DATeCH 2017 – PoCoTo Workshop – PoCoTo

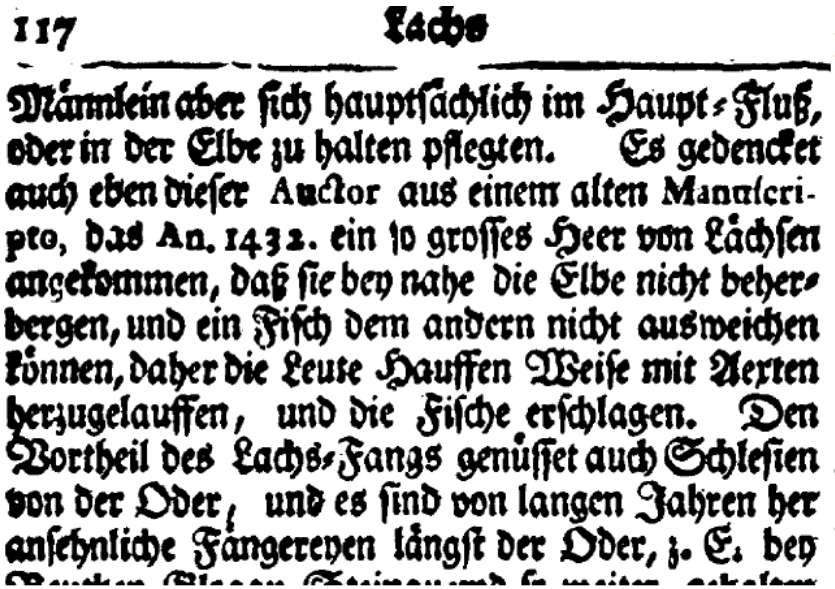
Florian Fink & Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München (LMU)

May 30, 2017

In the resent years a lot of historical documents have been scanned and OCR'ed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents has constantly improved.
- In some cases the quality can be further improved, by further adapting the original images and OCR engines.
- But still the quality of the recognition is not good enough for deeper scientific studies on the documents.


117 Lachs

Männlein aber sich hauptsächlich im Haupt-Fluß, oder in der Elbe zu halten pflegten. Es gedencket auch eben dieser Auctor aus einem alten Manuscripto, das An. 1432. ein so grosses Heer von Lächsen angekommen, daß sie bey nahe die Elbe nicht beherbergen, und ein Fisch dem andern nicht ausweichen können, daher die Leute Hauffen Weise mit Aexten herzugelauffen, und die Fische erschlagen. Den Vortheil des Lachs-Fangs genüsset auch Schlesien von der Oder, und es sind von langen Jahren her ansehnliche Fängereyen längst der Oder, z. E. bey

xi7_ Kchs
Männleinccher sich hauptsächlich im Haupt-Fluss, öderm der Gbe zu halten pflegten. Es gedencktt auch eben dieser Auctor aus einem alten Mannferi-pto, das An. 1431. ein 1o grosses Heer von Lachsen angekommen, daß sie bey nahe die Elbe nicht beher-bergen, und ein Fisih dem andern nicht auSweichm können, daher die Leute Haussen Weise mit Aexem bcr;ugelauffen, und die Zische erschlagen. Den Vortheil des LachS-Fangs gmüsset auch Schlesim von der Obtti und es sind von langen Jahren her ansehnliche Fangereyen längst der Oder, 5. & bey

Example of the OCR results of a snippet of the *BSB Zedlersches Universallexikon: article about salmon.*

Year	Language	ABBYY FR 11.1	Tesseract 3.03	OCRopus 0.7
1544	lat.	83,14	70,32	74,59
1649	lat.	88,07	84,87	78,98
1746	dt.	97,00	91,48	95,70
1779	lat.	82,13	80,77	75,46
1871	dt.	98,12	95,94	97,40

The results of the text recognition must be manually improved:

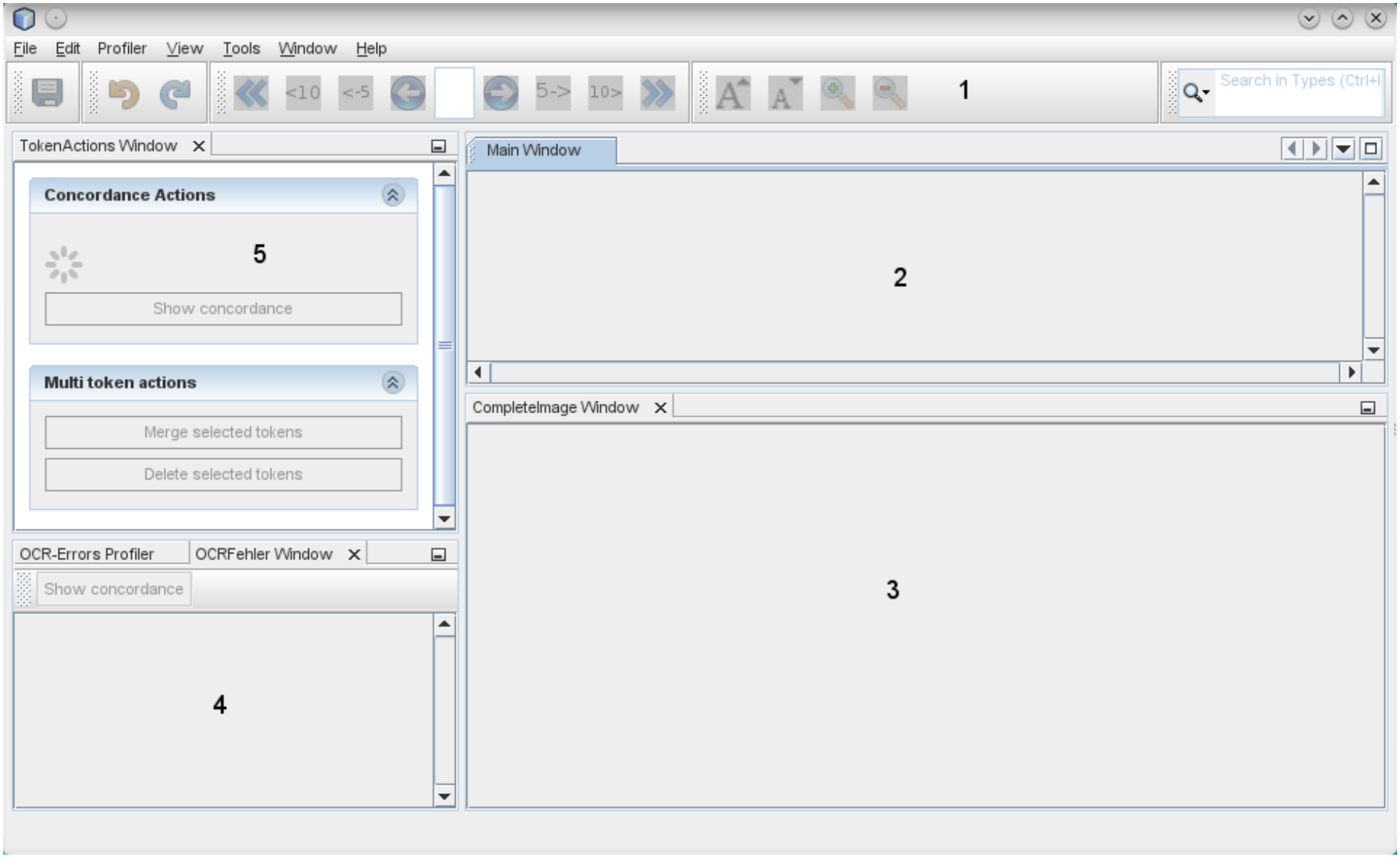
- Manual (double) keying of the original sources is expensive.
- Interactive postcorrection can be used examine the results of the OCR.
- Interactive postcorrection can be used to improve the results of the OCR.

- PoCoTo is a tool for the interactive post-correction of OCR'ed text:
- It was developed as part of the EU founded project IMPACT.
- It is open source and hosted on github.
- It contains linguistic and visual aids to support the post-correction.
- It contains aids to automatically correct systematic errors in the documents.
- You find its documentation in the PoCoTo manual (included in this workshop's data package).

- PoCoTo has an automatic update mechanism – once installed, it is automatically kept up to date.
- The recognition results are visualized with the images of the original documents.
- The concordance views enable to examine different errors and error pattern over the whole document.
- A specialized profiling web-service can be used to get correction suggestions for unknown words and frequent error patterns in the document.
- Different formats can be read, manually corrected and written back.

- You can download the application data file `ocrcorrection.zip` from this link or use the version that is part of this workshop's data package.
- Extract the archive to a convenient place
- Go to `ocrcorrection/bin` in the extracted directory and double click on the executable file `ocrcorrection` (Linux) or `ocrcorrection.exe` (Windows).
- You can create a link to this executable on your desktop for easier access.

- PoCoTo has an automatic updating mechanism.
- PoCoTo can be kept up to date without having to install it again.
- Whenever PoCoTorecognizes a newer version, it shows an *updates available* button in its lower right corner.
- To check for updates go to `Help->Check for updates`.
- To control the update go to `Tools->Plugins`.


File
Edit
Profiler
View
Tools
Window
Help
<10
<-5
5->
10>
1
Search in Types (Ctrl+I
TokenActions Window
Concordance Actions
5
Show concordance
Multi token actions
Merge selected tokens
Delete selected tokens
Main Window
2
CompleteImage Window
3
OCR-Errors Profiler
OCRFehler Window
Show concordance
4

PoCoTo is composed by 5 main areas. The size of each area can be freely adjusted:

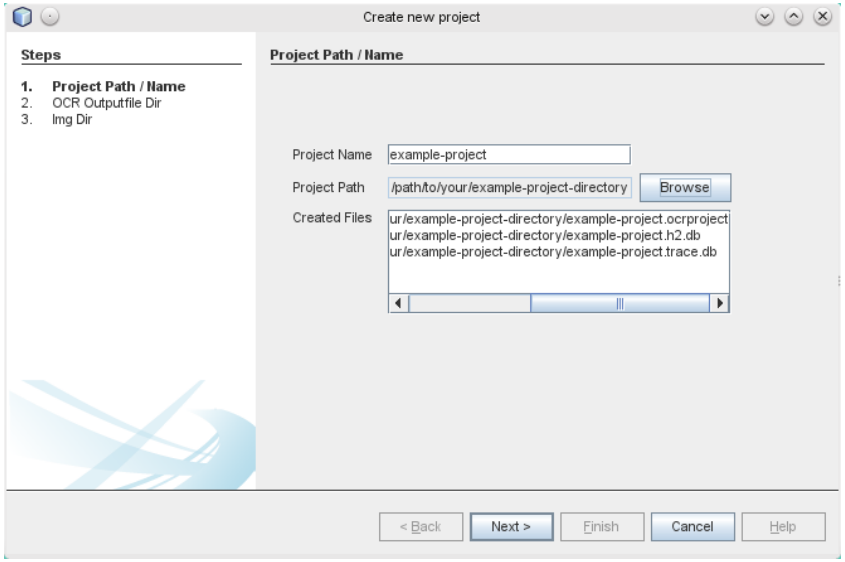
1. The menu area contains various commands for navigation and project maintenance.
2. The main view area shows tokens and offers the main correction possibilities.
3. The complete image area displays the page of the current active (selected) token.
4. The error area lists error frequency lists of common word or pattern errors.
5. The token actions area lets you create concordance views an helps you to split and merge tokens.

- PoCoTo handles your input documents as separate projects
- Each project is constructed over a set of different files:
  - The XML output files of your OCR engine.
  - The image input files of your documents – the same that you used for your OCR.
- PoCoTo expects those files to be organized in a specific way:
  - All the XML files for your project should be in one folder
  - All the image files for your project should be in another folder.
  - Each image file should have the same name as its corresponding XML file, except for the file's extension (`.xml`, `.png`, ...).
- It is more convenient to have the two folders for your XML and image files together in one place and use this folder as base path for your project.

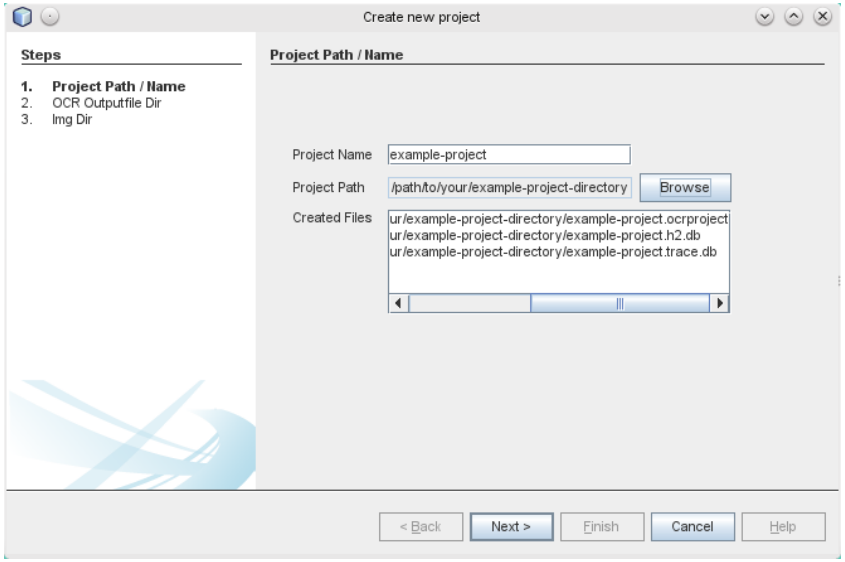
PoCoTo understands three different XML file formats, that you can use to create new projects.

1. The character based ABBYY-XML format.
2. The hOCR file format.
3. Ocropus-Directories.

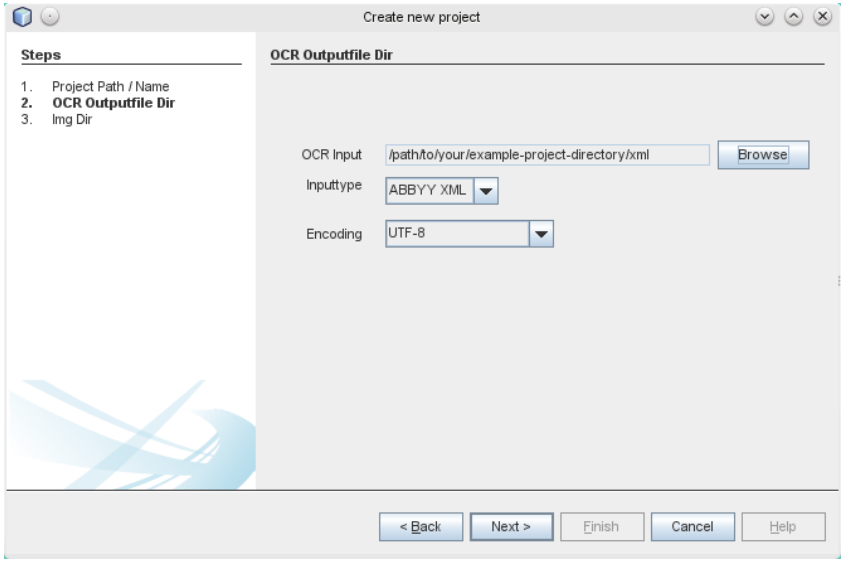
PoCoTo uses the information of the ABBYYX-XML file format directly to mark *suspicious* words. It will generate an error frequency list for you. If you use the hOCR format or Ocropus, PoCoTo is not able to generate such an error frequency list for you.



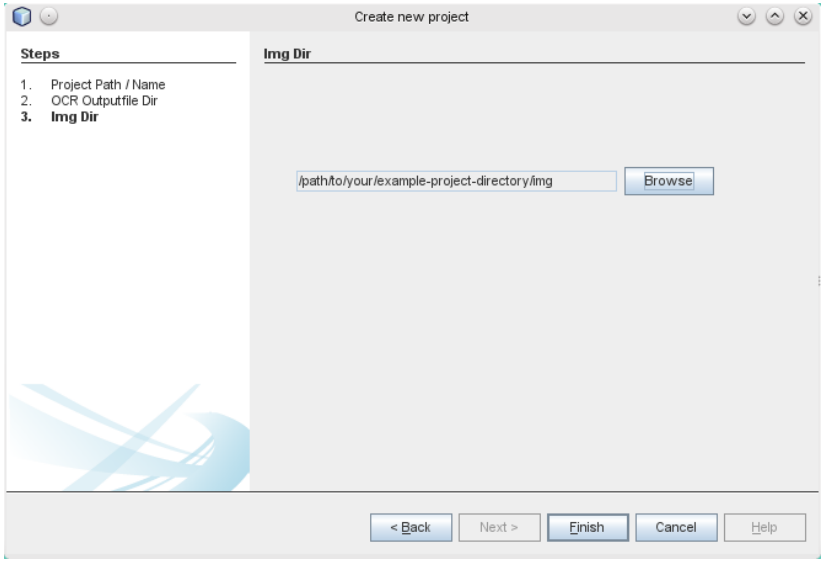
1. You can create new projects using the project wizard. Click to `File->New Project` and the first frame of the project wizard open.



1. You can create new projects using the project wizard. Click to `File->New Project` and the first frame of the project wizard open.
2. Insert a name and a path for your project. Click `Next`.

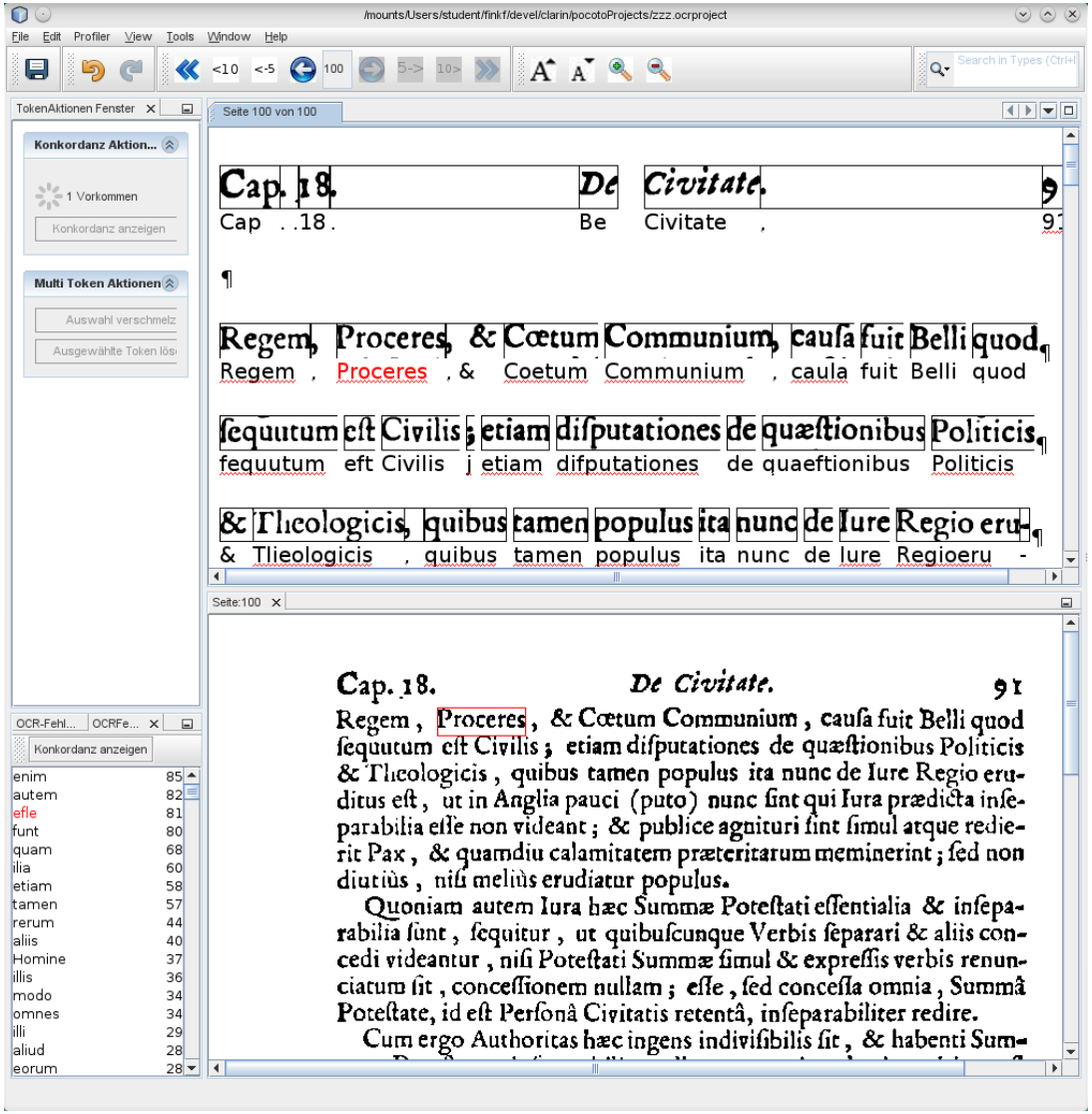


1. You can create new projects using the project wizard. Click to `File->New Project` and the first frame of the project wizard open.
2. Insert a name and a path for your project. Click `Next`.
3. Insert the path of your folder, that contains the XML files and select the type of your XML files. Click `Next`.

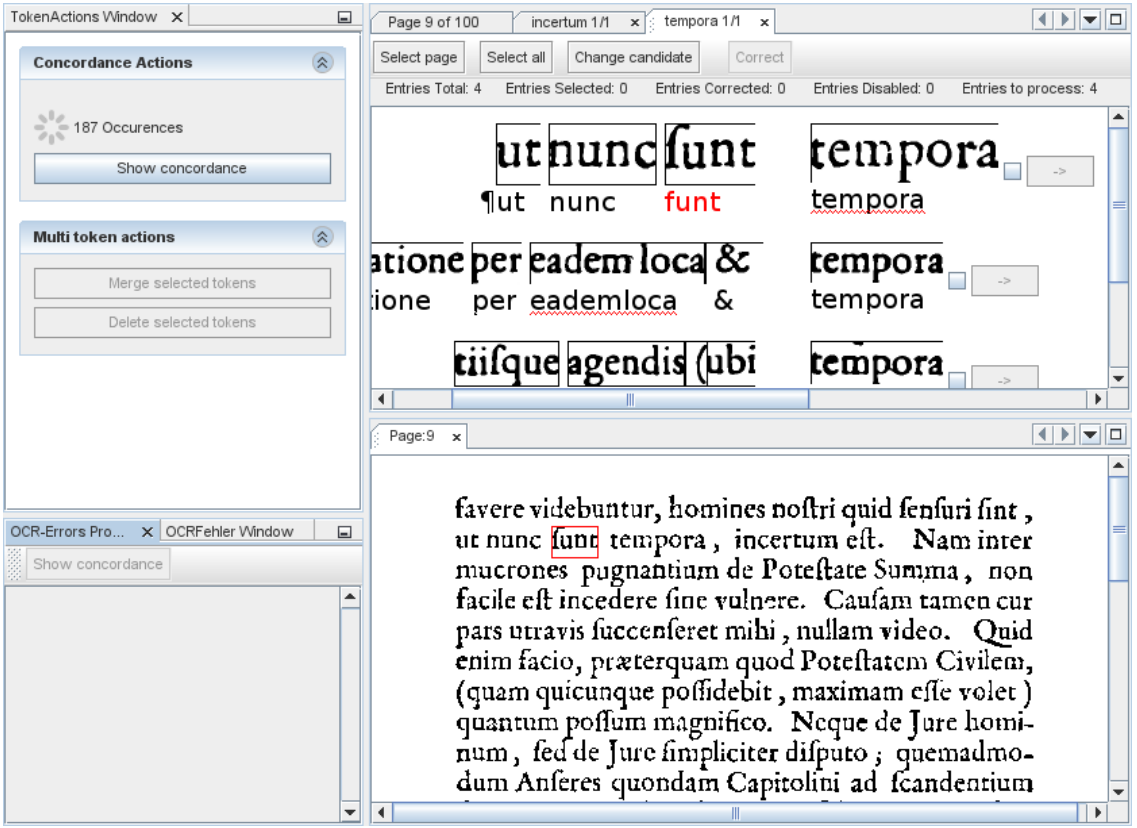


1. You can create new projects using the project wizard. Click to `File->New Project` and the first frame of the project wizard open.
2. Insert a name and a path for your project. Click `Next`.
3. Insert the path of your folder, that contains the XML files and select the type of your XML files. Click `Next`.
4. Select the path to the folder, that contains your image files. Click `Finish`.

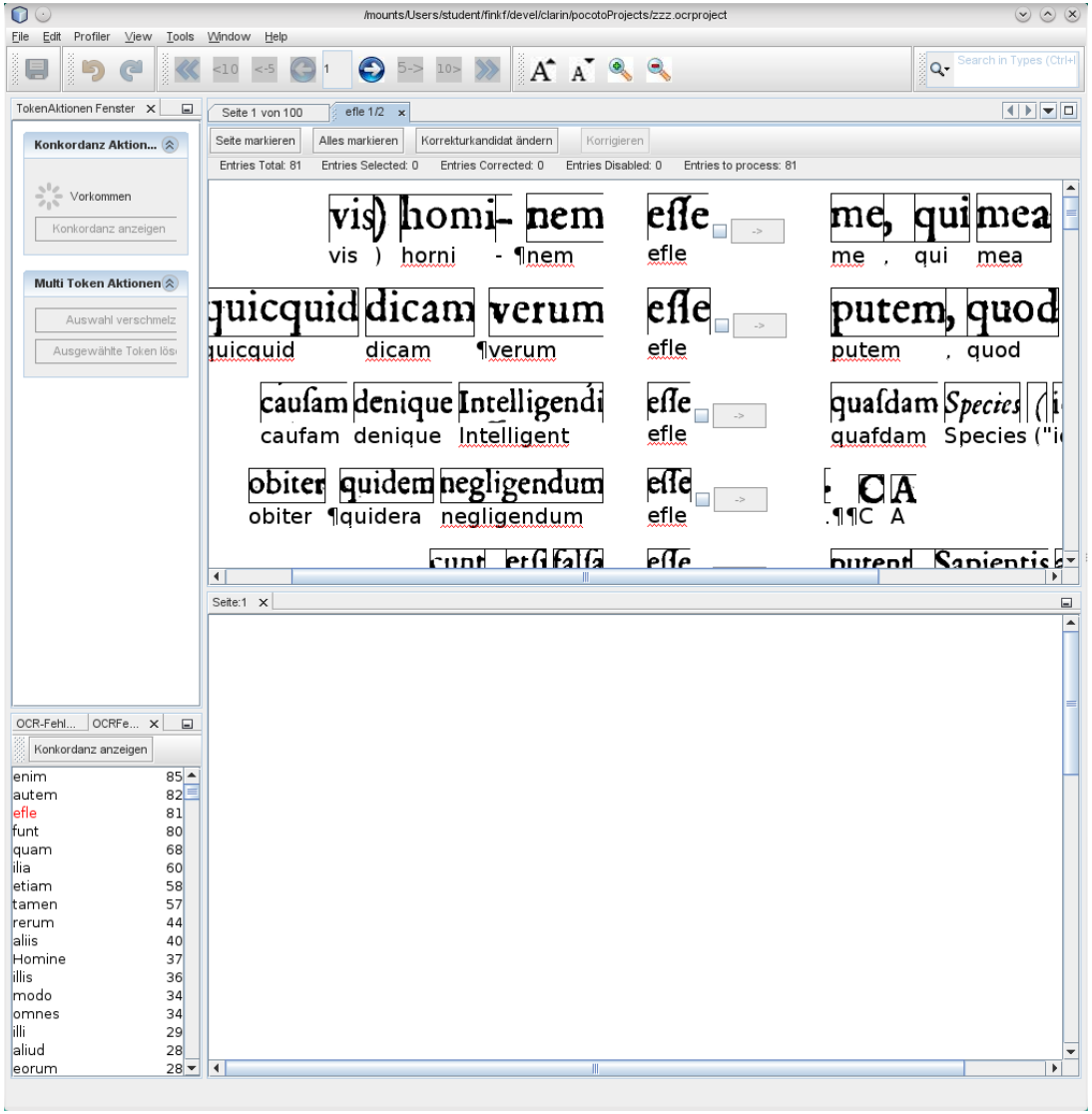
- After you have created a project, you will see the first page of your document opened.
- You can go to other pages, using the buttons in the tool bar.
- You can jump 1, 5 or 10 pages forward or backward at once or go to the first or last page of your document.
- You can navigate within a page, using your mouse wheel or the scroll bars in the areas.
- You can select or activate single token by simply clicking on them.
- You can increase or decrease the sizes of the different areas using your mouse pointer.



- The token of the text are displayed along with their image details.
- The page context shows the active token on the original page.
- Error frequencies – based on the confidence values of the OCR engine – are shown.

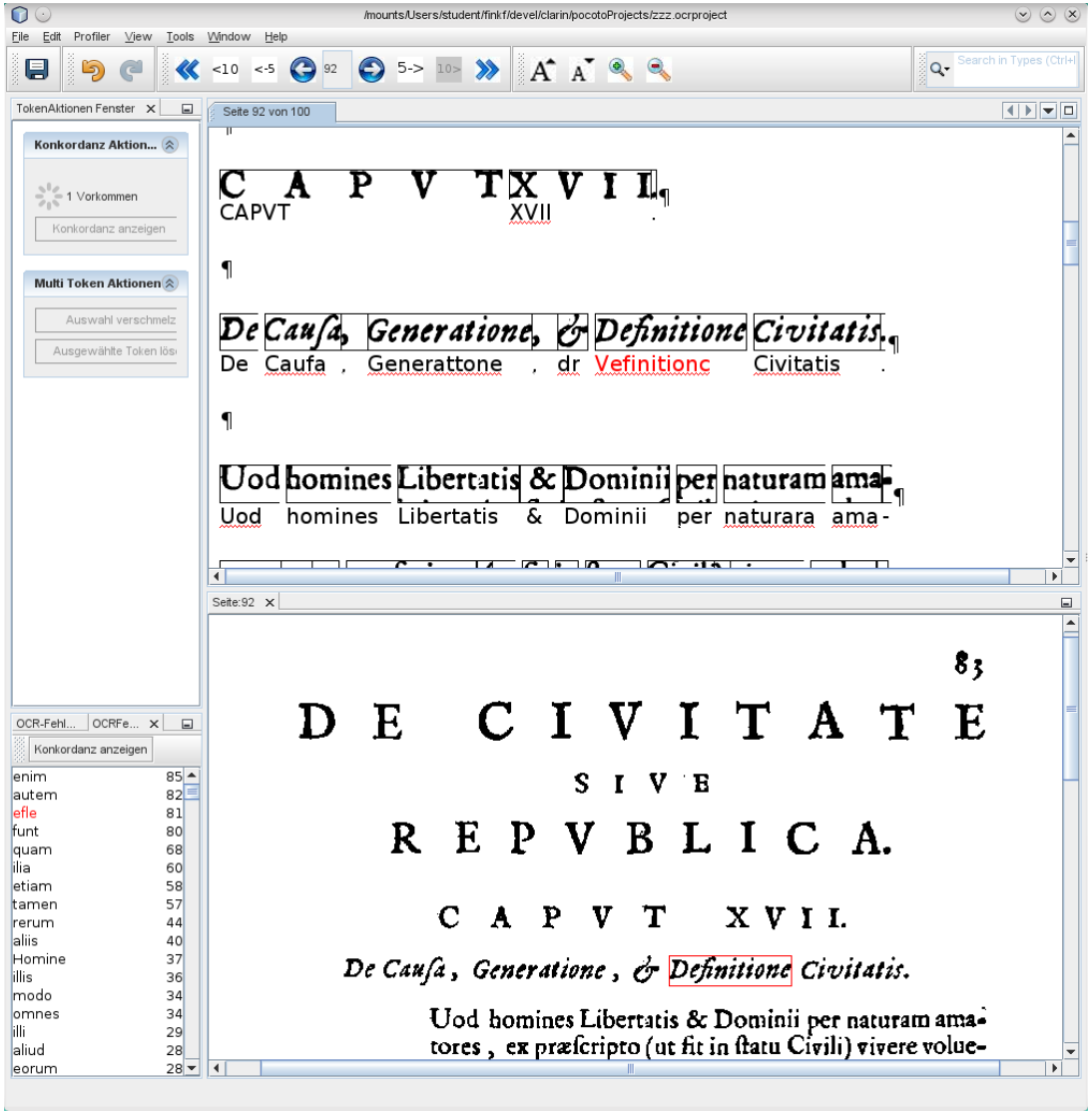


1. You can activate any token and if there exists any similar other token you can click to the `Show concordance view` button in the token action area
2. You can click on any entry in the two error frequency lists in the error area.

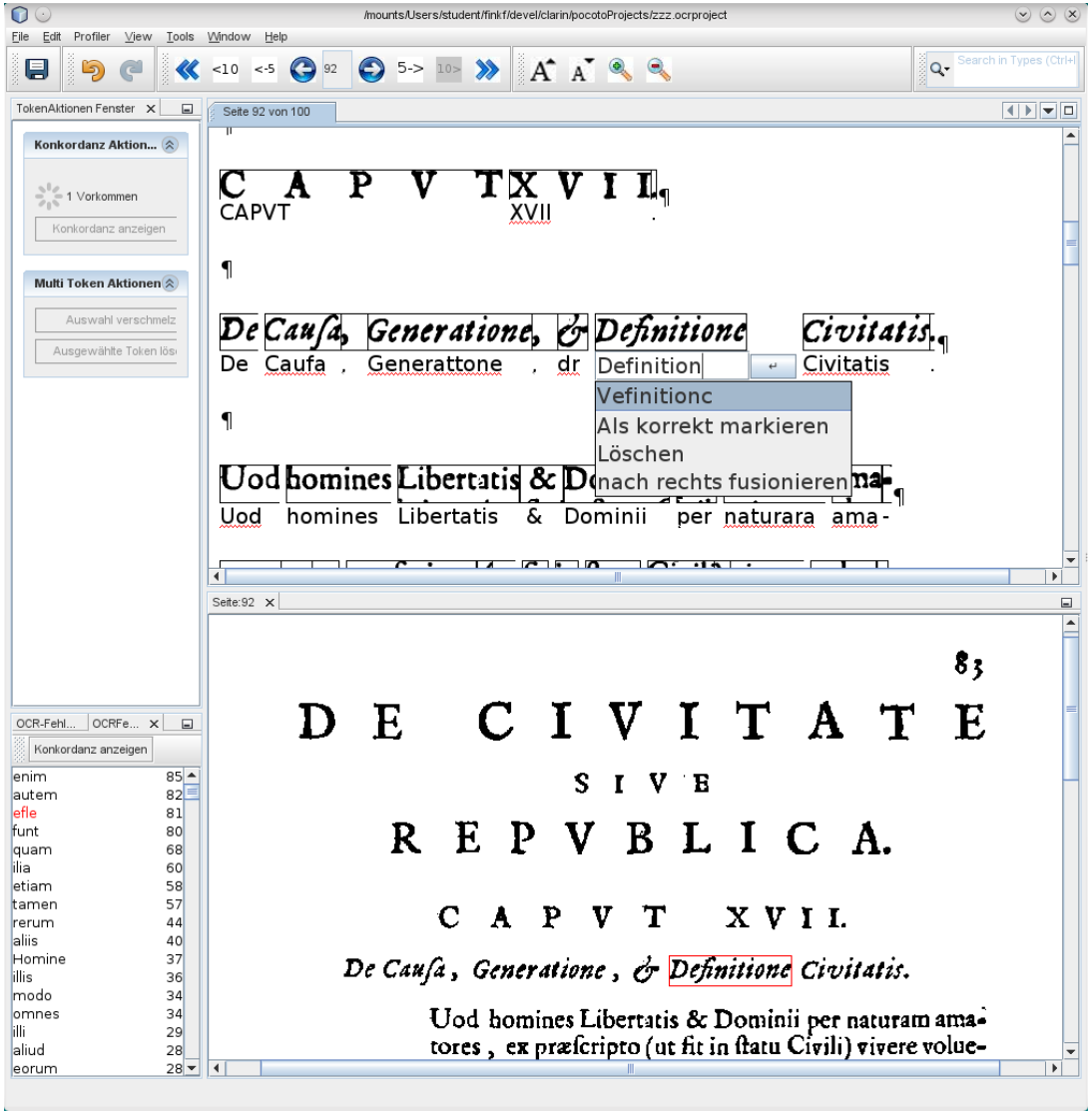


- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

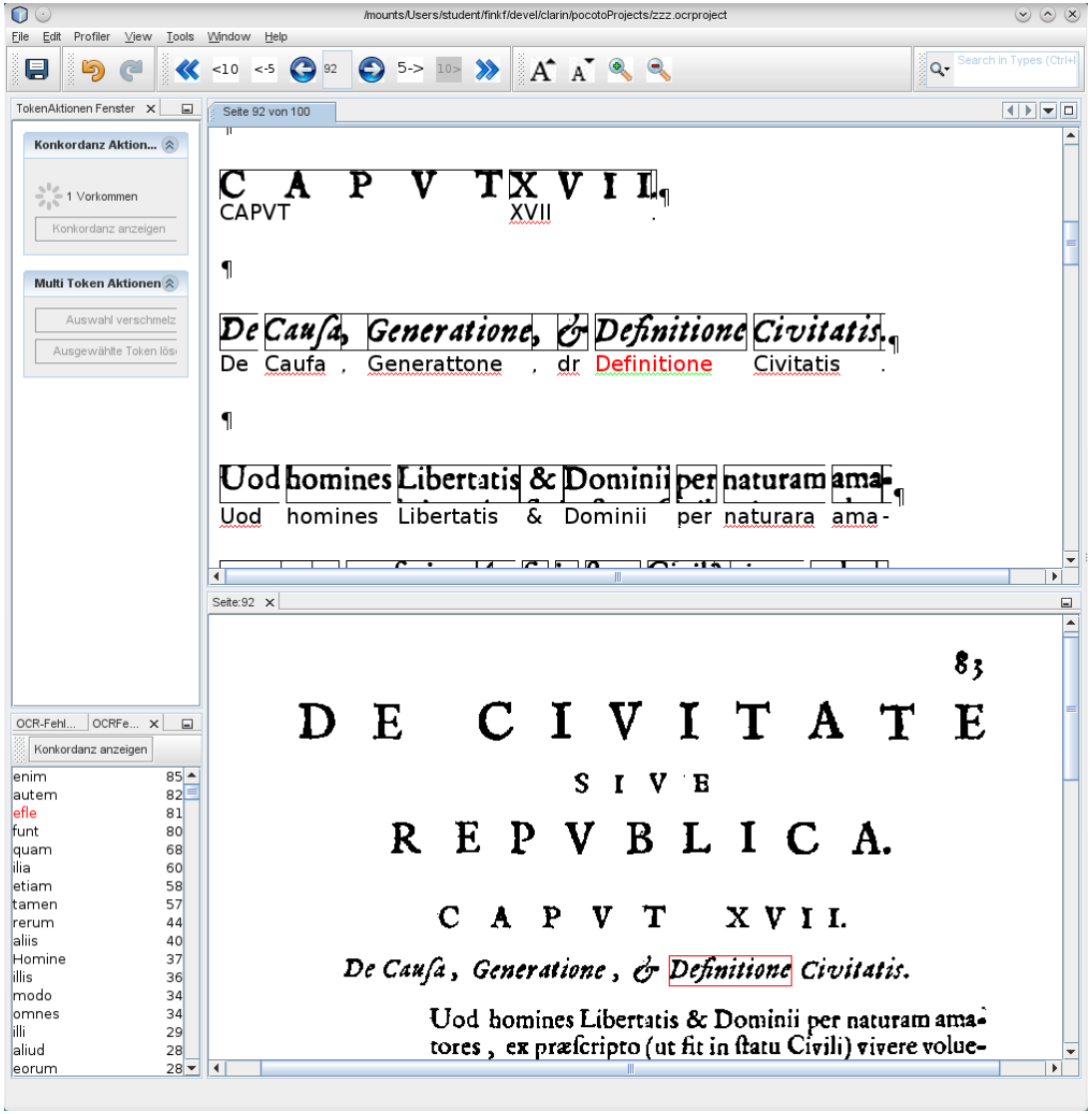
- PoCoTo automatically tokenizes the document on whitespace and punctuation.
- Each token can be examined in its page image.
- PoCoTo supports the correction of single tokens.
- Multiple occurrences of errors and (error patterns) can be corrected with concordance views.
- Split tokens (Splits) can be merged together.
- Merged tokens (Merges) can be split.



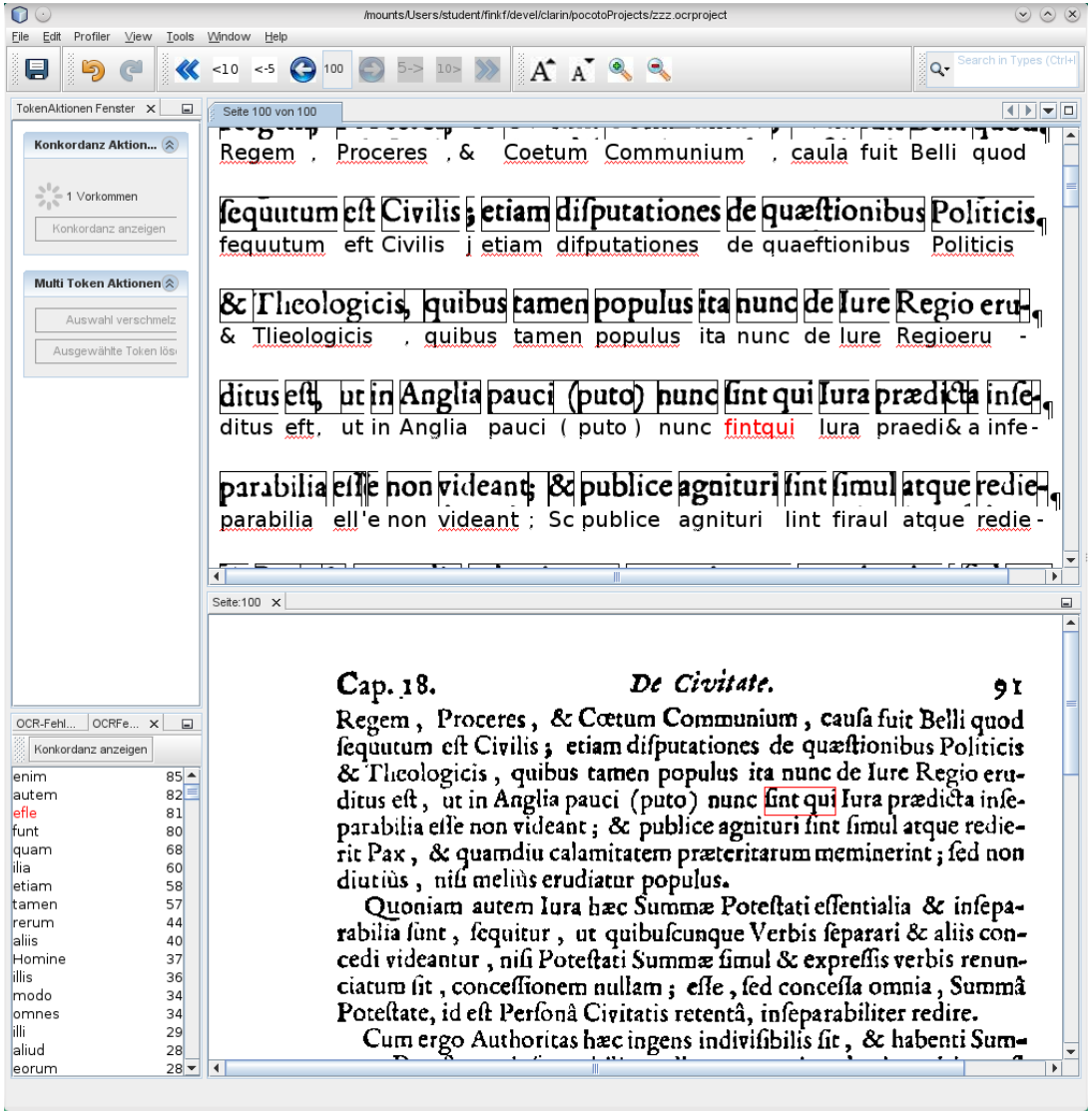
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neighbours.
- Words can be corrected manually in the window.



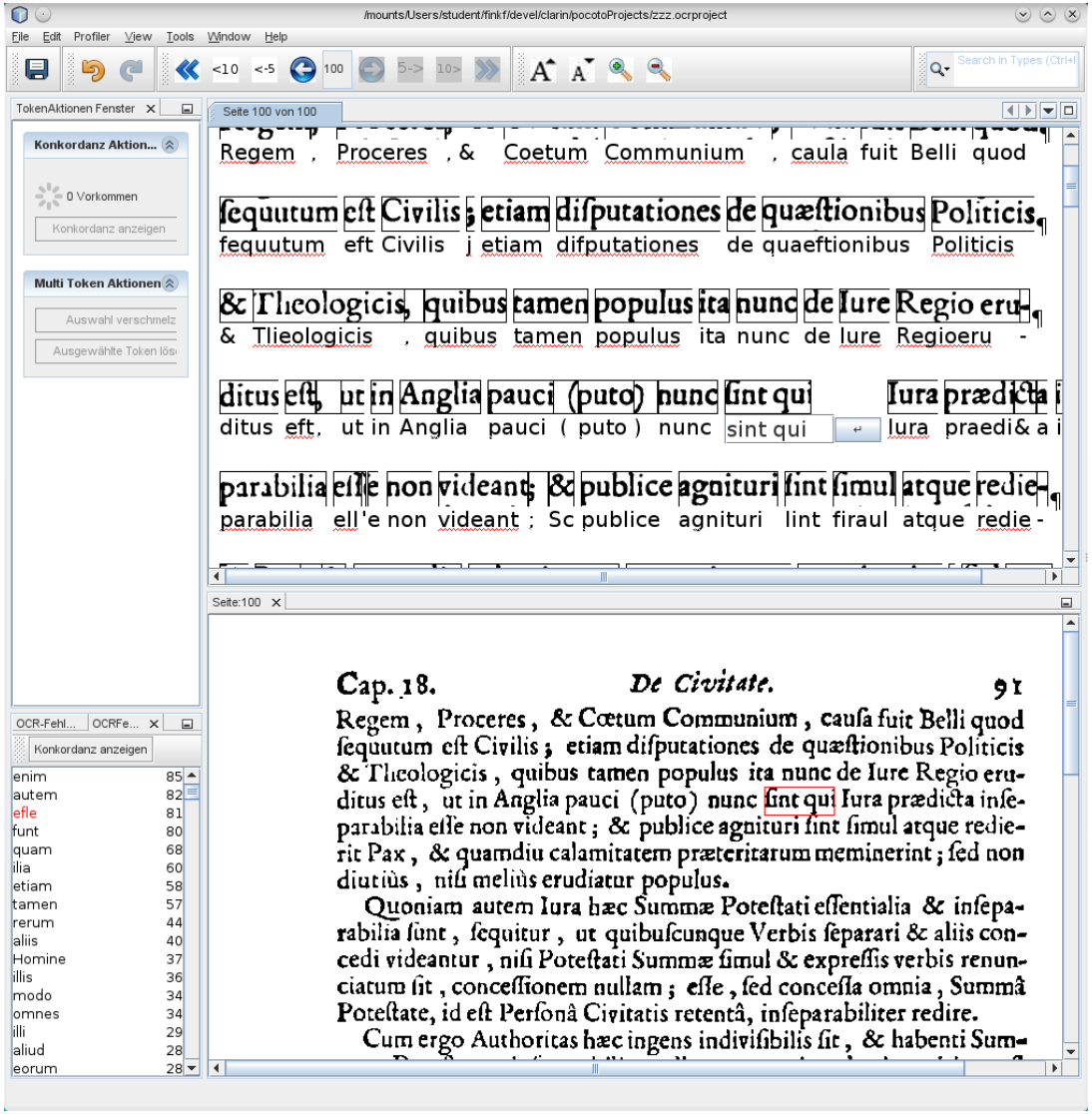
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neighbours.
- Words can be corrected manually in the window.



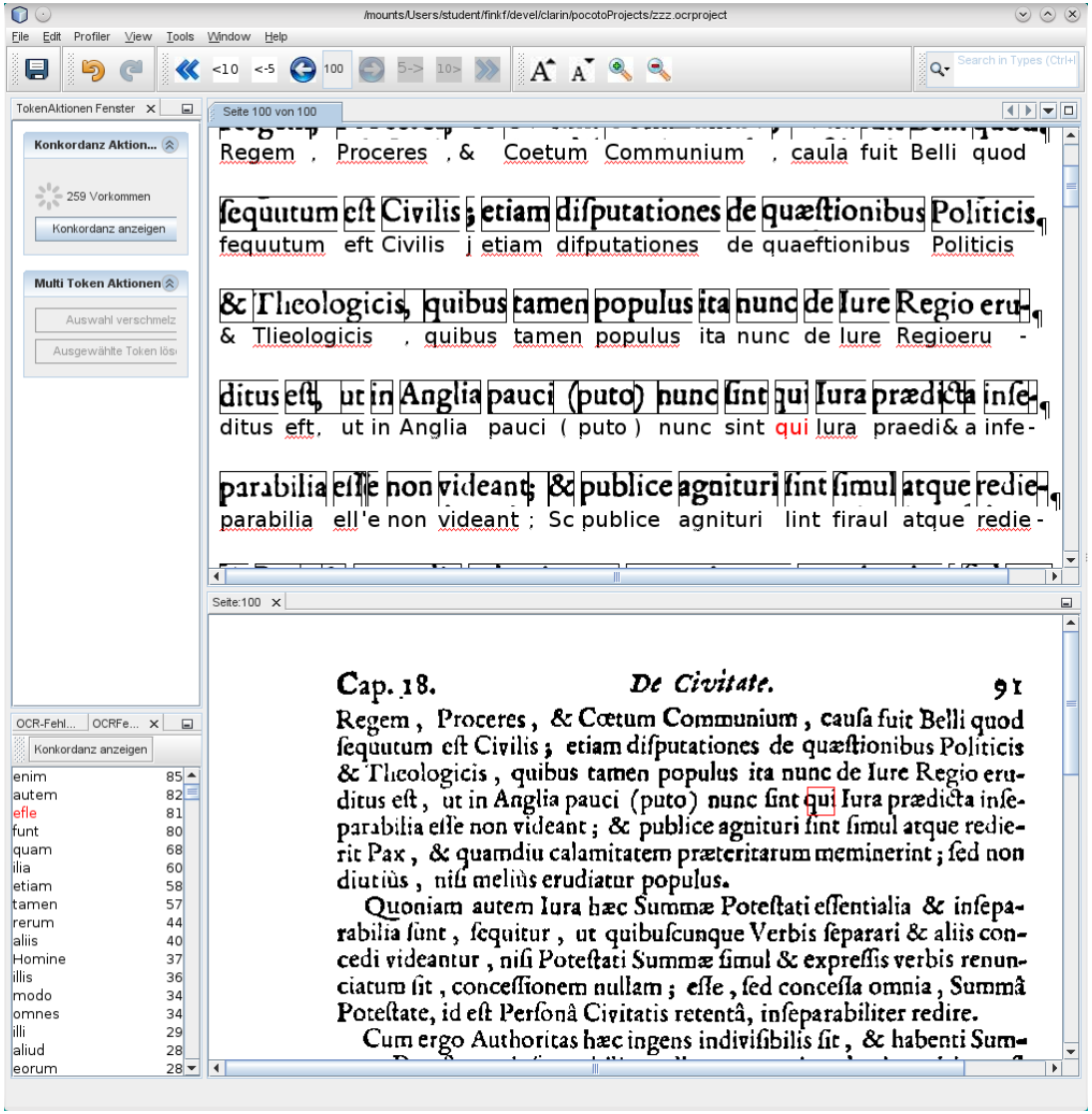
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neighbours.
- Words can be corrected manually in the window.



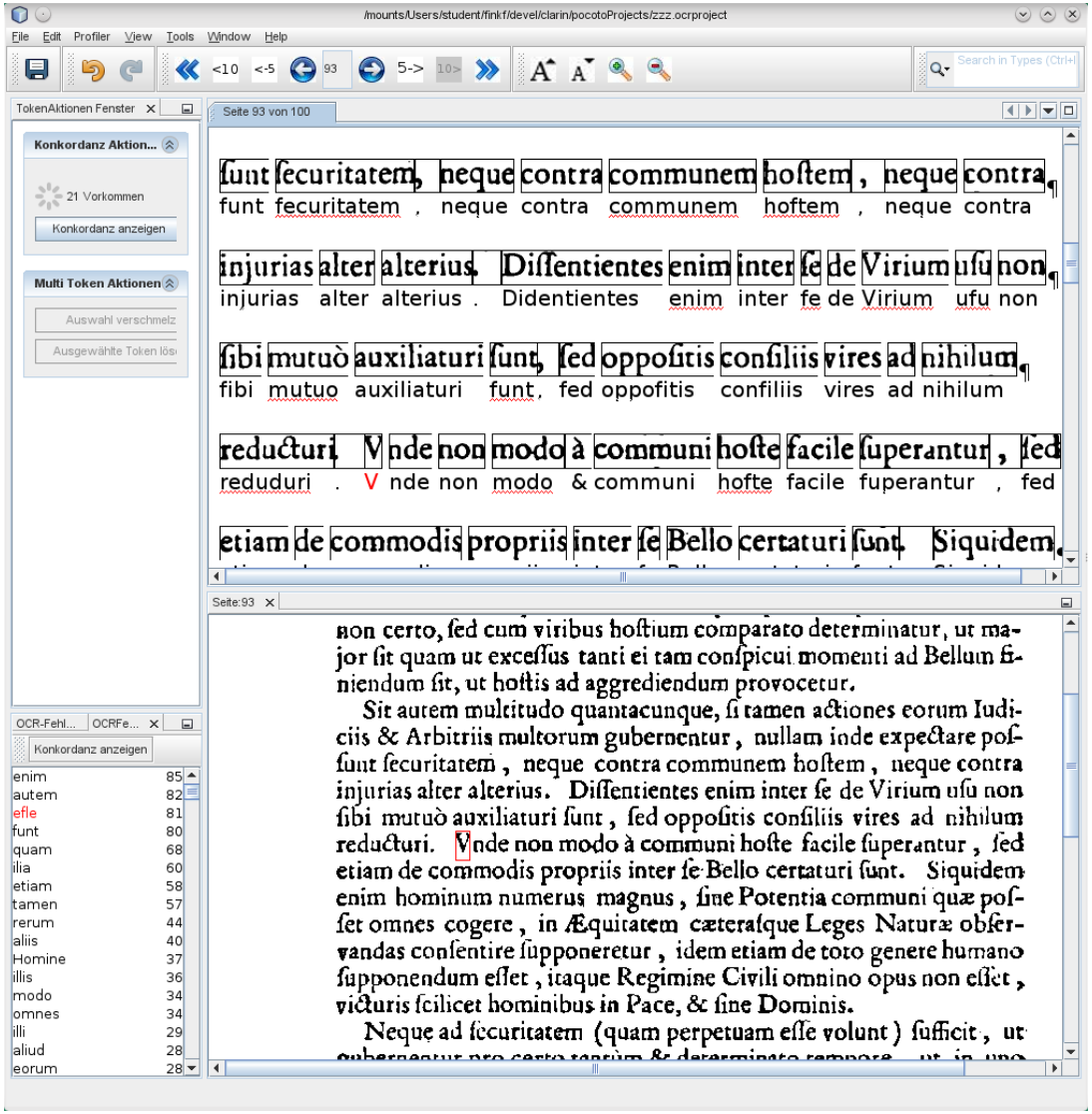
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



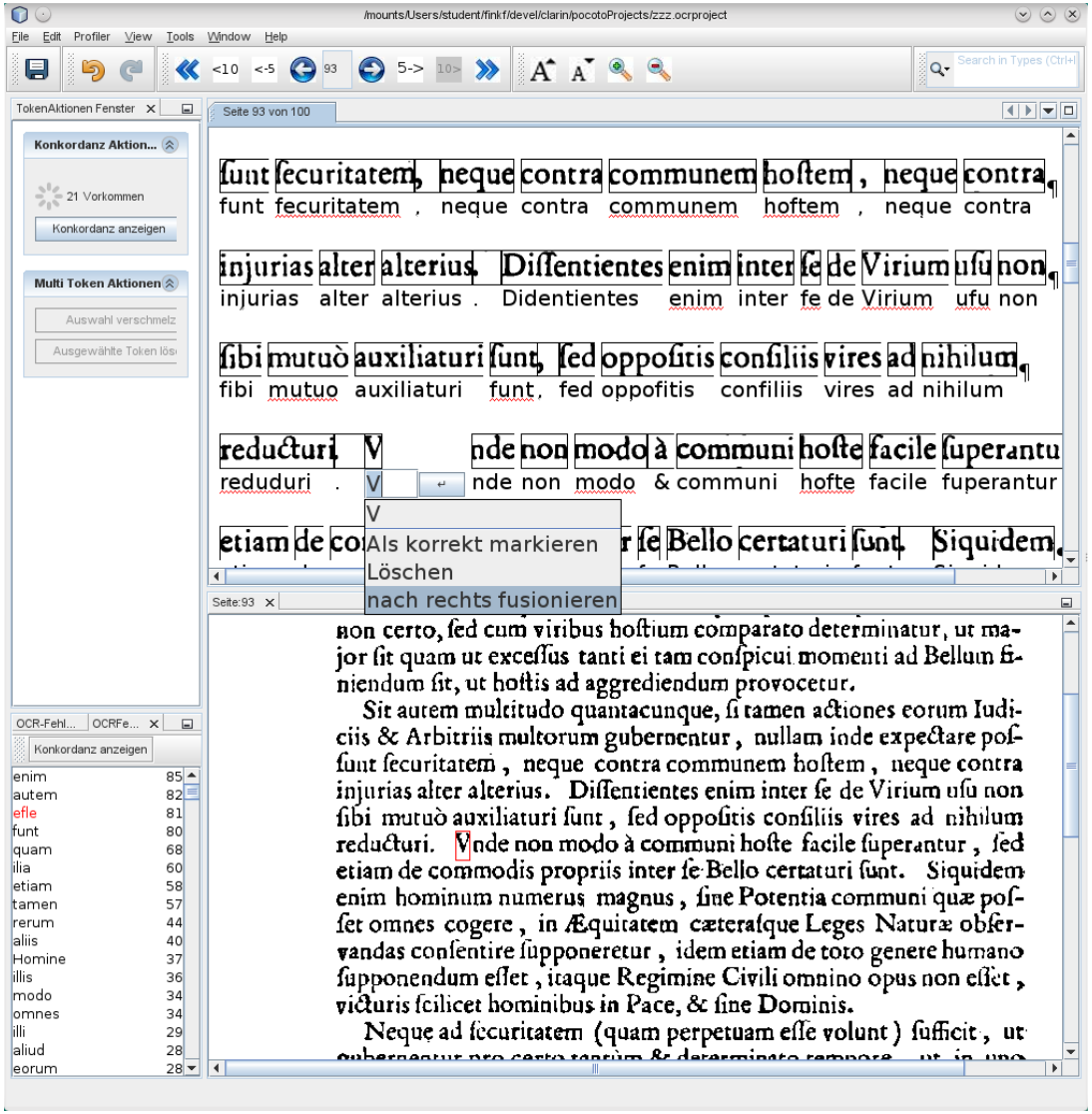
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



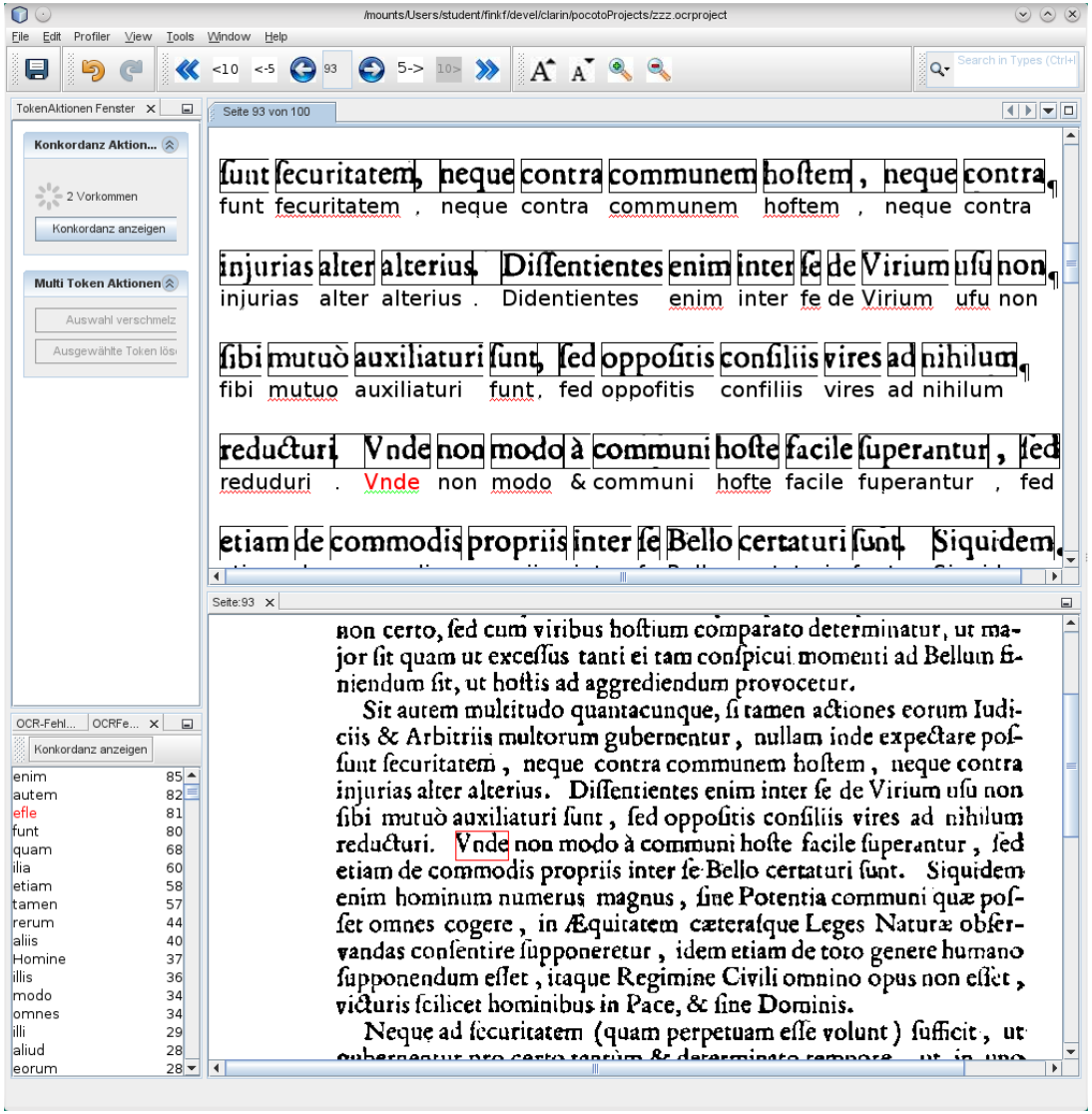
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



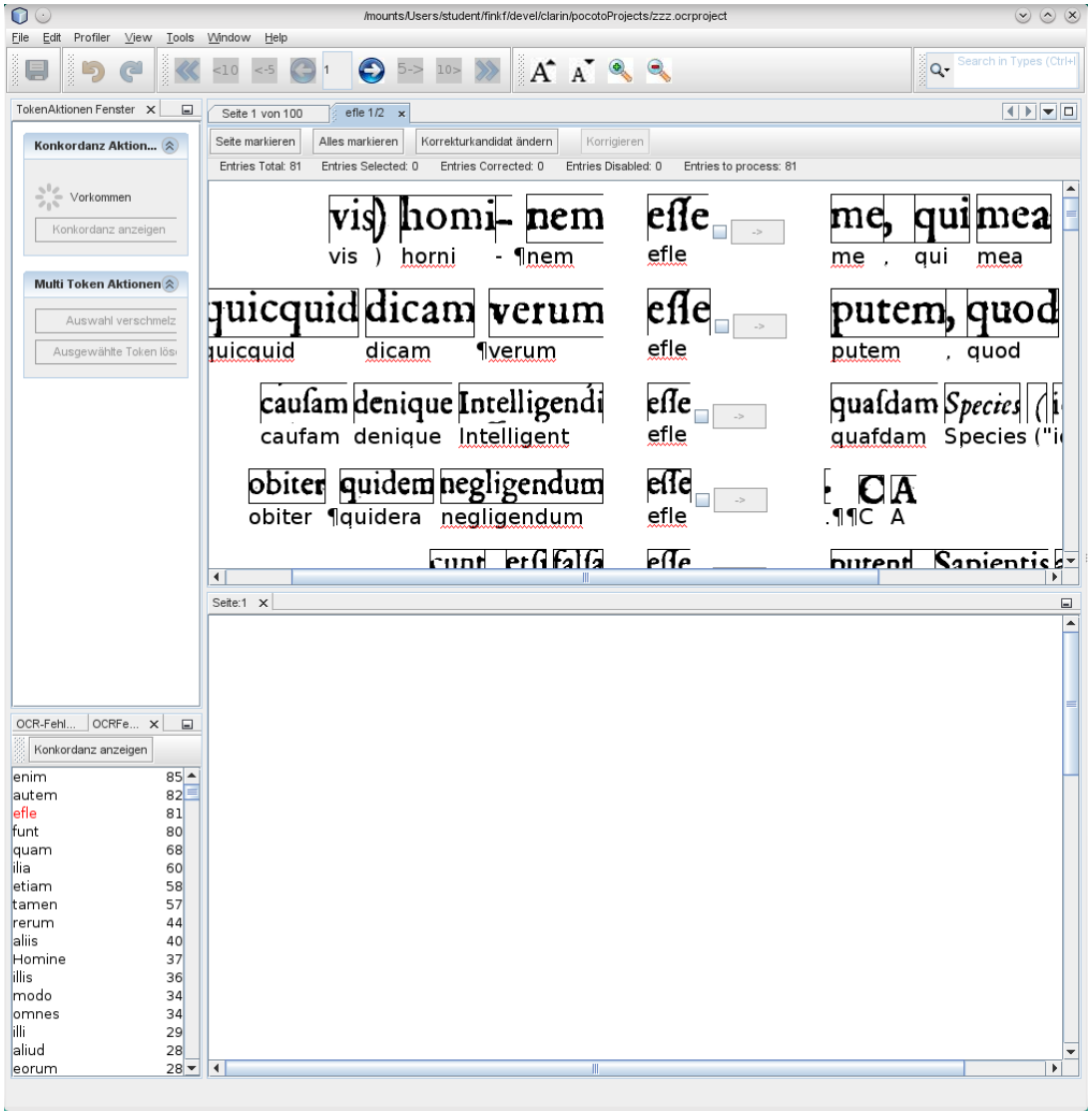
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



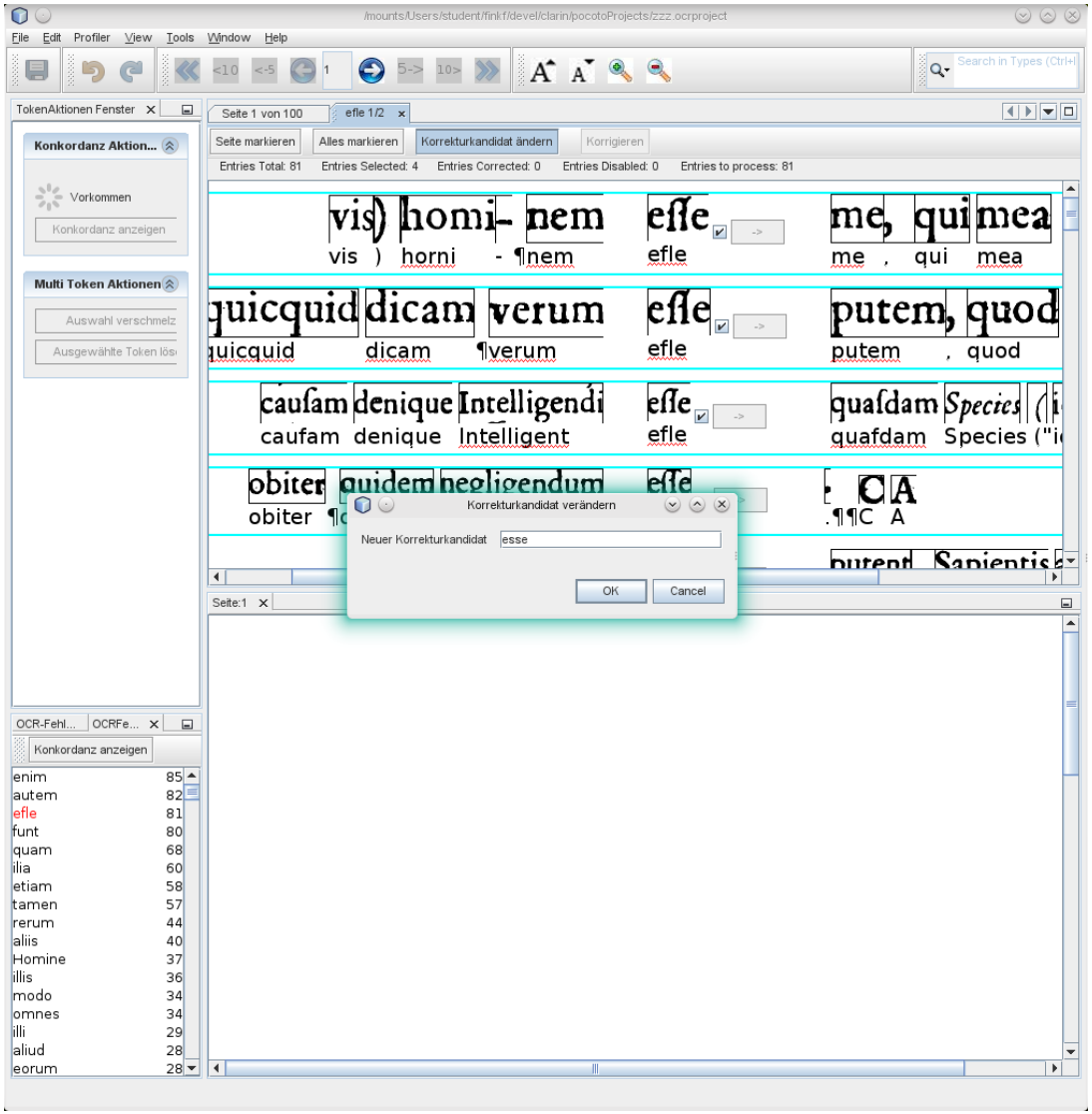
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



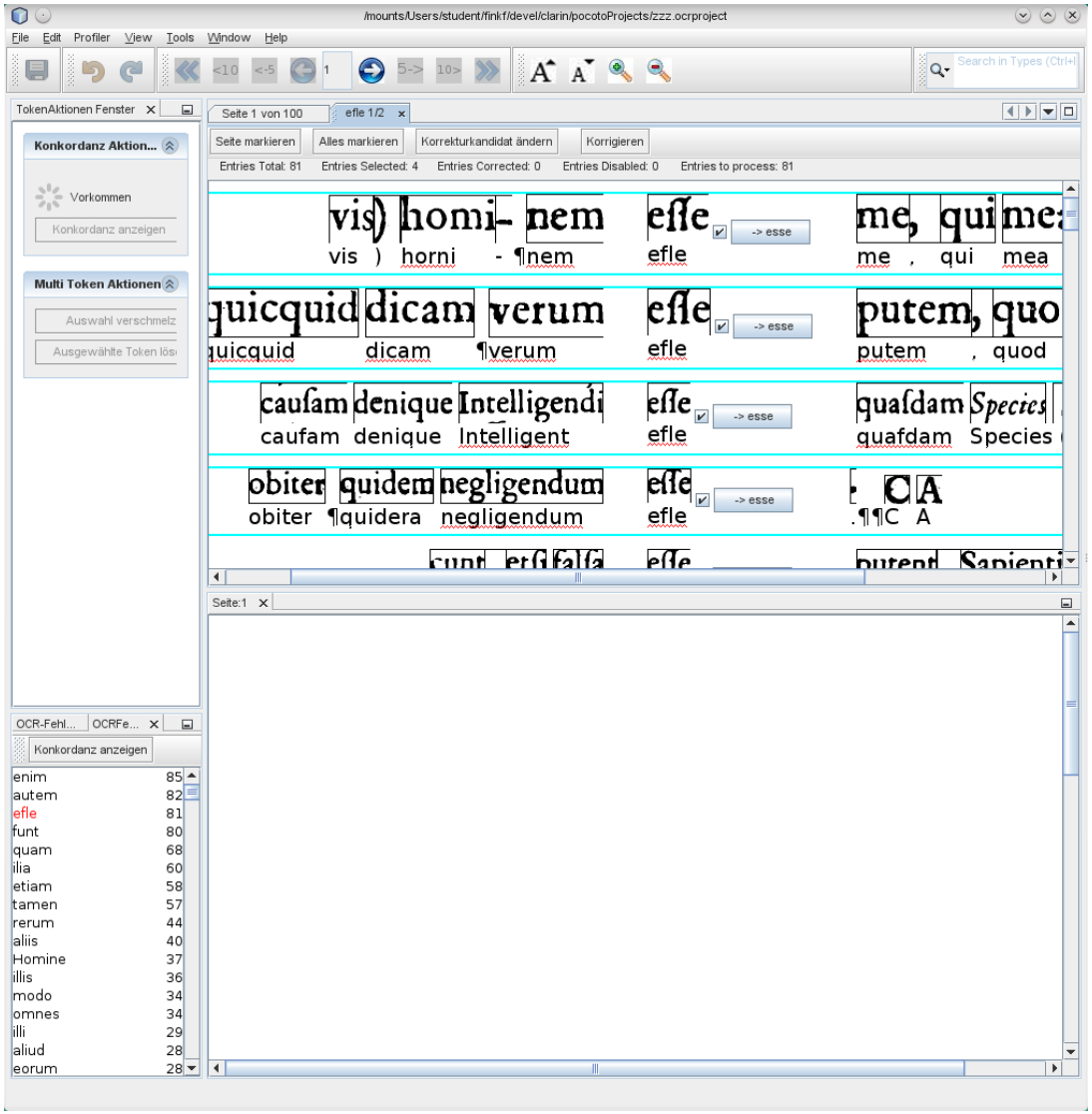
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.



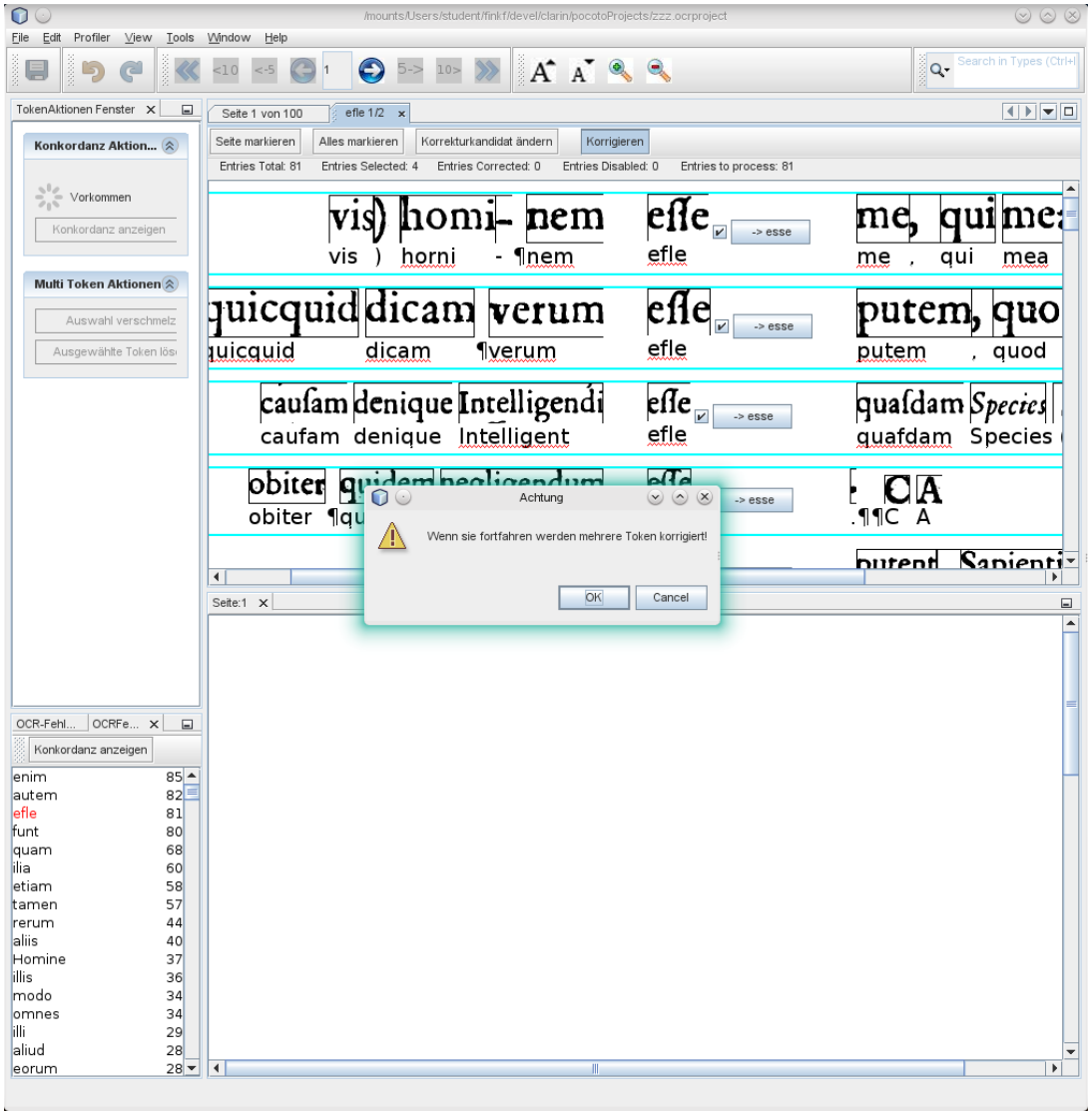
- Common errors and error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent errors can be easily selected and corrected in one step.



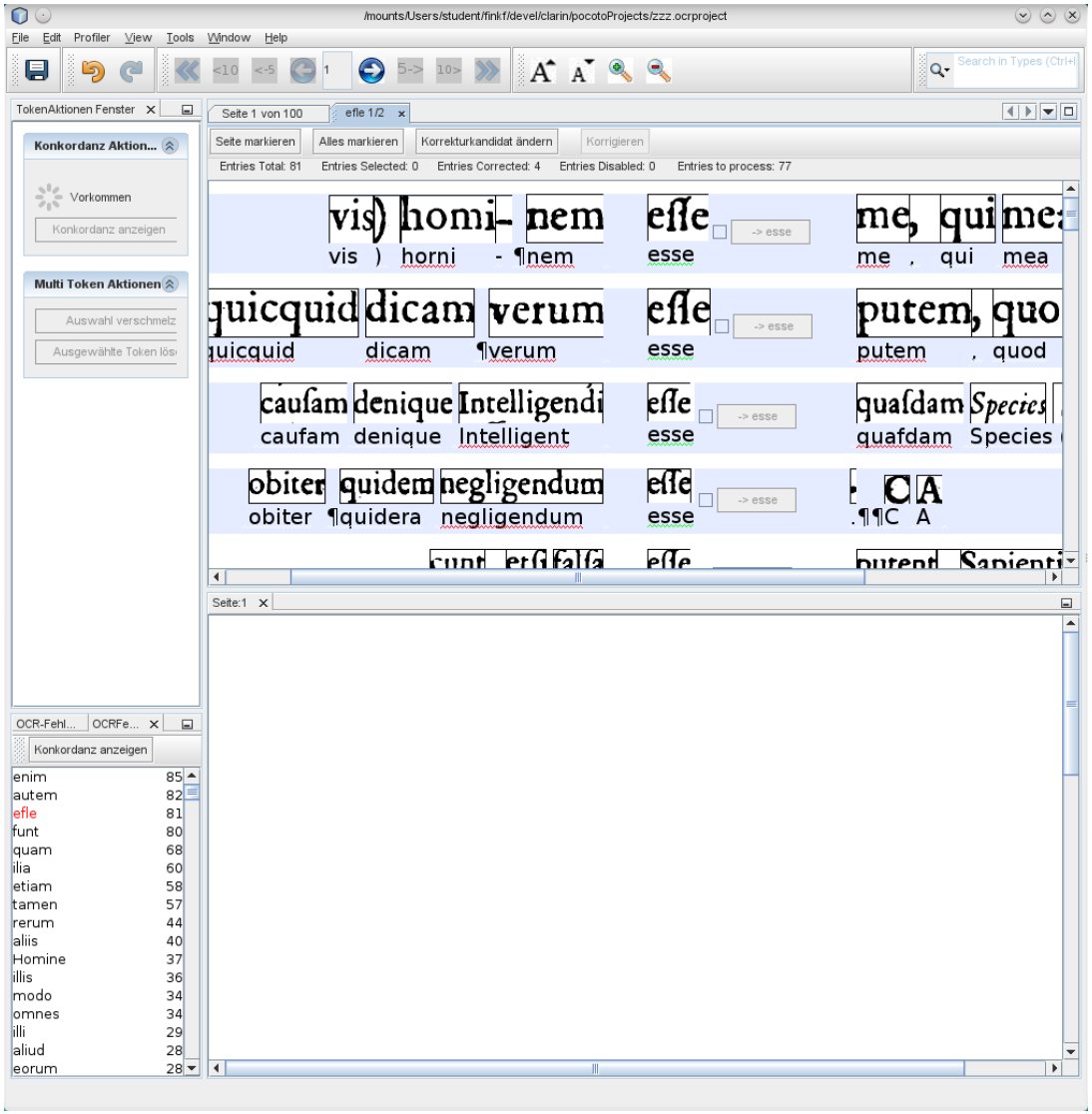
- Common errors and error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent errors can be easily selected and corrected in one step.



- Common errors and error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent errors can be easily selected and corrected in one step.



- Common errors and error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent errors can be easily selected and corrected in one step.



- Common errors and error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent errors can be easily selected and corrected in one step.

Thanks for your attention!