be edited in the overall visual context of the page layout. Use the "Switch to Text View" option to switch back to the "Text View".

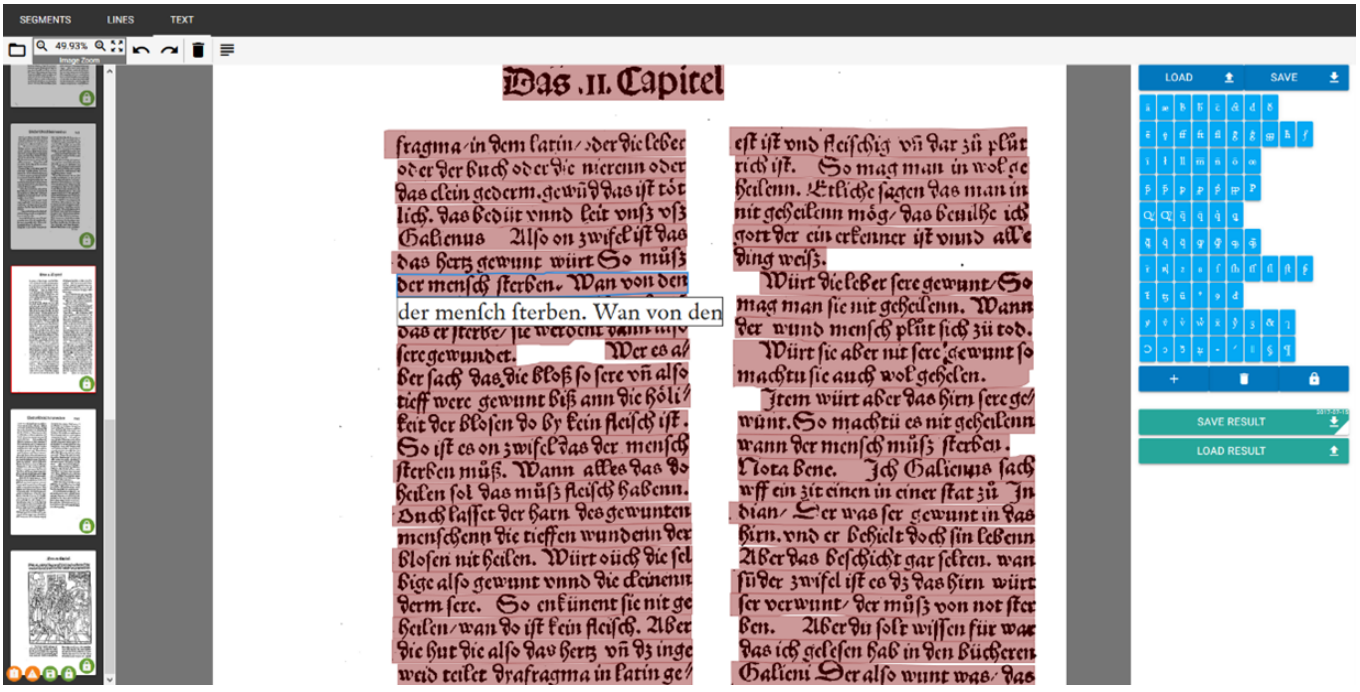

Fig. 37: Ground Truth Production with "Page View".

the left by simply clicking on them according to the position of the cursor. To add characters to the keyboard, simply click on the plus icon and copy and paste the corresponding character into the form that opens and confirm by pressing the "Save" button. If you want to delete characters from the keyboard, just drag them with the mouse onto the trash can icon of the delete option. Once all desired changes have been made, the keyboard is saved by clicking on "Save" and then locked with "Lock".

- Ready-made virtual keyboards can be selected using the 'Preset' button.
- In order to correct individual lines in the case of incorrect recognition within the "Text View", click in the corresponding line. The line that is then vertically centered can now be edited. If you are within the "Page View", the associated line text can be displayed by left-clicking on the corresponding line. Changes to the line text can also be made in the text field that is now open. To select the next line, press the "Tabulator" key. The further work steps are equivalent within both displays. If you have carried out all interventions and there is a correspondingly error-free line, press the "Enter" key. The line just edited turns green, ie: After saving the edited page via the "SAVE RESULT" button (shortcut: Ctrl + S) within OCR4all, this line is now automatically saved as Ground Truth. It can now be used with all other corrected lines as a training basis for plant-specific models and for evaluating the OCR models used, or is automatically output when you generate your final results (see below).
- If you come across erroneous line images during your correction work (e.g. halved lines, lines separated at the wrong place or even double line images), leave the corresponding text line empty and do not store any ground truth in it, since this text information in combination with incorrect line images causes problems during training could lead.
- If, during the correction of a work using Ground Truth Production, the user determines that the degree of recognition by mixed models is not yet sufficient due to various factors to carry out a manual, final text correction without too much time expenditure, OCR4all offers the option of Training of factory-specific models. Plant-specifically, these generally have higher recognition rates than mixed models.

4.8 Evaluation

Input : line-based OCR texts and corresponding ground truth

Output : error statistics

- In order to generate these, all those scans are selected in the right sidebar that were recognized using this current model and then corrected in "Ground Truth Production". If the user clicks on "EXECUTE" and leaves all the settings unchanged, a table is displayed in the console: At the top of the output are the error rate and the total number of errors ("errs") as a percentage. The errors found are displayed below – listed in a table by comparing the output text of the recognition and the ground truth created during the correction. The corrected text ("GT") can be seen in the first column, the text originally recognized by the model ("PRED") in the second column,

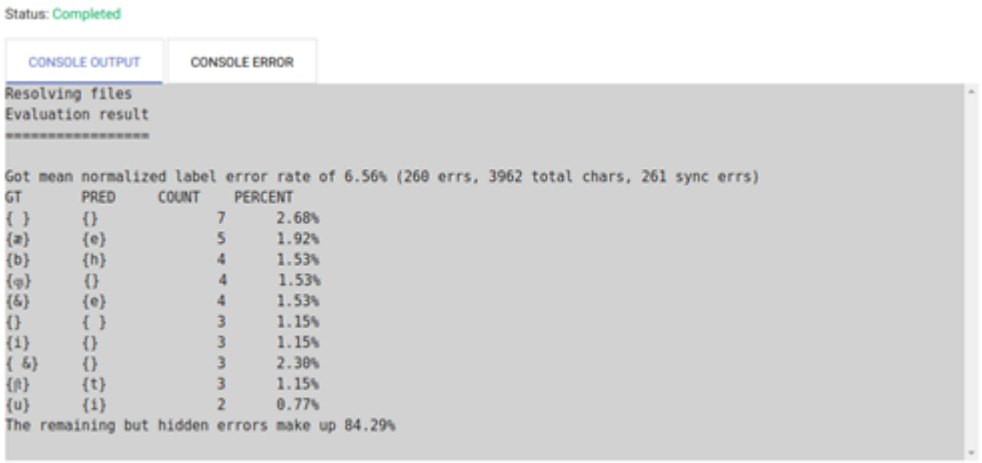

Fig. 38: Evaluation result with total error rate and the ten most common errors and their percentage of the total number of errors.

- Using this tabular listing and the recognition rate (100% - error rate), the user can now assess the usefulness of (renewed) training of plant-specific models.

4.9 Training

Input : line images with the corresponding ground truth and optionally already existing OCR models, which are used as so-called pre-training and data basis for model training.
Output : one or more OCR models

In general, the goal must be to get a text that is as error-free as possible!
But why then the creation of work-specific models using the training module instead of simple, final text correction?

correction progresses and thus reduce the correction effort for the pages of the work that still have to be corrected to a minimum.

- Within the training tool, work-specific models or ensembles can be trained on the basis of all lines of Ground Truth available for a work. For this purpose, the following values are entered in the general settings:
  - " **The number of folds** (= the number of models) **to train** ": **5** → In the following, a model ensemble consisting of five individual models is trained.
  - " **Only train a single fold** (= a single model)": *Do not enter anything!* → All five individual models are trained instead of just one.
  - " **Number of models to train in parallel** ": **-1** → All models of the ensemble are trained simultaneously.
  - " **Keep all characters loaded from the last model** ": Select if all characters contained in the "Pretraining" models should be retained in the model to be trained, i.e. added to its whitelist.
  - " **Whitelist characters to keep in the model** ": List of characters that are considered during training and in the resulting model. All characters outside of this "whitelist" are ignored.
  - "Pretraining":

    " **Train each model based on different existing models** " (Five drop-down lists open below; each one is populated with one of the mixed models of the model ensemble that was used as recommended for the first recognition of text in the present work; doesn't matter at which training iteration the user is: Even if, for example, the third plant-specific model is already being trained - the five basic mixed models used at the beginning are still always entered)

    OR

    " **Train all models based on one existing model**" (If the first text recognition was carried out on the basis of a single mixed model, then only one model is entered; however, it also applies here that this mixed model must be specified again for each iteration).
  - " **Data augmentation** ": *Do not enter anything.* → But: describes the number of data extensions per line. A value, e.g. 5, can be specified here in order to increase the amount of training material that is trained on. This can lead to the creation of better models, but requires significantly more training time.
  - " **Skip retraining on real data only** ": *Do not select!*
- The advanced settings remain unchanged.

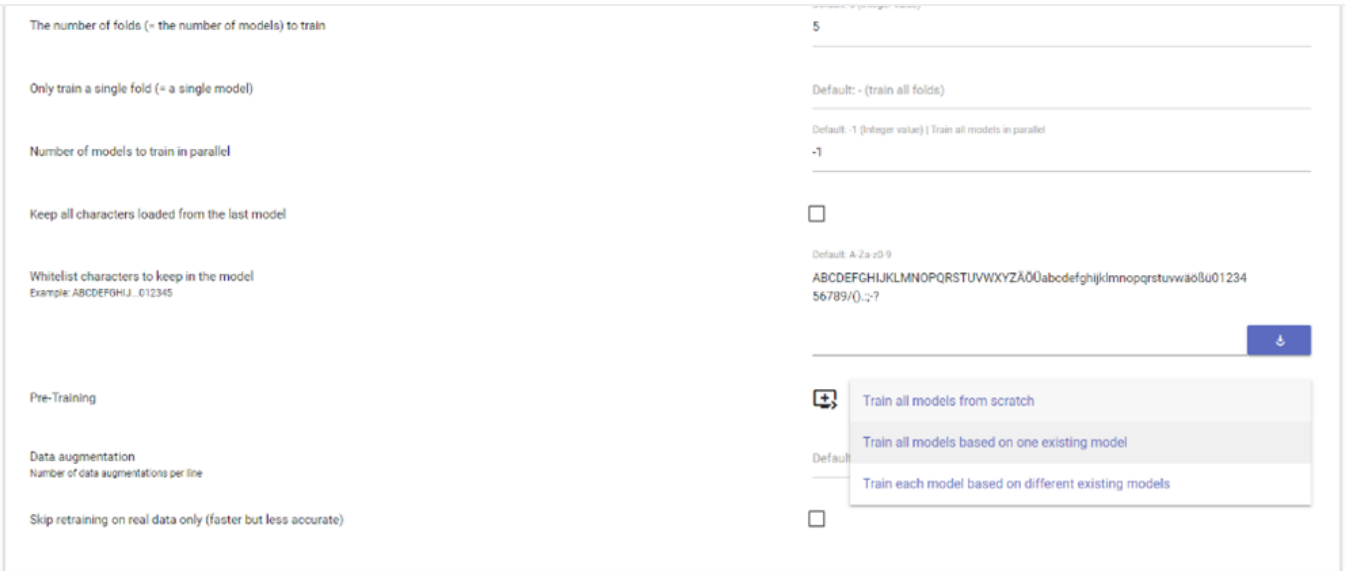

Fig. 39: Settings for training plant-specific model ensembles.

- The training is started with "EXECUTE". In the following, the training of the console can be understood. The training times vary depending on the total number of lines of Ground Truth available.
- According to the settings above, the training creates a work-specific model ensemble consisting of five individual models, which is saved in ocr4all/models/work title/0. The model ensemble therefore bears the name "0". It can now be used to recognize new text pages for further work on the present work and to improve the recognition within the menu item "Recognition" and the column of selectable models. If a second plant-specific model ensemble is to be created, with the help of which, for example, possible weaknesses of the first can be eliminated, the procedure described here is repeated. The designation "1" is then automatically assigned to the new factory-specific model. The designations of other model ensembles continue according to this scheme.

4.10 Post Correction

Input : Segmentation information for preprocessed images and associated text
Output : Corrected segmentation information and text

Under the "Post Correction" menu item, the segmentation information and texts created in the previous sub-modules can be manually adjusted and corrected. The sub-module is divided into three levels:

available for this. Note that changes at this level also affect the following levels. For example, removing a region and saving that change will result in the loss of associated lines and text.

- The "LINES" tab allows manual adjustment of the automatic line detection. Similar to the previous marking of the regions, individual lines can be added, their shape and position can be changed or they can be removed. The reading order can also be adjusted manually at row level. As with LAREX, these actions are performed using various tools from the toolbar and sidebar.

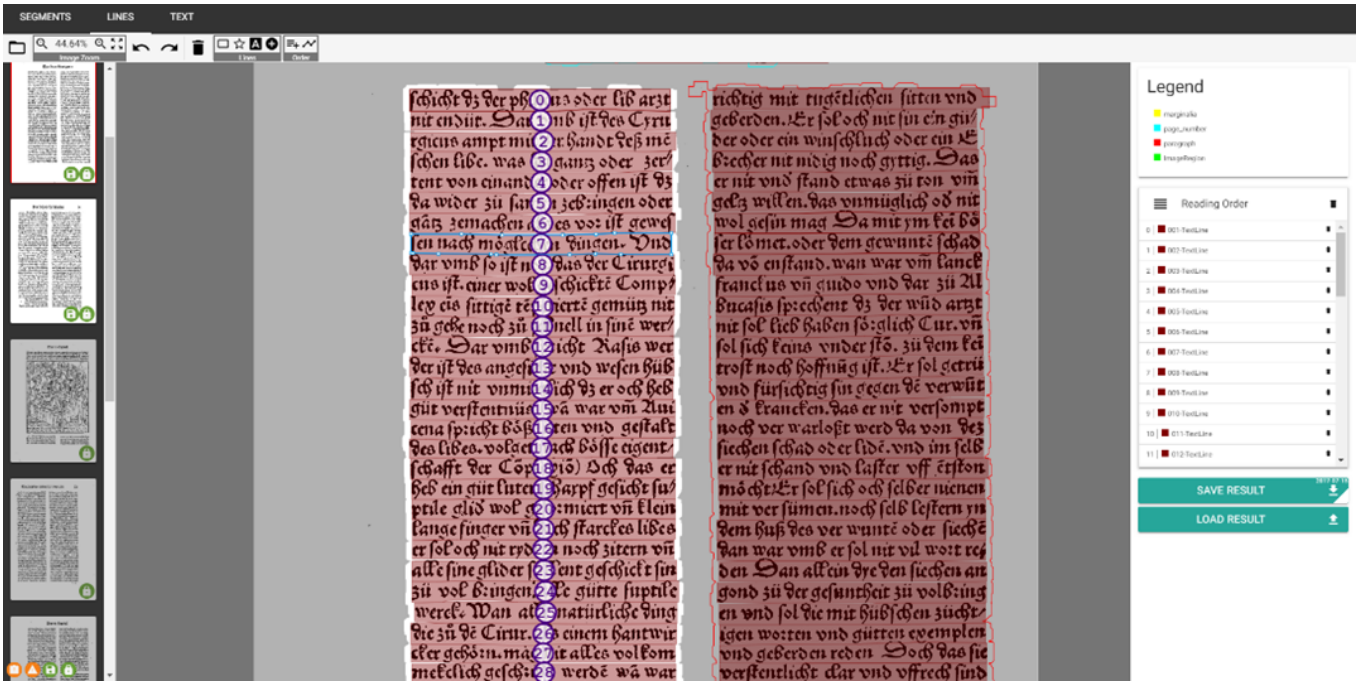

Fig. 40: Adjustment of the line-based Reading Order in the "Post Correction".

- Under "TEXT" you can find the "Ground Truth Production" sub-module (see above), which can be used to correct the texts assigned to the lines.

4.11 Result generation

Input : OCR results on a line basis, optional ground truth (if available) and additional data from segmentation (LAREX) and line segmentation
Output : final output as text (individual lines of text combined into pages and full text) and PageXML on a page basis

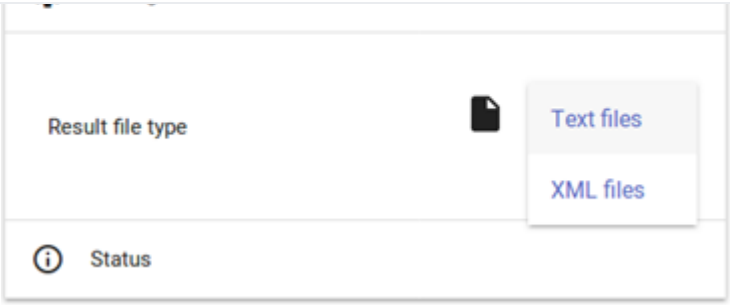

Fig. 41: Result generation.

- If the recognition and/or correction work on a work is completed from the user's point of view, results can be generated in the form of TXT and XML files. They are saved under ocr4all/data/results.
- Under "Settings" you can select whether text or PageXML files should be created. In the case of text files, a single TXT is created for each scanned page, as well as a coherent one containing the entire text of the edited work.
- The PageXML files are output on a scan page basis and contain information on the creation date, the last file changes, metadata of the scan page relating to them, the page size, the layout elements contained on the page including their exact coordinates, the reading order of the existing layout elements, the individual lines of text and the text of the lines themselves.

←3.2 Project Overview

5. Mistakes, common problems and how to avoid them→