

DATECH 2017 – PoCoTo Workshop – PoCoTo

Florian Fink

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



May 30, 2017

In the recent years a lot of historical documents have been scanned and OCR'ed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents is constantly improved.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

117

Lachs

xi7_Kchs

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

Männleinccher sich hauptsächlich im Haupt-Fluss, öderm der Gbe zu halten pflegten. Es gedencktt auch eben dieser Auctor aus einem alten Mannferipto, das An. 1431. ein 1o grosses Heer von Lachsen angekommen, daß sie bey nahe die Elbe nicht beherrbergen, 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.

Improving Access to Text

IMPACT

- 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).

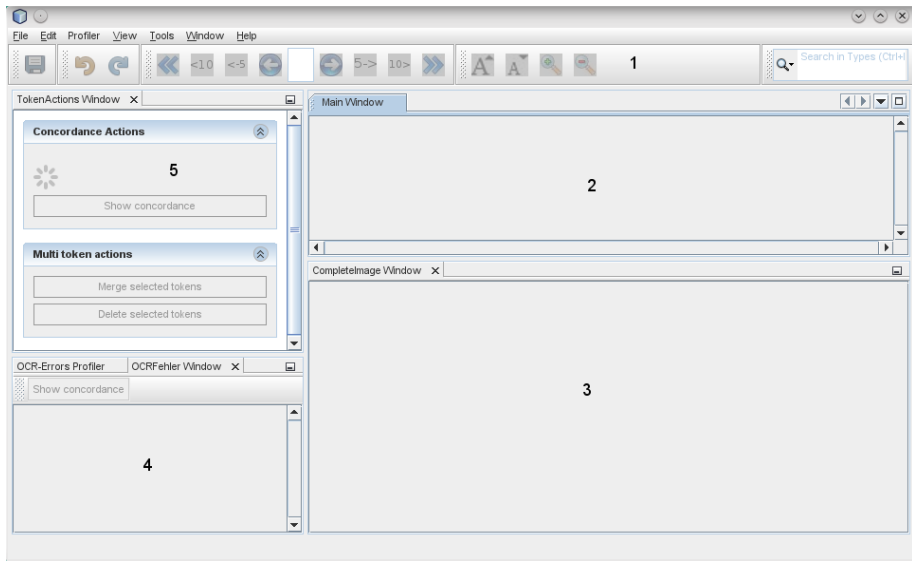
- PoCoTo has an automatic update mechanism – once installed, it is automatically kept up to date.
- The recognition results are visualized with the images of the original documents.
- The concordance views enable to examine different errors and error pattern over the whole document.
- A specialized profiling 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.

PoCoTo supports various formats of different OCR engines:

- hOCR used by Tesseract and Ocropus.
- ABBYY-XML used by the ABBYY FineReader.
- Ocropus-Directories used by Ocropus.
- DocXML used by the language profiler (next module).

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

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



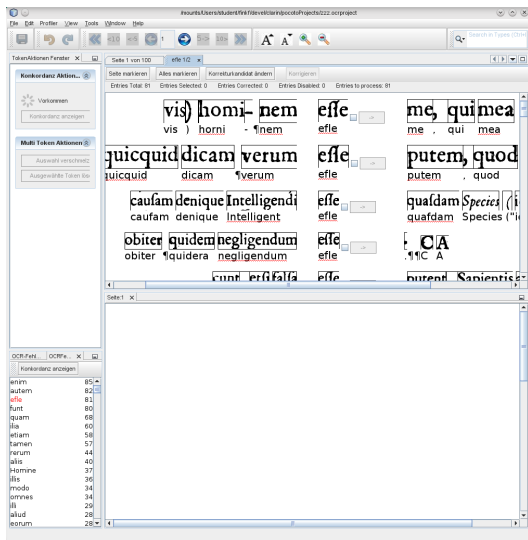
The screenshot displays the PoCoTo Workshop graphical user interface. The main window shows a Latin text from Cicero's *De Civitate*, specifically Chapter 18. The text is segmented into tokens, and the interface highlights the active token "Regem" in red. The text is displayed in a large, serif font, and the segmentation is visible below the main text. The interface includes a sidebar on the left with a search bar, a list of tokens, and a table of token frequencies. The table lists tokens and their corresponding frequencies, such as "Regem" with a frequency of 85. The table also includes a column for the token's position in the text, with values ranging from 1 to 28. The interface is designed to allow users to explore the tokenization of a text and visualize the results.

Tokenization details shown in the interface:

- Page: 100 von 100
- Text: **Cap. 18. De Civitate.**
- Text content: **Regem, Proceres, & Cœtum Communium, causa fuit Belli quod, sequutum est Civilis, etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura prædicta inseparabilia esse non videant; & publice agniti sunt simul atque redierit Pax, & quamdiu calamitatem præteritarum meminerint; sed non diutius, nisi melius erudiatur populus.**
- Tokenization: **Regem, Proceres, & Cœtum Communium, causa fuit Belli quod, sequutum est Civilis, etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura prædicta inseparabilia esse non videant; & publice agniti sunt simul atque redierit Pax, & quamdiu calamitatem præteritarum meminerint; sed non diutius, nisi melius erudiatur populus.**
- Token frequencies table:

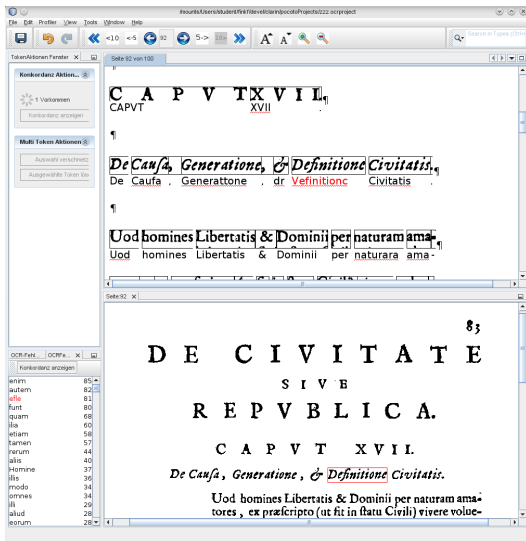
Token	Frequency
enim	85
autem	82
esse	81
sunt	80
quam	68
ita	60
etiam	58
tamen	57
perum	44
alio	40
homine	37
illis	36
modo	34
omnes	34
illi	29
aliud	28
eorum	28

- 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.

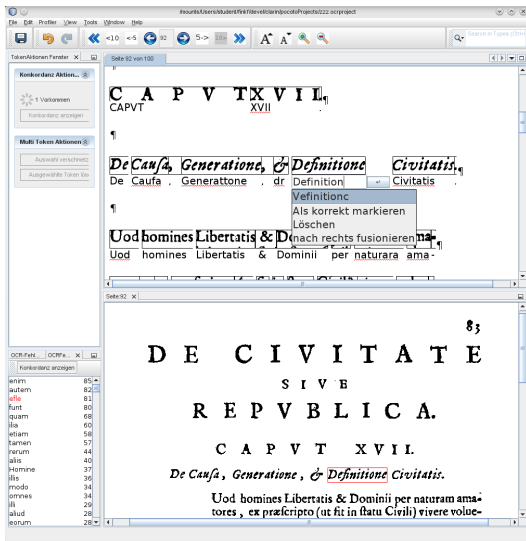


- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

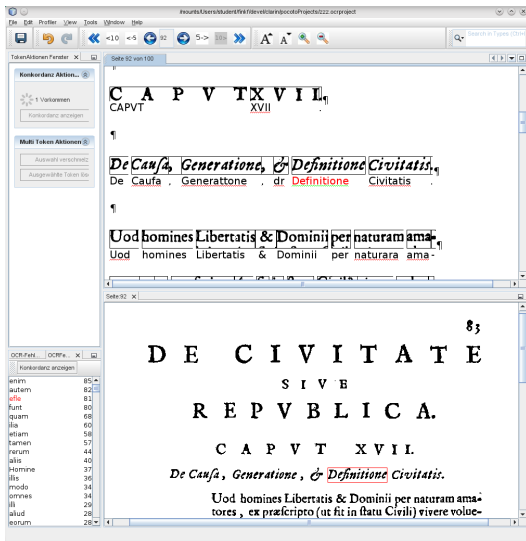
- PoCoTo supports the correction of tokens.
- Split tokens (Splits) can be merged together.
- Merged tokens (Merges) can be split.



- *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.

The screenshot shows the OC4J (Open Corpus J) software interface. The main window displays a Latin text document with the following content:

Regem, Proceres, & Coetum Communium, causa fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura praedicta inseparabilia esse non videant; & publice agniti sunt simul atque redierunt in Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus.

Quoniam autem lura haec Summae Potestati essentialia & inseparabilia sunt, sequitur, ut quibuscunque Verbis separari & aliis concedi videantur, nisi Potestati Summae simul & expressis verbis renunciatur sit, concessionem nullam; esse, sed concessa omnia, Summa Potestate, id est Personae Civitatis retenta, inseparabiliter redire.

Cum ergo Autoritas haec ingens indivisibilis sit, & habenti Summae

The interface includes a sidebar on the left with a search bar, a list of tokens, and a table of token counts. The table shows the following data:

Token	Count
enim	85
autem	82
esse	81
sunt	80
quam	68
ita	60
etiam	58
tamen	57
eorum	44
alio	40
homine	37
illis	36
modo	34
omnes	34
illi	29
aliud	28
eorum	28

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

The screenshot shows the OC4J editor interface. The main text area displays a Latin passage from Cicero's *De Re Publica*, Book 1, Chapter 18. The text is: "Regem, Proceres, & Coetum Communium, causa fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura praedicta inseparabilia esse non videant; & publice agniti sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus." The text is tokenized with red brackets. On the left, there are two panels: "Konkordanz Aktionen..." and "Multi Token Aktionen...". The "Multi Token Aktionen..." panel shows a list of tokens and their positions. The "Konkordanz Aktionen..." panel shows a list of tokens and their positions. The "Multi Token Aktionen..." panel shows a list of tokens and their positions. The "Konkordanz Aktionen..." panel shows a list of tokens and their positions.

Regem, Proceres, & Coetum Communium, causa fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura praedicta inseparabilia esse non videant; & publice agniti sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus.

Cap. 18. De Civitate. 91

Regem, Proceres, & Coetum Communium, causa fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura praedicta inseparabilia esse non videant; & publice agniti sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus.

Quoniam autem lura haec Summae Potestati essentialia & inseparabilia sunt, sequitur, ut quibuscunque Verbis separari & aliis concedi videantur, nisi Potestati Summae simul & expressis verbis renunciatur sit, concessionem nullam; esse, sed concessa omnia, Summa Potestate, id est Personae Civitatis retenta, inseparabiliter redire.

Cum ergo Autoritas haec ingens indivisibilis sit, & habenti Summae

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

The screenshot shows the PoCoTo Workshop interface. The main window displays a Latin text from a manuscript, with tokens highlighted in red. The text is: "Regem, Proceres, & Coetum Communium, caula fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, & Theologicis, quibus tamen populus ita nunc de lure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui lura praedicta inseparabilia esse non videant; & publice agniti sunt simul atque redierunt in Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus." The interface includes a sidebar on the left with a list of tokens and their frequencies, and a top menu bar with options like File, Edit, Profile, View, Tools, Window, and Help.

Tokenization actions:

- Konkordanz Aktionen...
- 259 Vorkommen
- Konkordanz anzeigen
- Multi Token Aktionen...
- Auswahl verschönern
- Ausgewählte Token löschen

Token list (left sidebar):

Token	Count
enim	85
autem	82
et	81
sunt	80
quam	68
ita	60
etiam	58
tamen	57
eorum	44
alio	40
Homine	37
illis	36
modo	34
omnes	34
illi	29
aliud	28
eorum	28

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

The screenshot shows the OCCToTol software interface. The main window displays a Latin text with tokens highlighted in red and blue. The text is: "funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra communem hostem, neque contra, injurias alter alterius. Dissidentes enim inter se de Virium usu non, injurias alter alterius. Dissidentes enim inter se de Virium usu non, sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum, fibi mutuo auxiliaturi funt, fed oppofitis confiliis vires ad nihilum, reducturi. Vnde non modo à communi hoste facile superantur, fed reduduri. Vnde non modo & communi hoste facile superantur, fed etiam de commodis propriis inter se Bello certaturi sunt. Siquidem, non certo, fed cum viribus hostium comparato determinatur, ut major sit quam ut excessus tanti ei tam conspicui momenti ad Bellum finendum sit, ut hostis ad aggrediendum provocetur. Sit autem multitudo quantacunque, si tamen actiones eorum Iudicii & Arbitrii multorum gubernentur, nullam inde expectare possunt securitatem, neque contra communem hostem, neque contra injurias alter alterius. Dissidentes enim inter se de Virium usu non sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum reducturi. Vnde non modo à communi hoste facile superantur, sed etiam de commodis propriis inter se Bello certaturi sunt. Siquidem enim hominum numerus magnus, sine Potentia communi quæ posset omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas confectire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimine Civili omnino opus non esset, victis scilicet hominibus in Pace, & sine Dorminis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo regum & determinato tempore, ut in uno

On the left side, there is a sidebar with the following sections:

- Konkordanz Aktionen...**
 - 21 Varianten
 - Konkordanz anzeigen
- Multi Token Aktionen**
 - Auswahl verschmälern
 - Ausgewählte Token löschen
- OCCToTol - OCCToTol**
 - Konkordanz anzeigen

Below the sidebar, there is a list of tokens with their frequency:

Token	Frequenz
enim	85
autem	82
esse	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
aliis	40
Homine	37
illis	36
modo	34
omnes	34
illi	29
aliud	28
eorum	28

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

The screenshot shows the PoCoTo software interface. The main window displays a Latin text with tokens highlighted in different colors (blue, red, green, yellow). The text is: "funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra communem hostem, neque contra, injurias alter alterius. Dissidentes enim inter se de Virium usu non, injurias alter alterius. Dissidentes enim inter se de Virium usu non, sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum, fibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum, reducturi. Vnde non modo à communi hoste facile superantur, reducturi. Vnde non modo à communi hoste facile superantur, etiam de commodis propriis inter se Bello certaturi sunt. Siquidem, etiam de commodis propriis inter se Bello certaturi sunt. Siquidem, non certo, fed cum viribus hostium comparato determinatur, ut major sit quam ut excessus tanti ei tam conspicui momenti ad Bellum finendum sit, ut hostis ad aggrediendum provocetur. Sit autem multitudo quantacunque, si tamen actiones eorum Iudicii & Arbitrii multorum gubernentur, nullam inde expectare possunt securitatem, neque contra communem hostem, neque contra injurias alter alterius. Dissidentes enim inter se de Virium usu non sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum reducturi. Vnde non modo à communi hoste facile superantur, sed etiam de commodis propriis inter se Bello certaturi sunt. Siquidem enim hominum numerus magnus, sine Potentia communi quæ posset omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas confectire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimine Civili omnino opus non esset, victuris scilicet hominibus in Pace, & sine Dorminis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo regum & determinato tempore, ut in uno

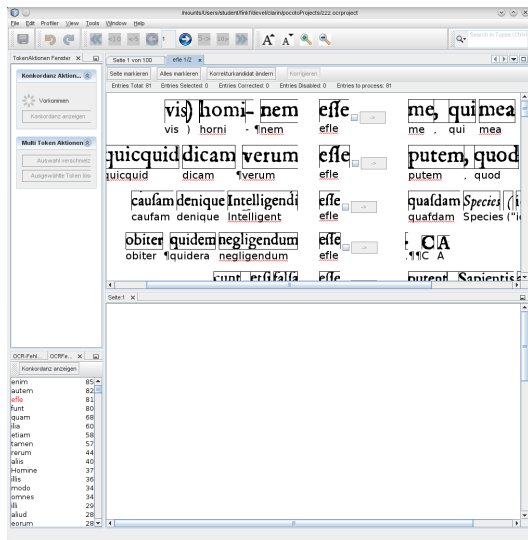
On the left, there is a sidebar with a search bar and a list of tokens. The list shows the frequency of various tokens in the text, such as "enim" (85), "autem" (82), "esse" (81), "funt" (80), "quam" (68), "ita" (60), "etiam" (58), "tamen" (57), "rerum" (44), "aliis" (40), "homine" (37), "illis" (36), "modo" (34), "omnes" (34), "illi" (29), "aliud" (28), and "eorum" (28).

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

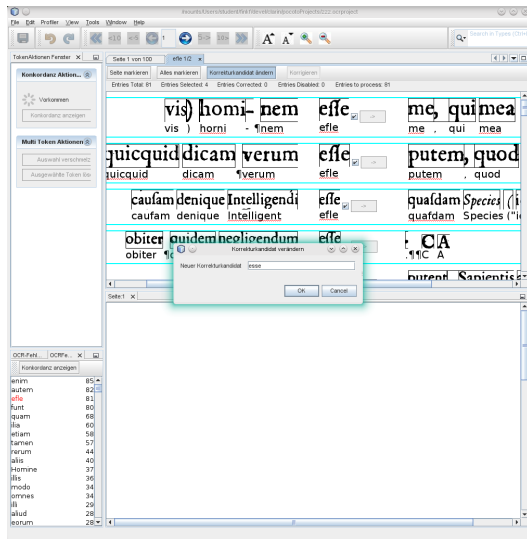
The screenshot shows the OcrFeit application interface. The main window displays a Latin text with tokens highlighted in different colors (blue, red, green) to indicate corrections or splits. The text is: "funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra communem hostem, neque contra, injurias alter alterius. Dissidentes enim inter se de Virium usu non, injurias alter alterius. Dissidentes enim inter se de Virium usu non, sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum, fibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum, reducturi. Vnde non modo à communi hoste facile superantur, fed reduduri. Vnde non modo & communi hoste facile superantur, fed etiam de commodis propriis inter se Bello certaturi sunt. Siquidem, non certo, fed cum viribus hostium comparato determinatur, ut major sit quam ut excessus tanti ei tam conspicui momenti ad Bellum finendum sit, ut hostis ad aggrediendum provocetur. Sit autem multitudo quantacunque, si tamen actiones eorum Iudicii & Arbitrii multorum gubernentur, nullam inde expectare possunt securitatem, neque contra communem hostem, neque contra injurias alter alterius. Dissidentes enim inter se de Virium usu non sibi mutuo auxiliaturi sunt, fed oppofitis confiliis vires ad nihilum reducturi. Vnde non modo à communi hoste facile superantur, sed etiam de commodis propriis inter se Bello certaturi sunt. Siquidem enim hominum numerus magnus, sine Potentia communi quæ posset omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas confectire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimine Civili omnino opus non esset, victis scilicet hominibus in Pace, & sine Dorminis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo regum & determinato tempore, ut in uno

On the left side, there is a sidebar with a search bar and a list of tokens. The list shows the frequency of various tokens in the text, such as "enim" (85), "autem" (82), "esse" (81), "funt" (80), "quam" (68), "ita" (60), "etiam" (58), "tamen" (57), "eorum" (44), "illis" (40), "homine" (37), "illis" (36), "modo" (34), "omnes" (34), "illi" (29), "aliud" (28), and "eorum" (28).

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



- 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.

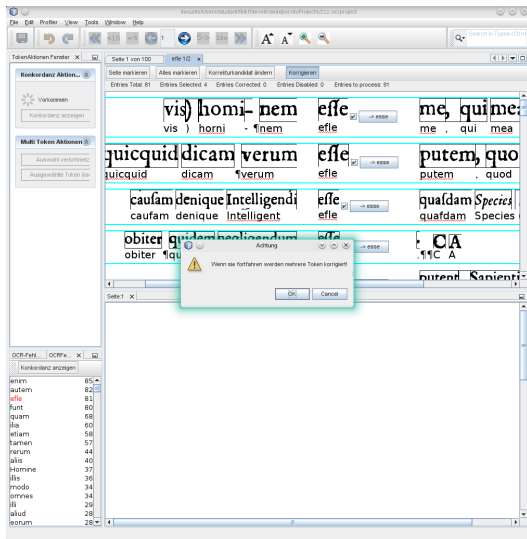


- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

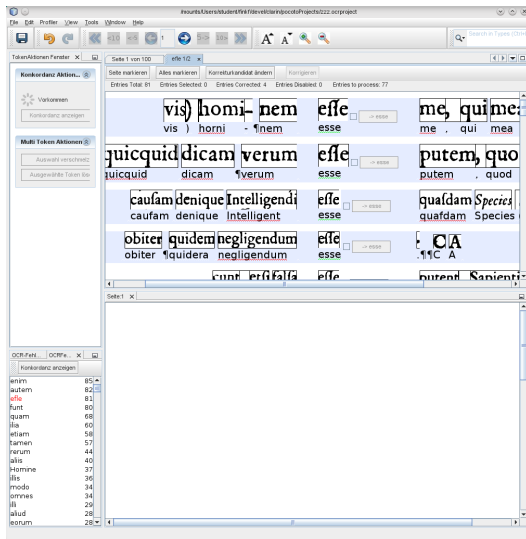
The screenshot shows the PoCoTo software interface. The main window displays a concordance view for the word 'esse'. The concordances are listed in a table with columns for the word, its frequency, and a list of concordances. The interface includes a menu bar, a toolbar, and a sidebar with various options.

Word	Frequency	Concordances
enim	85	
autem	82	
esse	81	
fuit	80	
quam	68	
ita	60	
etiam	58	
tamen	57	
perum	44	
alio	40	
homine	37	
illis	36	
modo	34	
omnes	34	
illi	29	
alud	28	
eorum	28	

- 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.




- 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.



- 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.

- 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.

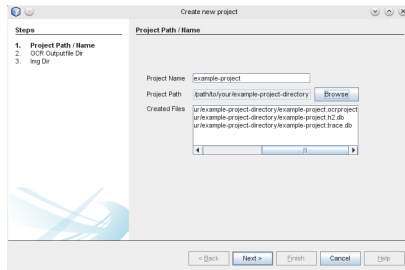
- PoCoTo has an automatic updating mechanism.
- PoCoTo can be kept up to date without having to install it again.
- Whenever PoCoTo recognizes a newer version, it shows an *updates available*  button in its lower right corner.
- To check for updates go to Help->Check for updates.
- To control the update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

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

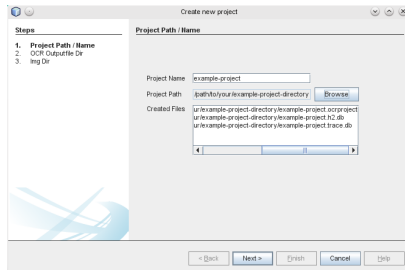
- ① The character based Abbyy XML format.
- ② The HOCR file format.
- ③ Ocropus-Directories.

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 or Ocropus, PoCoTo is not able to generate such an error frequency list for you.



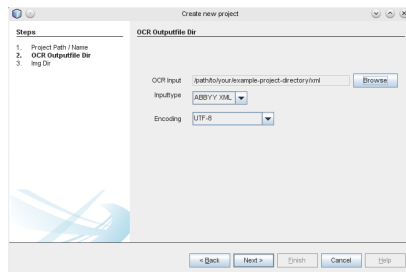
You can create new projects using the project wizard. Click to File->New Project and the first frame of the project wizard open.

- ➊ Insert a name and a path for your project. Click Next.
- ➋ Insert the path of your folder, that contains the XML files and select the type of your XML files. Click Next.
- ➌ Select the path to the folder, that contains your image files. Click Finish.



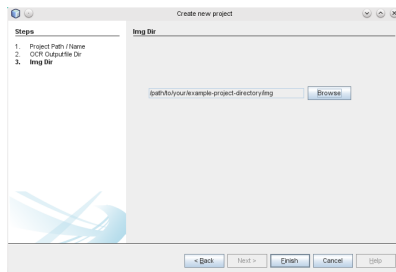
You can create new projects using the project wizard. Click to File->New Project and the first frame of the project wizard open.

- ➊ Insert a name and a path for your project. Click Next.
- ➋ Insert the path of your folder, that contains the XML files and select the type of your XML files. Click Next.
- ➌ Select the path to the folder, that contains your image files. Click Finish.



You can create new projects using the project wizard. Click to File->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 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.

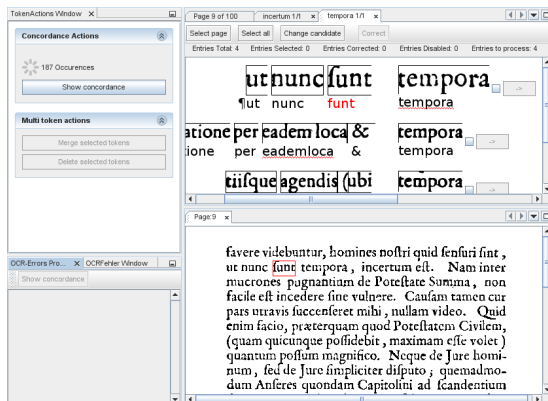


You can create new projects using the project wizard. Click to File->New Project and the first frame of the project wizard open.

- ❶ Insert a name and a path for your project. Click Next.
- ❷ Insert the path of your folder, that contains the XML files and select the type of your XML files. Click Next.
- ❸ 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.



- 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.

Thanks for your attention!