IEG
Leibniz-Institut für Europäische Geschichte

Preparing a Text Corpus with Transkribus

Pre-Processing, Training, Post-Processing

Dr. Markus Müller, Milena Ramirez
Leibniz-Institut für Europäische Geschichte, Mainz

19. November 2019

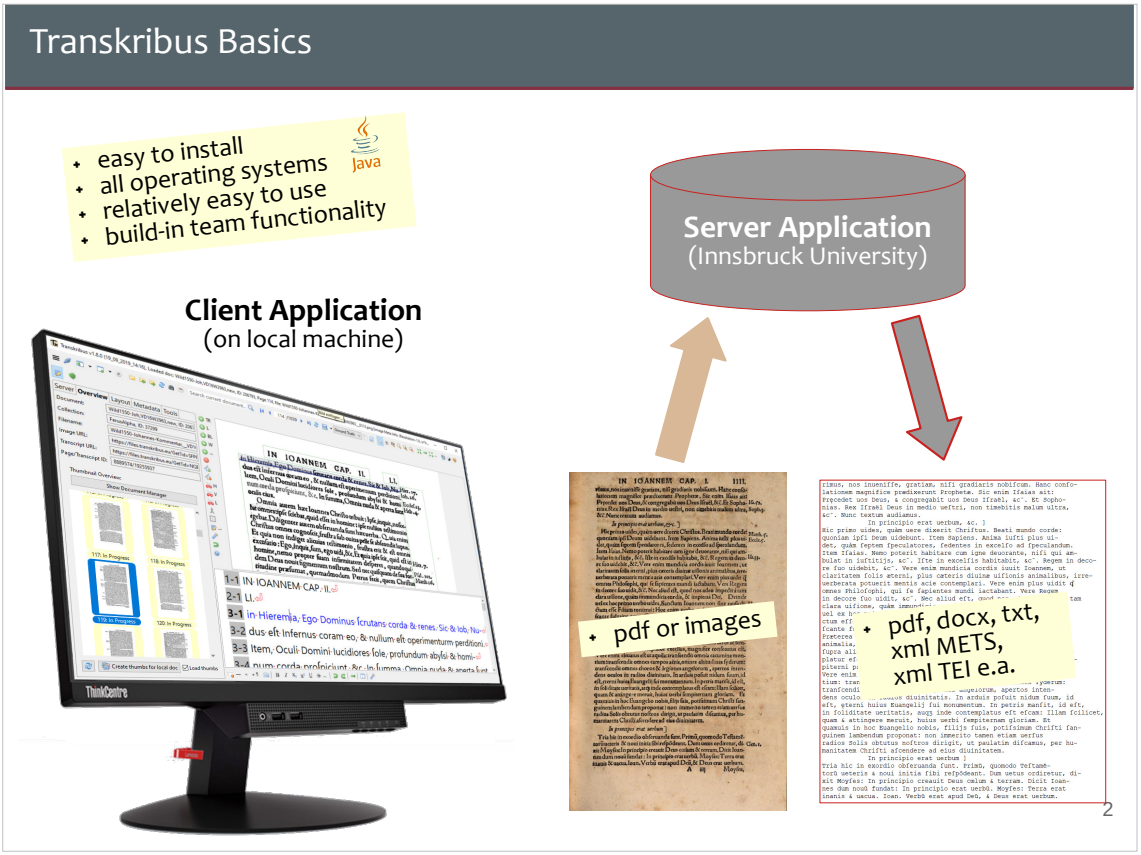
Initial problem:

In the 16th century, the inquisition in Spain, Italy and France censored the books of the Mainz cathedral preacher Johann Wild OFM. Interestingly, the actual words and sentences that were deleted or modified or inserted, differed between the different inquisitions, thus various censored versions of the same book were printed.

What exactly did the different censors change in the text? What were exactly are the differences between Spanish, Italian and French censorship?

In 2018, I began to search manually for these differences, but very soon, it became clear that I would never come to a reliable result because manual comparision of texts is extremely time-consuming and error-prone.

Therefore, I decided to invest my time into up-to-date OCR technology instead of manual search.

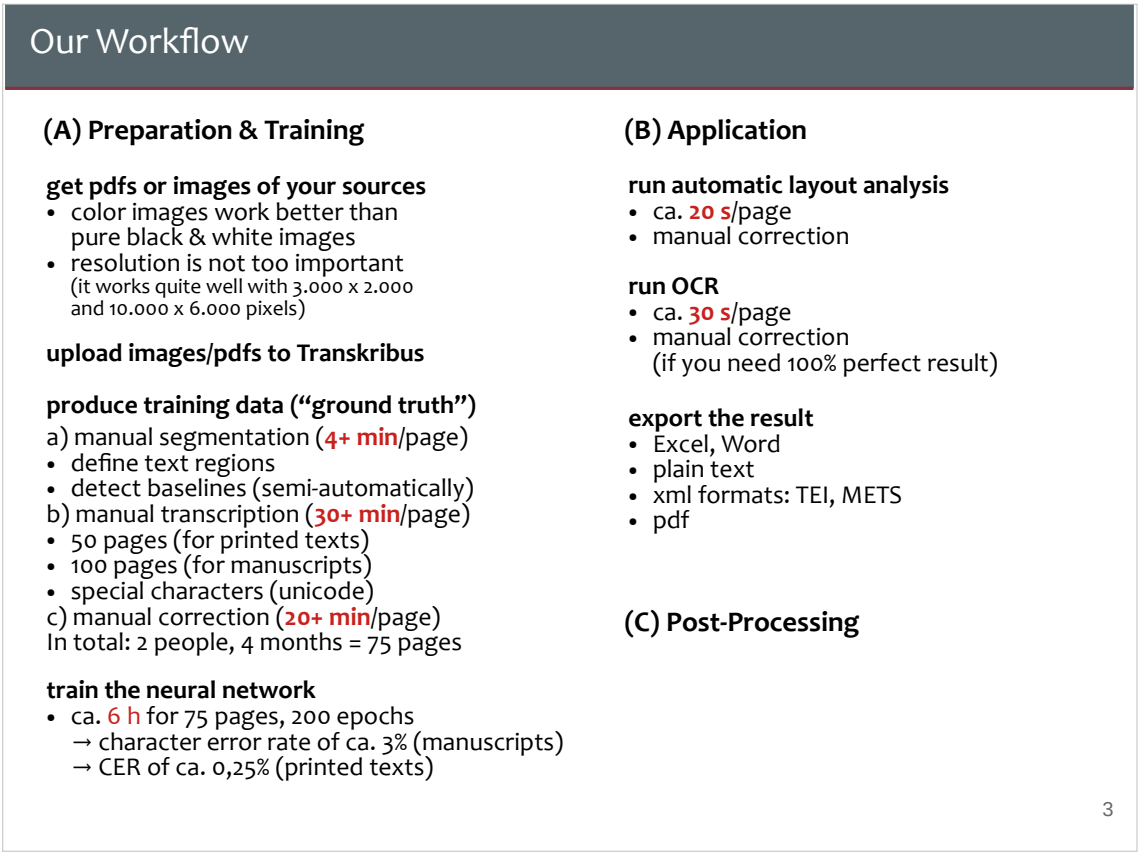


What is Transkribus?

- a server application for handwritten text recognition (HTR) running on servers at Innsbruck University.
- a very nice and easy to use client software for all steps of the OCR process.
- client software needs Java Runtime Environment
  therefore, it works on all operating systems without installation.
- Originally made for handwritten text recognition (HTR), i.e. manuscripts from the archives, it also works very well with printed texts.
- Basic facts about the Transkribus project as such on https://de.wikipedia.org/wiki/Transkribus

What do you need to get started?

- a bunch of images of a handwritten or printed text.
- a lot of time and patience.



The whole process took a lot of time and energy:

April–August 2019: build a workflow and manual transcription of 76 pages in a team of two with and mutual revision of the transcriptions

30 August 2019: training of a first model: >> character error rate: 0,23%

September/October 2019:transcription of another 20 pages of a second book and struggeling with the relatively poor layout analysis

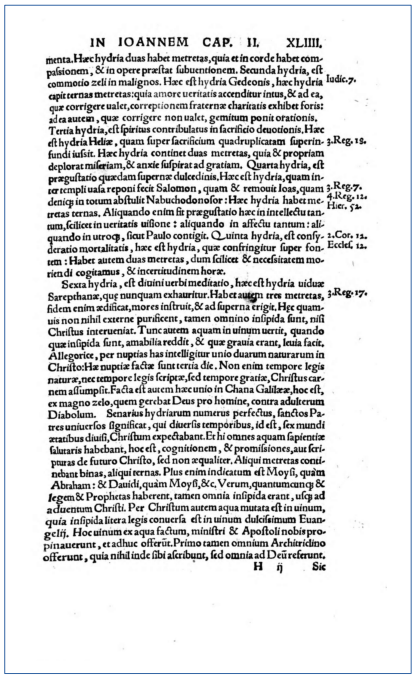
November 2019: second training with 20 pages: >> character error rate: 1,60%

Roadmap for the next months: transcribe and train more books of the same author and build a stable post-processing workflow, i.e. invest a lot of time into programming Python scripts.

„character error rate of xy%" = of all the characters on the page, xy% are wrong.

In 2018 , the same training still needed ca. 20 h and produced an error rate of ca. 10%. Thanks to better hardware and an improved architecture of the underlying neural network, the Innsbruck Transkribus team could reduce the time needed for training to ca. 6 h and improve the character error rate to under ca. 3% for manuscripts and under 1% for printed texts (2019).

Images



IN IOANNEM CAP. II. XLIIII.

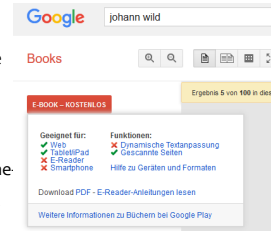
menta. Hæc hydria duas habet metretas, quia et in cor de habet compassionem, & in opere præstat subuentionem. Secunda hydria, est commotio zeli in malignos. Hæc est hydria Gedeonis, hæc hydria capit ternas metretas: quia amore ueritatis accenditur intus, & ad ea, quæ corrigere ualet, correptionem fraternæ charitatis exhibet foris: ad ea autem, quæ corrigere non ualet, gemitum ponit orationis. Tertia hydria, est spiritus contribulatus in sacrificio deuotionis. Hæc est hydria Heliæ, quam super sacrificium quadruplicatam superfundi iussit. Hæc hydria continet duas metretas, quia & propriam deplorat miseriam, & anxie suspirat ad gratiam. Quarta hydria, est prægustatio quædam supernæ dulcedinis. Hæc est hydria, quam inter templi uasa reponi fecit Salomon, quam & remouit Ioas, quam denicq; in totum abstulit Nabuchodonosor: Hæc hydria habet metretas ternas. Aliquando enim sit prægustatio hæc in intellectu tantum, scilicet in ueritatis uisione: aliquando in affectu tantum: aliquando in utroq;, sicut Paulo contigit. Quinta hydria, est consideratio mortalitatis, hæc est hydria, quæ confringitur super fontem: Habet autem duas metretas, dum scilicet & necessitatem morien di cogitamus, & incertitudinem horæ.

Sexta hydria, est diuini uerbi meditatio, hæc est hydria uiduæ Sarepthanæ, quę nunquam exhauritur. Habet autem tres metretas, fidem enim ædificat, mores instruit, & ad superna erigit. Hę quamuis non nihil externe purificent, tamen omnino insipida sunt, nisi Christus interueniat. Tunc autem aquam in uinum uertit, quando quæ insipida sunt, amabilia reddit, & quæ grauia erant, leuia facit. Allegorice, per nuptias has intelligitur unio duarum naturarum in Christo: Hæ nuptiæ factæ sunt tertia die. Non enim tempore legis naturæ, nec tempore legis scriptæ, sed tempore gratiæ, Christus carnem assumpsit. Facta est autem hæc unio in Chana Galilææ, hoc est, ex magno zelo, quem gerebat Deus pro homine, contra adulterum Diabolum. Senarius hydriarum numerus perfectus, sanctos Patres uniuersos significat, qui diuersis temporibus, id est, sex mundi ætatibus diuisi, Christum expectabant. Et hi omnes aquam sapientiæ saluatoris habebant, hoc est, cognitionem, & promissiones, aut scripturas de futuro Christo, sed non æqualiter. Aliqui metretas continebant binas, aliqui ternas. Plus enim indicatum est Moysi, quàm Abraham: & Dauidi, quàm Moysi, &c. Verum, quantumcunq; & legem & Prophetas haberent, tamen omnia insipida erant, usq; ad aduentum Christi. Per Christum autem aqua mutata est in uinum, quia insipida litera legis conuersa est in uinum dulcissimum Euangelij. Hoc uinum ex aqua factum, ministri & Apostoli nobis propinauerunt, et adhuc offerūt. Primo tamen omnium Architriclino offerunt, quia nihil inde sibi ascribunt, sed omnia ad Deū referunt.

Iudic. 7.
3. Reg. 18.
3. Reg. 7.
4. Reg. 12.
Hier. 52.
2. Cor. 12.
Eccles. 12.
3. Reg. 17.

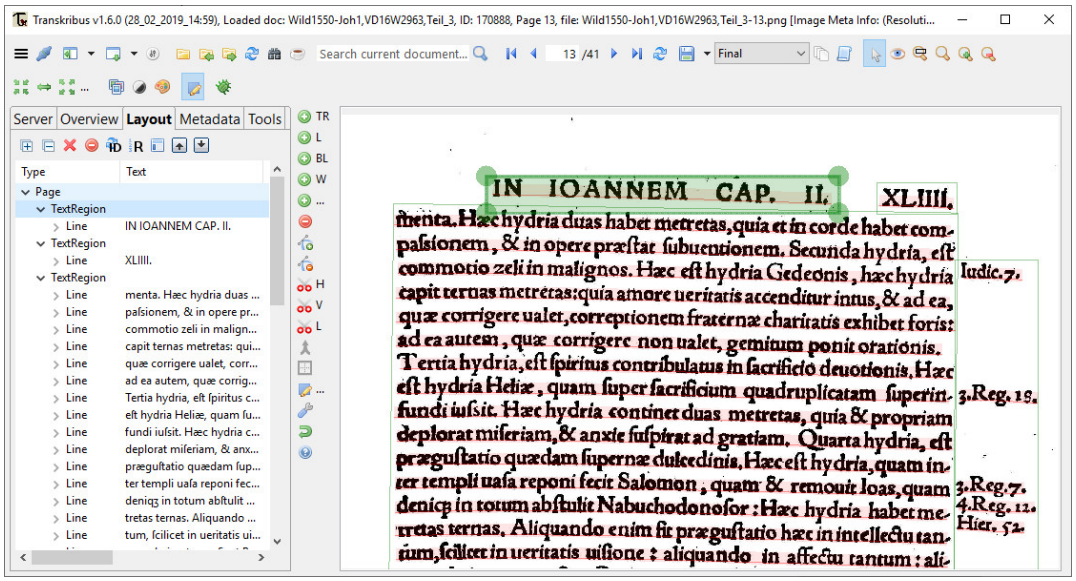
H ij Sic

- Download pdfs from Google Books (only available for copyright-free stuff)
  (Hover with the cursor over the „e books“ button on the left. A context dialog opens up. Click on „Download PDF“ in the context dialog. The pdf will be downloaded immediately and opens in your browser or in a local pdf reader.)
  Pro: super easy.
  Contra: only black & white, sometimes bad image quality.



- Download high quality pdfs or images from library websites
  (e.g. https://www.digitale-sammlungen.de/)
  Pro: best image quality available
  Contra: often sophisticated user interfaces, sometimes the same images as in Google Books.
- Produce your own digital images.
  (e.g. with your smartphone or even with specialized tools like https://scantent.eu/en/)
  Pro: maximum control etc.
  Contra: you need the technical skills an the time to do it well; may violate copyright rules or personal rights.

Segmentation Mode

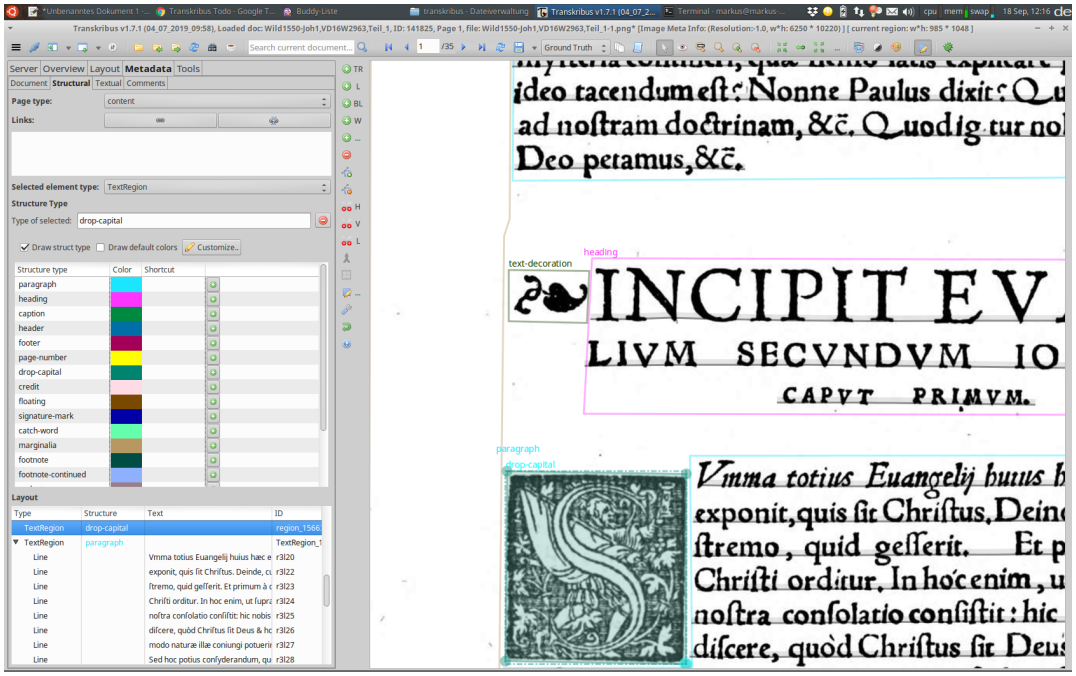


5

There are two maior „modes“ in Transkribus:

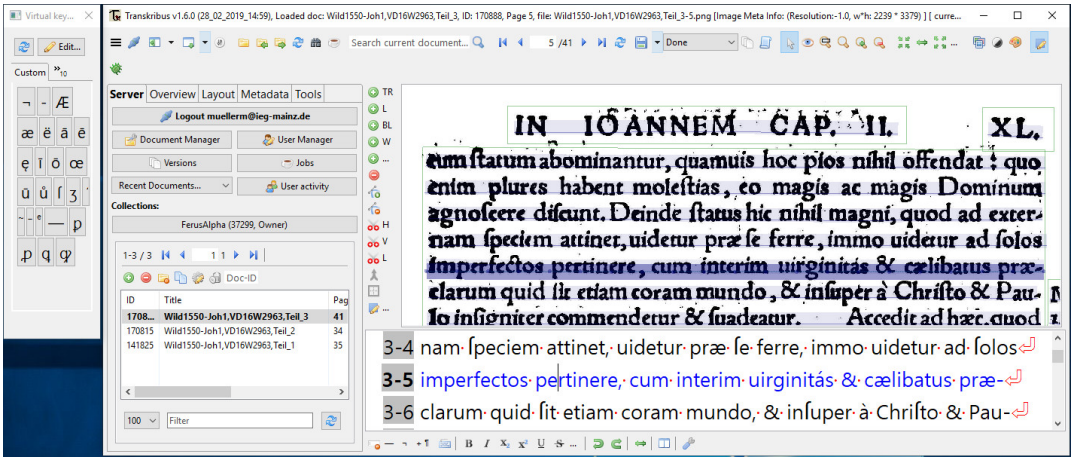
(1) Segmentation mode: By drawing rectangles and lines, you tell the software where the different text blocks are situated on the page.

Structural Metadata: Text Regions



You can add metadata to the different text regions (as well as to many other parts of the dataset, e.g. metadata of the document itself, metadata for documentation of the transcription standards etc.).

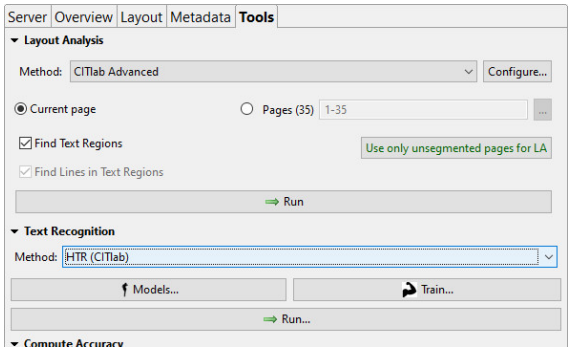
Transcription Mode



(2) Transcription mode: before the software can recognize text, it has to be trained with sample data. That means, you have to transcribe a lot of pages manually (ca. 50 pages for printed texts, 100 pages for manuscripts).

The Transkribus interface makes transcription very easy: The corresponding line is highlighted on the photo, you can define shortcuts for special characters etc.

Tools for Layout Analysis, Training and Text Recognition



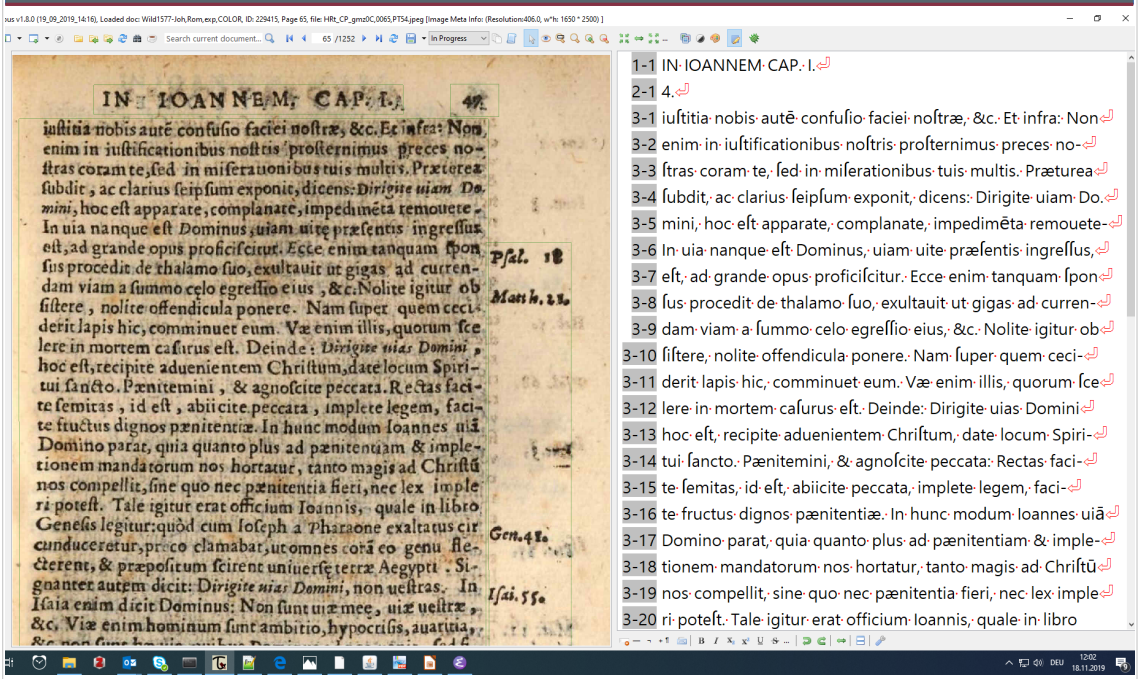
As soon as the transcription phase is completed, you can **train a „model“**, i.e. the software will try to figure out how the images and the given text („ground truth“) are connected. As soon as the software has found a reliable and stable way to come from the image to the desired text, the „training“ process is completed. The set of rules that the software produced is called „model“.

If you provide the software with another image of a similar kind that it has never „seen“ before, it should be able to „read“ the text on the image without any „help“ or intervention by the user.

Having finished the training, you can „run“ the OCR with the pre-trainied model on the rest of the book and Transkribus will automatically transcribe the text (in my case: another 1500 pages).

Unfortunately, the automatic layout analysis does not recognize the marginalia as an separate text region, thus still have to invest 3–5 min per page to correct the layout before you can acutally run the OCR.

Result: A (nearly) Perfect Text!



IN IOANNEM. CAP. I. 47

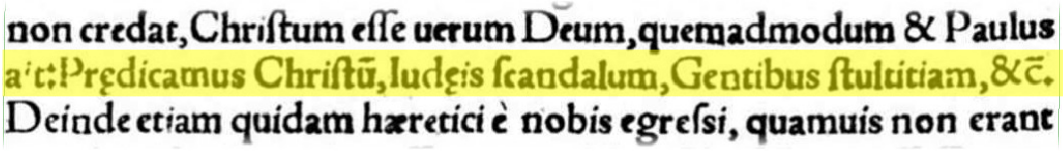
iuſtitia nobis autē confuſio faciei noſtræ, &c. Et infra: Non enim in iuſtificationibus noſtris proſternimus preces noſtras coram te, ſed in miſerationibus tuis multis. Præterea ſubdit, ac clarius ſeipſum exponit, dicens: *Dirigite uiam Domini*, hoc eſt apparate, complanate, impedimēta remouete. In uia nanque eſt Dominus, uiam uitę præſentis ingreſſus eſt, ad grande opus proficiſcitur. Ecce enim tanquam ſponſus procedit de thalamo ſuo, exultauit ut gigas ad currendam viam a ſummo cælo egreſſio eius, &c. Nolite igitur obſiſtere, nolite offendicula ponere. Nam ſuper quem ceciderit lapis hic, comminuet eum. Væ enim illis, quorum ſcelere in mortem caſurus eſt. Deinde: *Dirigite uias Domini*, hoc eſt, recipite aduenientem Chriſtum, date locum Spiritui ſancto. Pœnitemini, & agnoſcite peccata. Rectas facite ſemitas, id eſt, abiicite peccata, implete legem, facite fructus dignos pœnitentiæ. In hunc modum Ioannes uiā Domino parat, quia quanto plus ad pœnitentiam & impletionem mandatorum nos hortatur, tanto magis ad Chriſtū nos compellit, ſine quo nec pœnitentia fieri, nec lex impleri poteſt. Tale igitur erat officium Ioannis, quale in libro Geneſis legitur: quòd cum Ioſeph a Pharaone exaltatus circumduceretur, præco clamabat, ut omnes corā eo genu flecterent, & præpoſitum ſcirent uniuerſę terræ Aegypti. Signanter autem dicit: *Dirigite uias Domini*, non ueſtras. In Iſaia enim dicit Dominus: Non ſunt uiæ meę, uiæ ueſtræ, &c. Viæ enim hominum ſunt ambitio, hypocriſis, auaritia,

Pſal. 18
Matth. 23.
Gen. 42.
Iſai. 55.

```
1-1 IN·IOANNEM·CAP.·I.
2-1 4.
3-1 iuſtitia·nobis·autē·confuſio·faciei·noſtræ,·&c.·Et·infra:·Non
3-2 enim·in·iuſtificationibus·noſtris·proſternimus·preces·no-
3-3 ſtras·coram·te,·ſed·in·miſerationibus·tuis·multis.·Præterea
3-4 ſubdit,·ac·clarius·ſeipſum·exponit,·dicens:·Dirigite·uiam·Do.
3-5 mini,·hoc·eſt·apparate,·complanate,·impedimēta·remouete-
3-6 In·uia·nanque·eſt·Dominus,·uiam·uite·præſentis·ingreſſus,
3-7 eſt,·ad·grande·opus·proficiſcitur.·Ecce·enim·tanquam·ſpon
3-8 ſus·procedit·de·thalamo·ſuo,·exultauit·ut·gigas·ad·curren-
3-9 dam·viam·a·ſummo·celo·egreſſio·eius,·&c.·Nolite·igitur·ob
3-10 ſiſtere,·nolite·offendicula·ponere.·Nam·ſuper·quem·ceci-
3-11 derit·lapis·hic,·comminuet·eum.·Væ·enim·illis,·quorum·ſce
3-12 lere·in·mortem·caſurus·eſt.·Deinde:·Dirigite·uias·Domini
3-13 hoc·eſt,·recipite·aduenientem·Chriſtum,·date·locum·Spiri-
3-14 tui·ſancto.·Pœnitemini,·&·agnoſcite·peccata:·Rectas·faci-
3-15 te·ſemitas,·id·eſt,·abiicite·peccata,·implete·legem,·faci-
3-16 te·fructus·dignos·pœnitentiæ.·In·hunc·modum·Ioannes·uiā
3-17 Domino·parat,·quia·quanto·plus·ad·pœnitentiam·&·imple-
3-18 tionem·mandatorum·nos·hortatur,·tanto·magis·ad·Chriſtū
3-19 nos·compellit,·ſine·quo·nec·pœnitentia·fieri,·nec·lex·imple
3-20 ri·poteſt.·Tale·igitur·erat·officium·Ioannis,·quale·in·libro
```

The resulting text produced by Transkribus has only very few errors. The model used here has been trained with 30 pages of the same and 75 pages of a similar book.

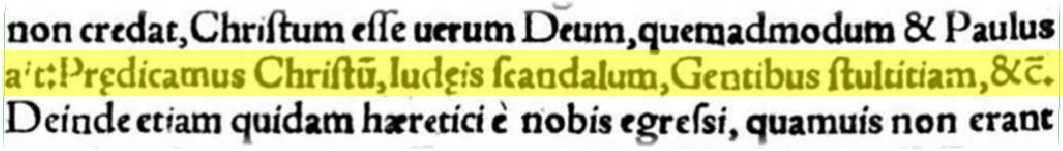
Pitfalls During Transcription


non credat, Chriſtum eſſe uerum Deum, quemadmodum & Paulus
a't: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c̄.
Deinde etiam quidam hæretici è nobis egreſsi, quamuis non erant

a't: Prędicamus Chirſtū, Iudeis ſcandalum, Gentibus stultitiam, &c̄.

How many errors can you spot in this transcription?

Pitfalls During Transcription


non credat, Chriſtum eſſe uerum Deum, quemadmodum & Paulus
a't: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c.
Deinde etiam quidam hæretici è nobis egreſsi, quamuis non erant

x a't: Prędicamus Chirſtū, Iudeis ſcandalum, Gentibus stultitiam, &c.

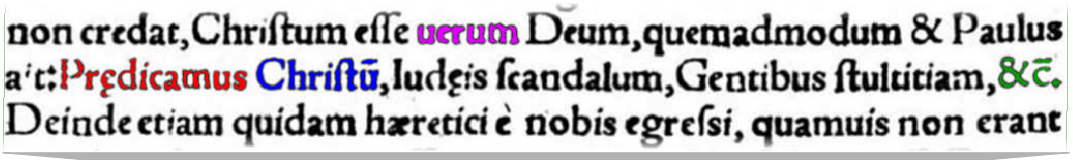
√ ait: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c.

Nota bene: **Errors in the transcription will scale up!**
→ the neural net will 'learn' every single mistake you make
→ OCR will produce erroneous output

It would be so nice to have a Latin spell-checker (for early modern Latin, including all the abbreviations ;-) …

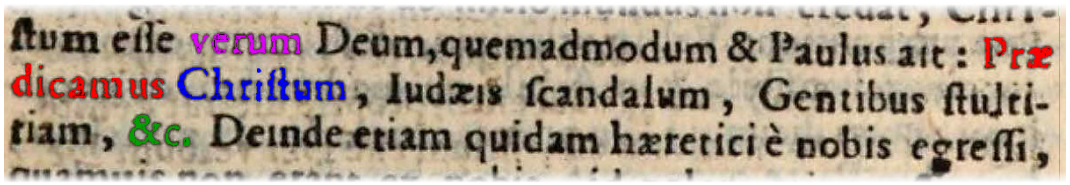
Why OCR alone is not enough…

Version A printed in Mainz, 1550


non credat, Chriſtum eſſe uerum Deum, quemadmodum & Paulus
ait: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c.
Deinde etiam quidam hæretici è nobis egreſsi, quamuis non erant

ait: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c̄.

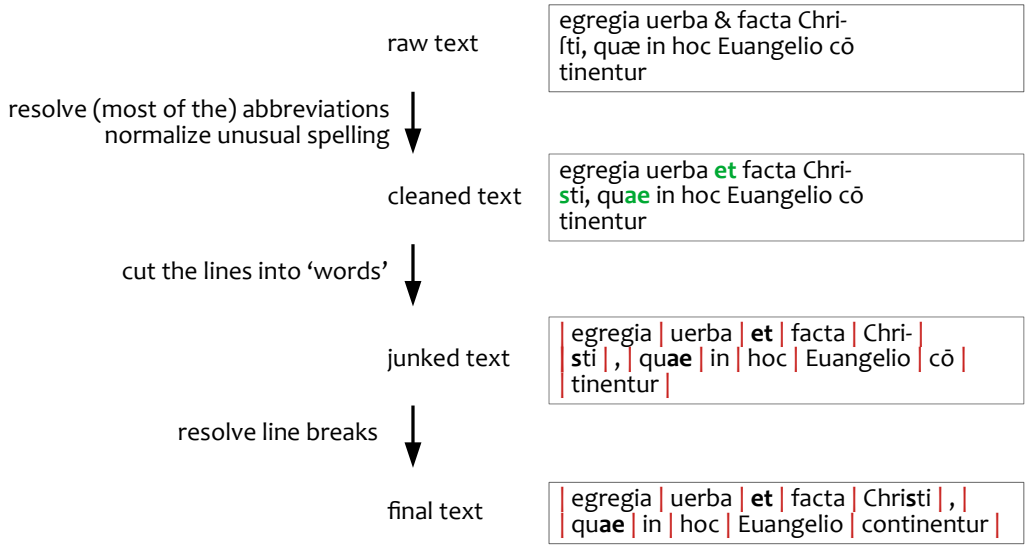
ait: Præ
dicamus Chriſtum, Iudæis ſcandalum, Gentibus ſtulti-
tiam, &c.


ſtum eſſe verum Deum, quemadmodum & Paulus ait: Præ
dicamus Chriſtum, Iudæis ſcandalum, Gentibus ſtulti-
tiam, &c. Deinde etiam quidam hæretici è nobis egreſſi,

Version B printed in Rome, 1577

I want to compare one version of the text with another version. Due to the different usage of abbreviations and different spelling, the text produced by Transkribus has be post-processed in order to be comparable.

Post-Processing to make the texts comparable

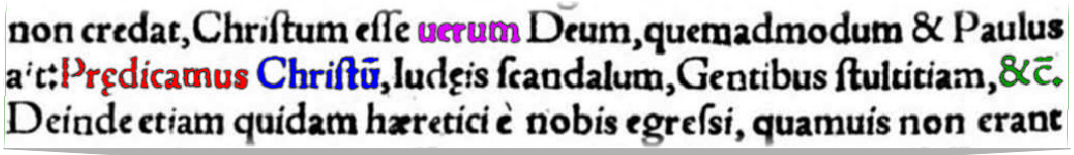


ca. 800 lines of python code

Post-processing cannot be done with Transkribus. At the moment, I try to tackle the problem with a Python script.

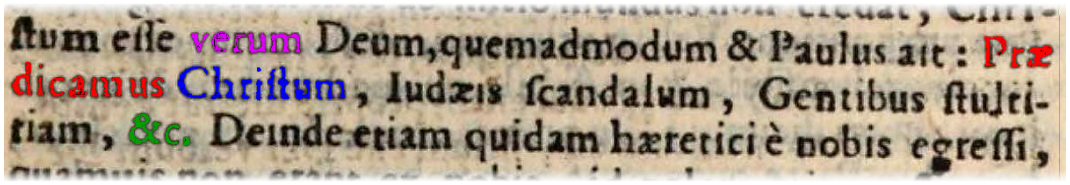
After post-proceſſing: ꝑfectly ſearchable & cōparable texts

Version A printed in Mainz, 1550


non credat, Chriſtum eſſe uerum Deum, quemadmodum & Paulus
ait: Prędicamus Chriſtū, Iudęis ſcandalum, Gentibus ſtultitiam, &c.
Deinde etiam quidam hæretici è nobis egreſsi, quamuis non erant

non credat, Christum esse **uerum** Deum, quemadmodum et Paulus ait: **Praedicamus** **Christum**, Iudaeis scandalum, Gentibus stultitiam, **etc.** Deinde etiam quidam haeretici e nobis egressi, quamuis non erant

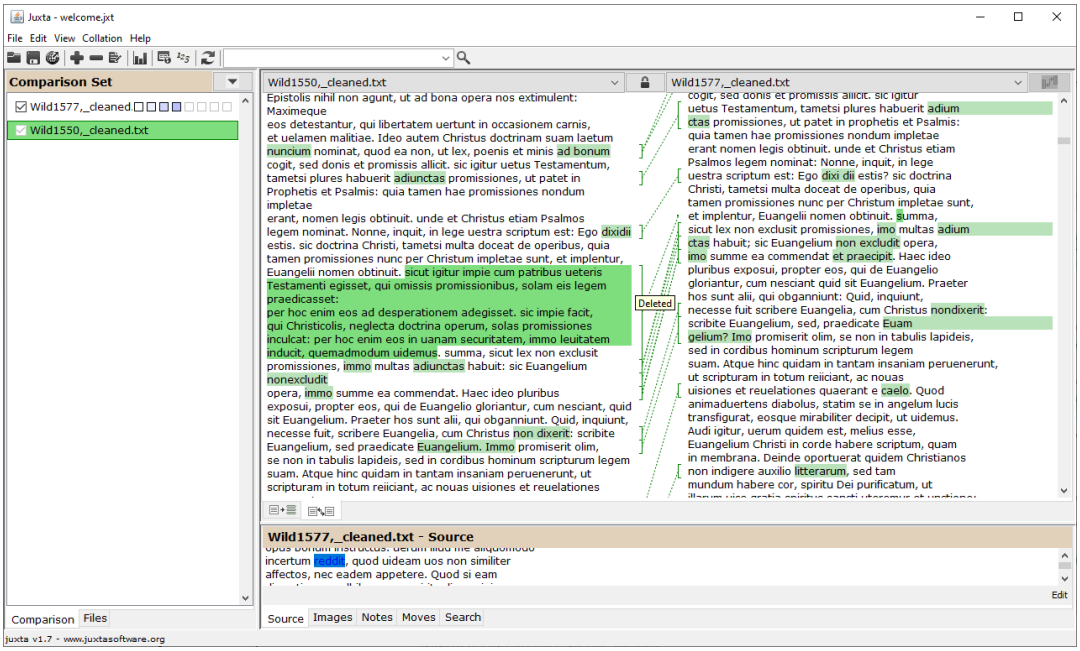
esse **uerum** Deum, quemadmodum et Paulus ait: **Praedicamus** **Christum**, Iudaeis scandalum, Gentibus stultitiam, **etc.** Deinde etiam quidam haeretici e nobis egressi,


ſtum eſſe verum Deum, quemadmodum & Paulus ait: Præ
dicamus Chriſtum, Iudæis ſcandalum, Gentibus ſtulti-
tiam, &c. Deinde etiam quidam hæretici è nobis egreſſi,

Version B printed in Rome, 1577

The Python script resolves the abbreviations and normalizes the text and, finally, the data for text comparison is ready.

Use Case: Compare Text Versions to Find Censorship



Comparision of two versions of the same text produced with Transkribus to find differences, especially censorship. (Using plain text files with the free Juxta tool.)

As you can see, there are still a lot of „false positives“, i.e. words that are marked as different in both texts not because they were acutally censored but because they were spelled differently, or due to an error in the transcription/OCR process, or due to an error made by the printer in the 16th century (believe me, they made quite a lot of errors!).

There is still a lot of work to be done to improve the post-processing as well as the comparison itself…

Further Information

Information

- Transkribus website: https://transkribus.eu/Transkribus/
- Transkribus wiki: English: https://transkribus.eu/wiki/index.php/Main_Page
  German: https://transkribus.eu/wikiDe/index.php/Hauptseite
- Article on medieval manuscript OCR with Transkribus (in German): https://mittelalter.hypotheses.org/21828

Ideas

- Random selection of the pages to be transcribed will cover more 'special cases' and exceptions than transcribing whole chapters.
- Use state-of-the-art spell checking technology to reduce errors during transcription and to speed up the post-processing of the OCR output of Transkribus.
- Further automation: Use the Transkribus REST api to download OCR data for post-processing and to upload the results of automated post-processing.