

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In the recent years a lot of historical documents have been scanned and OCR'ed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents has constantly improved.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

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 beherbergen, 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 Fangereyen 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 beherbergen, 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

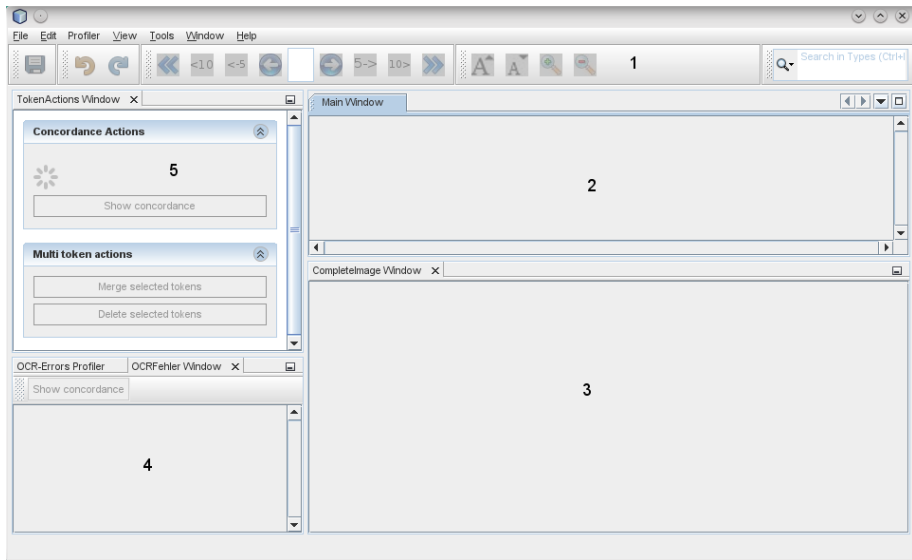
- PoCoTo is a tool for the interactive post-correction of OCR'ed text:
- It was developed as part of the EU founded project IMPACT.
- It is open source and hosted on [github](#).
- It contains linguistic and visual aids to support the post-correction.
- It contains aids to automatically correct systematic 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.

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

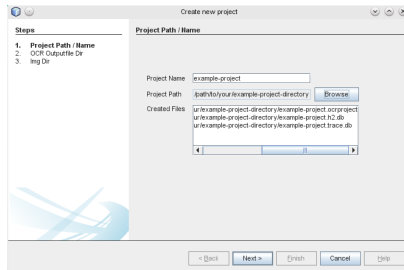
- ➊ The menu area contains various commands for navigation and project maintenance.
- ➋ The main view area shows tokens and offers the main correction possibilities.
- ➌ The complete image area displays the page of the current active (selected) token.
- ➍ The error area lists error frequency lists of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

- 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.

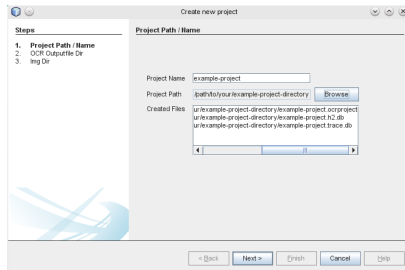
PoCoTo understands three different XML file formats, that you can use to create new projects.

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

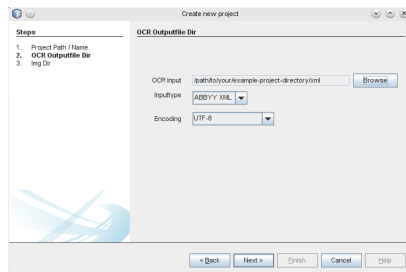
PoCoTo uses the information of the ABBYYX-XML file format directly to mark *suspicious* words. It will generate an error frequency list for you. If you use the hOCR format or Ocropus, PoCoTo is not able to generate such an error frequency list for you.



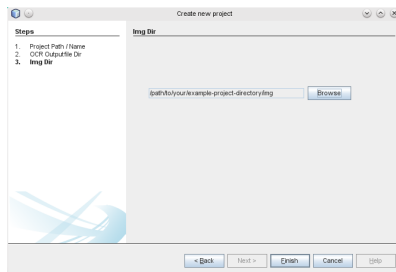
- 1 You can create new projects using the project wizard. Click to File->New Project and the first frame of the project wizard open.



- 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.



- ❶ You can create new projects using the project wizard. Click to File->New Project and the first frame of the project wizard open.
- ❷ Insert a name and a path for your project. Click Next.
- ❸ Insert the path of your folder, that contains the XML files and select the type of your XML files. Click Next.
- ❹ 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 screenshot shows the PoCoTo software interface. The top window displays the text "Cap. 18. De Civitate. 91" with tokens highlighted in red and blue. The bottom window shows the original text with the active token highlighted. The left sidebar contains a list of tokens and their frequencies.

**Konkordanz Aktion...**

1 Variationen

Konkordanz anzeigen

**Multi Token Aktionen**

Auswahl verschönern

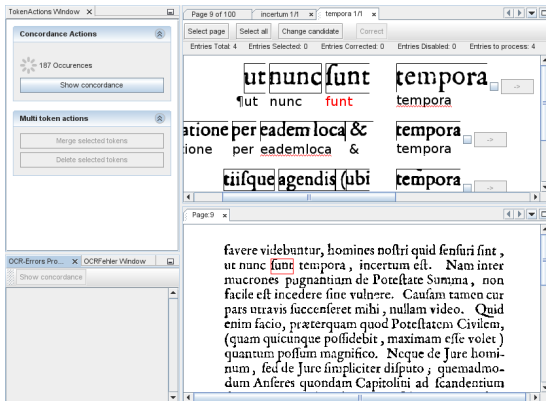
Ausgewählte Token löschen

OCRF... OCRF... x

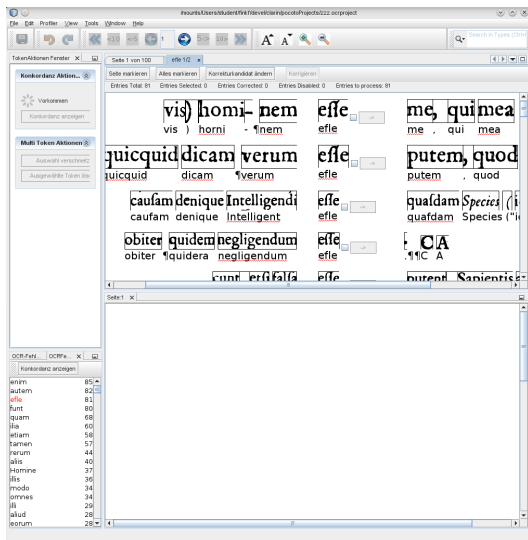
Konkordanz anzeigen

enim	85
autem	82
et	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.

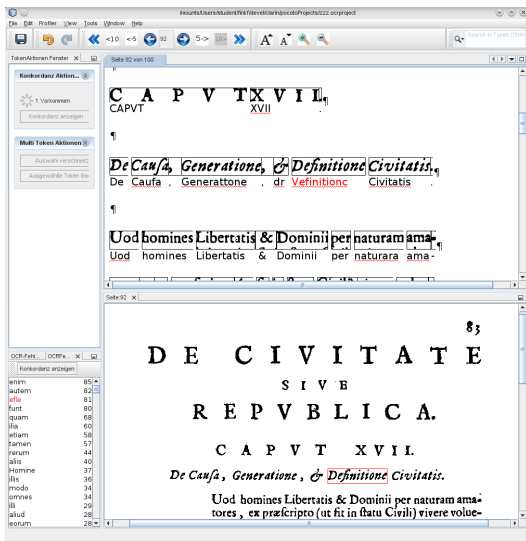


- 1 You can activate any token and if there exists any similar other token you can click to the Show concordance view button in the token action area
- 2 You can click on any entry in the two error frequency lists in the error area.

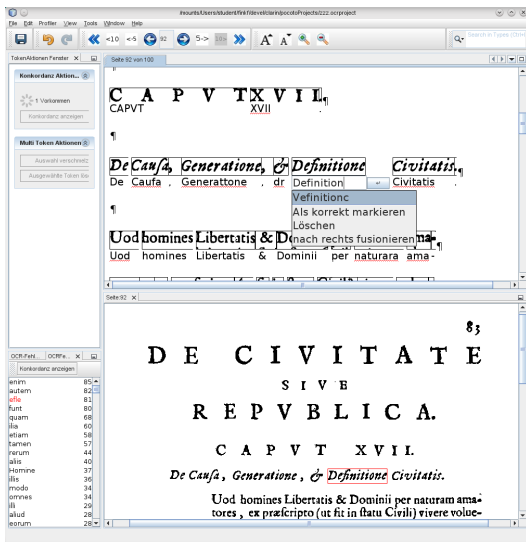


- Common error patterns in the document can be examined using the so-called concordance view.
- The concordance view lists similar words and patterns encountered in the document.
- Consistent error patterns can be easily selected and corrected in one step.

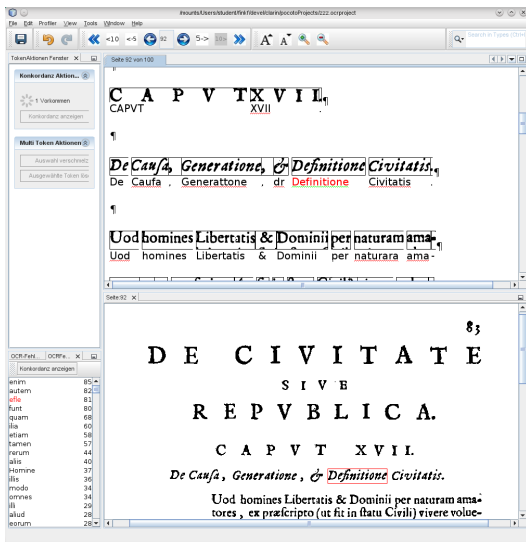
- PoCoTo automatically tokenizes the document on whitespace and punctuation.
- Each token can be examined in its page image.
- PoCoTo supports the correction of single tokens.
- Multiple occurrences of errors and (error patterns) can be corrected with concordance views.
- Split tokens (Splits) can be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



The screenshot displays the PoCoTo software interface. The main window shows a Latin text with tokens highlighted in red and blue. The text is: "Regem, Proceres, & Coetum Communium, caula fuit Belli quod sequutum est Civilis; etiam disputaciones 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 interface includes a sidebar with "Konkordanz anzeigen" and "Multi Token Aktionen" buttons. A table on the left lists tokens and their positions. The bottom window shows the text "Cap. 18. De Civitate. 91" and the same Latin text.

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

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

The screenshot displays the PoCoTo software interface. The main window shows a Latin text with various tokens highlighted in red and blue. 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 sunt simul atque redierit 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.

Token Actions Fenster: Seite 100 von 100

Konkordanz anzeigen

Multi Token Aktionen

Auswahl verschönern

Ausgewählte Token löschen

OCRFest - OCRFest - x

Konkordanz anzeigen

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

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 sunt 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 displays the PoCoTo software interface. The main window shows 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 redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus." The text is split into tokens, and some tokens are merged or split. The left sidebar shows a list of tokens and their frequency, and the bottom panel shows a list of tokens and their frequency.

Tokenization window: **Konkordanz Aktions...** (259 Vorkommen, Konkordanz anzeigen), **Multi Token Aktionen** (Auswahl verschönern, Ausgewählte Token löschen).

Text window: **Seite 100 von 100**. The text is split into tokens, and some tokens are merged or split. 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 redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus."

Bottom panel: **Cap. 18. De Civitate. 91**. The text is split into tokens, and some tokens are merged or split. 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 redierit Pax, & quamdiu calamitatem praeteritarum meminerint; sed non diutius, nisi melius erudiat populus. Quoniam autem lura haec Summæ Potestati essentialia & inseparabilia sunt, sequitur, ut quibuscunque Verbis separari & aliis concedi videantur, nisi Potestati Summæ simul & expressis verbis renunciatur sit, concessionem nullam; esse, sed concessa omnia, Summæ Potestate, id est Personæ Civitatis retentâ, inseparabiliter redire. Cum ergo Autoritas hæc ingens indivisibilis sit, & habenti Summæ Potestate..."

- 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 Latin text with token boundaries highlighted. The text is:
   
funt securitatem, neque contra communem hostem, neque contra,
   
funt securitatem, neque contra communem hostem, neque contra
   
inurias alter alterius. Dissidentes enim inter se de Virium usu non,
   
inurias 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
   
reducturi. Vnde non modo à communi hoste facile superantur, fed
   
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
   
fit quam ut excessus tanti ei tam conspicui momenti ad Bellum fi-
   
niendum fit, ut hostis ad aggrediendum provocetur.
   
Sic autem multitudo quantacunque, si tamen actiones eorum Iudi-
   
ciis & Arbitriis multorum gubernentur, nullam inde expectare pos-
   
sunt securitatem, neque contra communem hostem, neque contra
   
inurias 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, fed
   
etiam de commodis propriis inter se Bello certaturi sunt. Siquidem
   
enim hominum numerus magnus, sine Potentia communi quæ pos-
   
set omnes cogere, in Æquitate cæteraque Leges Naturæ obser-
   
vandas confectare 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) fufficit, ut
   
gubernentur non certo regum & determinato tempore, ut in uno

On the left, there is a sidebar with a list of tokens and their frequencies:

Token	Frequency
enim	85
autem	82
ita	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
illis	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, victoris scilicet hominibus in Pace, & sine Dorminis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo regum & determinato tempore, ut in uno

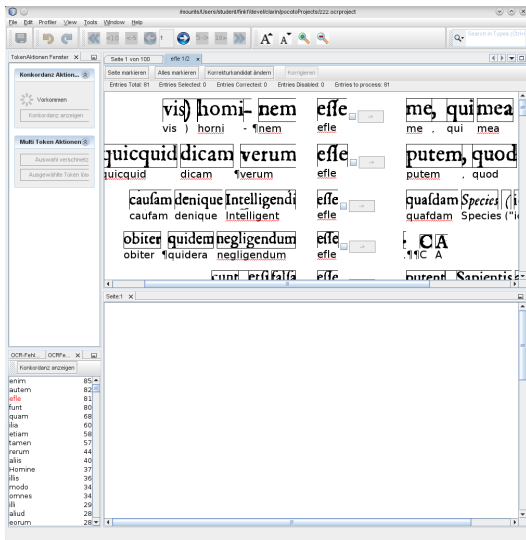
On the left, there is a sidebar with a search bar and a list of tokens. The list shows the frequency of various tokens in the text, such as "enim" (85), "autem" (82), "ita" (81), "funt" (80), "quam" (68), "ita" (60), "etiam" (58), "tamen" (57), "rerum" (44), "aliis" (40), "homine" (37), "illis" (36),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

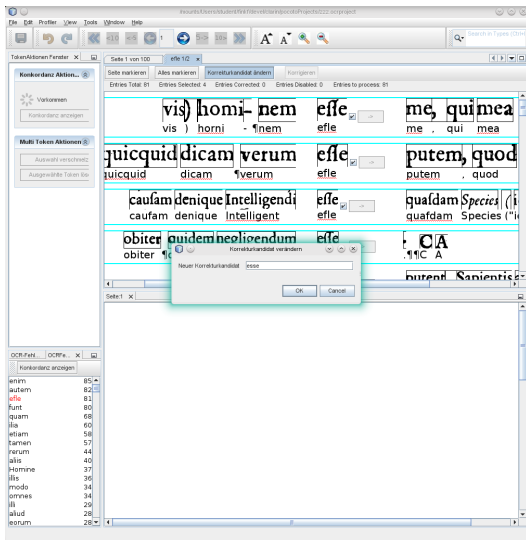
The screenshot displays the PoCoTo software interface. The main window shows a Latin text snippet with tokens highlighted in different colors (blue, red, green) to indicate corrections or splits. The text is: "funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra 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 confecture 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 includes: "enim", "autem", "ita", "funt", "quam", "ita", "etiam", "tamen", "rerum", "illis", "Homine", "modo", "omnes", "illi", "aliud", "eorum". The list is numbered 85 to 28. Below the list, there is a section titled "OCCURRENCE" with a table showing the frequency of each token. The table has two columns: "Token" and "Frequency". The tokens listed are: "enim", "autem", "ita", "funt", "quam", "ita", "etiam", "tamen", "rerum", "illis", "Homine", "modo", "omnes", "illi", "aliud", "eorum". The frequencies are: 85, 82, 81, 80, 68, 60, 58, 57, 44, 40, 37, 36, 34, 34, 29, 28, 28.

- 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.



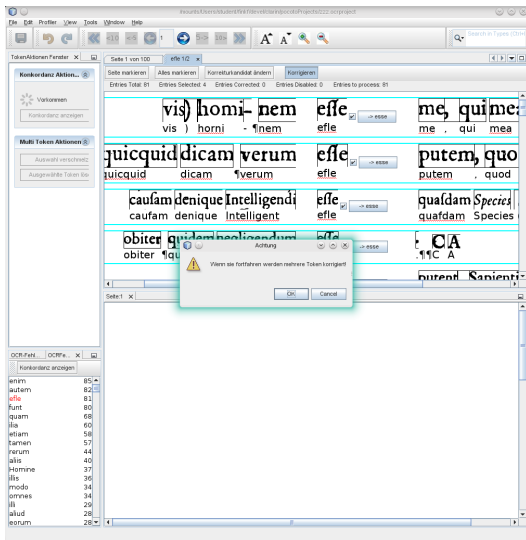
The screenshot shows the PoCoTo software interface. The main window displays a concordance view with the following text segments:

vis) homi- nem	effe	me, quime:
vis ) horni - finem	effe	me , qui mea
quicquid dicam verum	effe	putem, quo
quicquid dicam verum	effe	putem , quod
causam denique intelligendi	effe	quafdam Species
causam denique intelligent	effe	quafdam Species
obiter quidem negligendum	effe	. CA A
obiter quidera negligendum	effe	. CA A
cunt et falfa	effe	purend Sanienti

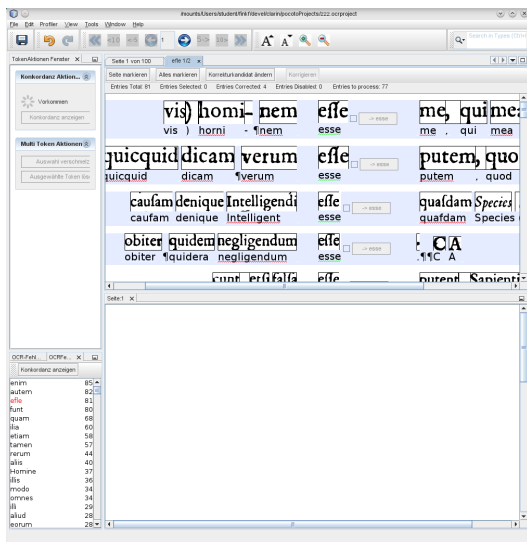
The left sidebar contains a 'Konkordanz Aktions...' panel with a list of words and their frequencies:

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

- 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!