

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In the recent years a lot of historical documents have been scanned and OCRed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents is constantly improved.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Text recognition on historical documents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

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 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 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, j. E. bey

Männlein~~c~~her 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.

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

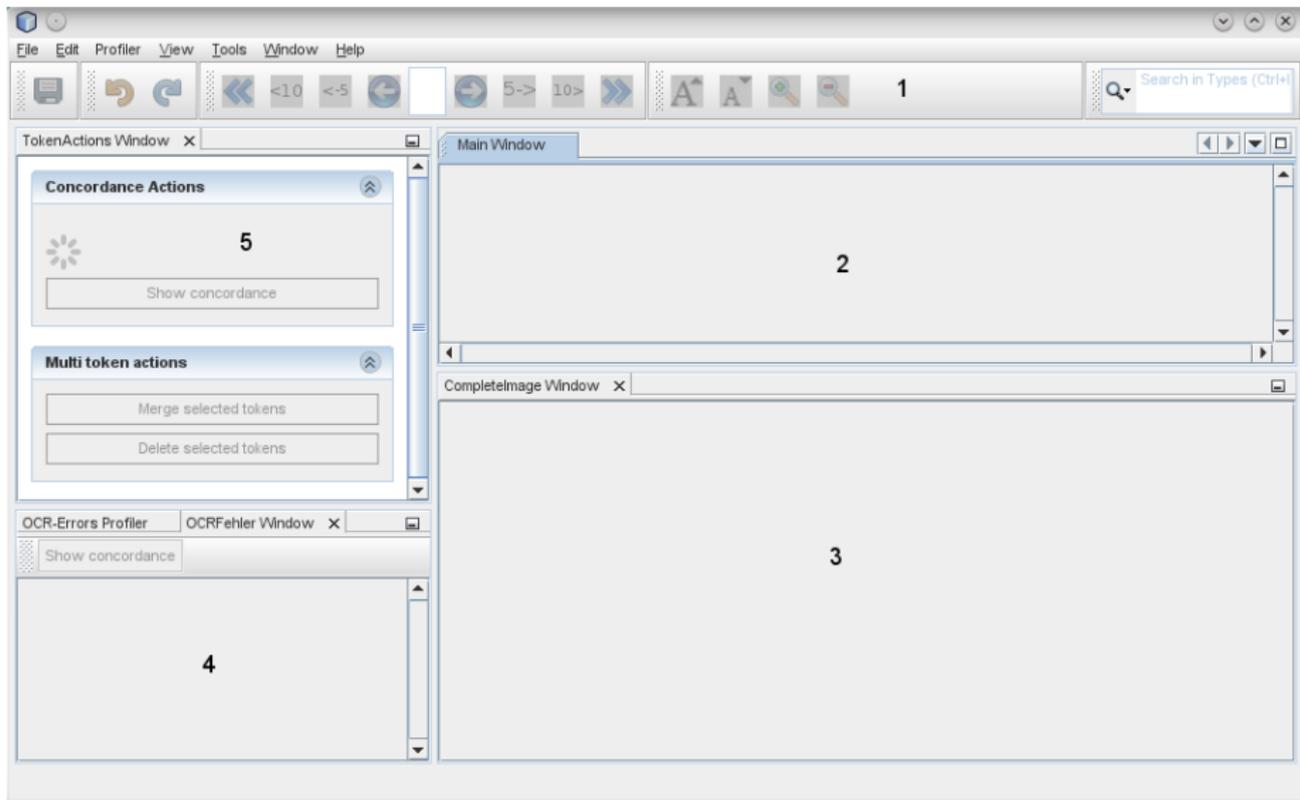
PoCoTo – The PostCorrectionTool

Overview

Improving Access to Text **IMPACT**

- PoCoTo is a tool for the interactive postcorrection of OCR'd 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 and helps you to split and merge tokens.

The visual UI

The screenshot displays the PoCoTo application window. The main text area shows a Latin document with the following text:

Cap. 18. De Civitate. 91
 Cap. 18. De Civitate. 91
 Regem, Proceres, & Cœtum Communium, causa fuit Belli quod
 Regem, Proceres, & Coetum Communium, caula fuit Belli quod
 sequutum est Civilis; etiam disputationes de quaestionibus Politicis,
 fequutum est Civilis; etiam difputationes de quaestionibus Politicis
 & Theologicis, quibus tamen populus ita nunc de Iure Regio eru
 & Theologicis, quibus tamen populus ita nunc de Iure Regioeru

Below the main text, a smaller version of the same text is shown with error frequencies:

Cap. 18. De Civitate. 91
 Regem, Proceres, & Cœtum Communium, causa fuit Belli quod
 sequutum est Civilis; etiam disputationes de quaestionibus Politicis
 & Theologicis, quibus tamen populus ita nunc de Iure Regio eru
 ditus est, ut in Anglia pauci (puto) nunc sint qui Iura prædicta inseparabilia esse non videant; & publice agniti sunt simul atque redierit Pax, & quamdiu calamitatem præteritarum meminerint; sed non diutius, nisi melius erudiatur populus.
 Quoniam autem Iura hæc Summæ Potestati essentialia & inseparabilia sunt, sequitur, ut quibuscunque Verbis separari & aliis concedi videantur, nisi Potestati Summæ simul & expressis verbis renunciatum sit, concessionem nullam; esse, sed concessa omnia, Summæ Potestate, id est Personæ Civitatis retentâ, inseparabiliter redire.
 Cum ergo Autoritas hæc ingens indivisibilis sit, & habenti Sum-

The interface includes a sidebar on the left with a search bar, a list of actions (Konkordanz Aktionen), and a list of OCR frequencies (OCR-Freq.). The OCR-Freq. list shows the following items:

OCR-Freq.	OCR-Freq.
enim	85
autem	82
esse	81
sunt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
alio	40
Homine	37
illis	36
modo	34
omnes	34
in	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.

Interactive postcorrection: correcting single tokens

The screenshot displays the PoCoTo application window. At the top, the title bar reads 'AnuritaUsers\student\F8\Filevelcarin\pocoto\projects\222_ocrproject'. The main window shows two pages of OCR output. The top page, 'Seite 92 von 100', contains the title 'CAPVT XVII' and the text 'De Caufa, Generatione, & Definitione Civitatis.' Below this, a 'Multi Token Aktionen' panel offers options like 'Auswahl verschieben' and 'Ausgewählte Token löschen'. The text 'De Caufa, Generatione, & Definitione Civitatis.' is shown with individual tokens highlighted in boxes, and 'Definitione' is highlighted in red. The bottom page, 'Seite 93', shows the title 'DE CIVITATE SIVE REPUBLICA. CAPVT XVII.' and the text 'Uod homines Libertatis & Domini per naturam amatores, ex præcripto (ut fit in ftatu Civili) vivere volue-'. A 'Konkordanz anzeigen' panel on the left lists words and their positions, such as 'enim' at 85 and 'autem' at 82.

- 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.

Interactive postcorrection: correcting single tokens

The screenshot shows the PoCoTo interface with two windows. The top window displays the original Latin text: "CAPVT XVII, De Cauſa, Generatione, & Definitione Civitatis, Uod homines Libertatis & Dominii per naturara amatores, ex præſcripto (ut fit in ſtatu Civili) vivere volue-". Several words are highlighted in red, indicating they are suspicious. A context menu is open over the word "Definitione", showing options: "Vefinitionc", "Als korrekt markieren", and "Löschen". The bottom window shows the corrected text: "DE CIVITATE SIVB REPUBLICA. CAPVT XVII. De Cauſa, Generatione, & Definitione Civitatis. Uod homines Libertatis & Dominii per naturam amatores, ex præſcripto (ut fit in ſtatu Civili) vivere volue-". The word "Definitione" has been manually corrected to "Definition".

- 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.

Interactive postcorrection: correcting single tokens

The screenshot displays the PoCoTo application window. The main area shows OCR output for a Latin document. At the top, the title 'CAPVT XVII' is shown with individual characters in boxes. Below, the text 'De Cauſa, Generatione, & Definitione Civitatis.' is displayed with red boxes around 'Definitione' and 'Civitat'. A second line of text, 'Uod homines Libertatis & Domini per naturam ama...', also has red boxes around 'ama'. On the left side, there are two panels: 'Konkordanz Aktionen...' with a '1 Vorkommen' indicator, and 'Multi Token Aktionen...' with 'Auswahl verschoben' and 'Ausgewählte Token löschen' buttons. At the bottom left, an 'OCR-Fehl...' list shows a table of errors and their positions.

OCR-Fehl...	OCRF...
enim	85
autem	82
elle	81
funt	80
quam	68
ita	69
etiam	58
tamen	57
rerum	44
alis	40
Homine	37
ilis	36
modo	34
omnes	34
is	29
aliud	28
eorum	28

- 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.

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo interface with a Latin text editor. The main window displays the text: "Regem, Procere, & Coetum Communium, caula fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, sequutum est Civilis; etiam disputationes de quaestionibus Politicis & Theologicis, quibus tamen populus ita nunc de Iure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui Iura praedicta inseparabilia esse non videant; & publice agniture sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminert; sed non diutius, nisi melius erudiatu populus." The text is annotated with red boxes around individual tokens, and some tokens are split into smaller segments. A sidebar on the left shows a list of tokens and their positions, with "elle" highlighted in red. The interface includes a menu bar, a toolbar, and a search bar.

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

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo interface with a Latin text editor. The main window displays the text: "Regem, Procere, & Coetum Communium, caula fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis, fequutum est Civilis; etiam disputationes de quaestionibus Politicis & Theologicis, quibus tamen populus ita nunc de Iure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui Iura praedicta inseparabilia esse non videant; & publice agniture sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminere; sed non diutius, nisi melius erudiatu populus." The text is annotated with red boxes around individual tokens, and some tokens are split into smaller segments. A sidebar on the left shows a list of tokens with their corresponding line numbers, and a search bar at the top right.

Cap. 18. *De Civitate.* 91
 Regem, Procere, & Coetum Communium, caula fuit Belli quod sequutum est Civilis; etiam disputationes de quaestionibus Politicis & Theologicis, quibus tamen populus ita nunc de Iure Regio eruditus est, ut in Anglia pauci (puto) nunc sint qui Iura praedicta inseparabilia esse non videant; & publice agniture sint simul atque redierit Pax, & quamdiu calamitatem praeteritarum meminere; sed non diutius, nisi melius erudiatu populus.
 Quoniam autem Iura haec Summae Potestati essentialia & inseparabilia sunt, sequitur, ut quibuscunque Verbis separari & aliis concedi videantur, nisi Potestati Summae simul & expressis verbis renunciatum sit, concessionem nullam; esse, sed concessa omnia, Summa Potestate, id est Persona Civitatis retenta, inseparabiliter redire.
 Cum ergo Autoritas haec ingens indivisibilis sit, & habenti Sum-

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

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo interface with a Latin text document. The text is displayed in two views: a top view showing individual tokens with red boxes around them, and a bottom view showing the text as a continuous paragraph. The text is from a historical document, likely a constitution or law, discussing the powers of the crown and the rights of the people.

Top view (tokens split):

Regem , Proceres , & Coetum Communium , caula fuit Belli quod
 fequutum eft Civilis ; etiam difputationes de quaeflionibus Politicis,
 fequutum eft Civilis ; etiam difputationes de quaeflionibus Politicis
 & Theologicis , quibus tamen populus ita nunc de Iure Regio eru-
 & Theologicis , quibus tamen populus ita nunc de Iure Regioeru
 ditus eft , ut in Anglia pauci (puto) nunc fint qui Iura praedi&a infe-
 ditus eft , ut in Anglia pauci (puto) nunc sint qui Iura praedi&a infe-
 parabilia effe non videant ; & publice agniture fint fimul atque redie-
 parabilia ell'e non videant : Sc publice agniture lint firaul atque redie-

Bottom view (tokens merged):

Cap. 18. *De Civitate.* 91
 Regem , Proceres , & Coetum Communium , caula fuit Belli quod
 fequutum eft Civilis ; etiam difputationes de quaeflionibus Politicis
 & Theologicis , quibus tamen populus ita nunc de Iure Regio eru-
 ditus eft , ut in Anglia pauci (puto) nunc fint qui Iura praedi&a infe-
 parabilia effe non videant ; & publice agniture fint fimul atque redie-
 rit Pax , & quamdiu calamitatem praeteritarum meminerint ; fed non
 diutius , nifi melius erudiatuſ populus.
 Quoniam autem Iura haec Summae Poteflatis effentialia & infe-
 parabilia funt , fequitur , ut quibufcunqve Verbis ſeparari & aliis con-
 cedi videantur , niſi Poteflatis Summae ſimul & exprefſis verbis renun-
 ciatum ſit , conceffionem nullam ; effe , fed conceſſa omnia , Summa
 Poteflatis , id eſt Perſona Civitatis retenta , infeſeparabiliter redire.
 Cum ergo Authoritas haec ingens indiviſibilis ſit , & habenti Sum-

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

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo software interface. The main window displays a Latin text with several tokens highlighted in red and split into smaller segments. The text is: "funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra communem hostem, neque contra, injurias alter alterius. Diffidenties enim inter se de Virium usu non, injurias alter alterius. Dididenties enim inter se de Virium usu non, sibi mutuo auxiliari funt, fed oppositis consiliis vires ad nihilum, reducturi. Vnde non modo à communi hoste facile superantur, fed reducendi. Vnde non modo & communi hoste facile superantur, fed etiam de commodis propriis inter se Bello certaturi funt. Siquidem, non certo, fed cum viribus hostium comparato determinatur, ut major sit quam ut excessus tanti ei tam conspicui momenti ad Bellum fiendum sit, ut hostis ad aggrediendum provocetur. Sit autem multitudo quantaunque, si tamen actiones eorum Iudicii & Arbitrii multorum gubernentur, nullam inde expectare possunt securitatem, neque contra communem hostem, neque contra injurias alter alterius. Diffidenties enim inter se de Virium usu non sibi mutuo auxiliari funt, fed oppositis consiliis vires ad nihilum reducturi. Vnde non modo à communi hoste facile superantur, fed etiam de commodis propriis inter se Bello certaturi funt. Siquidem enim hominum numerus magnus, sine Potentia communi quæ possit omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas consentire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimie Civili omnino opus non esset, victuris scilicet hominibus in Pace, & sine Dominis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo certum & determinatum tempore, ut in usu

On the left side, there is a "Konkordanz Aktion..." panel with a search bar and a list of tokens. The list shows the following tokens and their counts:

Token	Count
enim	85
autem	82
elle	81
funt	80
quam	68
ita	69
etiam	58
tamen	57
rerum	44
alis	40
Homine	37
illis	36
modo	34
omnes	34
in	29
aliud	28
eorum	28

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

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo interface with a Latin text. Tokens are split (e.g., 'funt securitatem, neque contra communem hostem, neque contra, funt securitatem, neque contra communem hostem, neque contra, injurias alter alterius. Diffidenties enim inter se de Virium usu non, injurias alter alterius. Diffidenties enim inter se de Virium usu non, sibi mutuo auxiliari funt, sed oppositis consiliis vires ad nihilum, reducturi. Vnde non modo à communi hoste facile superantur, reduci. Vnde non modo & communi hoste facile superantur, etiam de commodis propriis inter se Bello certaturi funt. Siquidem, Als korrekt markieren, Löschen, nach rechts fusionieren. non certo, sed cum viribus hostium comparato determinatur, ut major sit quam ut excessus tanti ei tam conspicui momenti ad Bellum fiendum sit, ut hostis ad aggrediendum provocetur. Sit autem multitudo quantaunque, si tamen actiones eorum Iudicii & Arbitrii multorum gubernentur, nullam inde expectare possunt securitatem, neque contra communem hostem, neque contra injurias alter alterius. Diffidenties enim inter se de Virium usu non sibi mutuo auxiliari funt, sed oppositis consiliis vires ad nihilum reducturi. Vnde non modo à communi hoste facile superantur, sed etiam de commodis propriis inter se Bello certaturi funt. Siquidem enim hominum numerus magnus, sine Potentia communi quæ possit omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas consentire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimie Civili omnino opus non esset, victuris (sicut hominibus in Pace, & sine Dominis. Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo certum & determinatum tempore, ut in usu').

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

Interactive postcorrection: splits and merges

The screenshot shows the PoCoTo software interface. The main window displays a Latin text with several 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. Diffidenties enim inter se de Virium usu non,

injurias alter alterius. Diffidenties enim inter se de Virium usu non,

sibi mutuo auxiliaturi sunt, fed oppositis consiliis vires ad nihilum,

fibi mutuo auxiliaturi sunt, fed oppositis consiliis vires ad nihilum,

reducturi. Vnde non modo à communi hoste facile superantur, fed

reduhuri. 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 fi-

niendum sit, ut hostis ad aggrediendum provocetur.

Sic 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. Diffidenties enim inter se de Virium usu non sibi mutuo auxiliaturi sunt, fed oppositis consiliis 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 possit omnes cogere, in Æquitatem cæteraque Leges Naturæ observandas consentire supponeretur, idem etiam de toto genere humano supponendum esset, itaque Regimie Civili omnino opus non esset, victuris scilicet hominibus in Pace, & sine Dominis.

Neque ad securitatem (quam perpetuam esse volunt) sufficit, ut gubernentur non certo certum & determinatum tempore, ut in usu

The interface also shows a toolbar at the top with various navigation and search icons. On the left side, there are panels for 'Konkordanz Aktions...' and 'Multi Token Aktions...'. At the bottom left, there is a list of tokens with their corresponding counts:

Token	Count
enim	85
autem	82
elle	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
alis	40
Homine	37
ilis	36
modo	34
omnes	34
is	29
aliud	28
eorum	28

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

Interactive postcorrection: concordances

The screenshot shows the PoCoTo interface with the following elements:

- Top Panel:** 'Konkordanz Aktions...' with a 'Vorankommen' button and 'Konkordanz anzeigen'.
- Multi Token Aktions...** with 'Auswahl verschieben' and 'Ausgewählte Token löschen'.
- Main View:** A grid of text fragments. The first row shows 'vis) homi- nem' and 'me, qui mea'. The second row shows 'quicquid dicam verum' and 'putem, quod'. The third row shows 'causam denique intelligendi' and 'quafdam Species ("i'. The fourth row shows 'obiter quidem negligendum' and 'CA'. The fifth row shows 'obiter quidam negligendum' and 'arent Sapientis'.
- Bottom Left Panel:** 'OCR-Feh... OCRFe...' with a 'Konkordanz anzeigen' button and a list of words and their counts:

enim	85
autem	82
effe	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
alio	40
Homine	37
illis	36
modo	34
omnes	34
in	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.

Interactive postcorrection: concordances

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

vis) homi- nem	effe	me, qui mea
vis) horni - nem	effe	me . qui mea
quicquid dicam verum	effe	putem, quod
quicquid dicam verum	effe	putem . quod
causam denique intelligendi	effe	qualdam Species
causam denique intelligendi	effe	quafdam Species ("i
obiter quidem neoliondum	effe	CA
obiter quidem neoliondum	effe	CA
		turendi Sanientis

A dialog box titled "Neuer Korrekturkandidat" is open, showing the word "esse" and buttons for "OK" and "Cancel".

The sidebar on the left shows a list of words and their frequencies:

enim	85
autem	82
effe	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
illis	40
Homine	37
illis	36
modo	34
omnes	34
in	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.

Interactive postcorrection: concordances

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

vis) homi- nem	effe	me, quime
vis) horni - nem	effe	me . qui mea
quicquid dicam verum	effe	putem, quo
quicquid dicam verum	effe	putem . quod
causam denique intelligendi	effe	quafdam Species
causam denique intelligendi	effe	quafdam Species
obiter quidem negligendum	effe	CA
obiter quidera negligendum	effe	.11C A
Sunt erifala	effe	putend Sanienti

The sidebar on the left shows a list of words and their frequencies:

enim	85
autem	82
effe	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
alio	40
Homine	37
illis	36
modo	34
omnes	34
in	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.

Interactive postcorrection: concordances

The screenshot shows the PoCoTo software interface. The main window displays a concordance view for the word 'esse'. The text is organized into rows, with the word 'esse' highlighted in blue. The concordance entries are:

- vis) homi- nem esse me, quime
- vis) horni - nem esse me . qui mea
- quicquid dicam verum esse putem, quo
- quicquid dicam verum esse putem . quod
- causam denique intelligendi esse quaedam Species
- causam denique intelligendi esse quaedam Species
- obiter quidem negligendum esse CA
- obiter quidam negligendum esse .11C A
- esse putend Sanienti

The sidebar on the left shows a list of words and their frequencies:

Word	Frequency
enim	85
autem	82
esse	81
funt	80
quam	68
ita	60
etiam	58
tamen	57
rerum	44
alio	40
Homine	37
illis	36
modo	34
omnes	34
in	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.

Interactive postcorrection: correction suggestions

The image shows a three-line snippet of a historical document with a correction menu overlaid on the second line. The original text is shown in a box above the corrected text. The word 'Zranckreich' is highlighted in red, and a context menu is open over it, listing correction options.

Original text (top row): cherung geben liessen/ daß Franckreich niemahls einige praeten

Corrected text (middle row): cherung geben liessen/ daß **Zranckreich** niemahls einige praeten - ¶

Original text (bottom row): tion auf sie formiren wolte; so wenig sie / würden

Corrected text (bottom row): tion auf sie kormiren wolte ; so wenig sie / würden ¶

Original text (third row): Franckreich / Engelland und andere benachbarte Staaten still ¶

Corrected text (third row): Franckreich / Engelland und andere benachbarte Staaten still ¶

Correction menu options:

- Als korrekt markieren
- Franckreich
- Frankreich
- Ranckreich
- Vranckreich
- nach rechts fusionieren
- Silbentrennung beheben

- 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 file¹.
- 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.

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

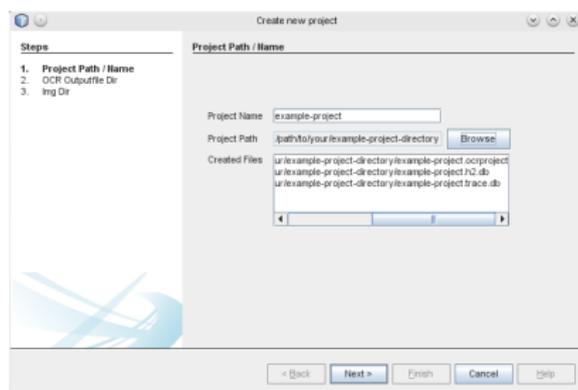
PoCoTo's file formats

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

- 1 The character based Abbyy XML format.
- 2 The HOOCR 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 HOOCR format, PoCoTo is not able to generate such an error frequency list for you.

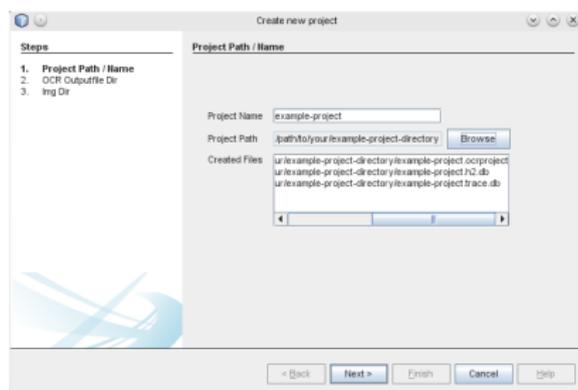
Creating a new project



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**.

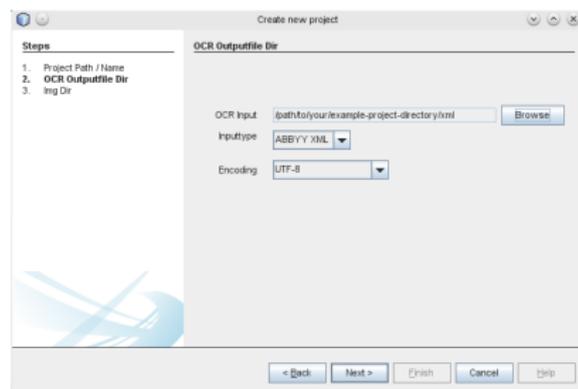
Creating a new project



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`.

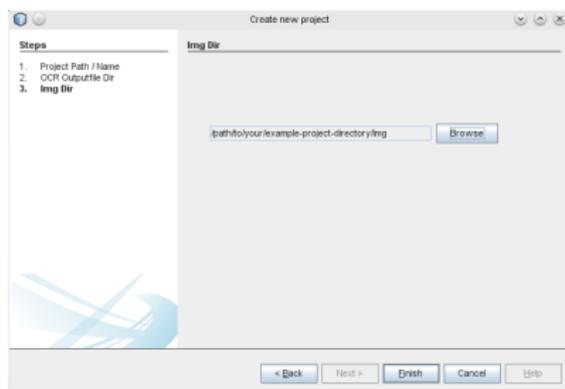
Creating a new project



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

The screenshot shows the PoCoTo interface with a concordance view. The main window displays a Latin text with highlighted tokens and their occurrences in a concordance table. The concordance table shows the token 'funt' highlighted in red, and its occurrences in the text: 'ut nunc funt tempora', 'atione per eadem loca & tempora', and 'tiisque agendis (ubi tempora'. The left sidebar contains a 'Concordance Actions' panel with a 'Show concordance' button and a 'Multi token actions' panel with 'Merge selected tokens' and 'Delete selected tokens' buttons. The bottom window shows the original Latin text with the token 'funt' highlighted in red.

- 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!