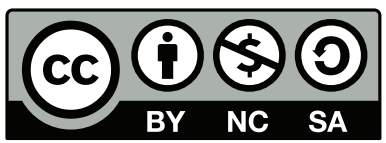
Module 10 PoCoTo: Theory

Florian Fink

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



2015-09-15

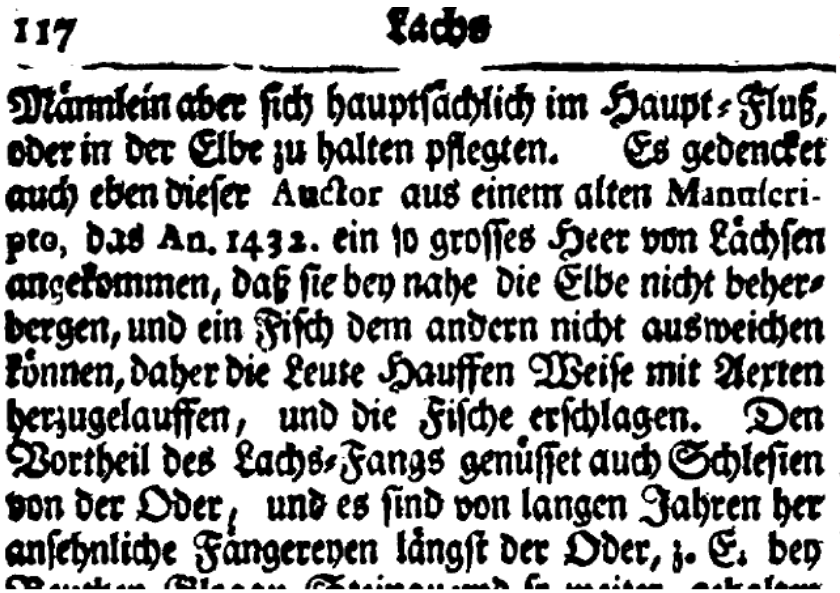
Introduction to postcorrection

Motivation

In the resent years a lot of historical documents have been scanned and OCRed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents is constantly improved.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.

Text recognition on historical documents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 Lachſen 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*.

Characterwise recognition rates

Jahr	Sprache	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:

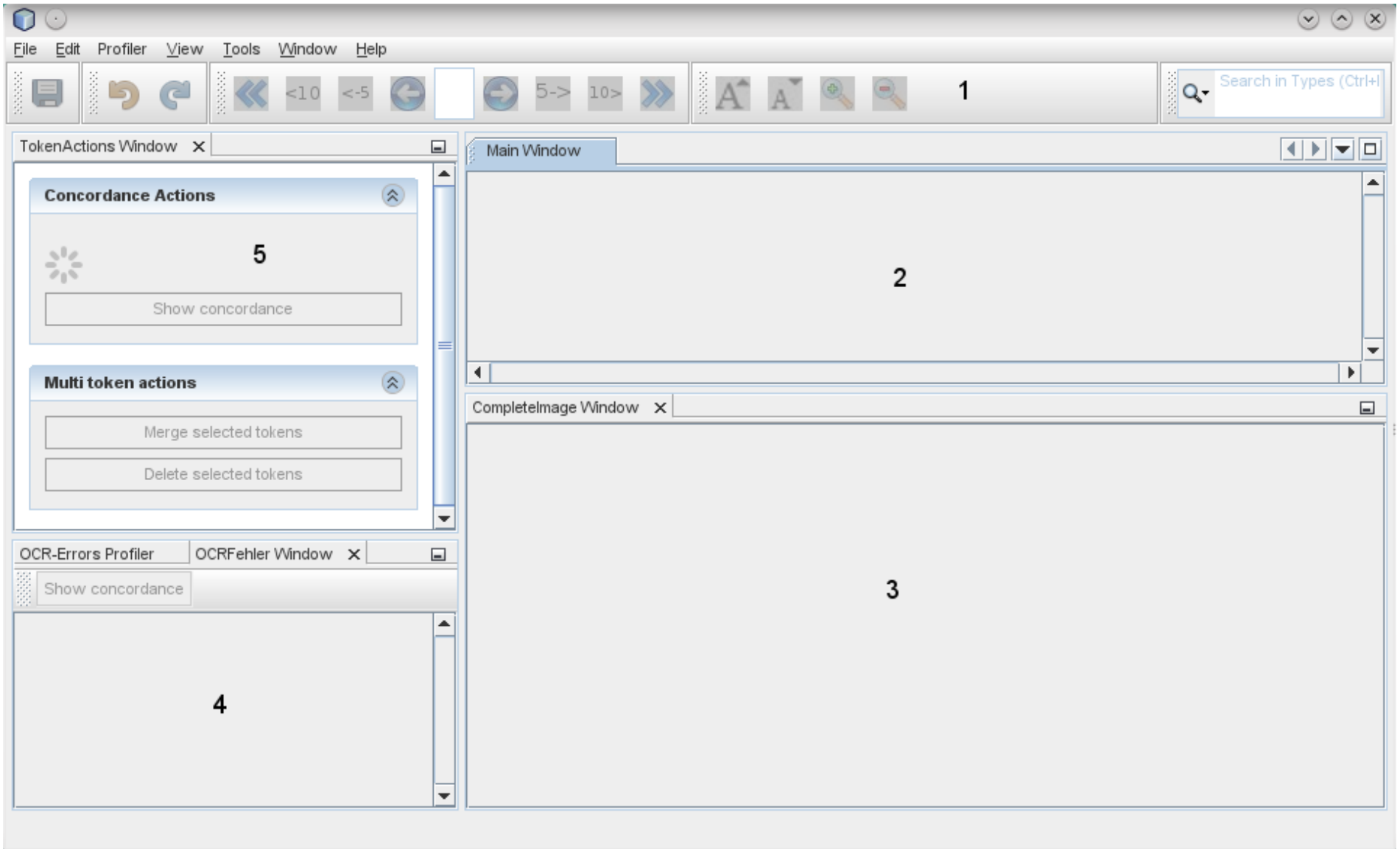
- manual (double) keying of the original sources is expensive
- interactive postcorrection can be used examine the results of the OCR
- interactive postcorrection can be used to improve the results of the OCR

PoCoTo – The PostCorrectionTool

Overview

- PoCoTo is a tool for the interactive postcorrection of OCRed text.
- It was developed as part of the EU founded project IMPACT.
- It is open source and anyone can help to improve it.
- Currently it is under active development.
- It contains aids to automatically correct systematic errors
- It contains linguistic and visual aids to support the postcorrection.
- You find its documentation in the PoCoTo manual.

The main areas

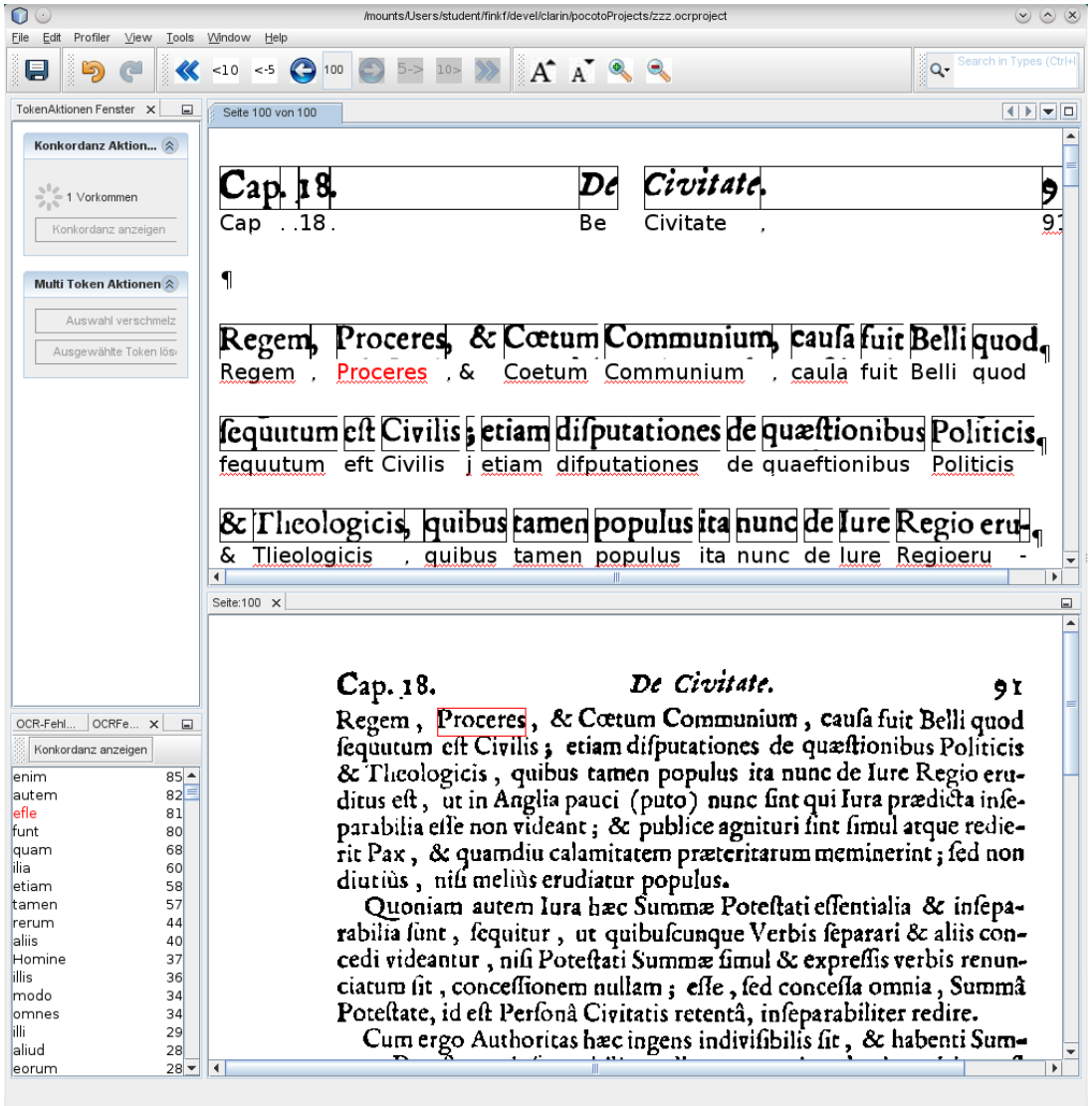


The 5 main areas of PoCoTo

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

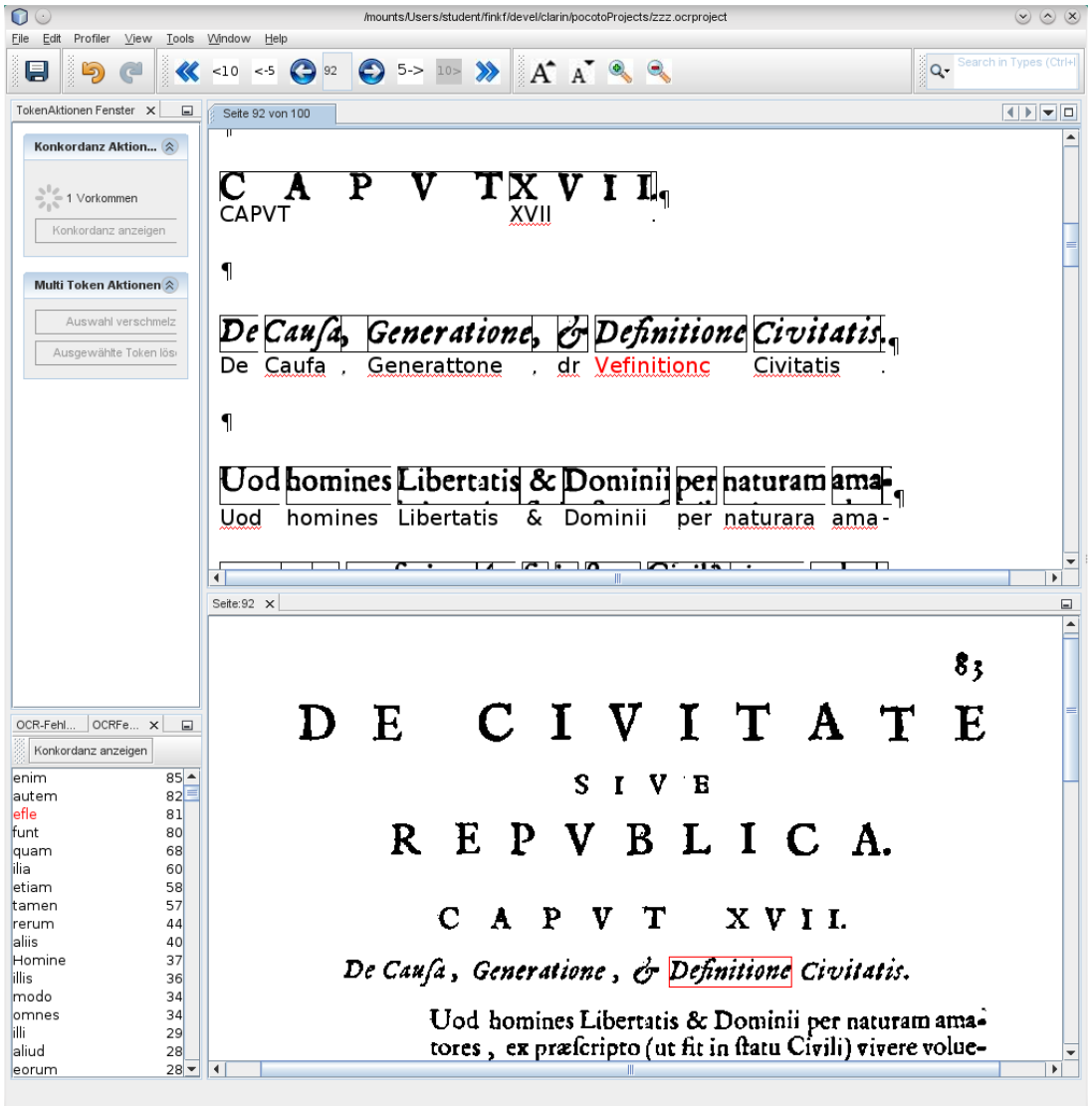
1. The menu area contains various commands for navigation and project maintenance.
2. The main view area shows tokens and offers the main correction possibilities.
3. The complete image area displays the page of the current active (selected) token.
4. The error area lists error frequency lists of common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

The visual UI



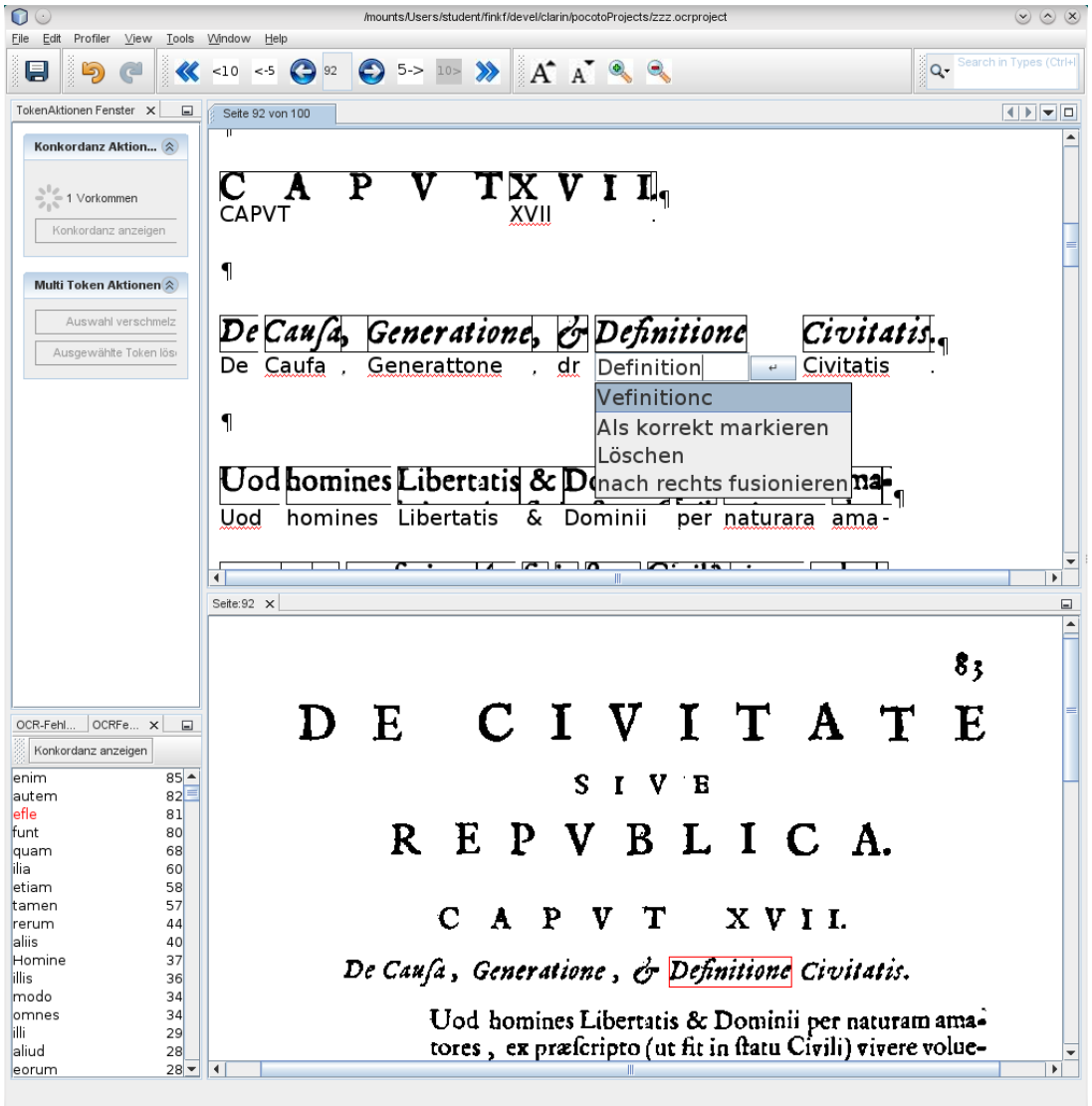
- The token of the text are displayed along with their image details.
- The page context shows the active token on the original page.
- Error frequencies – based on the confidence values of the OCR engine – are shown.

Interactive postcorrection: correcting single tokens



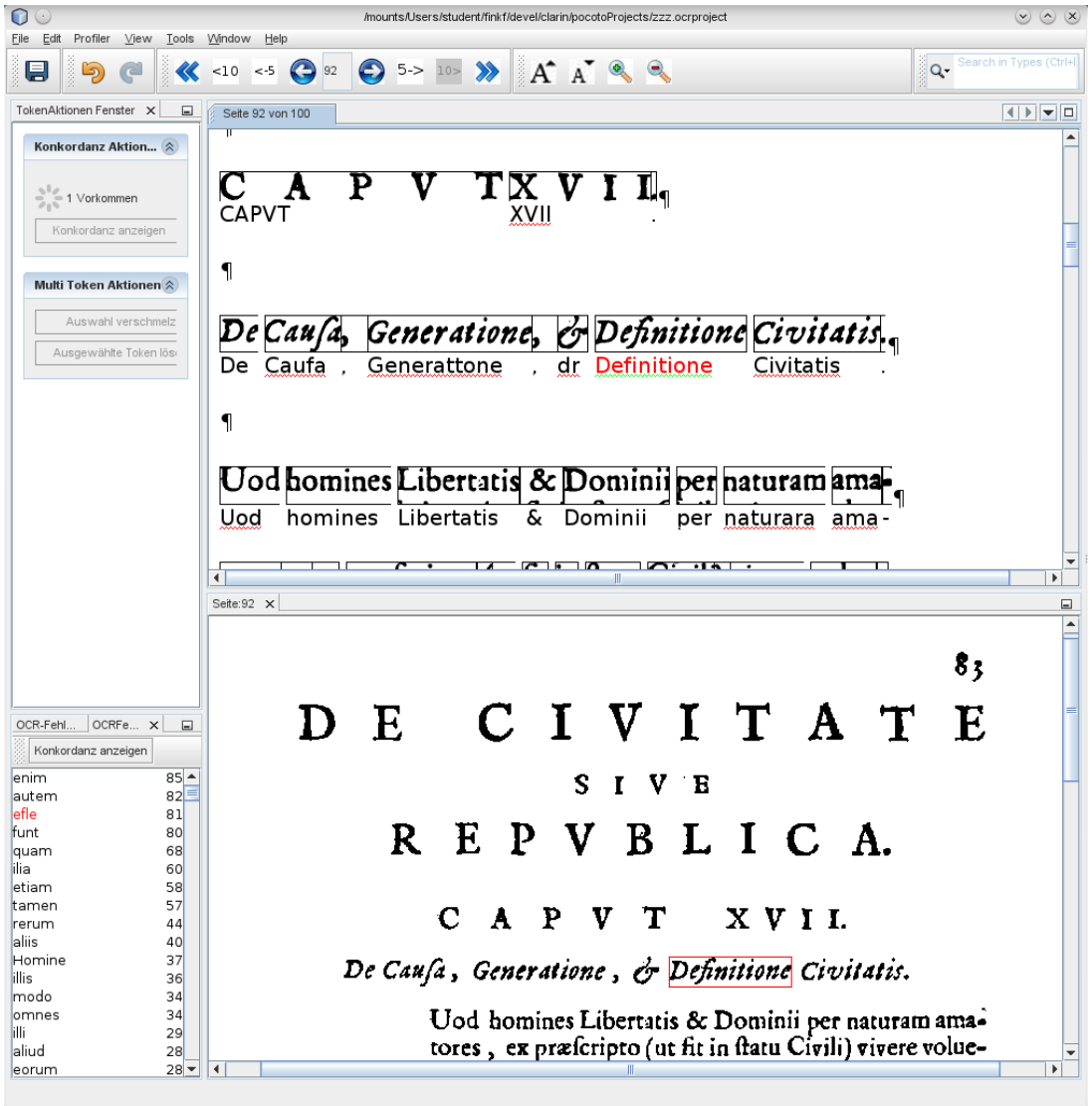
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neigbours.
- Words can be corrected manually in the window.

Interactive postcorrection: correcting single tokens



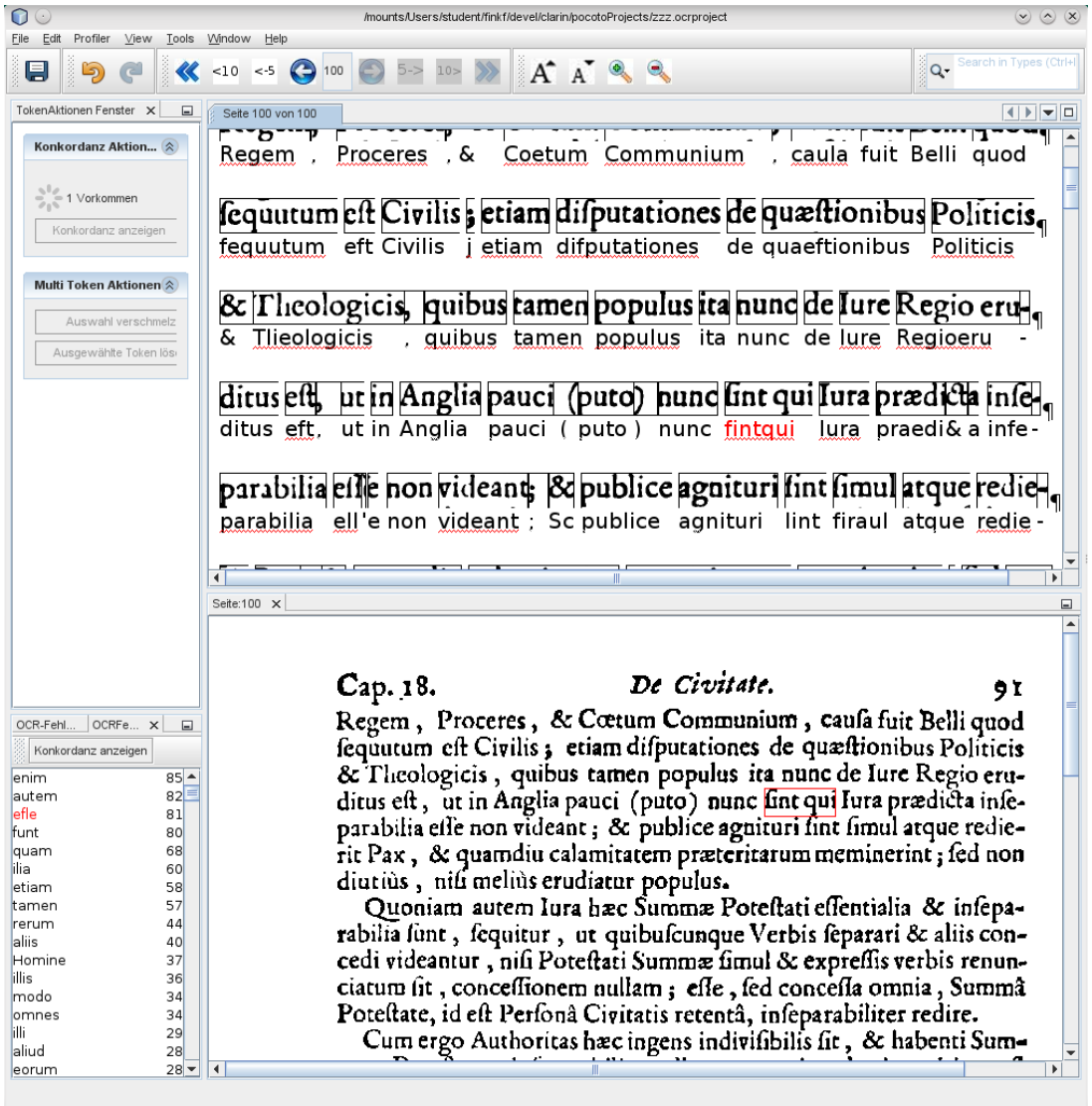
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neigbours.
- Words can be corrected manually in the window.

Interactive postcorrection: correcting single tokens



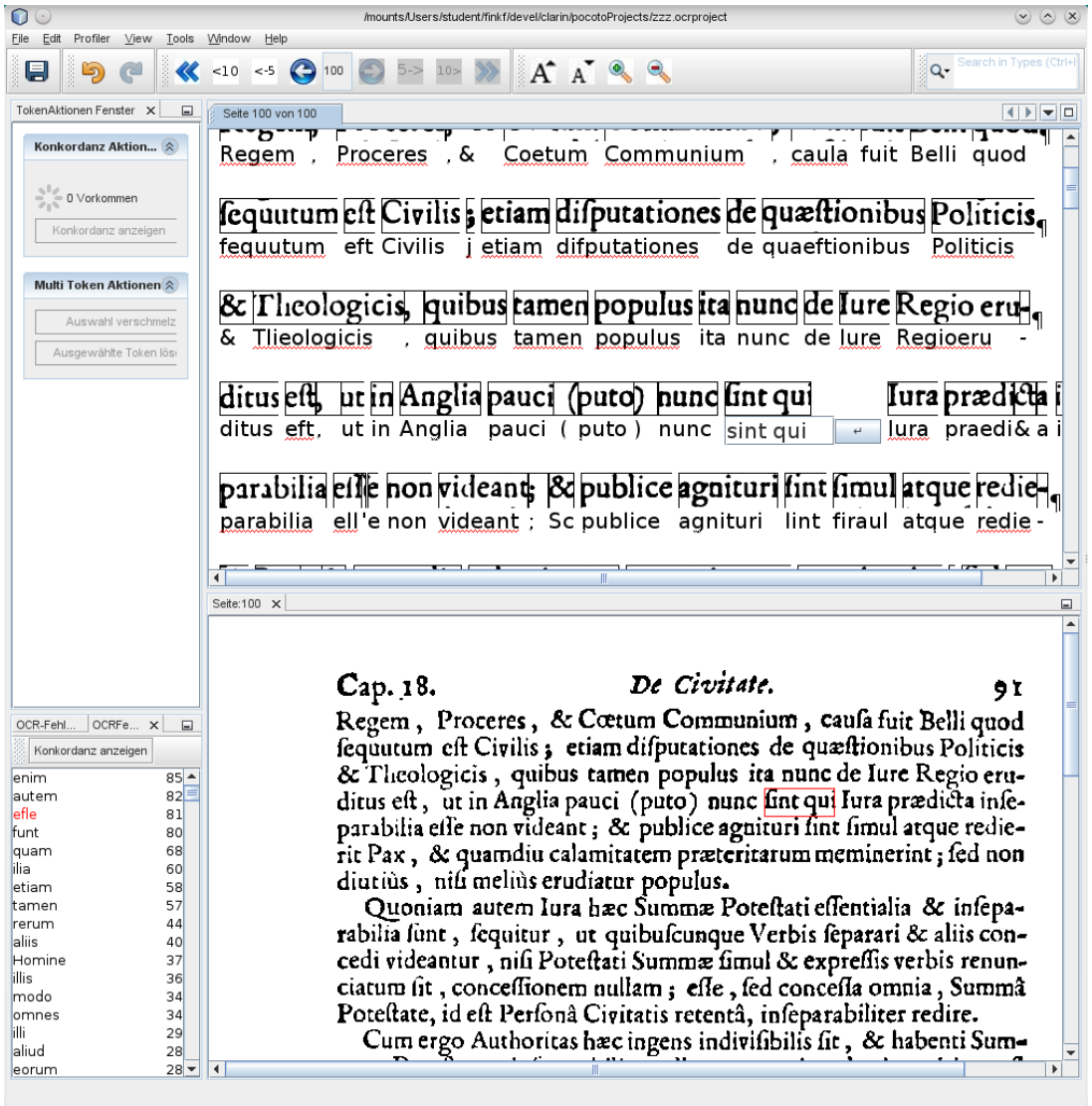
- *Suspicious* words are marked in the text.
- Words can be marked as correct.
- Words can be merged with their right neigbours.
- Words can be corrected manually in the window.

Interactive postcorrection: splits and merges



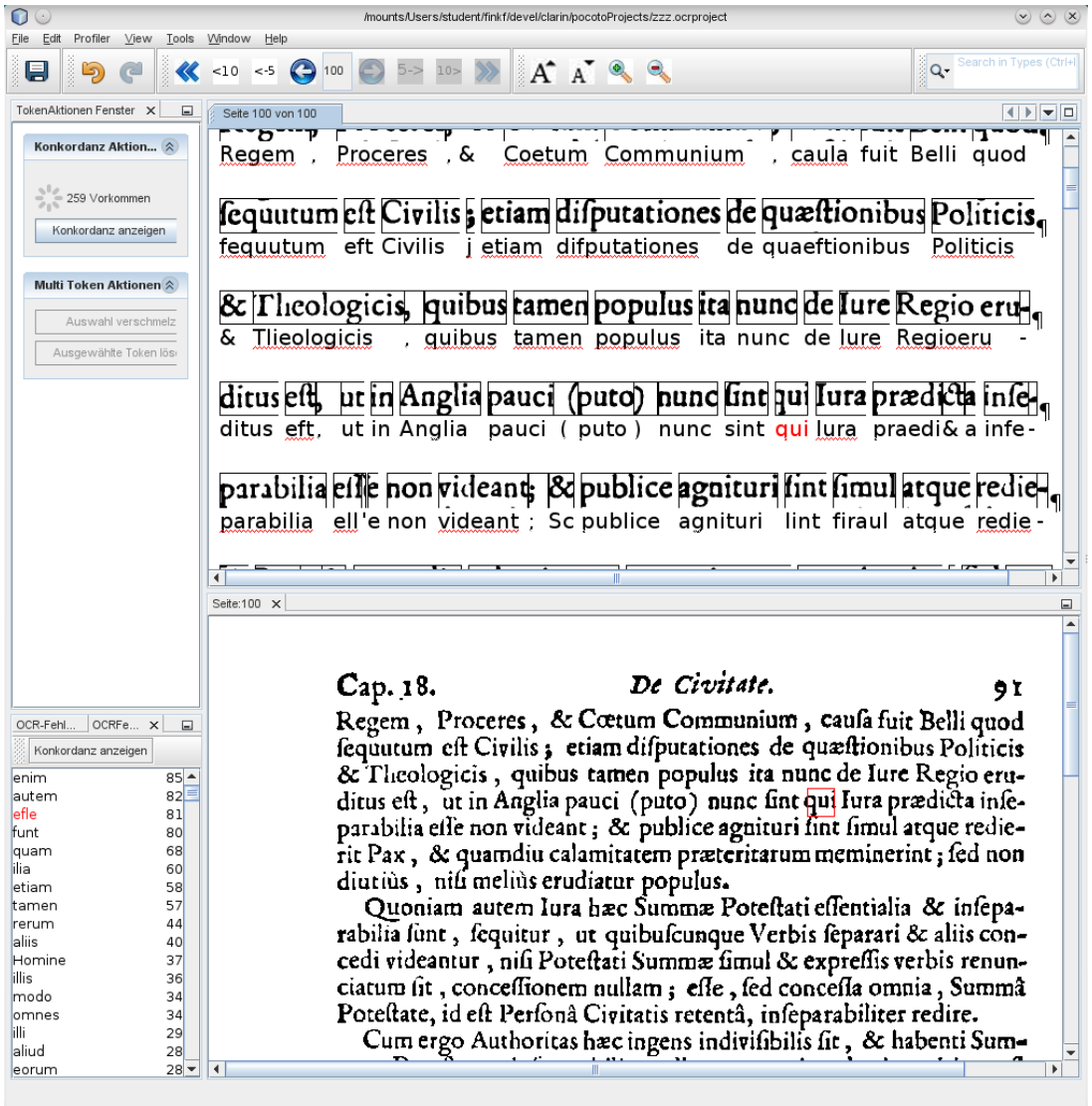
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: splits and merges



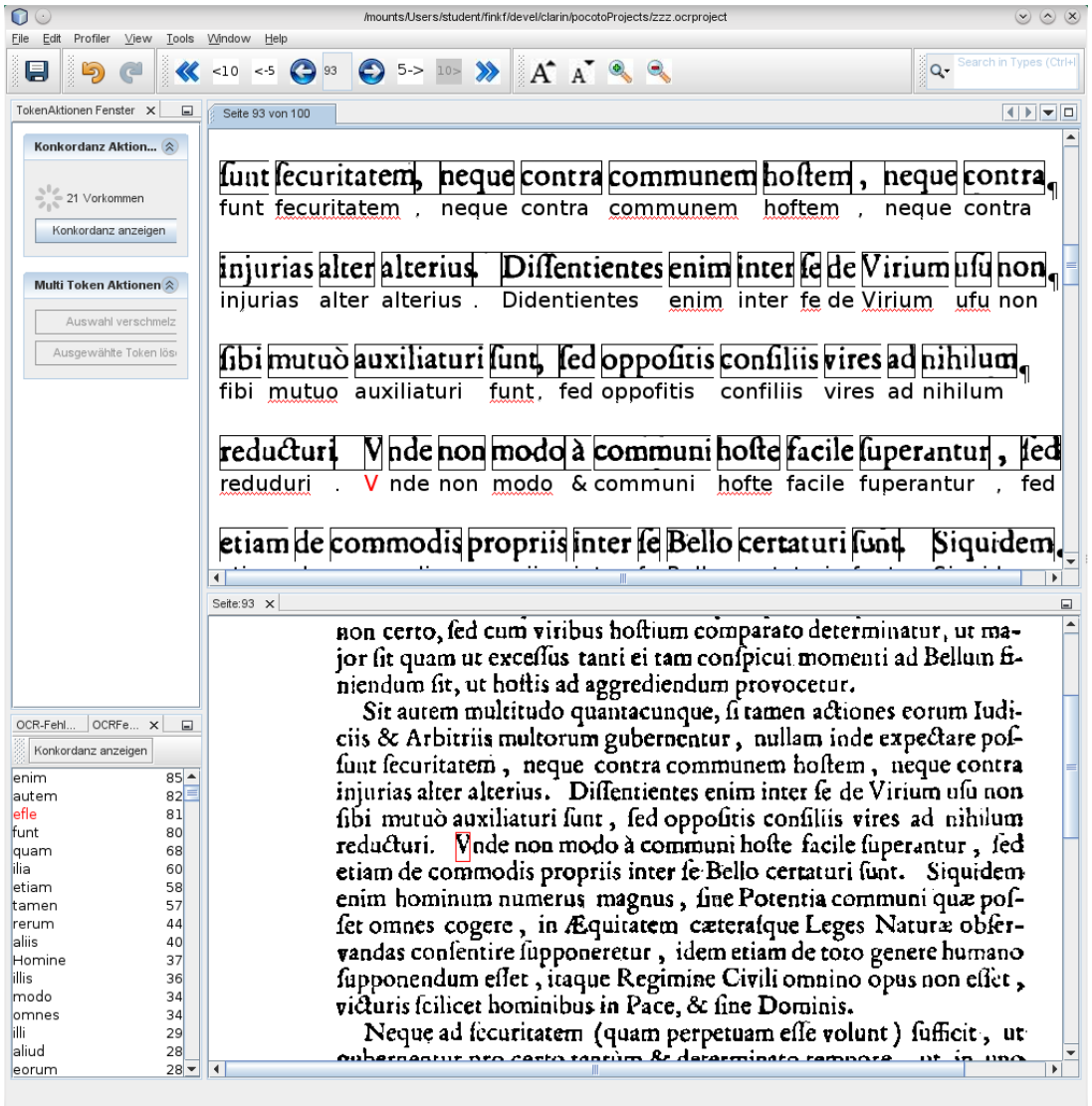
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: splits and merges



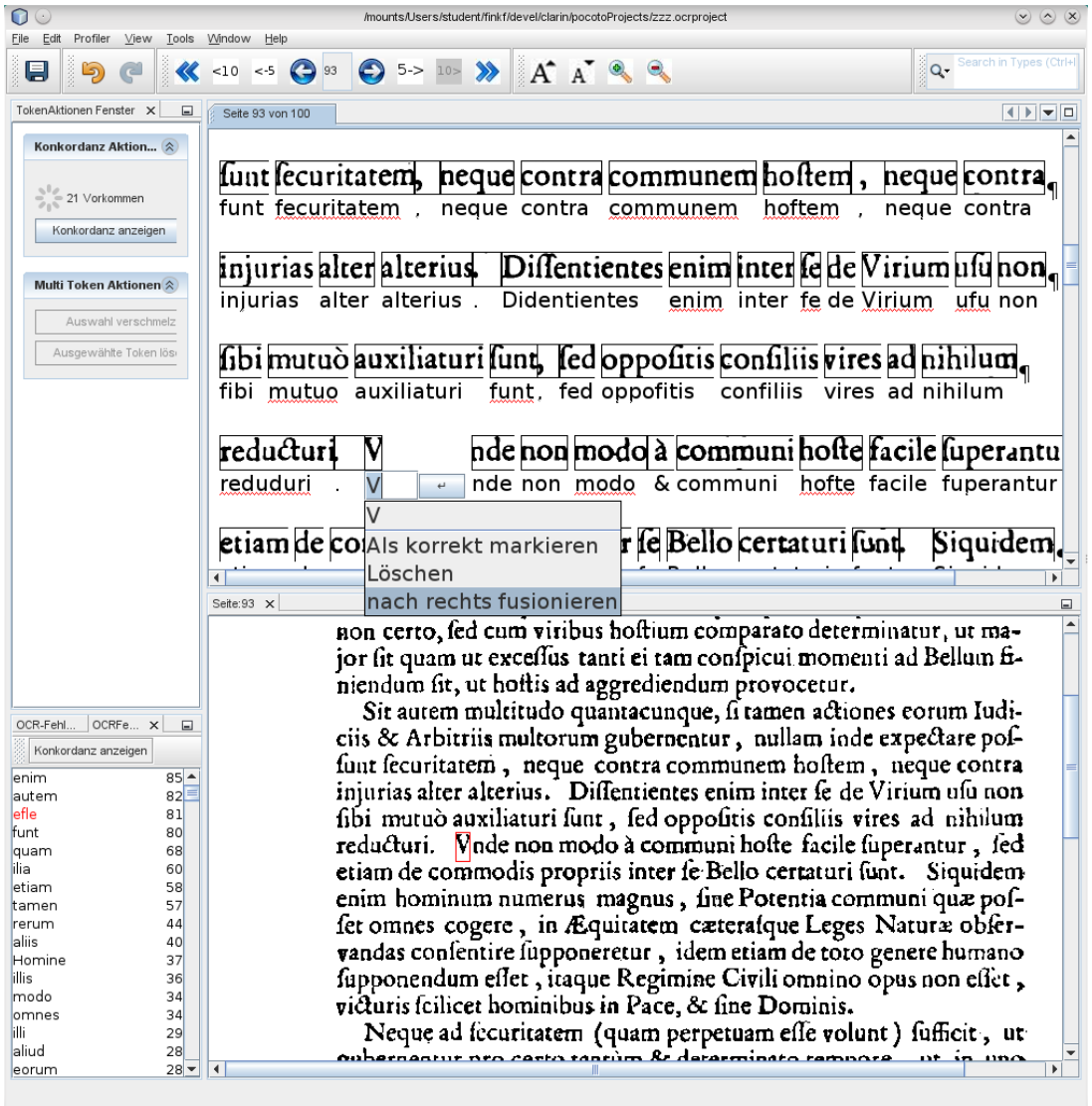
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: splits and merges



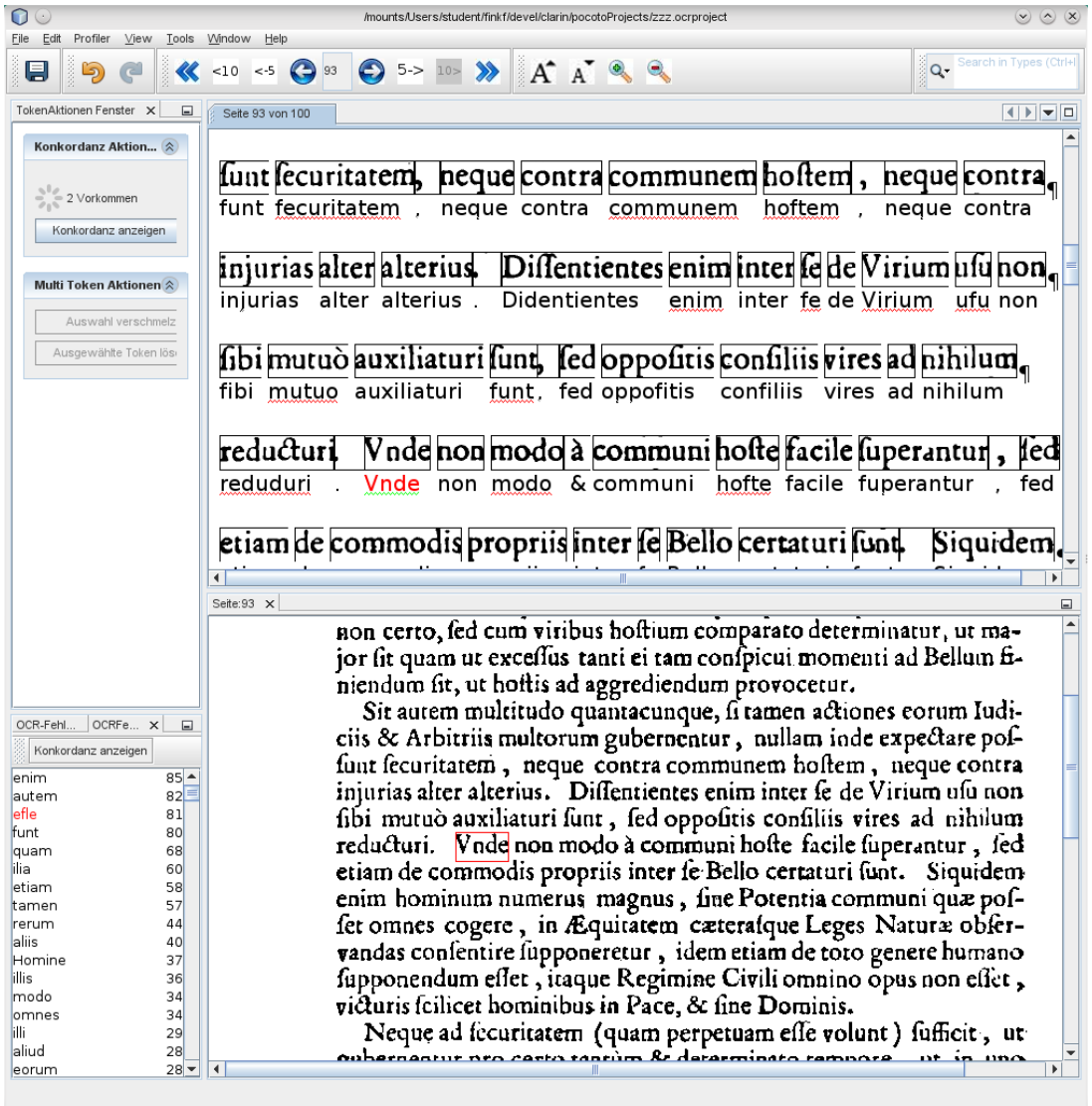
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: splits and merges



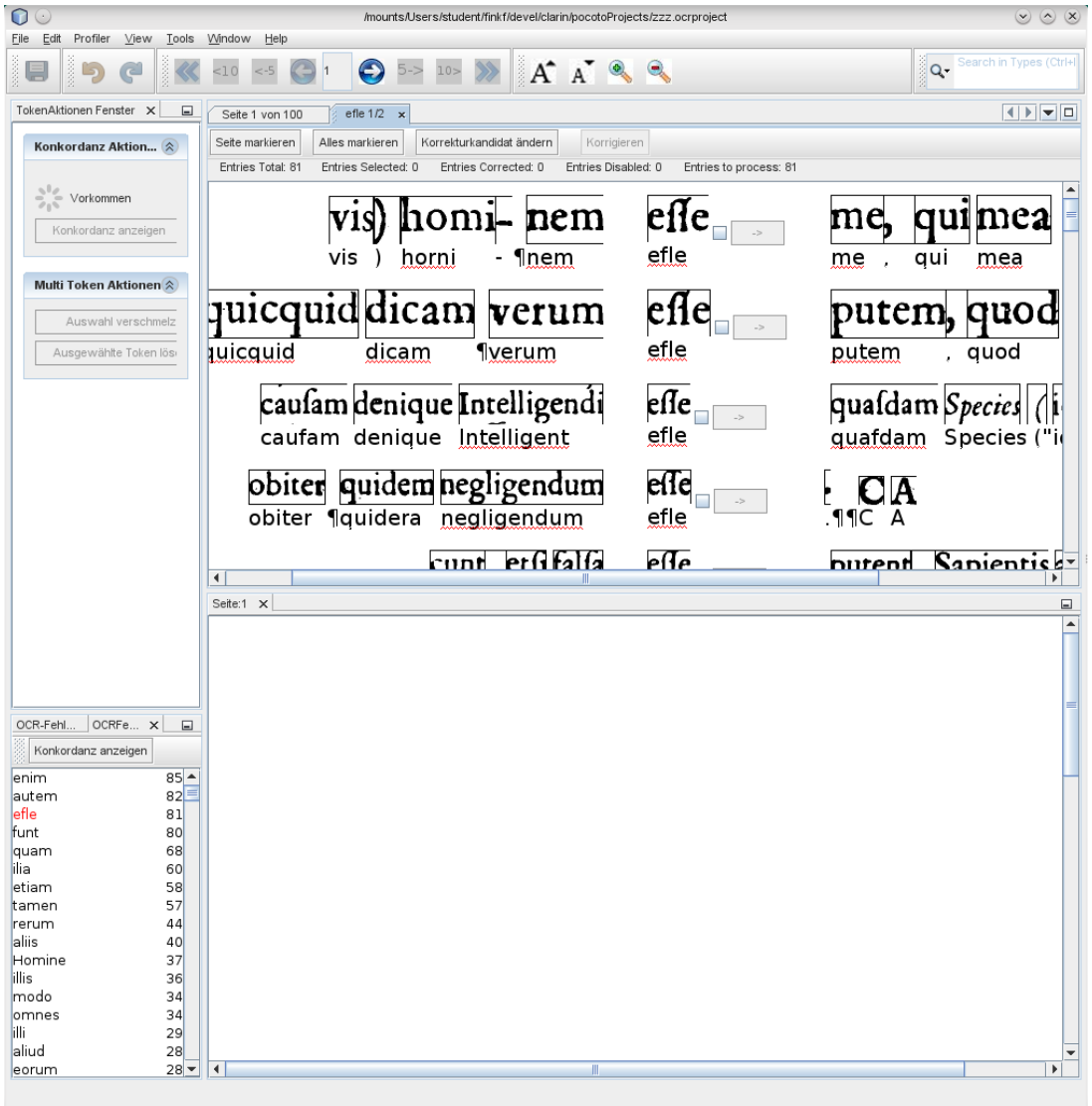
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: splits and merges



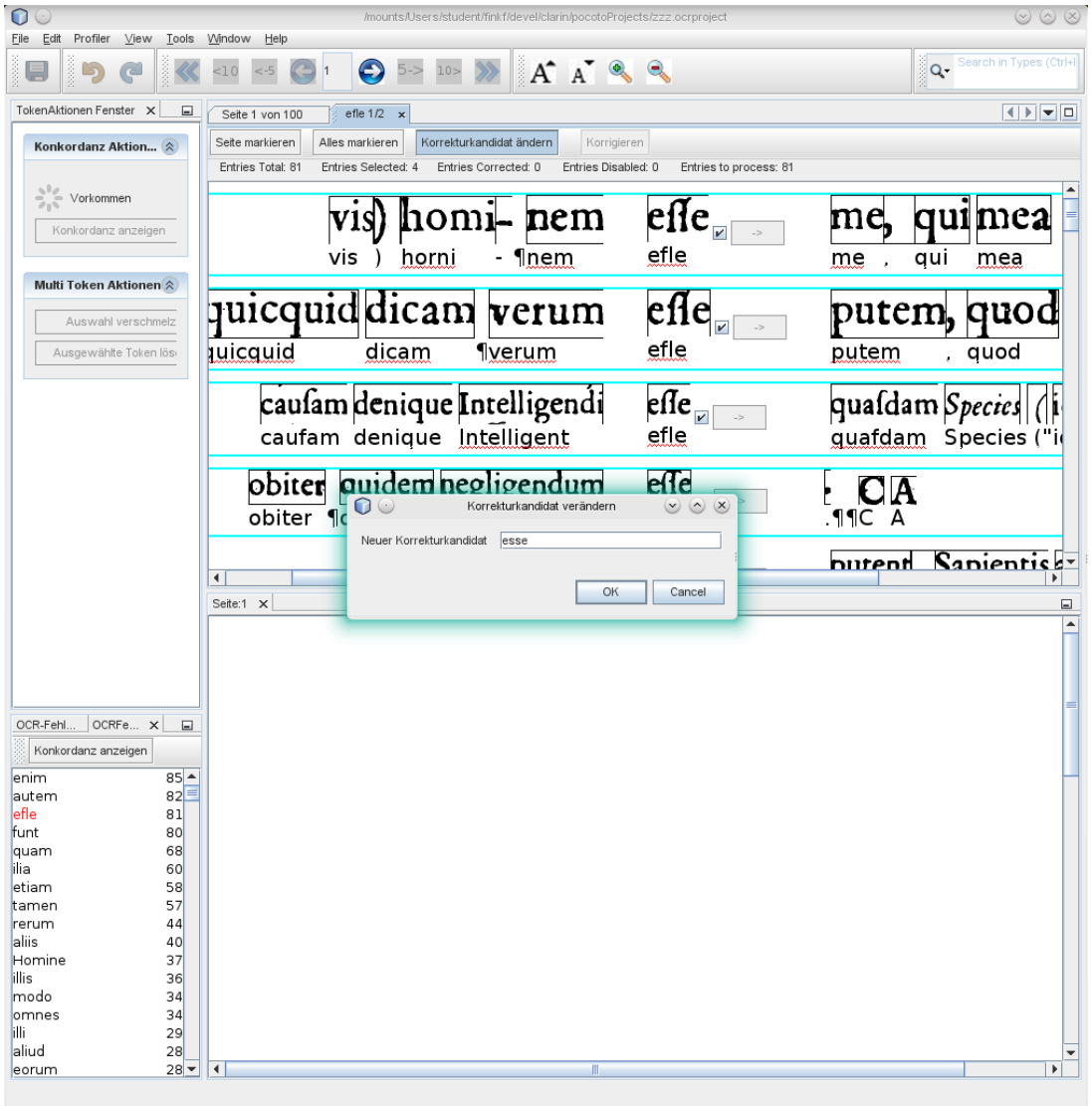
- Merged token can be easily split.
- Multiple, split token can be easily merged back together.

Interactive postcorrection: concordances



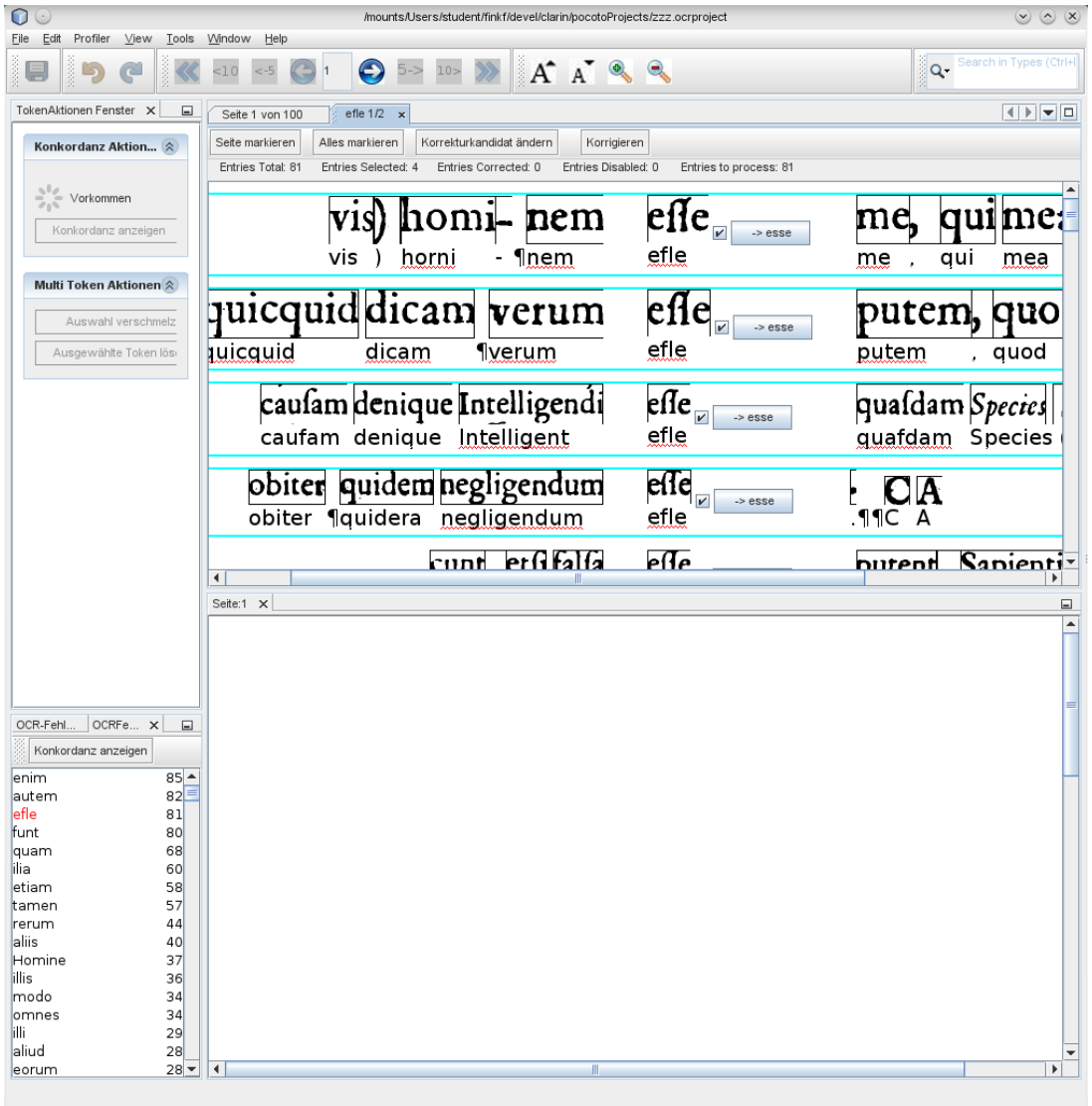
- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

Interactive postcorrection: concordances



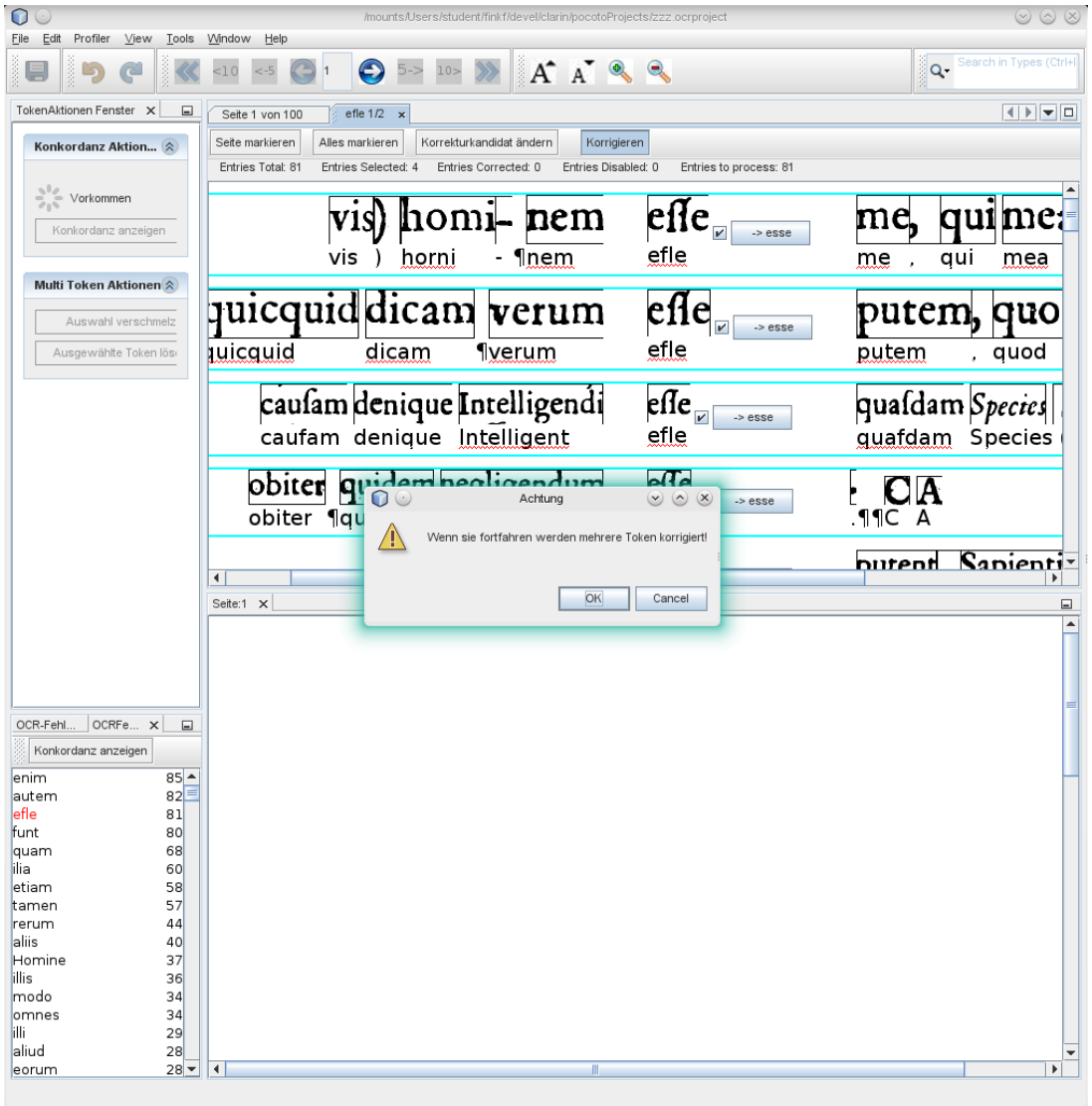
- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

Interactive postcorrection: concordances



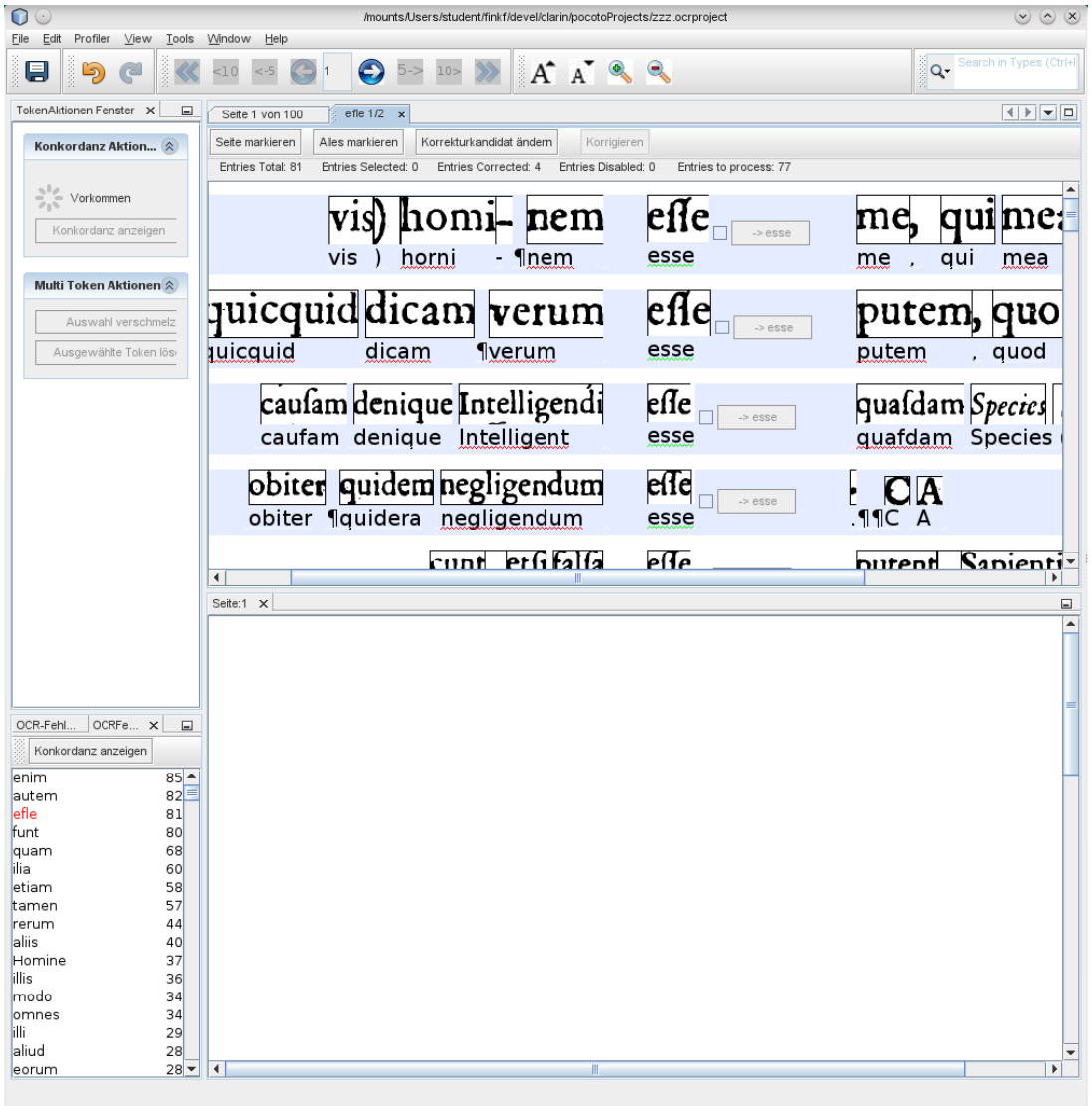
- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

Interactive postcorrection: concordances



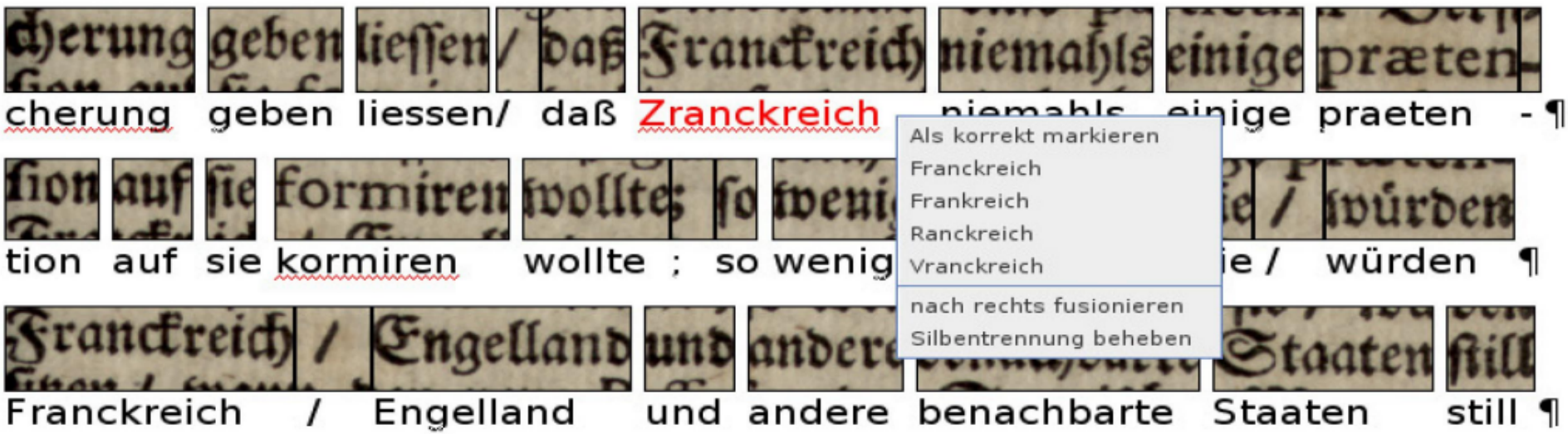
- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

Interactive postcorrection: concordances



- 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.

Interactive postcorrection: correction suggestions



- PoCoTo uses an external language profiler to generate correction suggestions.
- Common OCR and historical error pattern are both recognized.
- There are a lot of different language available for the profiler.

PoCoTo – Projects and profiles

PoCoTo installation

- You can download the application from this link
  - After the download has finished you should see a file called `ocrcorrection.zip` in your download folder.
  - Copy or move this file to a convenient place and extract the archive.
  - You will see a folder called `ocrcorrection`.
  - Navigate into the directory `ocrcorrection/bin` and double click on the executable file `ocrcorrection` or `ocrcorrection.exe`
  - You can create a link to this executable on your desktop for easier access.
  - After you double clicked on the file, PoCoTo should start.
  - You can create a link to this executable on your desktop for easier access.
- The downloading and installation of the tool will be covered in more depth in the next module

PoCoTo project structure

- 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[1].
- 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.

[1] save for the file's file extension (`.xml`, `.png`, ...)

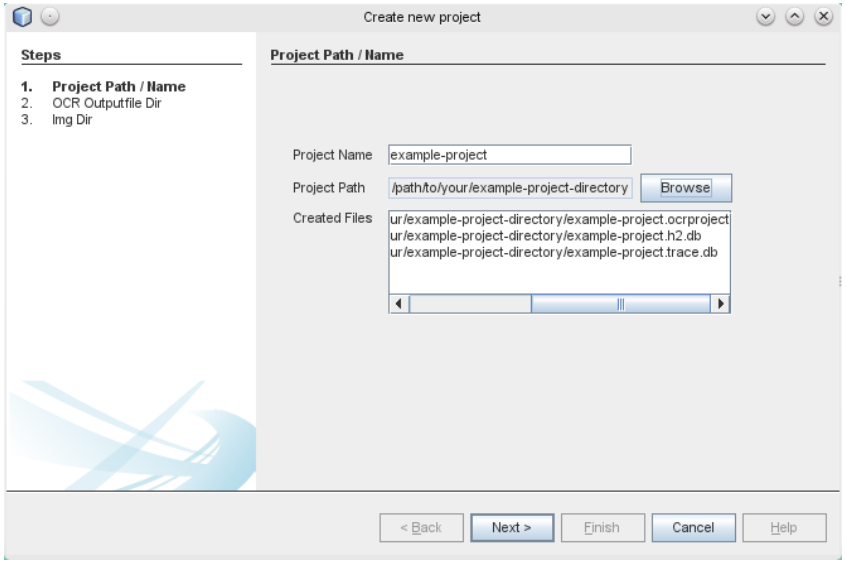
PoCoTo's file formats

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

1. The character based Abbyy XML format.
2. The HOCR file format.

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

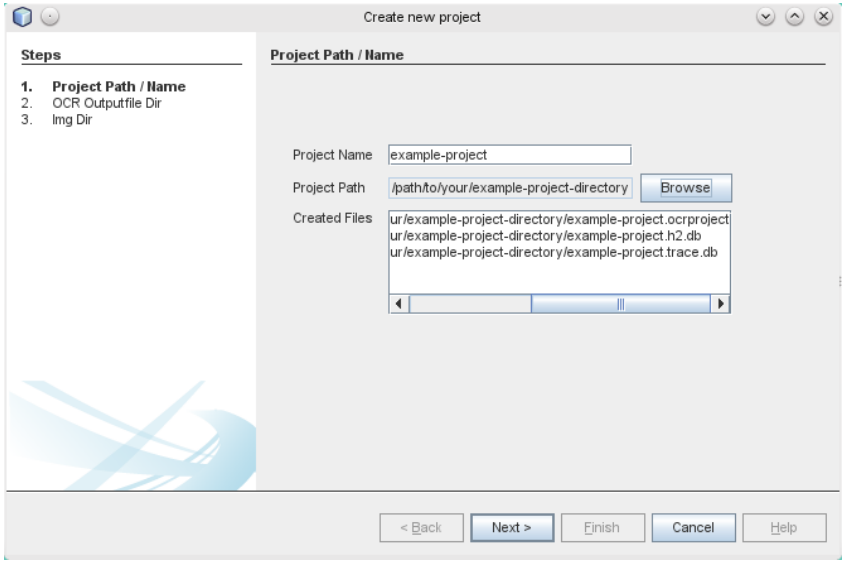
Creating a new project



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

1. Insert a name and a path for your project. Click `next`.
2. Insert a the path of your folder, that contains the XML files and select the type of your XML files. Click `next`.
3. Select the path to the folder, that contains your image files. Click `finish`.

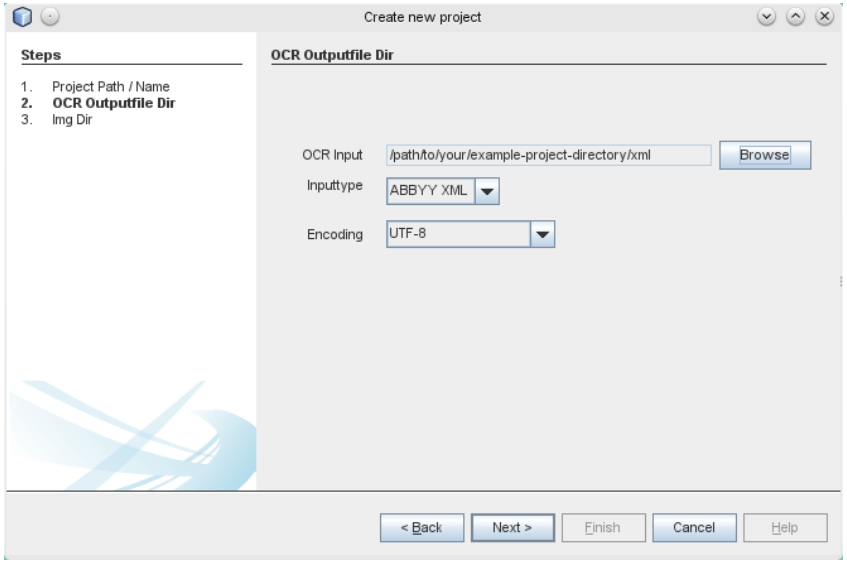
Creating a new project



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

1. Insert a name and a path for your project. Click `next`.
2. Insert a the path of your folder, that contains the XML files and select the type of your XML files. Click `next`.
3. Select the path to the folder, that contains your image files. Click `finish`.

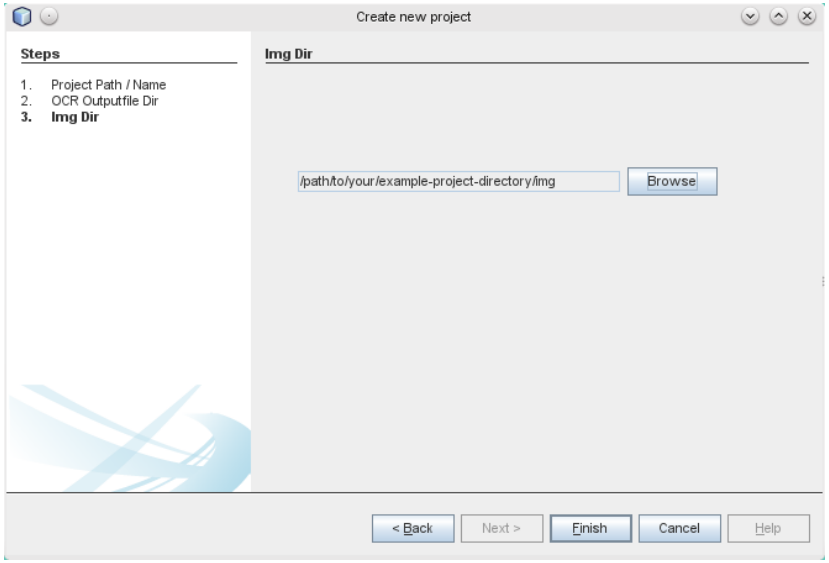
Creating a new project



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

1. Insert a name and a path for your project. Click `next`.
2. Insert a the path of your folder, that contains the XML files and select the type of your XML files. Click `next`.
3. Select the path to the folder, that contains your image files. Click `finish`.

Creating a new project



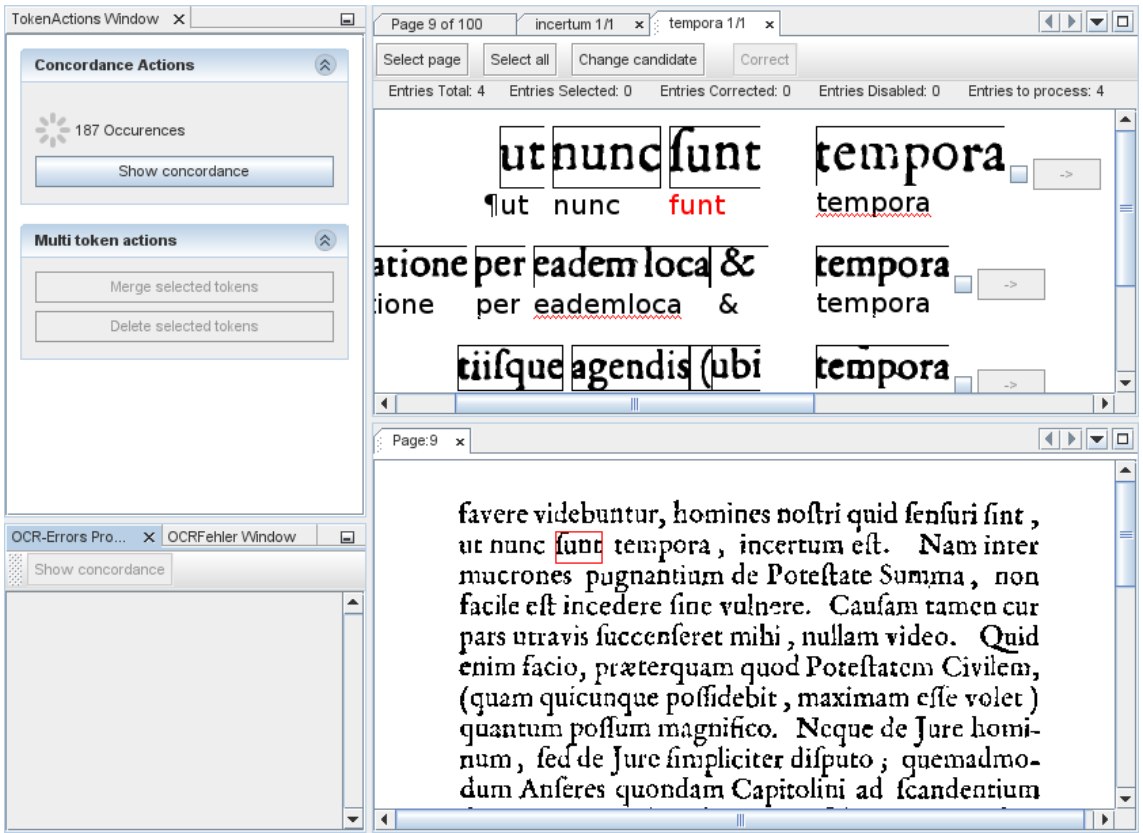
You can create new projects using the project wizard. Click to `file->create new project` and the first frame of the project wizard open.

1. Insert a name and a path for your project. Click `next`.
2. Insert a the path of your folder, that contains the XML files and select the type of your XML files. Click `next`.
3. Select the path to the folder, that contains your image files. Click `finish`.

Navigation in the project

- 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.

Creating a concordance view



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.

Profiling a project

- If you want to profile a document, make sure that you have configured a valid profiler web service url (see the profiler manual for more information).
- You can always use the default profiler url of `PoCoTo`.
- You can always profile your current project by clicking `profiler->order document profiler` in the menu area:
  - If the url is valid and the profiler web service is running, you will see a window, which lets you choose which language profile to use.
  - Select a language and click to `order document profile`.
  - Do as `PoCoTo` says and get your self some coffee.
- After the profiling has stopped, you now will have access to the common error pattern tab in the error area and you will get a list of correction suggestions if you try to correct a token.

Thanks for your attention!