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In the resent years a lot of historical documents have been scanned and OCRed.

- The overall quality of the character recognition on historical documents is in general good.
- The performance of the OCR engines even on historical documents is constantly improved.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Text recognition on historical documents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

T4dbs

xi7 Kchs

oder in der Elbe ju halten offeaten. angefommen, daß fie ben nahe Die Elbe nicht beherbergen, und ein Rifch dem andern nicht ausweichen tonnen, daber die Leute hauffen Deife mit Merten berjugelauffen, und die Fische erschlagen. Den Bortheil des Lachse Fangs genuffet auch Schlefien von der Oder, und es find von langen Jahren ber ansehnliche Fangereven langft der Oder, j. E. bey

Manniein aber fich hauptfachlich im Saupt-Stuf, Männleinccher sich hauptsächlich im Haupt-Fluss, Es gebender öderm der Gbe zu halten pflegten. Es gedencktt aud eben diefer Auctor aus einem alten Manuferi- auch eben dieser Auctor aus einem alten Mannferipto, bas An. 1432. ein to groffes Seer von Eachfer pto, das An. 1431. ein 10 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

- 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:

- The menu area contains various commands for navigation and project maintenance.
- In the main view area shows tokens and offers the main correction possibilities.
- The complete image area displays the page of the current active (selected) token.
- The error area lists error frequency lists of common word or pattern errors.
- 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

| 00 | Anounts/Users/student/fini/fidevel/clain/spocotoProjects/zzz.orproject | |
|--|--|----------------------------|
| E Dat House Dem Toos | (10 <5 (3 12 (5) 5.> 105)) A A A | Q. Search in Types (Ctrief |
| Token Aktionen Fernster 🗙 🔚 | Selle 92 von 100 | |
| Konkordanz Aktion | | - |
| -S ¹ / ₂₁ 1 Varkammen Konkordenz anzeigen | CAPVT PVTXVII | 1 |
| Malti Token Aktionen 🛞 | 1 | - |
| Auswahl verschnetz Ausgewählte Token kin | De Caufa, Generatione, & Definitione Civitati | <u>न</u> ् |
| | Vefinitionc Als korrekt markieren | |
| | Löschen Uod homines Libertatis & Denach rechts fusionieren ma | |
| | Uod homines Libertatis & Dominii per naturara ama | |
| | | ~ |
| | Sets 12 × | 8 |
| | | 83 |
| OCR-Fehl. OCRFe. x | DE CIVITAT | E |
| enim 85- | SIVE | _ |
| efle 81 funt 80 quam 68 ilia 60 | REPVBLICA. | |
| etiam 58 tamen 57 rerum 44 alis 40 | CAPVT XVII. | |
| Homine 37 ilis 36 | De Caufa, Generatione, & Definitione Civitatis. | |
| modo 34 omnes 34 Ili 29 aliud 28 | Uod homines Libertatis & Dominii per naturar tores, ex præfcripto (ut fit in flatu Civili) vivere | n ama- volue- |
| eorum 28 | | • |

- *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.

| 00 | | Anounta/Usershtudent/Ini/fileveliciarin/pocoto/hojects/zzz.orproject | ~ × |
|--|-------|---|-----|
| Ele Edit Profiler Wew | Tools | Muqaw Reb | |
| 90 | « | <10 <6 😮 100 🖾 100 🔊 100 🔊 🗛 🔥 🔍 | |
| TokenAktionen Ferster 🗙 | | Sele 100 von 100 | |
| Konkordanz Aktion | | Regem , Proceres , & Coetum Communium , caula fuit Belli quod | 1 - |
| ອັງສູ່ ອັງອີ 1 Vorkammen Kankardenz anzeige | 50 | fequutum eft Civilis etiam difputationes de quæftionibus Politicis fequutum eft Civilis i etiam difputationes de quaeftionibus Politicis | 1 |
| Multi Token Aktioner Auswahl verschr Ausgewählte Token | n 🛞 | & Theologicis, quibus tamen populus ita hund de fure Regio eru & Theologicis, quibus tamen populus ita nunc de lure Regioeru | n |
| | | ditus eft, ut in Anglia pauci (puto) hunc fint qui lura prædkta infe ditus eft, ut in Anglia pauci (puto) nunc fintqui lura praedika infe- | 4 |
| | | parabilia elle non videant & publice agnituri fint fimul atque redie parabilia ell'e non videant : Sc publice agnituri lint firaul atque redie | 7, |
| | | | |
| | | Sele 100 X | - |
| | | Cap. 18. De Civitate. 91 | |
| OCRUEN OCREA Y | 4 | Regern, Proceses, & Contum Communium, caufa fuir Belli good | 1 |
| Konkorstanz anzeigen | | fequatum eff Civilis : etiam diffutationes de quæftionibus Politicis | |
| enim | 85.4 | & Theologicis, quibus tamen populus ita nune de Jure Regio ent- | |
| autem | 82 | dirus eff. ut in Anglia pauci (puto) punc fint out fura prædicta infe- | |
| efle | 81 | parabilia elle non videant : & publice agnituri fint fimul atque redie- | |
| quam | 68 | rit Pax & quandiu calamitatem proteritatum meminerint : fed non | |
| ilia | 60 | dintiùs nifi meliùs erudiatur populus. | |
| etiam | 58 | Queriem euten luns han Summer Basefleti effentielle. Re infene | |
| rerum | 44 | adulta funz, faquina, un anthufungun Varhis fananzi Realita gan | |
| alis | 40 | rabinatione, requirer, ut quibulcunque verbis leparari ce anis con- | |
| Homine | 3/ | ceut viocantur, niu Poteitati Summæ timul & exprehis verbis renun- | |
| modo | 34 | ciatum int, concettionem nullam; elle, ied concella omnia, Summä | |
| omnes | 34 | Potettate, id ett Perlonä Civitatis retentä, infeparabiliter redire. | |
| alud | 28 | Cum ergo Authoritas hæc ingens indivitibilis fit, & habenti Sum- | * |
| eorum | 28 🕶 | | • |

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

| Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Regent / Process / 6. Coetum Communium / caula fuit Belli quod Bitus eff. ut in Anglia pauci (puto) hunc gint qui / lura prædifa a Parabilia eff pon videant / 6. publice ggnituri fint finul atque redier Regent / Process / 6. Coetum Communium , caufa fuit Belli quod Regent / Process / 6. Coetum Communium , caufa fuit Belli quod Regent / Process / 6. Coetum Communium , caufa fuit atque redier Regent / Process / 6. Coetum Communium , caufa fuit Belli quod Regent / Process / 6. Coetum Communium , caufa fuit guod Regent / Process / 6. Coetum Communium , caufa fuit guod Regent / Process / 6. Coetum Communium , caufa fuit guod Regent / Process / 6. Coetum Communium , caufa fuit guod Regent / Process / 6. Coetum communium , caufa fuit guod | 00 | Insunts/Users/student/Vni/Fideveliciaris/pocotoProjects/zzz.corproject | |
|--|--|--|-------------------------------|
| Image: Section of the secon of the secon of the section of the section of the se | Ele Edit Profiler Wew Tools | Window Unit | |
| Regern / Process . 6. Coeturi Communium . caula fuit belli quod Regern / Process . 6. Coeturi Communium . caula fuit belli quod Regurtum eff Civilis etiam difutationes de quaeffionibus Politicis Regern / Process . 6. Coeturi Communium . caula fuit belli quod Regurtum eff Civilis etiam difutationes de quaeffionibus Politicis Regern / Process . 6. Coeturi Communium . caula fuit belli quod Regern / Process . 6. Coeturi Communium . caula fuit belli quod Regern / Process . 9 Regern / Process . 9 Regern / Process . 6. Coeturi Communium . caula fuit percelis a Intersetti . 1000 (cis guibus fameri populus ita nunc de lure Regioeru- ditus eff. ut in Anglia pauci (puto) nunc sint qui _ lura prædika i parabilia effe non videant : 5c publice agnituri lint firmul atque redie- sette x Cap. 18. De Civilite. 91 Regern / Proces, & Cocum Communium , caula fuit Beli guod Regern / Proces & Cocum Communium , caula fuit Beli guod Regern / Proces & Cocum Commune populus ita nunc de lure Regio era- sette x 91 Regern / Proces & Cocum Commune populus ita nunc de lure Regio era- sette x 91 Regern / Proces & Cocum Commune populus ita nunc de lure Regio era- sette x 91 Regern / Pro | | <10 <5 🙄 100 💭 100 💹 100 💹 🔥 🐴 🐴 🔦 🔍 | |
| Regem Process 6. Coetum Communium caula fuit belli quod Image: Communium Caulationes de quaettionibus Politicis Image: Communium Caulationes de politice Image: Communium Cau | TokenAktionen Ferster X | Sette 100 von 100 | < > - |
| Image: Second | Konkordanz Aktion 🛞 | Regem , Proceres , & Coetum Communium , caula fuit Belli | quod |
| Main Manundia & Theologicis, quibus tamen populus ita nunc de lure Regio erui- ta Theologicis , quibus tamen populus ita nunc de lure Regio erui- ditus eft. ut in Anglia pauci (puto) nunc sint qui ura prædika ei ditus eft. ut in Anglia pauci (puto) nunc sint qui ura prædika ei parabila effe non videant; & publice agnituri fint fimul atque redie- parabila effe non videant; Sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fint fimul atque redie- de la parabila effe non videant; sc publice agnituri fint fimul atque redie- te de la parabila effe non videant; sc publice agnituri fimila effe de la parabila effe non videant; sc publice agnituri find aque redie- rito agnita effe non videant; sc publice agnituri find aque redie- rito agnita effe non videant; sc publice agnituri find agnita parabila effe agnita effe non videant; sc publice agnituri find agnita parabila effe agnita effe non videant; sc publice agnituri find agnita field on diutita, niti mellas erudarar populas. The find field erud field erud field on diuta ita, niti mellas erud agnitar pareitaran memeritaria field on diuta ita, niti mellas erud agnitar pareitaran memeritaria field on con ei videantur, niti potetari Suman find & Enferati erud videantur, niti potetari Suman find & Senferati erud sc anno erud nita concefinoem nultara, effe, fed concefin onnia, suman poceta, ei de Perioda Ciriatia recent, inforparabiliar erdir. | ່າງເລີ້ມ ມີເຊິ່ງເພື່ອ ປີ Vorkammen Kankardanz anzeigen | fequutum eft Civilis etiam difputationes de quæftionibus Pol fequutum eft Civilis j etiam difputationes de quaeftionibus Poli | liticis 1 |
| Bitus Eff, br.fri Anglia pauci (puto) hunc fint qui lura prædkås i ditus eft. ut in Anglia pauci (puto) hunc sint qui lura prædkås i parabilia effe non videant; & publice agnituri lint fimul atque redie- parabilia effe non videant; & publice agnituri lint fimul atque redie- sense si commente site | Multi Teken Aktionen 🛞 | & Tlicologicis, quibus tamen populus ita hunc de lure Regio & Tlicologicis, guibus tamen populus ita nunc de lure Regioe | eru- ru |
| Parabilia effe pon videant Sc publice agnituri fint fimul atque rediet parabilia effe pon videant Sc publice agnituri fint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia effe pon videant Sc publice agnituri lint fimul atque rediet parabilia fint fine parabilia effe parabilia fint fine parabilia fint fint fine parabilia fint fint fine parabilia fint fine parabilia fint fint fine parabilia fint fint fine parabilia fint fint fint fint fint fint fint fint | | ditus eft, ut in Anglia pauci (puto) hunc fint qui du lura pr ditus eft, ut in Anglia pauci (puto) nunc sint qui du lura pr | æd icta i aedi& a i |
| Cap. 18. De Civitate. 91 Regern, Proceres, & Cetum Communiom, caufa fuit Belli quod fequuam eff Civitis e ciam discussiones de quatitonibus Politicis atom en anti- tata effective effective effective effective effective effective atom effective effective effective effective atom effective effective effective atom effective at | | parabilia cile non videant; & publice agnituri fint fimul atque parabilia ell'e non videant : Sc publice agnituri lint firaul atque | redie redie |
| Cap. 18. Dr Civitet. 91 Cap. 18. Dr Civitet. 91 Regern, Proceres, & Cream Commun, caußa fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit Belli good fequuam eff Civitis șt eim difunctiones de quark fuit de lure Regio era- ditur eft, uit n Aglia pauci (punc) nanc fințui glut par pradultă inde- parabilă effe non vicleari, 26 publice agnitori fint final arque redie- rit Rax, & guandui calamittem prateritarum meminierint șt del non diurităs, nitit mellăs eruditare populas. Munima atem Iura Bez Gamanes fund âte caprefis rethă s renon- ciatum fit, conceflionem nullară, effe, fed concefla omaia, Summă Poceflare, i del Perefis Civitairis zencen, înfeparabiliser redire. | | | × |
| Cap. 18. De Civitet. 91 Regen, Proceres, & Cettum Commonium, caula fuit Belli god fequatum eff Cirilis, etiam difputationes de quaditonibus Politicis & Theologicis, quibus tatamen populas its nunc de lure Regio eru- ditus eff, uri n Anglia pauci (puto) nanc finz qui trar predicta indé- parabilia del non vileans, & publice agnitori fins final aque redie- rit Dax, & Quandu calamittem practicitaminerini, fed oon ditarias, niti arelias eruditare populas. Tuonian attem Iran text funza final experimentaria di final attam teres di constante di finale di calamittem practicitaminerini, fed oon ditarias, niti arelias eruditare populas. Tuonian attem Iran tex Camma Forefait effentiali & infepa- rito di videntari, nifi Portetati Summa final & experifi verbis ferante redi videntari, nifi Portetati Summa final & experifi verbis ferante redi videntari, nifi Portetati Summa final & experifi verbis feranteri. | | Selec 100 X | |
| Cap. 18. De Civilité. 91 Regera, Proceres, & Cetam Commun, caufa fuit Belli quod fequarum eff Civilis, stiam diputationes de quatitoni bus Politicis and the second second second second second second second second second fequarum eff Civilis, stiam diputationes de quatitoni bus Politicis & Tlecologicis, quibus tamen populas ita nunc de lure Regio era- ditu eft, uit a Anglia pauci (puoto nunc fincqui plura pradultà inde- parabilia effe non vicleant, & publice agnitori fint final aque redie- ria Pax, & quandi cualmittem prateriarum meminismin fed do non diuritàs, niti mellas eruditare populas. Tuoniana autem Iura hes Cammas Potefait effentialis & infepa- rità di dividenti r, figuitur, ur quibufcunque Verbis feparari & alis con- ciatum fit, conceffiorem lulara, effe, fed concefla onnia, Summà Poceltar, i del Perioda Civitatis recendi, infeparalibilier redire. | | | - |
| Cap. 18. De Cetalité. de Cetalité. De Cetalité. estis faite Belli goud renorme aussi a Regen, Proceres, & Cetaun Oromaniam, esufs fuit Belli goud fequatum et Civitis e time difputationes de queficionibus Politicis & El Itologicis. quibe tamen populai sin anuce de lure Regio era- ditar eft, ut in Anglia pauci (puro) anne fint qui lera pradictà infe- parabilis effe non riceat; & grubble segnituri fint finul aque realie- tiane et lurité, a lurité de lure Regio era- ditar eft, ut in Anglia pauci (puro) anne fint qui lera pradictà infe- parabilis effe non riceat; & grubble segnituri fint finul aque realie- tiane et lurité, infin de seruitarus polus. Quoniana autem Ira hez Summa Potefati effentislia & infepa- rabilia fint en lirà loctati Summa finul & experielli verbis ferant radia de riceatum it, concefilorem allam; fele, (ca coccifa onnia, Summa Potefate, i del Periodi Civitati rendi, infopatabiliter redire. | | Our and the Charlester | |
| Regern, Proceres, & Catum Communin, cuta fuit Belli quod momente unamo Regern, Proceres, & Catum Communin, cuta fuit Belli quod futura de l'units, citam d'inputationes de quarkinoitous Politicia attente de l'units, citam d'inputationes de quarkinoitous Politicia et Tiscologicis, quibus tamen populus ita nanc de lure Regio era- dita est, a citam dia paraci (puoto) nanc fint qui lura prachità nic- parabila elle non vicleant, & publice agnitori fint final atque redie- parabila elle non vicleant, & publice agnitori fint final atque redie- parabila elle non vicleant, & publice agnitori fint final atque redie- ner di atta e su di atta esta esta esta esta esta esta esta | | Cap. 18. De Civitaie. | 91 |
| fequetam eff. Civilis e citam diffuctiones de quaditonibus Políticis autom est- autom est- ditus eft, ut in Anglia pausi (puto) nune funçuj lura pradità inde- partabilis effe non citeart, & gubile segnituri fun funda arque redic- partabilis effe non citeart, & gubile segnituri fun funda arque redic- tian est atam est | OCR-FeN. OCRFe. X | Regem, Proceres, & Cœtum Communium, caufa fuit Belli | quod |
| ann an Article Construction of the second se | Konkordanz anzeigen | fequatum est Civilis; etiam disputationes de quæstionibus Pol | liticis |
| ditus eft, vit in Afglia pausi (pitto) nanc fint qui fura pradicità inde- parta di superiore di superi di superiore di sup | enim 85 * | & Theologicis, quibus tamen populus ita nunc de Iure Regio | o eru- |
| parabaixa elle non vilcania; sé publice agnitioni fain famal aque redie- rie Dax, sé quambia calamitatem paraetriarian meninerint je foi on diutiùs, niti mellois eraditatra populos. Quoniana autem Iara has Cammae Potefait ellentialia & infepa- rationa e cali rediante a quibalcanque Verbis feparari & aliti con- tento e cali vicenti più fordetta i Summa fund & experifik verbis renon- ciatum fit, concefionem nullam; elle, fed concella omnia, Summa Portos Potefaita (concelli ornatis, Summa | autern 82 | ditus eft, ut in Anglia pauci (puto) nunc lint qui lura prædicta | infe- |
| num rit Par, & gummdu calamitatem præteritarum meminerint ; fed non diuritiks , niti melinis rudiarur populus. Mom Quoniana autem lara hæc Summa Poreflati effentislis & infepa- an rabila funr , fequitar , quiblucinque etrolis féparari & a lilis con- reditive autem and foretfati Summa fund & experifis verbis renun- cedi videanur , nif Poreflati Summa fund & experifis verbis renun- roda Poreflate, i del Poreflati Summa fund atter presidenter redire. | etie 81 funt 90 | parabilia elle non videant : & publice agnituri fint fimul arque t | redie- |
| in an eigen eine einen eine einen eine eine | guam 68 | rit Pay & quandiu calamitatem proteritatum meminerint (e | dnon |
| atam Garage Sandara (1990) atam Garage Sandara (1990) at | ilia 60 | diutilite nife meliite anudiator populate | |
| Reconstant and a local state of the contrast contrast contrast and the contrast and the contrast and the contrast and the contrast of the contrast of the contrast of the contrast of the contrast and the contrast and the contrast of the contr | tamen 58 | Quoniam autem lura has Summa Potestati effentialia & in | fans- |
| ann 40 Hanna 40 Hanna 10 Hanna 10 | rerum 44 | schille fins families as mile formers Unkis formers i for ality | nepa- |
| Jennes 7 cedi videantur, nii Poteftari Summa imali & exprefis verbis renon- ciatum fit, conceffionem nullarn ; effe, fed conceffa omnia, Summå mods 4 Poteftare, id eft Perfonà Civitatis recentà, infeparabiliter redire. | alis 40 | ratina init, requitir, ut quibuicunque verbis ieparati oc anis | 0.001- |
| modo 34 ciatum (st, concellionem nullam ; effe, fed concella omnia, Summä mones 34 Potestate, id eft Persona Civitatis retenta, infeparabiliter redire. | Homine 37 | cedi videantur, nili Potettati Summæ timul & expretiis verbis r | enun- |
| omnes 34 Poteltate, id elt Persona Civitatis retenta, infeparabiliter redire. | modo 34 | ciatum ist, concelitonem nullam; elle, fed concella omnia, Su | mmä |
| | omnes 34 | Poteltate, id eft Perfonâ Civitatis retentâ, infeparabiliter redire | . |
| Cum ergo Authoritas hæc ingens indivifibilis fit, & habenti Sum- | 10 29 alud 28 | Cum ergo Authoritas hæc ingens indivifibilis fit, & habenti | Sum= |
| | eorum 28 * | | · · · |

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

| 00 | | Anounts/Users/student/Wnifideveliciarin/pocots/Projects/tzzz.corproject | 0 × |
|--|--------------|---|-------------------------|
| Ele Edt Profiler Wew | Tools | Muqaw Reb | |
| 🕒 🦻 (° | « | <10 <6 😋 100 😳 555 103 💓 🗛 🐴 🔦 🔍 | |
| TokenAktionen Penster 🗙 | | Sete 100 von 100 | |
| Konkordanz Aktion | ۲ | Regem , Proceres , & Coetum Communium , caula fuit Belli quod | - 1 |
| - ^{5,1} / ₂₍₁), 250 Vorkonmen Konkordenz anzeige | 20 | fequitum eft Civilis ; etiam difputationes de guæftionibus Politicis fequitum eft Civilis ; etiam difputationes de quaeftionibus Politicis | <u>s</u> _¶ = |
| Multi Token Aktionen Auswahl verschw Ausgewählte Token | netz 1020 | & Theologicis, quibus tamen populus ita nunc de fure Regio eru & Theologicis , quibus tamen populus ita nunc de lure Regioeru | J _¶ |
| | | ditus eft, ht in Anglia pauci (puto) hund fint qui lura prædičta infe ditus eft, ut in Anglia pauci (puto) nunc sint qui lura prædičta infe | 8, |
| | | parabilia elle non videant, & publice agnituri fint fimul atque redie parabilia elle non videant : Sc publice agnituri lint firaul atque redie | ₩ 1 |
| | | | |
| | | Sele 100 X | - |
| | | Cap. 18. De Civitate. 91 | |
| OCREAN OCREA X | | Regern . Proceres . & Cottum Communium . Caufa fuit Belli good | |
| Konkordanz anzeigen | | fequatum of Civilis ; etiam difputationes de quaftionibus Politicie | |
| enim | 85 + | & Theologicis, quibus tamen populus ita nune de Jure Regio ent- | _ L |
| autern | 82 | dirus eft, ut in Anglia pauci (puto) punc fint pui lura prædicta infe- | |
| efle | 81 | parabilia effe non videant : & publice agnituri fint fimul arque redie- | |
| quam | 68 | rit Pay & quandiu calamitatem proteritatum meminerint - fed non | |
| ilia | 60 | diatilis nifemelije erudiatar populas. | |
| etiam | 58 | Quantar autor lure has Summer Dataflati affantialia Re infant. | |
| rerum | 44 | subilia fina faquina un avibufamona Varbia francesi Praliis con | |
| alis | 40 | rabination, requirer, or quibulcunque verbis leparari oc anis con- | |
| Homine | 3/ | ceut viocantur, nui Potetrati Summæ timul & expretitis verbis renun- | |
| mode | 34 | ciatum ist, concettionem nullam; elle, fed concella omnia, Summä | |
| omnes | 34 | Poteltate, 1d ett Perlonä Civitatis retentå, infeparabiliter redire. | |
| alud | 28 | Cum ergo Authoritas hæc ingens indivifibilis fit , & habenti Sum- | |
| eorum | 28 🕶 | | • |

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

| 00 | | Anounts/Usershitudent/Fini/Idevelclarin/pocotoProjects/tzzz.orproject | • × |
|-------------------------|----------|--|-----|
| Ele Edt Profiler Vier | w Tools | Window (Sep | |
| Tokan Materian Terrater | × 🗆 | | |
| | | 5 5666 53 VOT 100 | |
| Konkordariz Aktion | n (8) | | |
| -1 - 21 Verkommer | | unt lecuritatem, neque contra communem noitem, neque contra | 9 |
| 49 | | funt fecuritatem , neque contra communem hoftem , neque contra | |
| Noritoritaria arce | a you | | |
| Multi Token Aktion | nun 🔍 | injurias alter alterius. Diffentientes eniminter le de Viriumufunon | 9 |
| | | injurias alter alterius . Didentientes enim inter fe de Virium ufu non | |
| | | | |
| Ausgewählte Tok | ben lós | fibi mutuò auxiliaturi funt, fed oppofitis confiliis vires ad nihilum | |
| | | fibi mutuo auxiliaturi funt, fed oppofitis confiliis vires ad nihilum | |
| | | | |
| | | reducturi Nnde non modo à communi hofte facile superantur, les | đ |
| | | reduduri V nde non modo & communi hofte facile fuperantur fee | đ |
| | | | |
| | | etiam de commodis propriis inter fe Bello certaturi fuot. Siguidem | 3 |
| | | trianing benning de behnd inter te bene bermitari bild | 1 . |
| | | Selection of the select | - |
| | | non certo, fed cum viribus hoftium comparato determinatur, ut ma- | - |
| | | jor fit quam ut excellus tanti ei tam confpicui momenti ad Bellum fi- | |
| | | niendum fit, ut hoftis ad aggrediendum provocetur. | |
| 000 5-11 00005 | | Sit autem multitudo quantacunque, fi tamen actiones corum Iudi- | |
| Kinchardson annaisen | | ciis & Arbitriis multorum gubernentur, nullam inde expectare pol- | |
| enim | 85 * | funt fecuritatem, neque contra communem hoftem, neque contra | - |
| autem | 82 | injurias alter alterius. Diffentientes enim inter fe de Virium ulu non | |
| funt | 80 | libi mutuo auxiliaturi lunt, led oppolitis confiliis vires ad minium | |
| quam | 68 | reducturi. Mnde non modo à communi holte facile luperantur, fed | |
| etiam | 58 | etiam de commodis propriis inter le Bello certaturi junt. Siguidem | |
| tamen | 57 | fai amos coorra in Revisiten contestante Lores Newsy chief | |
| afis | 40 | randas confentire funnomeretur, idem stiam de toro genere humano | |
| Homine | 37 36 | funnonendum effet, iraque Regimine Civili omnino onus non effet | |
| modo | 34 | niduris foilicet hominibus in Pace & fine Dominis | |
| omnes | 34 29 | Neone ad focuritatem (quam perpetuam effe volunt) fufficit, ut | |
| alud | 28 | aubernanue neo carto sancim & determinato remnore ut in uno | |
| eorum | 28 - | | 21 |

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

| 00 | | Anounts/Users/shtudent/Yin/fidevel/clarin/pocotoProjects/tzzz.scrproject | 8 0 8 |
|-------------------------|----------|--|------------|
| Ele Edit Profiler View | Tools | | |
| TokenAktionen Ferster X | | A A A | |
| Konkordanz Aktion | | | |
| Rocket Gine Patrone. | | funt fecuritatem heque contra communem hoffern heque cor | ntra |
| The State of Verkommen | | funt fecuritatem neque contra communem hoftem neque con | tra |
| Konkordenz enzeige | en. | | |
| | | injurias alter alterius. Diffentientes eniminter fe de Viriumufu | non. |
| Multi Token Aktioner | n 🛞 | injurias alter alterius . Didentientes enim inter fe de Virium ufu | non |
| Auswahl verschri | vetz | | |
| Ausgewählte Token | 1010 | fibi mutuò auxiliaturi funt, fed oppofitis confiliis vires ad nihilut | n, |
| | | fibi mutuo auxiliaturi funt, fed oppofitis confiliis vires ad nihilum | -1 |
| | | | |
| | | reducturi V nde non modo à communi hofte facile super | rantu |
| | | reduduri V - nde non modo & communi hofte facile fuper | antur |
| | | V | |
| | | etiam de colAls korrekt markieren r le Bello certaturi funt. Sigur | dem, |
| | | Löschen | • |
| | | Sate 53 x nach rechts fusionieren | 8 |
| | | non certo, fed cum viribus hoftium comparato determinatur, ut m | a- ^ |
| | | jor lit quam ut excellus tanti el tam compicul momenti ad Benum | E - |
| | | Sit aurem multitudo quantacunque, fi tamen actiones corum Inc | ti. 🗖 |
| OCR-FeN. OCRFe. X | | ciis & Arbitriis multorum gubernentur, nullam inde expectare po | of- |
| Konkordanz anzeigen | 07.0 | funt fecuritatem, neque contra communem hoftem, neque cont | ra - |
| autem | 82 | injurias alter alterius. Diffentientes enim inter fe de Virium ufu no | on |
| efle | 81 | fibi mutuò auxiliaturi funt, fed oppolitis confiliis vires ad nihilu | im I |
| quam | 68 | reducturi. Mnde non modo à communi holte facile luperantur, f | eđ |
| etiam | 58 | ettam de commonis propriis inter le Bello certaturi tunt. Siguide | an a |
| tamen | 57 44 | fer omnes cogere in & duiratem cateralque Leges Nature obs | |
| alis | 40 | vandas confentire fupponeretur, idem etiam de toto genere huma | no |
| ilis | 36 | fupponendum effet, itaque Regimine Civili omnino opus non effe | ι, |
| modo | 34 34 | victuris feilicet hominibus in Pace, & fine Dominis. | |
| di . | 29 | Neque ad fecuritatem (quam perpetuam effe volunt) fufficit, | ut 🗌 |
| eorum | 28 ¥ | aubernantur nro carto tantúm & determinato remnora ut in u | •••• |

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

| 00 | | Anounts/Aberts/htudent/Ynif/ideveliciarin(pocotoProjects/zzz.ocrproject | |
|-------------------------|----------|--|---------|
| Ele Edit Profiler Vie | w Iools | | |
| Token Aktionen Fernster | XE | | |
| | | | |
| Konkordanz Akbo | n 🛞 | funt facuritatent beque contra communem hoffent beque to | ontra |
| -total | | funciecuritaten, nequecontracommunem nonem, neque | ontra |
| Konkordeng enge | sigen | funt recuritatem, neque contra communem nottem, neque c | ontra |
| | - | Entering Development in the formation of | |
| Multi Token Aktion | nen 😒 | injurias alter alterius. Difentientes eniminter iene viriumui | unon |
| Accountly years | | injurias alter alterius. Didentientes enim inter fe de Virium uf | u non |
| | Lars Man | | |
| | | normatuo auxinaturi junt, jed oppolitis contilus vires ad nihi | um |
| | | fibi mutuo auxiliaturi funt, fed oppofitis confiliis vires ad nihilu | ım |
| | | | |
| | | reducturi Vndenon modo a communi holte facile luperantu | r, ied |
| | | reduduri . Vnde non modo & communi hofte facile fuperantur | , fed |
| | | | |
| | | etiam de commodis propriis inter le Bello certaturi funt Siq | uidem |
| | | | |
| | | Sets 53 × | |
| | | non certo, fed cum viribus holtium comparato determinatur, ut | ma |
| | | jor fit quam ut excellus tanti ei tam confpicui momenti ad Bellu | m #- |
| | | niendum ist, ut nottis ad aggrediendum provocetur. | Tadi 👘 |
| OCR-FeN. OCRFe | × E | ciis & Arbitriis multorum gubernentur pullam inde ernechter | nof |
| Konkordanz anzeige | n | funt fecuritatem - neque contra communem hoftem - neque co | ontra |
| enim autem | 85 * | injurias alter alterius. Diffentientes enim inter fe de Virium ufu | non |
| efle | 81 | fibi mutuò auxiliaturi funt, fed oppofitis confiliis vires ad nih | ilum |
| quam | 80 68 | reducturi. Vnde non modo à communi hofte facile fuperantur , | , fed 👘 |
| ilia otiom | 60 59 | etiam de commodis propriis inter fe Bello certaturi funt. Siqui | dem |
| tamen | 57 | enim hominum numerus magnus, fine Potentia communi quæ | pof- |
| rerum aliis | 44 | iet omnes cogere, in Æquitatem cæterafque Leges Natura o | bler- |
| Homine | 37 | vandas contentire lupponeretur, idem etiam de toto genere hun | nano |
| modo | 36 34 | supponendum etter, staque Regimine Civili omnino opus non e | ner, |
| omnes | 34 | Neave ad focuritatem (quam percetuam effe robust) fofficir | |
| alud | 28 | enternation neo casto transim & determinato rempore ut in | , ut |
| 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.

| Die Det Brother View T | Anounts/Users intudent Trial fide velicitar in pocotoProjects/222 corproject | 000 |
|--|--|----------------------------|
| 90 | 🕷 💷 🖾 🚺 😜 🔤 🔤 🔊 🗛 🔥 🔨 🗠 | Q. Search in Types (Ctrief |
| TokenAktionen Ferster | Sete 1 von 100 effe 1/2 x | < > - |
| Konkordanz Aktion | Sete martieren Alles martieren Korreiturianskist ändem Horrigieren Entries Total 31 Entries Setectet 4 Entries Corrected 0 Entries Disabled 0 Entries to process: 81 | |
| $\frac{\frac{n^2}{\sigma_1 n}}{Kankanders anseigen}$ | vis) horni - nem effe | me, qui mea |
| Multi Token Aktionen & | juicquid dicam verum effe | putem, quod |
| | caufam denique Intelligendi effe | quafdam Species (i |
| | obiter auidem neolioendum effe | |
| | Heuer Korrelazionskiel (reco | Dutent Sanientis |
| | | |
| CR-FeN. OCRFe. X | | |
| Konkordanz anzeigen nim 8 utem 8 | 5 A | - |
| ne s int S uam 6 a 6 | 11 0 8 0 | |
| tiam S amen S erum 4 | e 7 4 | |
| his 4 Iomine 3 is 3 nodo 3 | 07 77 86 44 | |
| mnes 3 i 2 iud 2 | 4 9 | Ţ |
| iorum 2 | 8 - 4 | • |

- 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.

| | Anounts/Users/student/Yn#Triteveliciarin/pocotoPro | ojecta/zzz.ocrproject | 808 |
|--|---|---|----------------------------|
| | 🕷 410 85 😋 1 🙄 555 105 356 🗛 A* A* 9 | < < | Q- Search in Types (Ctrivi |
| TokenAktionen Fernster | Sete 1 von 100 effe 1/2 x | | |
| Konkordanz Aktion 🛞 | Seite martieren Alles martieren Korreiturkansldat ändern Korrigi Ermies Total 81 Entries Selected, 4 Ermies Corrected, 0 Ermies Disc | eren Ibled: D Entries to process: 81 | |
| $\frac{h^2 d_{\rm eff}}{d_{\rm eff}^2} \ {\rm Varkammen} \\ \hline {\rm Konkordenc enceigen} \\$ | vis) homi- nem | effe efle | me, quime |
| Multi Token Aktionen 8 | juicquiddicam verum | effe | putem, quo |
| Ausgewählte Token kis | uicquid dicam Iverum | efle | putem , quod |
| | caufam denique Intelligendi caufam denique Intelligent | effe efle | qualdam Species |
| | obiter quidem negligendum obiter Iquidera negligendum | efie efie | |
| | cont etfifalfa | effe | putent Sanienți |
| | Sele:1 × | | 8 |
| | | | - |
| | | | |
| Konkorstanz anzeigen | | | |
| anim 8 | | | _ |
| autern 8. | 2 | | |
| unt 8 | â | | |
| quam 6 | | | |
| lia 6 | 2 | | |
| tamen 5 | 2 | | |
| erum 4 | 4 | | |
| alis 4 | 0 | | |
| Homine 3 | 2 | | |
| As 3 mode 3 | | | |
| nmnas 3 | | | |
| 11 2 | 9 | | |
| aiud 2 | в | | |
| | | | |

- 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.

| 0 | Anounts/Users/student/fini/fideveliciarin/pocotsPro | ojecta/222 ocrproject | * * * |
|---|--|--|------------------------------------|
| | 🕷 410 63 🕼 1 😜 50 100 💓 🗛 🛉 4 | | Q. Search in Types (Ctri+) |
| TokenAktionen Fenster | Sete 1 von 100 effe 1/2 x | | < > < D |
| Konkordanz Aktion | Seite markieren Alles markieren Korreiturkandidat ändern Korrigi Emtries Totat 81 Entries Selected: 0 Entries Corrected: 4 Entries Disa | oren bled: 0 Entries to process: 77 | |
| $\frac{h_{d}^{2}}{\sigma_{1}^{2}} \sim \text{Varkammen}$ Konkerstenz enzeigen | vis) homi- nem | effe | me, quime: |
| Multi Token Aktionen | | | |
| Auswahl verschme | juicquid dicam verum | efle | putem, quo |
| Ausgewählte Token lö | juicquid dicam iverum | esse | putem , quod |
| | caufam denique Intelligendi caufam denique Intelligent | effe esse | qualdam Species quafdam Species |
| | obiter quidem negligendum | effe esse | |
| | cunt erfifalfa | effe | putent Sanienti |
| | N Select V | | |
| | | | - |
| | | | |
| | | | |
| 010 F.H. 010F. H | 2 | | - |
| Kookardaan annainan | | | |
| | 5.4 | | |
| utern | 82 | | |
| fle i | 81 | | |
| unt | 80 | | |
| ia i | 20 | | |
| tiam | 58 | | |
| amen ! | 57 | | |
| erum - | 44 | | |
| alis | 40 | | |
| lomine | 37 | | |
| nodo | 24 | | |
| mnes | 34 | | |
| 6 | 29 | | |
| aiud | 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



- 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.

- The character based Abbyy XML 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.

| © 💿 | Cr | eate new project | \odot \odot \otimes |
|--|---|--|---------------------------|
| Steps | Project Path / Na | ime | |
| Project Path / Hame OCR CuportP Dr mg Dr | Project Name Project Path Created Files | Rumple or sjoct particle system particle syste | |
| | | | |
| | | < gack Next > Einish Cancel | Help |

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

| © 💿 | Cr | eate new project | \odot \odot \otimes |
|--|---|--|---------------------------|
| Steps | Project Path / Na | ime | |
| Project Path / Hame OCR Guputte Dr mg Dr | Project Name Project Path Created Files | numple project particulty anample project directiny in anample project directiny instanties project Anample with anample project directiny instanties project Ana. with anample project directiny instanties project Trace. db | |
| | | | |
| | | < gack Next > Einish Cancel | Help |

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

| 00 | C | reate new project | 8 O 8 |
|---|----------------|---|---------|
| Steps | OCR Outputfile | Dir | |
| Project Path / Name OCR Outputfile Dir J. Img Dir | | | |
| | OCR Input | (path.to/your/example-project-directory/xml | Browse |
| | Inputtype | ABBYY XML 💌 | |
| | Encoding | UTF-8 | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | < Back Next > Enish Canc | el Holp |

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

| 00 | Create new project | 8 0 S |
|---|--|-------|
| Steps 1. Project Path / Name | Img Dir | |
| OCR Outputfile Dir Img Dir | | |
| | (path/to/your/example-project-directory/img Browse | |
| | | |
| | | |
| | | |
| | < Back Next > Enish Cancel | Help |

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



- 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
- 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!