LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

uwe springmann
❧ scientia cum humanitate ❧
digital humanist

Wittscholarship 2014

- Latin Natural Language Processing
- OCR-Nachkorrektur-Tool PoCoTo

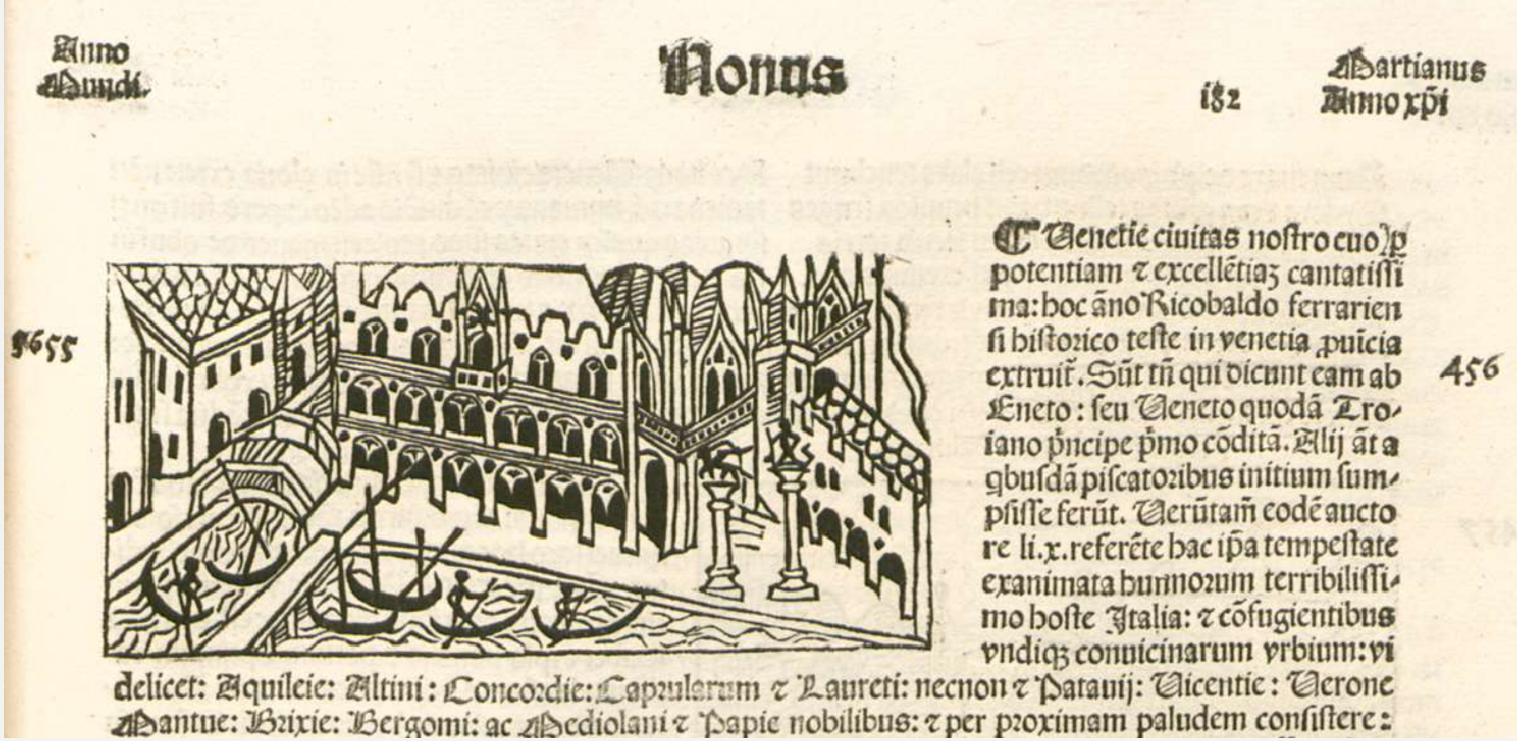
2014-06-06

Latin NLP?

- huge heritage: largest body of historical literary sources
- Latin publications dominate print production until about 1750
- many titles have never been reprinted
- either key or barrier to cultural heritage of the western world
- has been left out of the EU IMPACT (**Imp**roving **Ac**cess to **T**exts, 2008-12) project despite its importance

Foresti 1486, Supplementum Chronicarum

- color is gone, but “printing space economy” still holds (abbreviations)
- was heavily used by Schedel's Weltchronik (Latinist project @LMU)



Anno Mundi.

Nonus

182

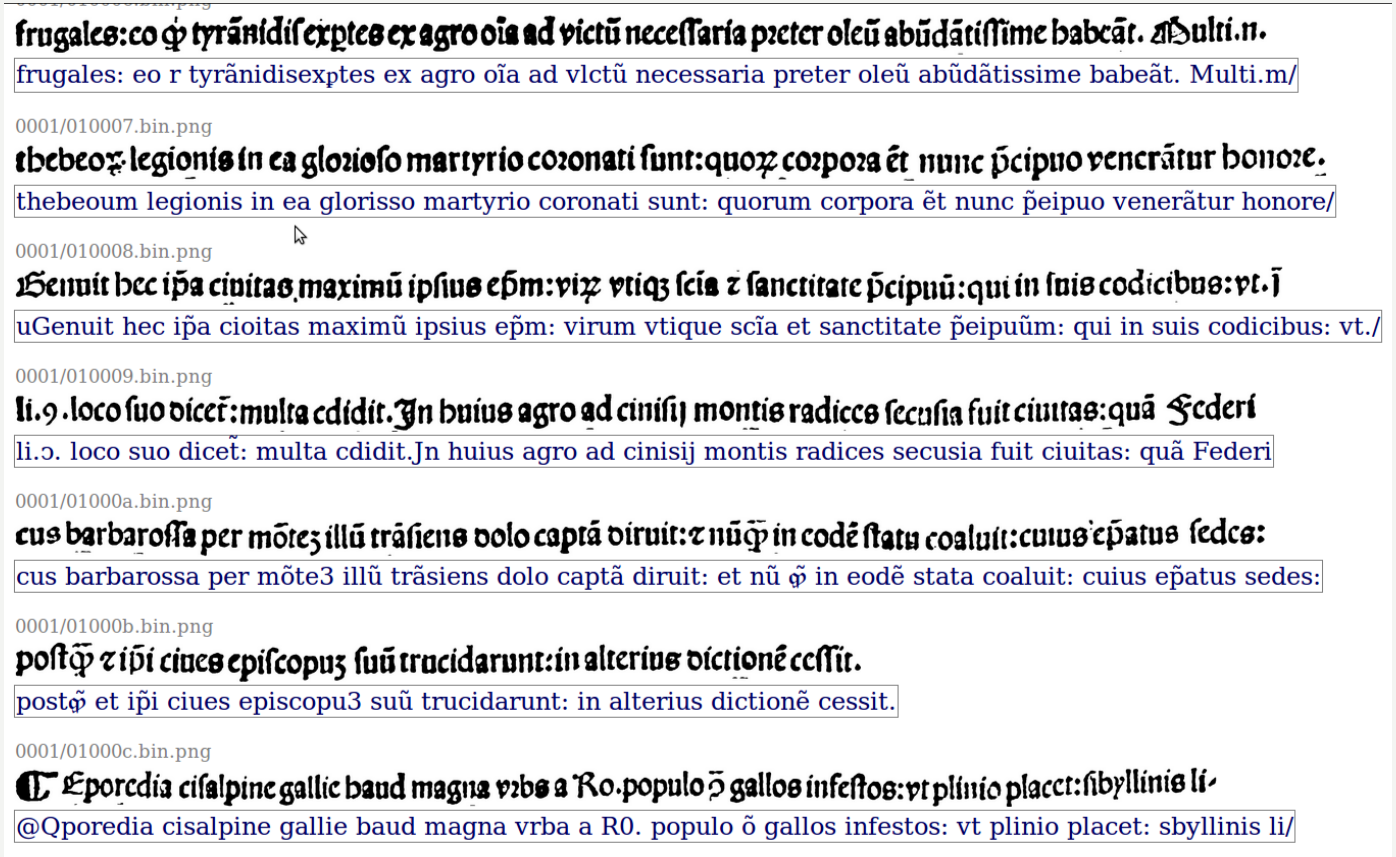
Martianus Anno xp̄i

5655

❡ Venetie ciuitas noſtro euo ꝑ potentiam ⁊ excellētiaʒ cantatiſſima: hoc āno Ricobaldo ferrarienſi hiſtorico teſte in venetia ꝓuīcia extruit̄. Sūt tñ qui dicunt eam ab Eneto: ſeu Veneto quodā Troiano p̄ncipe p̄mo cōditā. Alij āt a ꝗbuſdā piſcatoribus initium ſumpſiſſe ferūt. Verūtam̄ codē auctore li.x. referēte hac ip̄a tempeſtate exanimata hunnorum terribiliſſimo hoſte Italia: ⁊ cōfugientibus vndiq̄ʒ conuicinarum vrbium: videlicet: Aquileie: Altini: Concordie: Caprularum ⁊ Laureti: necnon ⁊ Patauij: Vicentie: Verone Mantue: Brixie: Bergomi: ac Mediolani ⁊ Papie nobilibus: ⁊ per proximam paludem conſiſtere:

456

Foresti 1486 (very modest training)



frugales: eo r tyrãnidisexꝑtes ex agro oĩa ad vlctũ necessaria preter oleũ abũdãtissime babeãt. Multi.m/

0001/010007.bin.png

thebeoum legionis in ea glorisso martyrio coronati sunt: quorum corpora ẽt nunc p̃eipuo venerãtur honore/

0001/010008.bin.png

uGenuit hec ip̃a cioitas maximũ ipsius ep̃m: virum vtique scĩa et sanctitate p̃eipuũm: qui in suis codicibus: vt./

0001/010009.bin.png

li.ɔ. loco suo dicet̃: multa cdidit.Jn huius agro ad cinisij montis radices secusia fuit ciuitas: quã Federi

0001/01000a.bin.png

cus barbarossa per mõte3 illũ trãsiens dolo captã diruit: et nũ ꝙ̃ in eodẽ stata coaluit: cuius ep̃atus sedes:

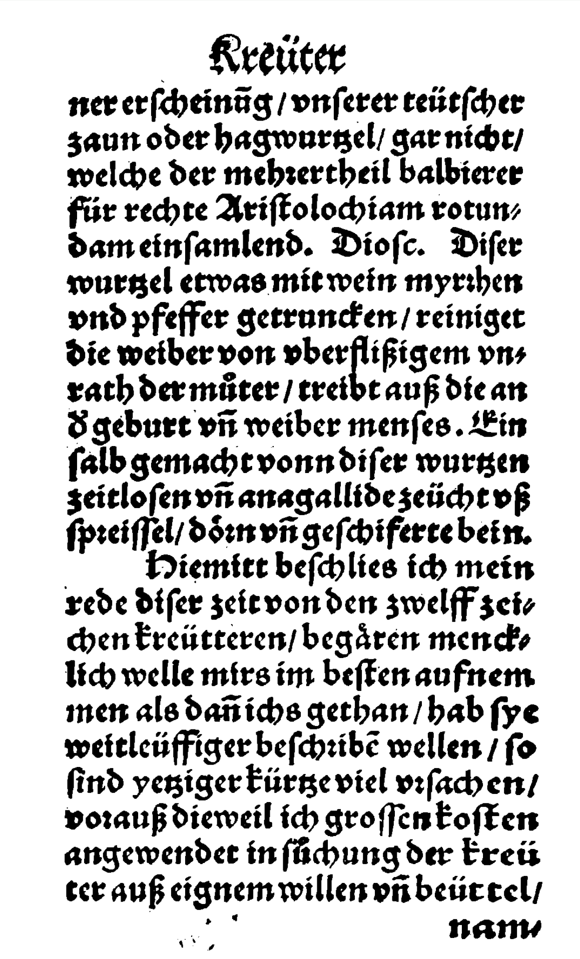
0001/01000b.bin.png

postꝙ̃ et ip̃i ciues episcopu3 suũ trucidarunt: in alterius dictionẽ cessit.

0001/01000c.bin.png

@Qporedia cisalpine gallie baud magna vrba a R0. populo õ gallos infestos: vt plinio placet: sbyllinis li/

Adam von Bodenstein, 1577



Kreüter
ner erſcheinũg / vnſerer teütſcher
zaun oder hagwurtzel/ gar nicht/
welche der mehꝛertheil balbierer
für rechte Ariſtolochiam rotun⸗
dam einſamlend. Dioſc. Diſer
wurtzel etwas mit wein myrꝛhen
vnd pfeffer getruncken / reiniget
die weiber von vberflißigem vn⸗
rath der můter / treibt auß die an
ð geburt vñ weiber menſes. Ein
ſalb gemacht vonn diſer wurtzen
zeitloſen vñ anagallide zeücht vß
ſpꝛeiſſel/ doͤꝛn vñ geſchiferte bein.
Hiemitt beſchlies ich mein
rede diſer zeit von den zwelff zei⸗
chen kreütteren/ begaͤren menck⸗
lich welle mirs im beſten aufnem
men als dañ ichs gethan / hab ſye
weitleüffiger beſchꝛibẽ wellen / ſo
ſind yetziger kürtze viel vꝛſachen/
voꝛauß dieweil ich groſſen koſten
angewendet in ſůchung der kreü
ter auß eignem willen vñ beüttel/
nam⸗

Kreüter
ner erſcheinüg/ vnſerer teütſcher
zaun oder hagwurtzel/ gar micht/
welche der mehrertheil balbierer
für rechte Ariſtolochiam rotun⸗
dam einſamlend. Dioſc. Diſer
wurtzel etwas mit wein myrrhen
vnd pfeffer getruncken/ reiniget
die weiber von vberfliſzigem vn⸗
rath der můter/ treibt auſz die an
d geburt vñ weiber menſes. Ein
ſalb gemacht vonn diſer wurtzen
zeitloſen vñ anagallide zeücht vſz
ſpreiſſel/ doͤ rnvñ geſchiferte bein.
Hiemitt beſchlies ich mein
rede diſer zeit von den zwelff zei⸗
chen kreütteren/ begaͤren menck⸗
lich welle mirs im beſten aufnem
men als dañ ichs gethan/ hab ſye
weitleüffiger beſchribẽ wellen/ ſo
ſind yetziger kürtze viel vrſach_en/
vorauſz dieweil ich groſſen koſten
angewendet in ſůchung der kreü
ter auſz eignem willen vñ beüt_tel/
ncn⸗

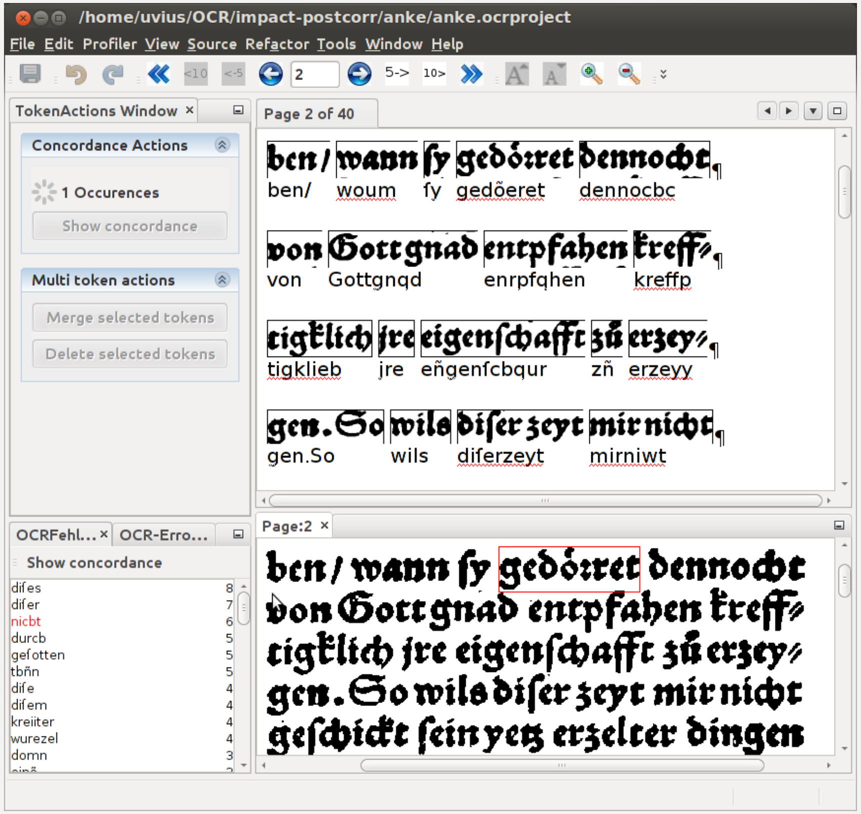
PostCorrectionTool PoCoTo

- locally installable Java package for postcorrection
- developed at CIS as part of the EU IMPACT project
- word synopsis: image + OCR with interactive correction
- error profiling: calculate statistical error model based on
  a) historical spelling (not an error)
  b) proper OCR errors
  and propose most probable correction candidate
- batch correction: rank errors according to frequency & error pattern
  and enable quick correction decision in concordance view

- try it out:

http://www.digitisation.eu/tools/browse/ocr-post-correction-and-enrichment/post-correction-tool/
https://github.com/thorstenv/PoCoTo

PoCoTo (developed at CIS)



den / wann ſy gedőꝛet dennocht
ben/ woum fy gedőeret dennocbc

von Gott gnad entpfahen kreff-
von Gottgnqd enrpfqhen kreffp

tigklich jre eigenſchafft zů erzey-
tigklieb jre eñgenfcbqur zñ erzeyy

gen. So wils diſer zeyt mir nicht
gen.So wils diferzeyt mirniwt

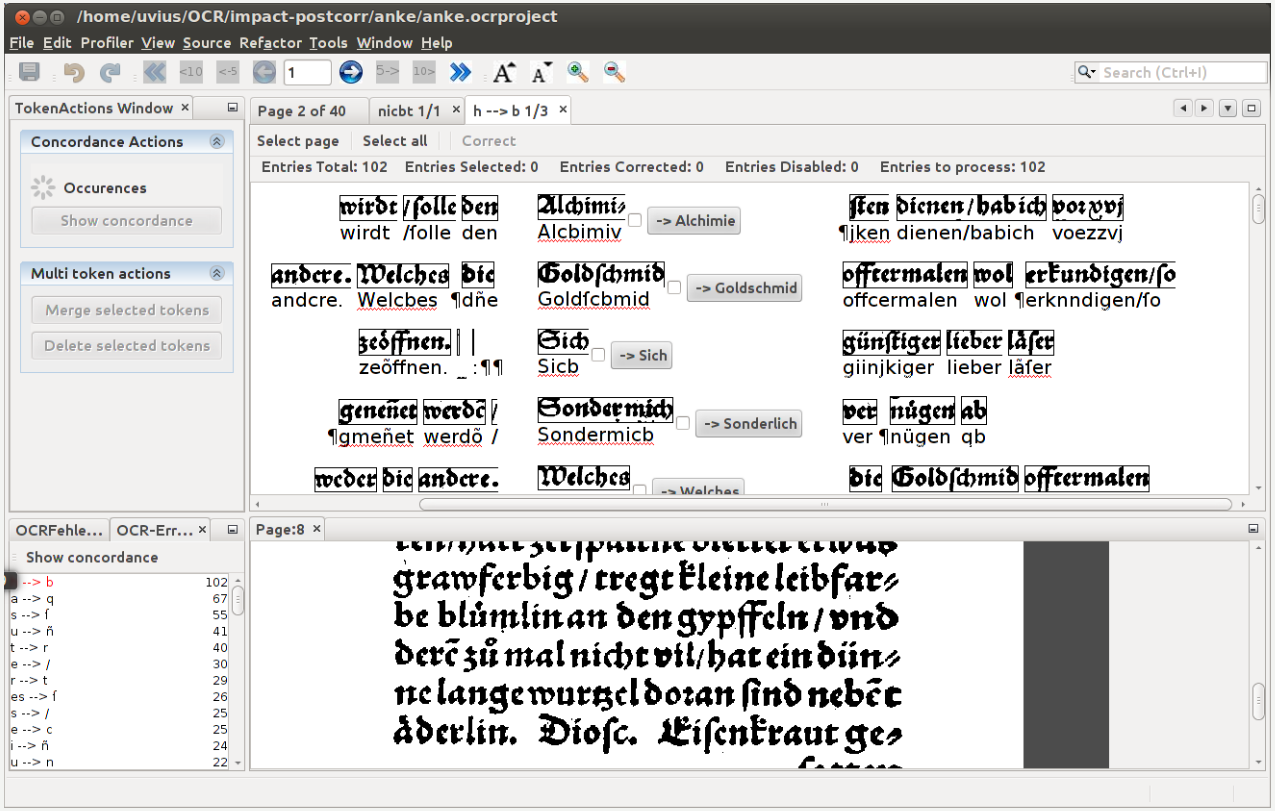
diſes	8
diſer	7
nicbt	6
durcb	5
geſotten	5
tbñn	5
diſe	4
diſem	4
kreiiter	4
wurezel	4
domn	3

den / wann ſy gedőꝛet dennocht
von Gott gnad entpfahen kreff-
tigklich jre eigenſchafft zů erzey-
gen. So wils diſer zeyt mir nicht
geſchickt ſein yetz erzelter dingen

- error frequency
- word synopsis (tesseract hocr output)
- page context

PoCoTo (developed at CIS)

error pattern concordance with batch correction



Entries Total: 102 Entries Selected: 0 Entries Corrected: 0 Entries Disabled: 0 Entries to process: 102

wirdt /folle den — Alcbimiv □ — ¶jken dienen/babich voezzvj

andcre. Welcbes ¶dñe — Goldfcbmid □ — offcermalen wol ¶erknndigen/fo

zeõffnen. _ :¶¶ — Sicb □ — giinjkiger lieber lâfer

¶gmeñet werdõ / — Sondermicb □ — ver ¶nügen qb

--> b	102
a --> q	67
s --> ſ	55
u --> ñ	41
t --> r	40
e --> /	30
r --> t	29
es --> ſ	26
s --> /	25
e --> c	25
i --> ñ	24
u --> n	22

grawferbig/tregt kleine leibfar=
be blůmlin an den gypffeln/vnd
derē zů mal nicht vil/hat ein dün=
ne lange wurtzel doran ſind nebē
äderlin. Dioſc. Eiſenkraut ge=

Thank you for your attention!