Digitization of written sources

Federico Boschetti
federico.boschetti@ilc.cnr.it
CNR-ILC & VeDPH
June 8, 2022 - University of Pisa
Summer School - Digital Tools for Humanists

Introduction

The importance of OCR and HTR

The Optical Character Recognition (OCR) is a bottleneck in many activities that need large quantities of legacy information:

- digital libraries
- corpus linguistics
- digital history
- …

The importance of OCR and HTR

Nowadays OCR can perform 99% of accuracy on recent, good quality printed editions and it can reach 98% of accuracy on challenging printed documents

The new field of Handwritten Text Recognition is very promising, so that libraries, universities and other institutions (such as state archives) are planning to acquire the digital text not only from printed documents but also from manuscripts

Acquisition and pre-processing of digital images

Digital images and digital texts

Scanning is the process of acquiring information from two-dimensional or three-dimensional objects, in order to create digital images

Different operations can be performed on digital images of a document and digital texts:

- crop an arbitrary part
- change brightness and contrast
- compare the high fidelity of the layout and of the figures to the original manuscript or printed edition
- ...

- copy and paste it
- search it
- tokenize it
- count the tokens
- make indexes
- …

Scanners

Various kinds of scanners are available, but a simple flatbed scanner can be enough, if the document is not fragile. The coplanarity of the written surface of the document with the moving carriage of the scanner has a high impact on the accuracy of the recognition

Scanners

Currently documents are acquired also by smartphones, but the quality of the acquisition is poor, compared to a flatbed scanner

DPI and PPI

DPI means Dots per Inch and PPI means Pixels per Inch. In order to have an accurate OCR, 600 DPI are optimal, but 300 DPI can be acceptable.

The preservation of the master images and metadata

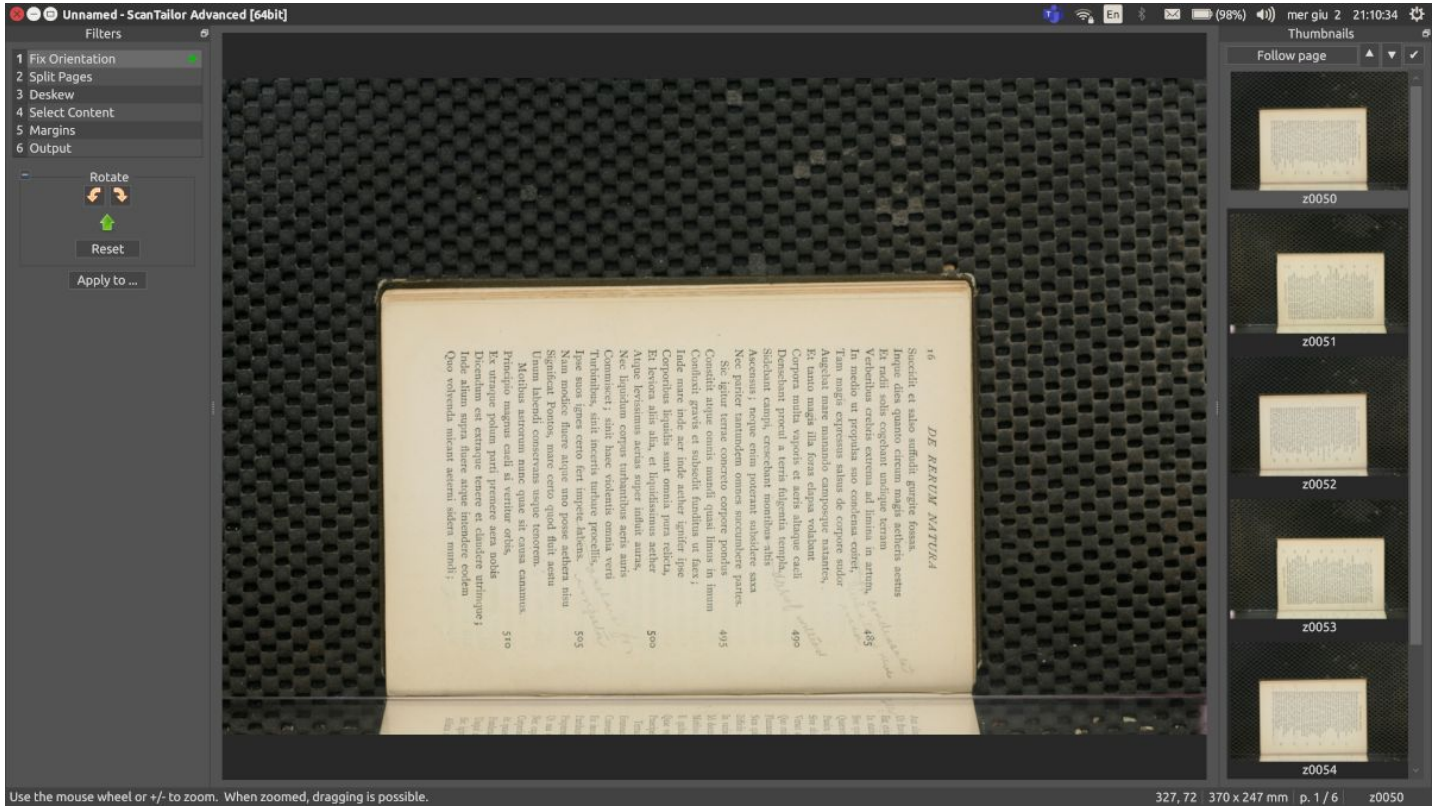
Along the digital text acquisition workflow one or more image elaborations are required. It is necessary to keep always the original images and possibly to preserve also the metadata related to the necessary transformations and use naming conventions for the files, with minimal metadata about dpi, color, etc.

OCR (or HTR) preprocessing on images

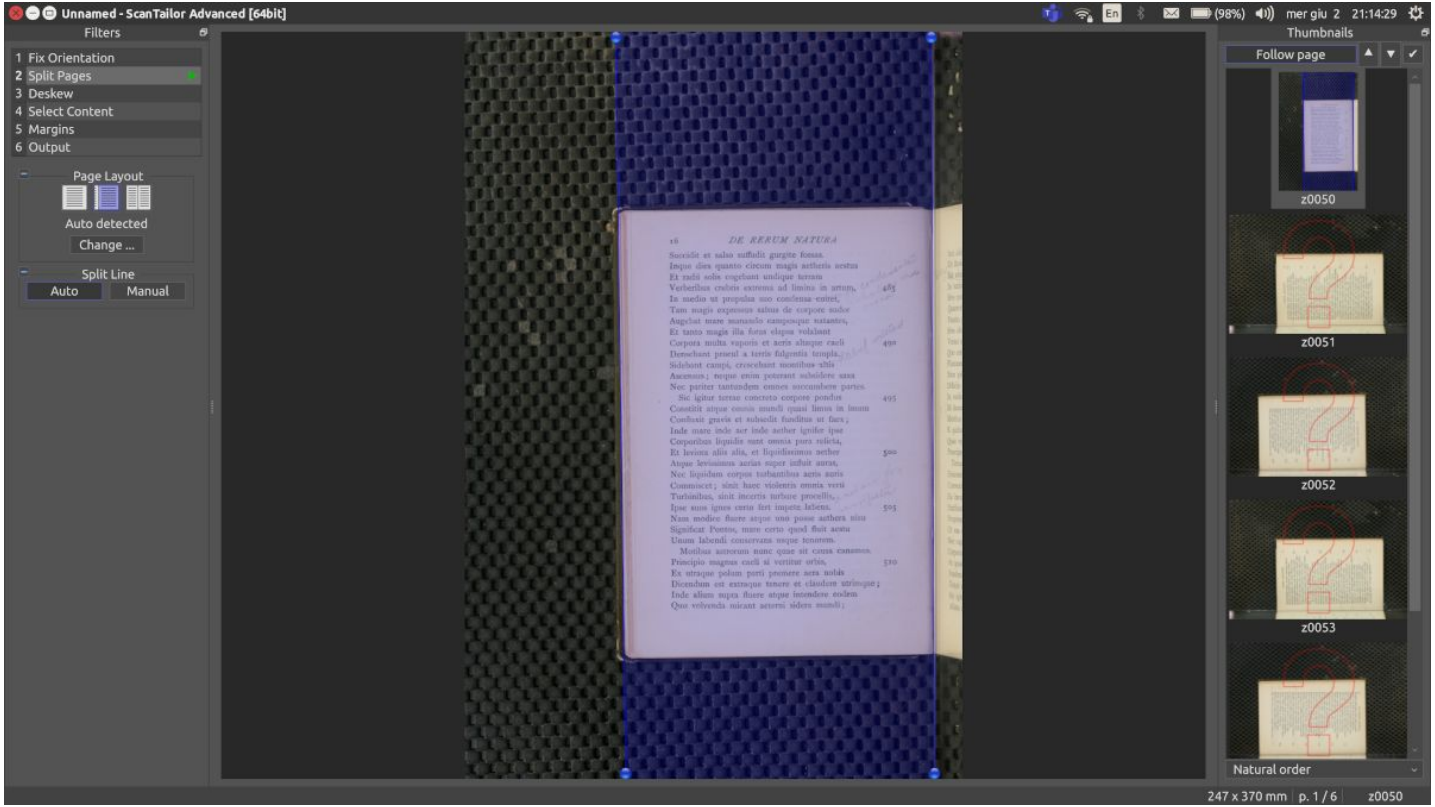
In order to improve the accuracy of OCR or HTR, images must be processed at least with the following operations:

- fixing orientation (if necessary)
- splitting pages (if two pages have been scanned together)
- deskewing (i.e. small rotation)
- selecting content
- adding margins
- change the output resolution (if necessary)
- binarization
- dewarping

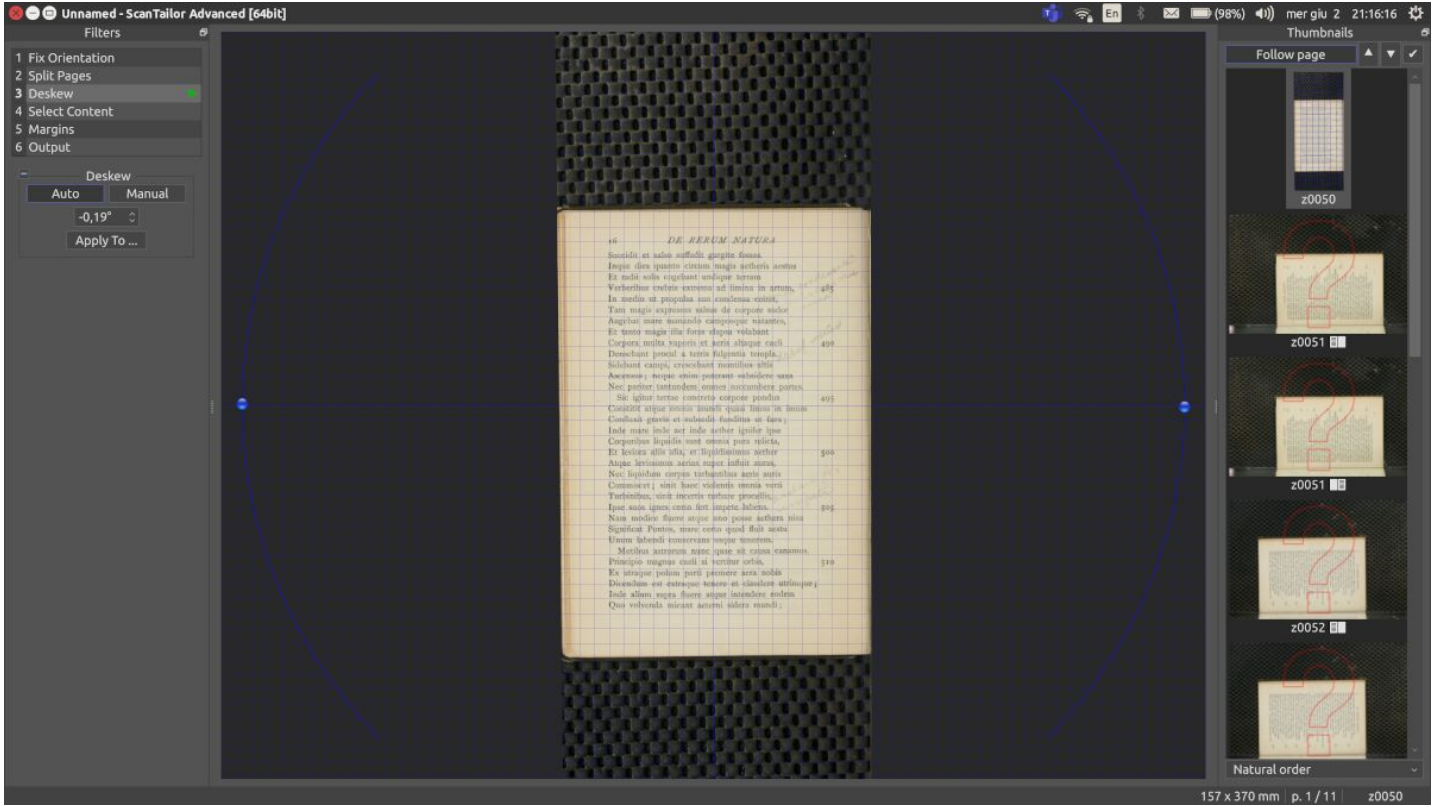
Fixing orientation



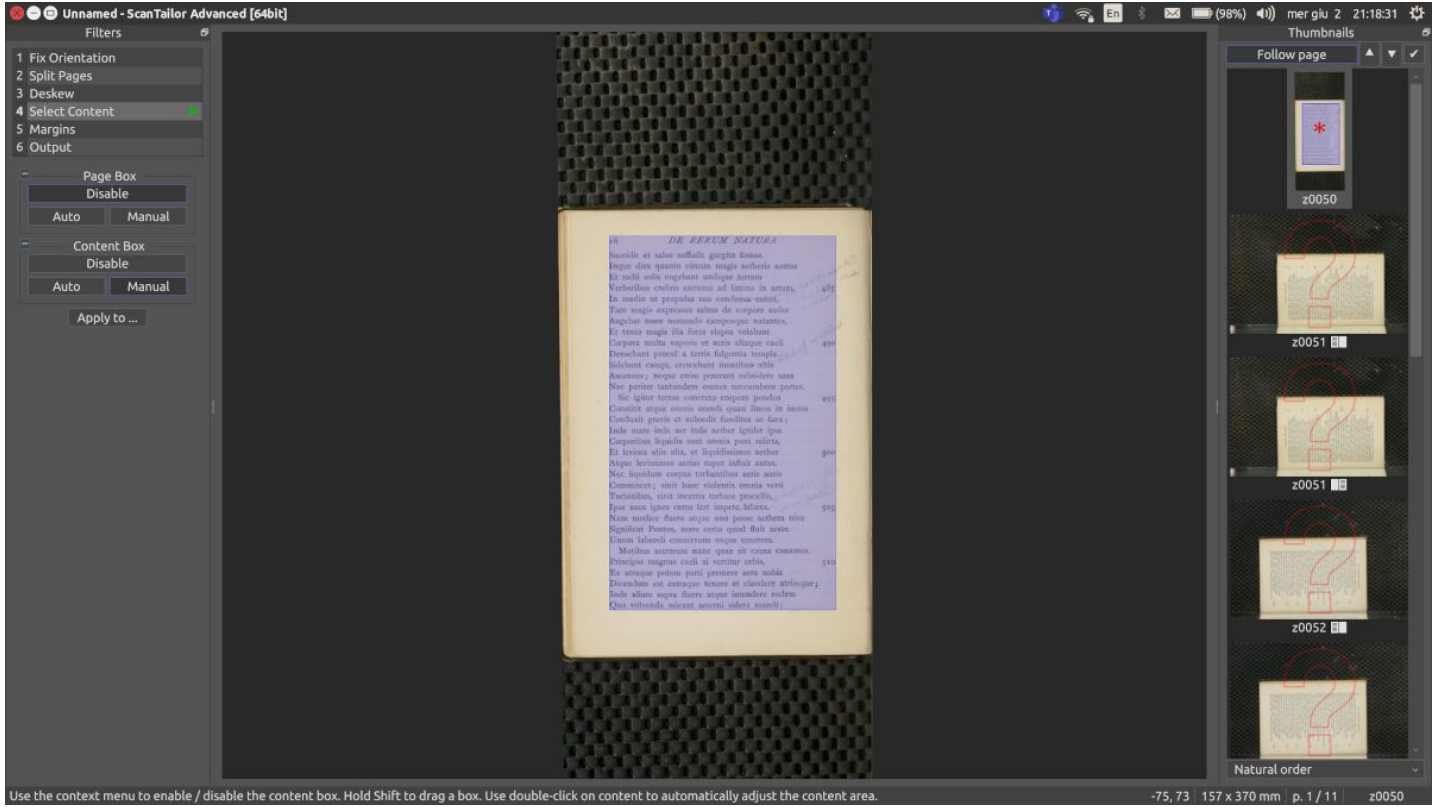
Splitting pages



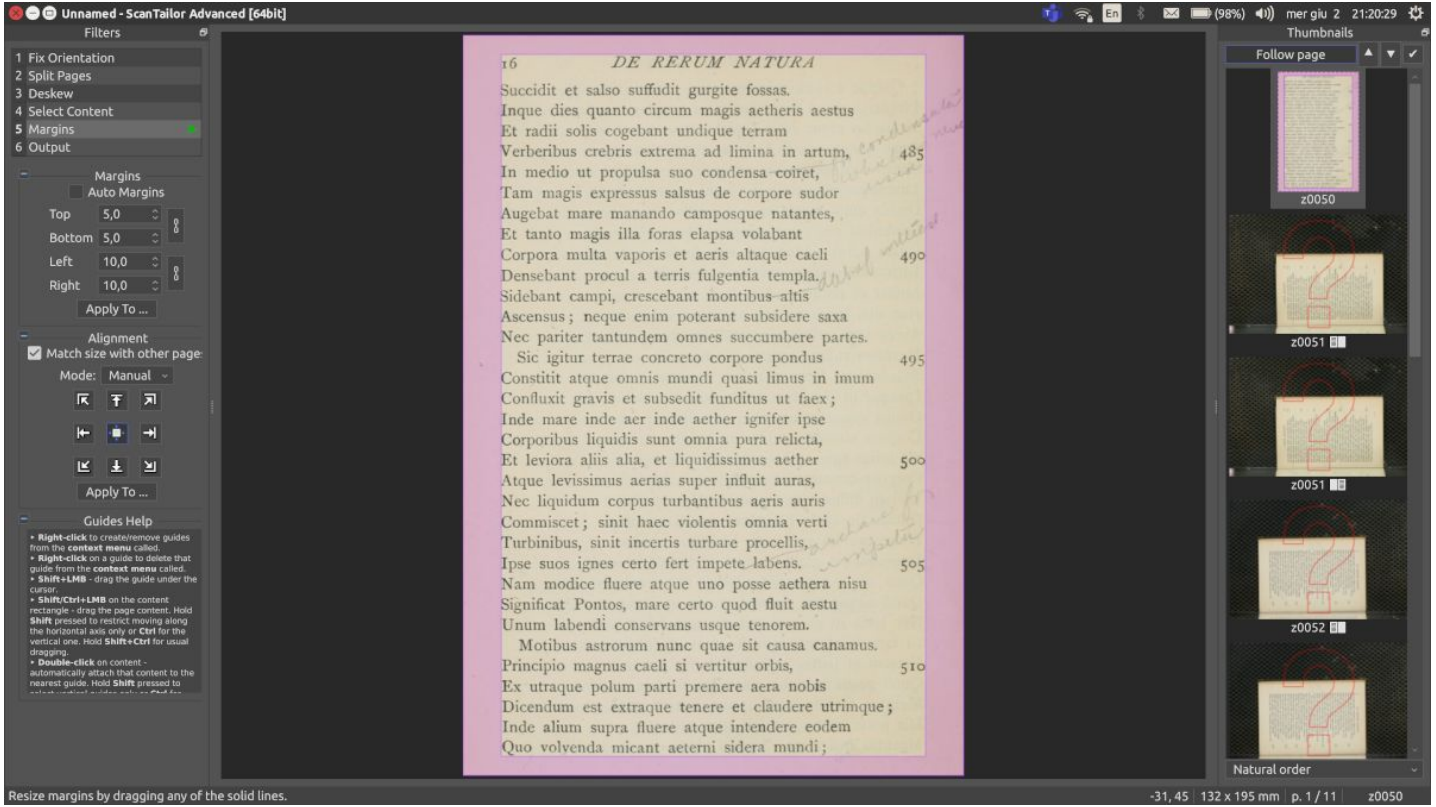
Deskewing


Unnamed - ScanTailor Advanced [64bit]
Filters
1 Fix Orientation
2 Split Pages
3 Deskew
4 Select Content
5 Margins
6 Output
Deskew
Auto
Manual
-0,19°
Apply To ...
Thumbnails
Follow page
z0050
z0051
z0051
z0052
Natural order
157 x 370 mm
p. 1 / 11
z0050

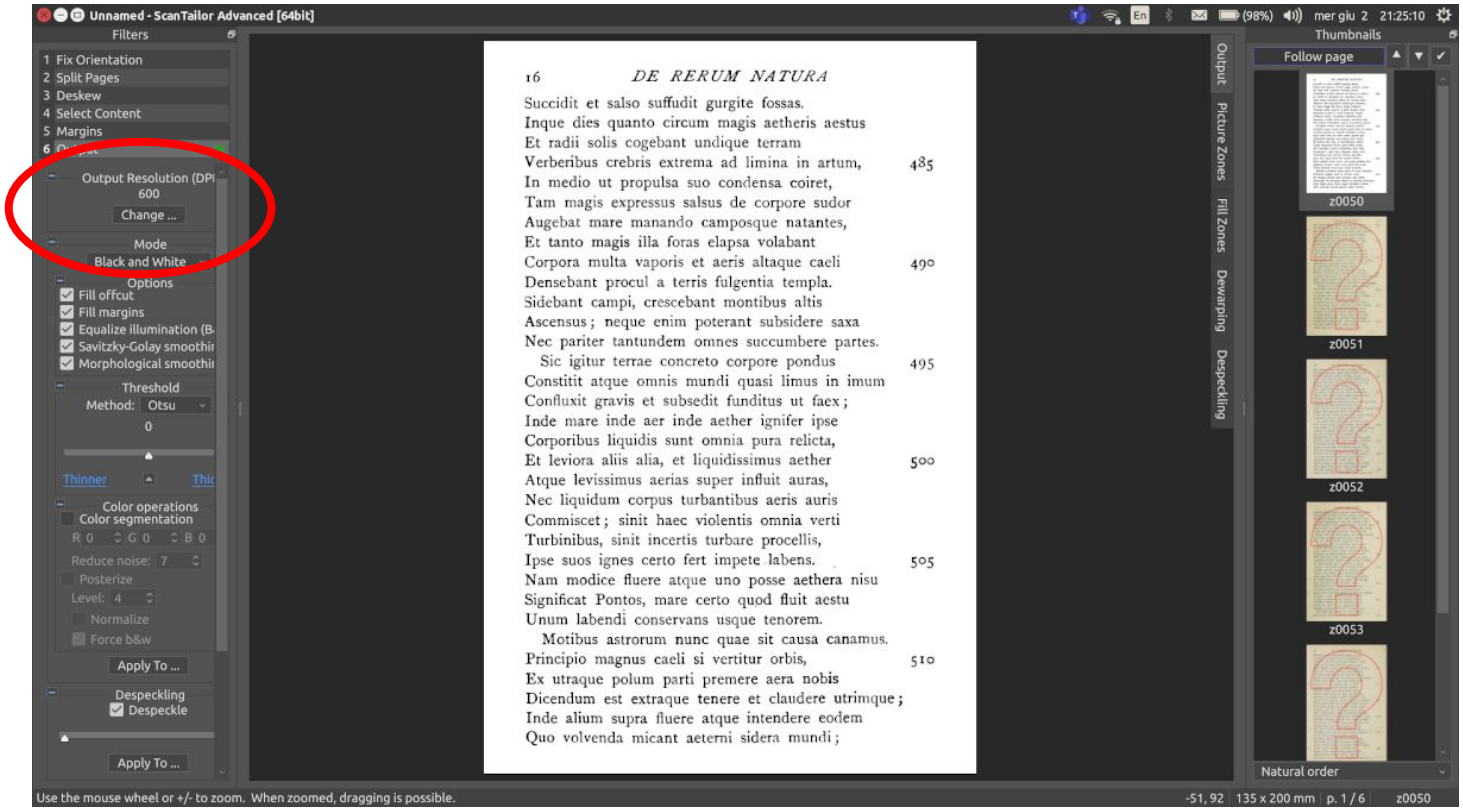
Selecting content



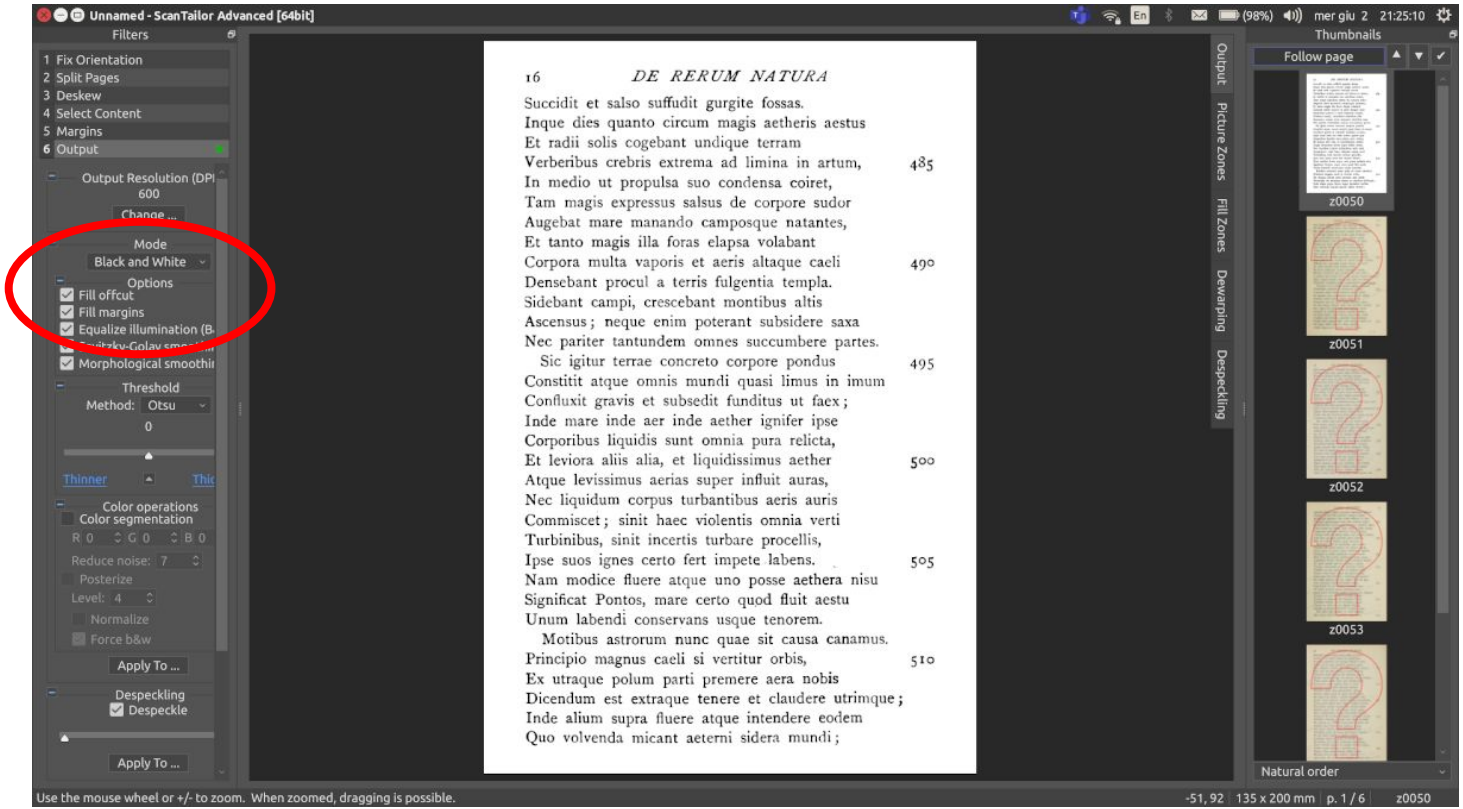
Adding margins



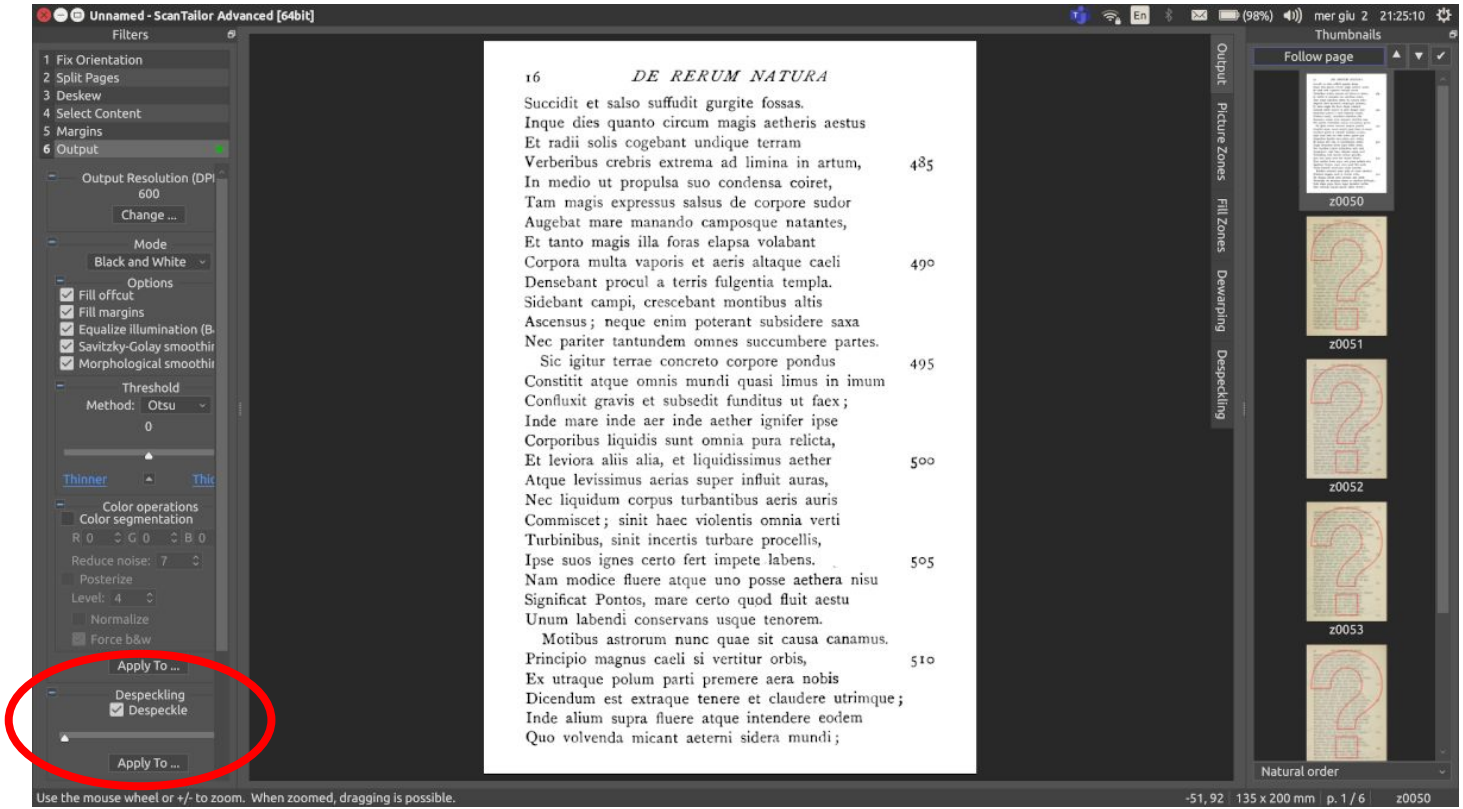
Changing resolution



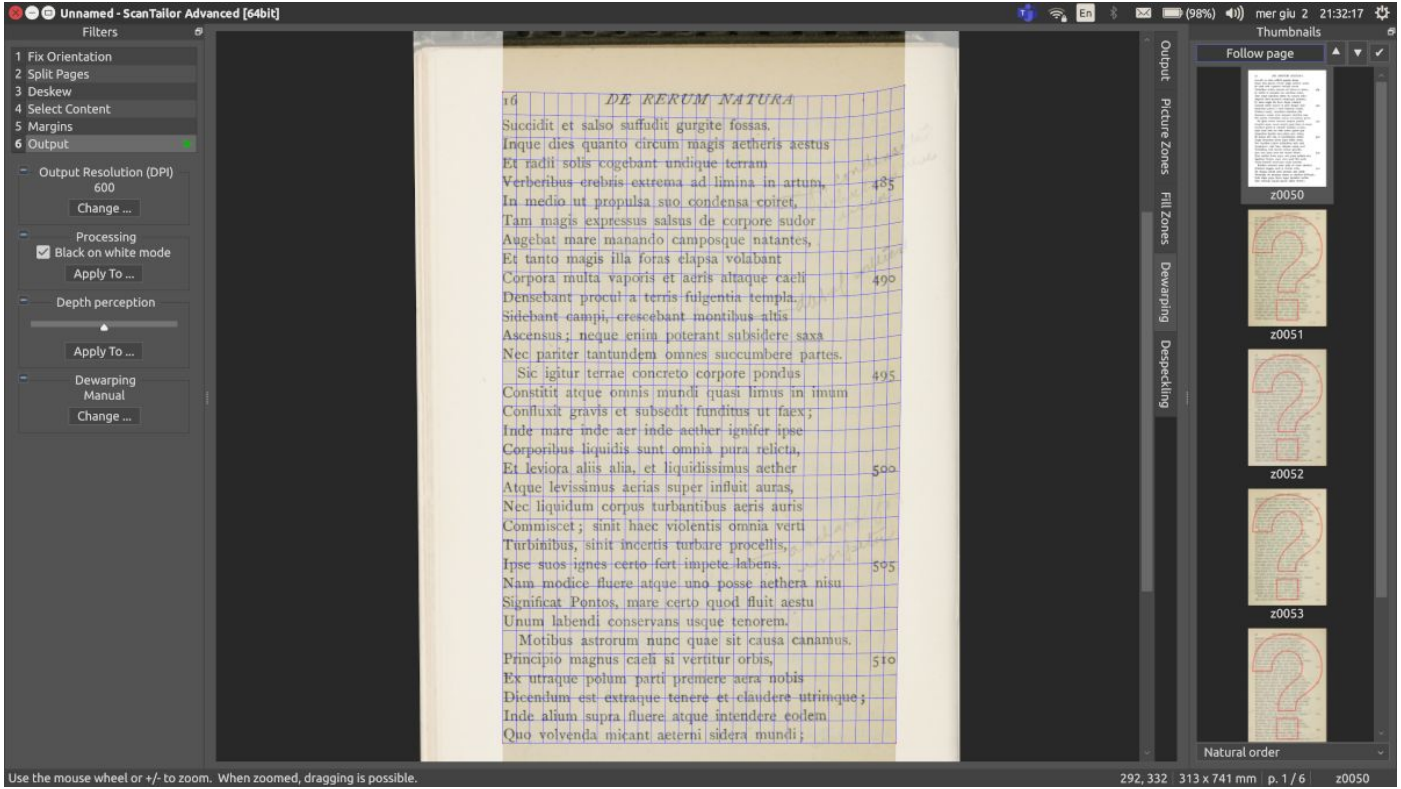
Binarization



Despeckling



Dewarping


Unnamed - ScanTailor Advanced [64bit]
Filters
1 Fix Orientation
2 Split Pages
3 Deskew
4 Select Content
5 Margins
6 Output
Output Resolution (DPI)
600
Change ...
Processing
Black on white mode
Apply To ...
Depth perception
Apply To ...
Dewarping
Manual
Change ...
16 DE RERUM NATURA
Succidit et salso suffudit gurgite fossas.
Inque dies quanto circum magis aetheris aestus
Et radii solis cogebant undique terram
Verberibus crebris extrema ad limina in artum, 485
In medio ut propulsa suo condensa coiret,
Tam magis expressus salsus de corpore sudor
Augebat mare manando camposque natantes,
Et tanto magis illa foras elapsa volabant
Corpora multa vaporis et aeris altaque caeli 490
Densebant procul a terris fulgentia templa.
Sidebant campi, crescebant montibus altis
Ascensus; neque enim poterant subsidere saxa
Nec pariter tantundem omnes succumbere partes.
Sic igitur terrae concreto corpore pondus 495
Constitit atque omnis mundi quasi limus in imum
Confluxit gravis et subsedit funditus ut faex;
Inde mare inde aer inde aether ignifer ipse
Corporibus liquidis sunt omnia pura relicta,
Et leviora aliis alia, et liquidissimus aether 500
Atque levissimus aerias super influit auras,
Nec liquidum corpus turbantibus aeris auris
Commiscet; sinit haec violentis omnia verti
Turbinibus, sinit incertis turbare procellis,
Ipse suos ignes certo fert impete labens. 505
Nam modice fluere atque uno posse aethera nisu
Significat Pontos, mare certo quod fluit aestu
Unum labendi conservans usque tenorem.
Motibus astrorum nunc quae sit causa canamus.
Principio magnus caeli si vertitur orbis, 510
Ex utraque polum parti premere aera nobis
Dicendum est extraque tenere et claudere utrimque;
Inde alium supra fluere atque intendere eodem
Quo volvenda micant aeterni sidera mundi;
Output
Picture Zones
Fill Zones
Dewarping
Despeckling
Thumbnails
Follow page
z0050
z0051
z0052
z0053
Natural order
Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.
292, 332
313 x 741 mm
p. 1 / 6
z0050

Manual or automated?

On small projects, this operations usually are performed manually; on massive projects, usually they are performed automatically.

Exercise

Compare processed and raw images on https://archive.org

Optical Character Recognition (OCR)

Commercial applications

There are many commercial applications for Optical Character Recognition, such as Abbyy FineReader, Adobe Acrobat Pro, etc. (Among many other comparative evaluations, see for example: https://www.adamenfroy.com/best-ocr-software)

There are also many solutions online (e.g. a new service on GoogleDrive to extract text from PDF of images)

And finally there are many apps to capture images with a smartphone and convert them into text or searcheable PDFs.

Commercial applications: strength points

The main advantages of commercial software are:

- simple to install on Windows and Mac
- easy to use
- graphical interface

Commercial applications: weakness points

The main issues of commercial software are:

- necessity to renew the license for new versions
- scalability (when you must pay per page recognized)
- languages and scripts (FineReader has additional packages for old or ancient scripts, such as Fraktur)

Open source applications for OCR

The most performant open source applications for OCR are:

- Tesseract (https://github.com/tesseract-ocr/tesseract)
- OCRopus (https://github.com/ocropus/ocropy) and its derivatives, listed below
- Kraken (http://kraken.re)
- Calamari (https://github.com/Calamari-OCR/calamari)

Another interesting OCR project is

- Gamera

Open source applications: strength points

The main advantages of these projects are:

- scalability (to process millions of pages)
- scientific research to process challenging documents (endangered languages, ancient languages and scripts, low quality paper and ink, damaged documents)
- support of the community

Open source applications: weakness points

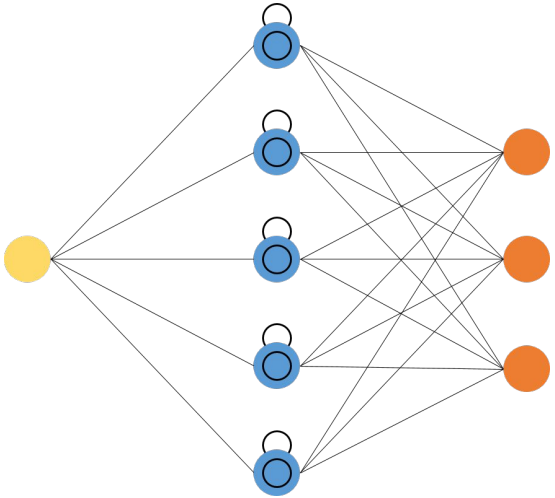
The main issues of these projects are:

- incompatible versions in quick evolution
- no graphical interface (only command line)
- not available for all the Operative Systems

What is an OCR engine?

In simple words, an OCR engine is a classifier, which assigns a **label** (i.e. a character or a sequence of characters) to an **image region**

For this reason, the most recent OCR engines are based on Neural Networks



datascience.stackexchange.com

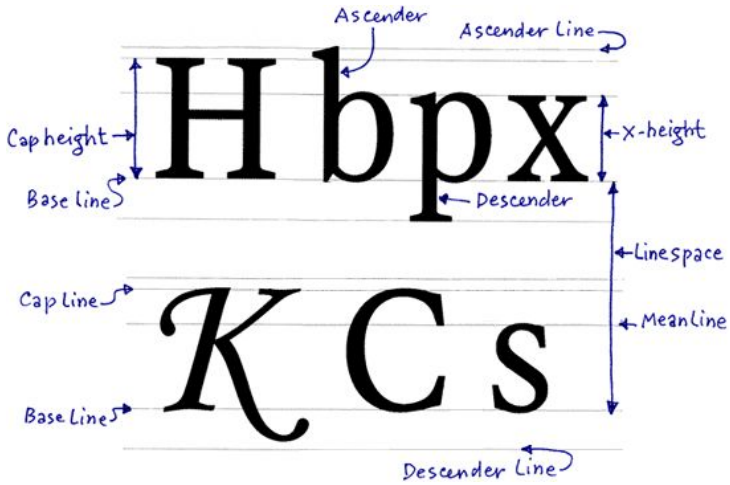
Layout analysis

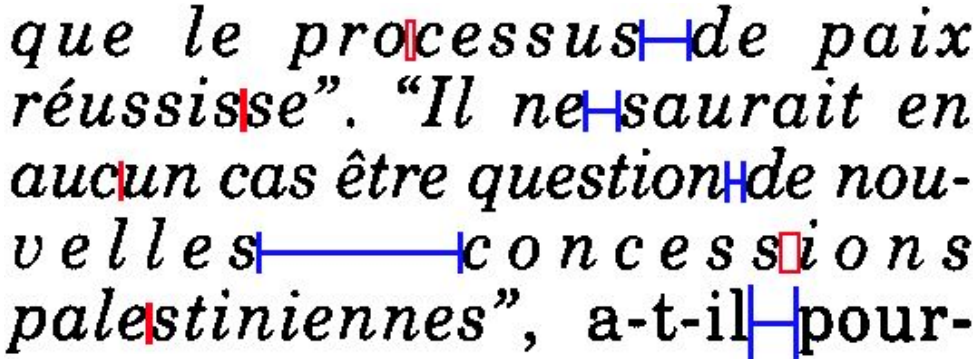
In order to assign a label, at the character level, to an image region, regions must be identified by layout analysis and segmentation

The **layout analysis** decomposes the page in its textual and graphical components (e.g. columns of text, illustrations, and tables)

Segmentation

Textual blocks are hierarchically segmented in lines, words, and characters





how-ocr-works.com

Segmentation issues

Bad segmentation causes bad OCR

Factors that must be taken into account:

- avoid artifacts during the image acquisition process, such as page warping (when it is possible, pages should be unbounded!)
- preprocess the images to reduce artifacts
- if an OCR engine makes a bad segmentation, try another one (for example, if Abbyy FineReader does not satisfy your needs, try tesseract or Kraken and vice versa)

Trained data sets

Both commercial and open source OCR applications are provided with pre-trained data sets

For this reason, we can perform the optical character recognition on a variety of languages and scripts, without taking care of the training phase

Training

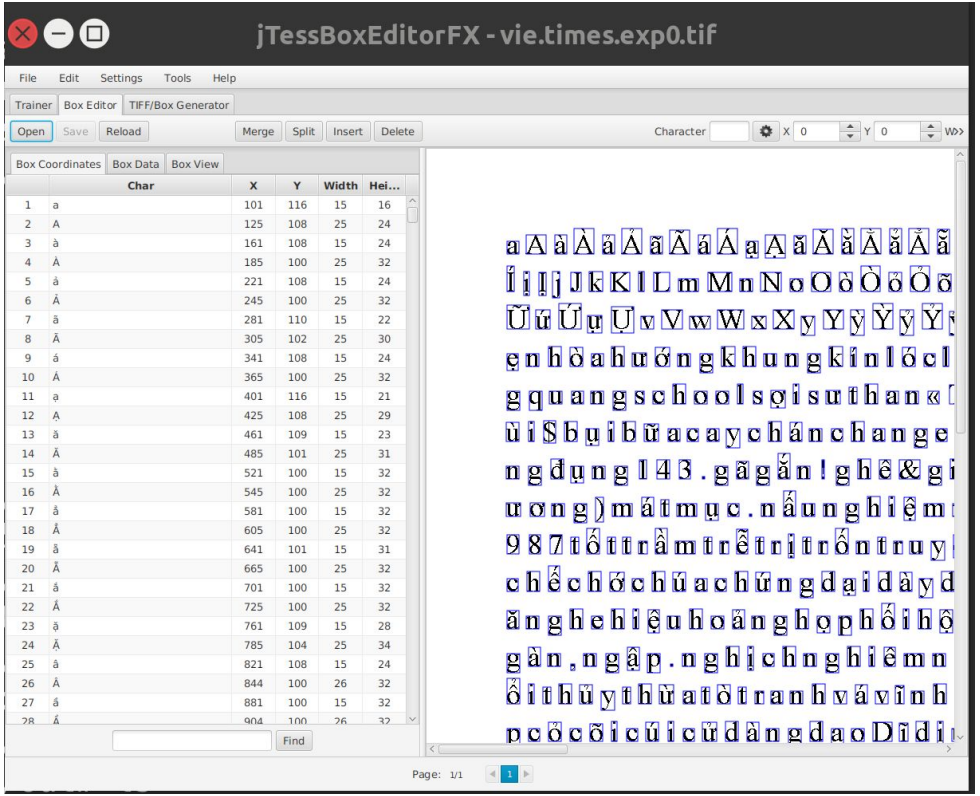
When the accuracy of the recognition is not satisfactory, it is necessary to train the system

Training is based on an **accurate** association between **text** and **image**

The text that exactly matches the image is called **ground truth**

Some OCR engines, such as tesseract, need a small amount of ground truth, some others on the contrary need a large amount.

Training: the case of tesseract



Performing OCR

tesseract -l <language(s)> <image> <output without suffix>

tesseract -l ita+lat img001.tiff doc001

(training data are available here: https://github.com/tesseract-ocr/tessdata)

hocr

tesseract -l <language(s)> <image> <output without suffix> hocr

tesseract -l ita+lat img001.tiff doc001 hocr

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
 <head>
  <title></title>
  <meta http-equiv="Content-Type" content="text/html;charset=utf-8"/>
  <meta name='ocr-system' content='tesseract 5.0.0-alpha-20210401-94-ga968' />
  <meta name='ocr-capabilities' content='ocr_page ocr_carea ocr_par ocr_line ocrx_word ocrp_wconf'/>
 </head>
 <body>
  <div class='ocr_page' id='page_1' title='image "z0053.tif"; bbox 0 0 12181 19222; ppageno 0'>
   <div class='ocr_carea' id='block_1_1' title="bbox 3910 479 11495 887">
    <p class='ocr_par' id='par_1_1' lang='ita' title="bbox 3910 479 11495 887">
     <span class='ocr_line' id='line_1_1' title="bbox 3910 479 11495 887; baseline -0.006 -81; x_size 400; x_descenders 93; x_ascenders 120">
      <span class='ocrx_word' id='word_1_1' title='bbox 3910 500 5473 805; x_wconf 68'>LIBER</span>
      <span class='ocrx_word' id='word_1_2' title='bbox 5838 479 8127 887; x_wconf 51' lang='lat'>QUINTUS.</span>
      <span class='ocrx_word' id='word_1_3' title='bbox 11121 576 11227 763; x_wconf 95' lang='lat'>I</span>
      <span class='ocrx_word' id='word_1_4' title='bbox 11311 573 11495 854; x_wconf 95' lang='lat'>9</span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_2' title="bbox 402 1167 11496 18639">
    <p class='ocr_par' id='par_1_2' lang='lat' title="bbox 485 1167 11496 6397">
     <span class='ocr_line' id='line_1_2' title="bbox 525 1167 10060 1598; baseline -0.004 -88; x_size 407; x_descenders 85; x_ascenders 117">
      <span class='ocrx_word' id='word_1_5' title='bbox 525 1182 3279 1598; x_wconf 89'>Quandoquidem</span>
      <span class='ocrx_word' id='word_1_6' title='bbox 3515 1198 4706 1505; x_wconf 89'>claram</span>
      <span class='ocrx_word' id='word_1_7' title='bbox 4944 1185 6378 1593; x_wconf 89'>speciem</span>
      <span class='ocrx_word' id='word_1_8' title='bbox 6610 1241 8492 1583; x_wconf 89'>certamque</span>
```

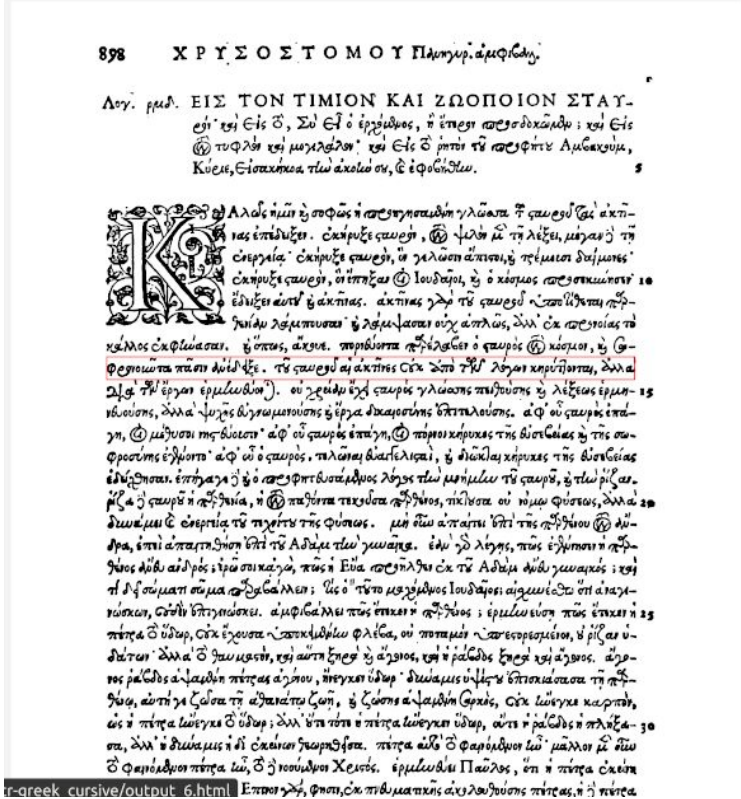
Early printed editions and Handwritten Text Recognition (HTR)

Early printed editions and manuscripts

Early printed editions and manuscripts are challenging:

- complex and/or irregular layout
- abbreviations
- ligatures
- irregular letters

Kraken on early printed editions



898 Χ Ρ Υ Σ Ο Σ Τ Ο Μ Ο Υ Πανηγυρ. ἀμφιβαλλ.

Λογ. ρμδ. ΕΙΣ ΤΟΝ ΤΙΜΙΟΝ ΚΑΙ ΖΩΟΠΟΙΟΝ ΣΤΑΥ-
ρόν· καὶ εἰς τὸ, Σὺ εἶ ὁ ἐρχόμενος, ἢ ἕτερον προσδοκῶμεν; καὶ εἰς
τὸν τυφλὸν καὶ μογιλάλον· καὶ εἰς τὸ ῥητὸν τοῦ προφήτου Αμβακοὺμ,
Κύριε, εἰσακήκοα τὴν ἀκοήν σου, καὶ ἐφοβήθην. 5

Αλῶς ἡμῖν καὶ σοφῶς ἡ προηγησαμένη γλῶσσα τοῦ σταυροῦ τὰς ἀκτῖ-
νας ἐπέδειξεν. ἐκήρυξε σταυρὸν, τὸν ψιλὸν μὲν τῇ λέξει, μέγαν δὲ τῇ
ἐνεργείᾳ· ἐκήρυξε σταυρὸν, ὃν γελῶσιν ἄπιστοι, καὶ τρέμουσι δαίμονες·
ἐκήρυξε σταυρὸν, ὃν ἔπηξαν οἱ Ιουδαῖοι, καὶ ὁ κόσμος προσεκύνησεν· 10
ἔδειξεν αὐτοῦ καὶ ἀκτῖνας. ἀκτῖνας γὰρ τοῦ σταυροῦ ὑποτίθεται παρ-
θενίαν λάμπουσαν· καὶ λάμψασαν οὐχ ἁπλῶς, ἀλλ' ἐκ προνοίας τὸ
κάλλος ἐκφήνασαν. καὶ ὅπως, ἄκουε. πορνεύοντα παρέλαβεν ὁ σταυρὸς τὸν κόσμον, καὶ σω-
φρονοῦντα πᾶσιν ἀνέδειξε. τοῦ σταυροῦ αἱ ἀκτῖνες οὐκ ἀπὸ τῶν λόγων κηρύττονται, ἀλλὰ
διὰ τῶν ἔργων ἑρμηνεύονται). οὐ χρείαν ἔχει σταυρὸς γλώσσης πειθούσης καὶ λέξεως ἑρμη- 15
νευούσης, ἀλλὰ ψυχῆς εὐγνωμονούσης καὶ ἔργα δικαιοσύνης ἐπιτελούσης. ἀφ' οὗ σταυρὸς ἐπά-
γη, οἱ μέθυσοι νηστεύουσιν· ἀφ' οὗ σταυρὸς ἐπάγη, οἱ πόρνοι κήρυκες τῆς εὐσεβείας καὶ τῆς σω-
φροσύνης ἐγένοντο· ἀφ' οὗ σταυρὸς, τελῶναι εὐαγγελισταὶ, καὶ διῶκται κήρυκες τῆς εὐσεβείας
ἐδείχθησαν. ἐπήγαγε δὲ καὶ ὁ προφητευσάμενος λόγος τὴν μνήμην τοῦ σταυροῦ, καὶ τὴν ῥίζαν.
ῥίζα δὲ σταυροῦ ἡ παρθενία, ἡ τὸν παθόντα τεκοῦσα παρθένος, τίκτουσα οὐ νόμῳ φύσεως, ἀλλὰ 20
δυνάμει καὶ ἐνεργείᾳ τοῦ τεχνίτου τῆς φύσεως. μὴ οὖν ἀπαίτει ἐπὶ τῆς παρθένου τὸν ἄν-
δρα, ἐπεὶ ἀπαιτηθήσῃ ἐπὶ τοῦ Αδὰμ τὴν γυναῖκα. ἐὰν γὰρ λέγῃς, πῶς ἐγέννησεν ἡ παρ-
θένος ἄνευ ἀνδρὸς; ἐρῶ σοι κἀγὼ, πῶς ἡ Εὖα προῆλθεν ἐκ τοῦ Αδὰμ ἄνευ γυναικός; καὶ
τί δεῖ σώματι σῶμα παραβάλλειν; τίς δ' τοῦτο μαχόμενος Ιουδαῖος; αἰσχυνέσθω ὅτι ἀναγι-
νώσκων, οὐδὲν ἐπιγινώσκει. ἀμφιβάλλει πῶς ἔτεκεν ἡ παρθένος; ἑρμηνεύσῃ πῶς ἔτεκεν ἡ 25
πέτρα τὸ ὕδωρ, οὐκ ἔχουσα ὑποκειμένην φλέβα, οὐ ποταμὸν ὑπεστορεσμένον, οὐ ῥίζαν ὑ-

How to train Kraken

Tutorial:

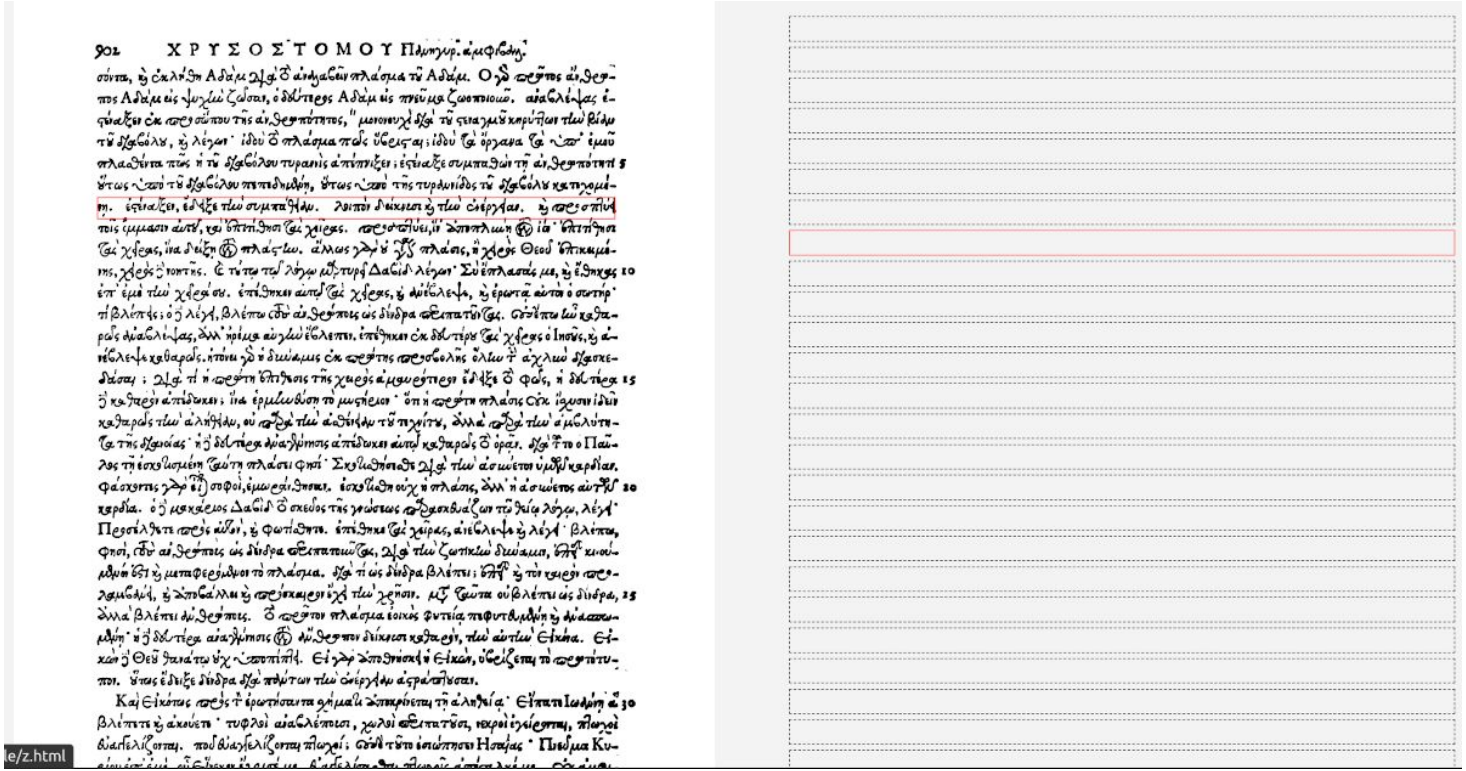
http://kraken.re/training.html#training

Training ancient Greek, early editions:

https://github.com/pharos-alexandria/ocr-greek_cursive/blob/91d72606e2a60593e5eccafe14e6c98493a90ce7/README.md

How to train Kraken

ketos transcribe -o train.html img001.tif img002.tif



Improving OCR and HTR

Accuracy

OCR is evaluated according to the **accuracy**, a measure that is expressed by the following formula

matches / (matches+mismatches+adds+dels)

according to the general formula

TP+TN / (TP+TN+FP+FN)

matches are the agreement between the OCR result and the ground truth

Techniques to improve OCR and HTR

OCR and HTR can be improved by **postprocessing**

A couple of strategies are worthy of attention:

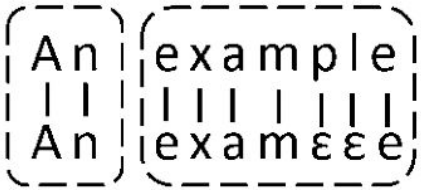
- alignment of multiple and independent OCR engines with efficient selecting criteria
- alignment to different editions of the same text, with criteria to distinguish between OCR errors to be corrected and genuine variants

Alignment

https://link.springer.com/article/10.1007/s10032-020-00359-9

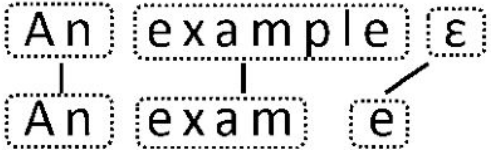
GT: An example
OCR: An exame

Char alignment first:



char errors: 2
word errors: 1

Direct word alignment:



word errors: 2

Exercise

Try to align two sequences of characters

https://bioboot.github.io/bimm143_W20/class-material/nw

Manual correction

WikiSource

Modifica di Pagina:Tragedie di Eschilo (Romagnoli) I.djvu/129

```
<poem>
[[file:Tragedie di Eschilo (Romagnoli) I-25.png|400px|center]]
{{Ct|v=2|t=2|PRIMO CANTO INTORNO ALL'ARA}}
{{vc|{{smaller|I vegliardi, compiendo intorno all'ara lente evoluzioni ritmiche, cantano, alternandosi i due semicori,
le seguenti coppie strofiche.}}}}
{{vc|{{sc|coro}}}}
''Strofe I''
A sterminio di città mosse l'esercito
del Gran Re, la terra invase che finitima
surge contro il suolo d'Asia:
su compagini di tronchi, su compagini di canapi,
superò d'Elle Atamàntide il tragitto{{Nota separata|Pagina:Tragedie di Eschilo (Romagnoli) I.djvu/351|14}},
poi che un giogo, un ponte tutto irto di cunei
del mar sopra la cervice ebbe confitto.

''Antistrofe I''
Il Signore dei frequenti asiaci popoli
furioso, da due bande spinse d'uomini
una greggia innumerevole
su la terra dei nemici, qua pedoni, là dal pelago.
</poem>
```

PRIMO CANTO INTORNO ALL'ARA

I vegliardi, compiendo intorno all'ara lente evoluzioni ritmiche, cantano, alternandosi i due semicori, le seguenti coppie strofiche.

CORO

Strofe I

A sterminio di città mosse l'esercito
del Gran Re, la terra invase che finitima
surge contro il suolo d'Asia:
su compagini di tronchi, su compagini di canapi,
superò d'Elle Atamàntide il tragitto,
poi che un giogo, un ponte tutto irto di cunei
del mar sopra la cervice ebbe confitto.

Antistrofe I

WikiSource

Aiuto:Stato di Avanzamento del Lavoro

Aiuto: Stato di Avanzamento del Lavoro

Manuale ▸ Guida del percorso di qualità dei testi ▸ **Stato di Avanzamento del Lavoro**

Lo **Stato di Avanzamento del Lavoro (SAL)** è il livello che indica la qualità dei testi che hanno intrapreso il *percorso di qualità* di Wikisource.

Nel namespace pagina [modifica]

	SAL 25%	*predefinito*
	SAL 50%	La pagina è problematica
	SAL 75%	La pagina è stata trascritta e formattata
	SAL 100%	La pagina è stata riletta da un utente diverso da quello che ha portato la pagina al SAL 75%

Il SAL 00% si usa per segnalare:

- pagine vuote
- pagine che contengono testo (es.: pubblicità di altri volumi della collana, *ex libris*, oppure i "giudizi della critica") o immagini (es.: timbri o etichette della biblioteca o del proprietario) che non fanno parte dell'opera in senso stretto
- pagine in lingue diverse dall'italiano (vedi {{iwpage}})

	SAL 00%	La pagina non necessita di trascrizione

Nel namespace indice [modifica]

	SAL 25%	*predefinito*
	SAL 50%	Tutte le pagine hanno raggiunto o superato il SAL 50% (escluse le pagine SAL 00%)
	SAL 75%	Tutte le pagine hanno raggiunto o superato il SAL 75% (escluse le pagine SAL 00%)
	SAL 100%	Tutte le pagine hanno raggiunto il SAL 100% (escluse le pagine SAL 00%)

Commerce Numérique

commerce 9 | Page 80: Show Image | Status: ★★★★ | Save

Index

OCR

— 80 —

— 80 —

suite creusé des lacs, songé à capter les eaux, à faire

suite creusé des lacs, songé à capter les eaux, à faire

établir des barrages, sans quoi (on pouvait faire le

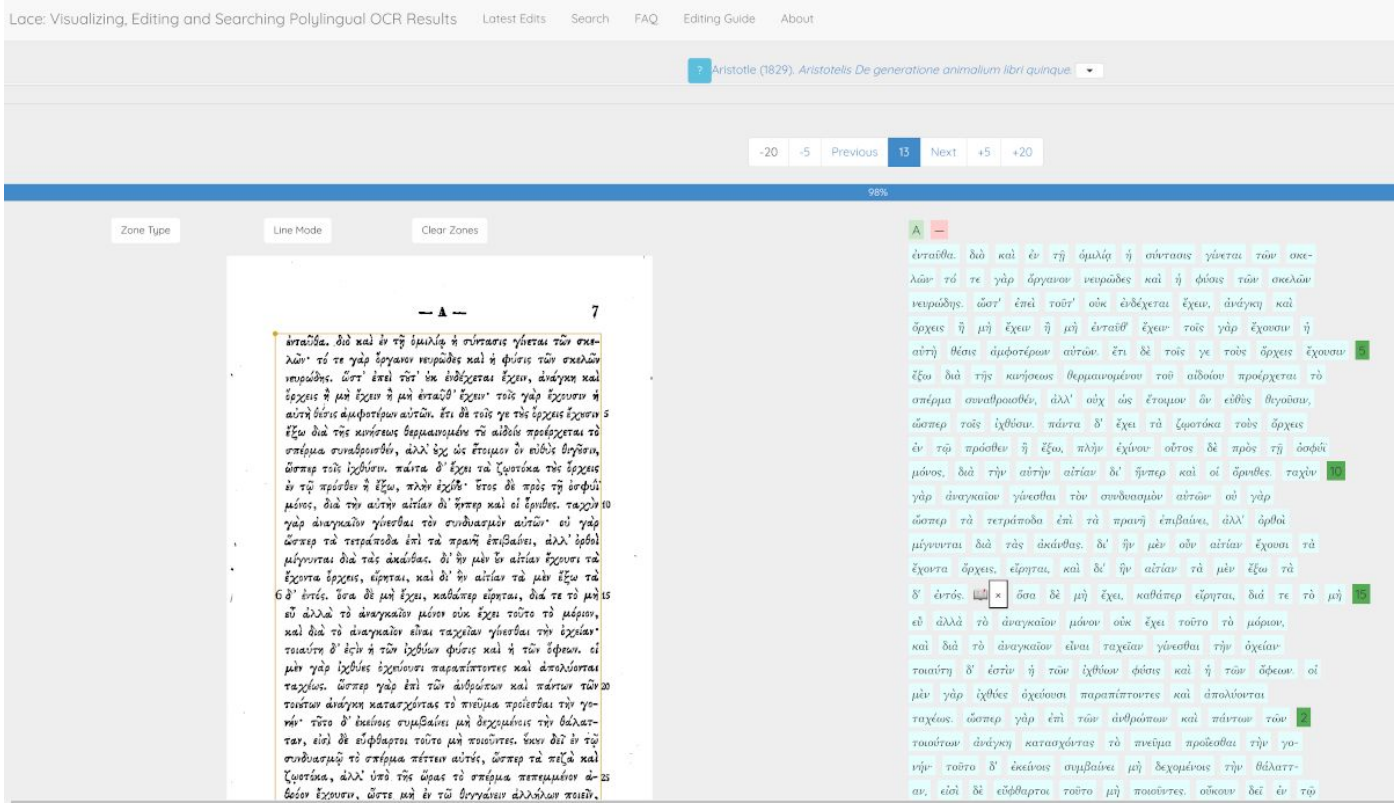
établir des barrages, sans quoi (on pouvait faire le

calcul) en trente-deux heures plus une goutte d'eau.

calcul) en trente-deux heures plus une goutte d'eau.

TEI

Lace (http://heml.mta.ca/lace)



HTR

From OCR to HTR

It is necessary to continue improving OCR accuracy, because it is the real bottleneck for computational linguistics and digital humanities

Currently OCR is very satisfactory on modern languages and documents with a simple layout, but it is challenging on early documents and old or ancient scripts

HTR is emerging with very promising results

HTR: new challenges

High quality digital images, **image enhancement techniques** (**preprocessing**), **larger amounts of training data**, and **better HTR engines** (**processing**) are crucial to increase the performance of handwritten text recognition systems, but also textual and linguistic analysis applied to the HTR output (**post-processing**) can be taken into account

Processing: segmentation (regions)

Il *Paraclitus* di Lelio Manfredi, composto fra il 1515 e il 1520 alla corte di Ferrara

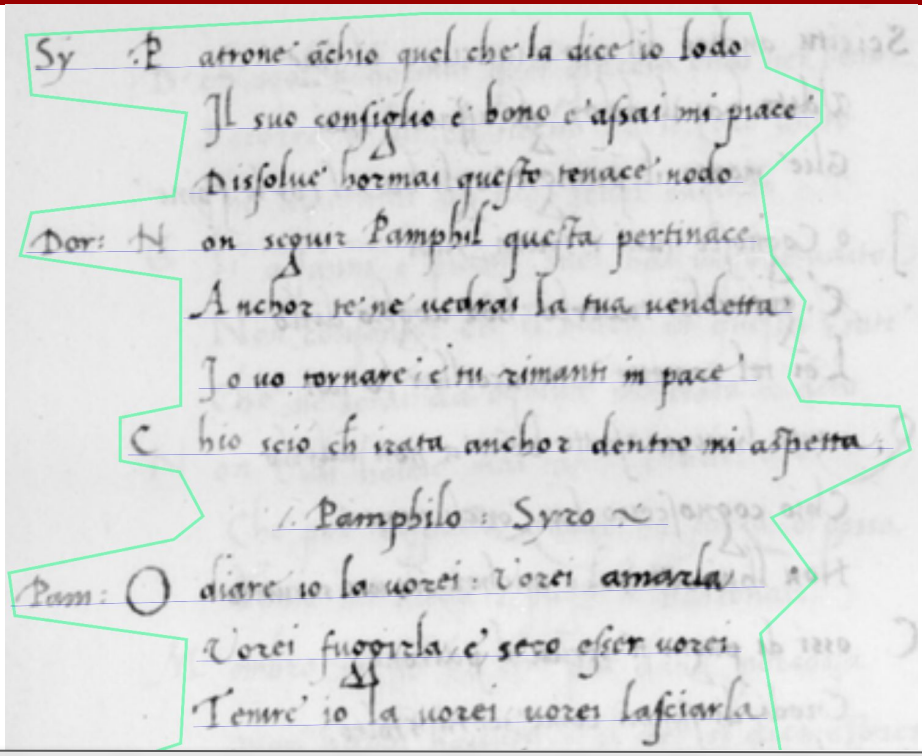


image from gallica.bnf.fr

Processing: segmentation (baselines and masks)

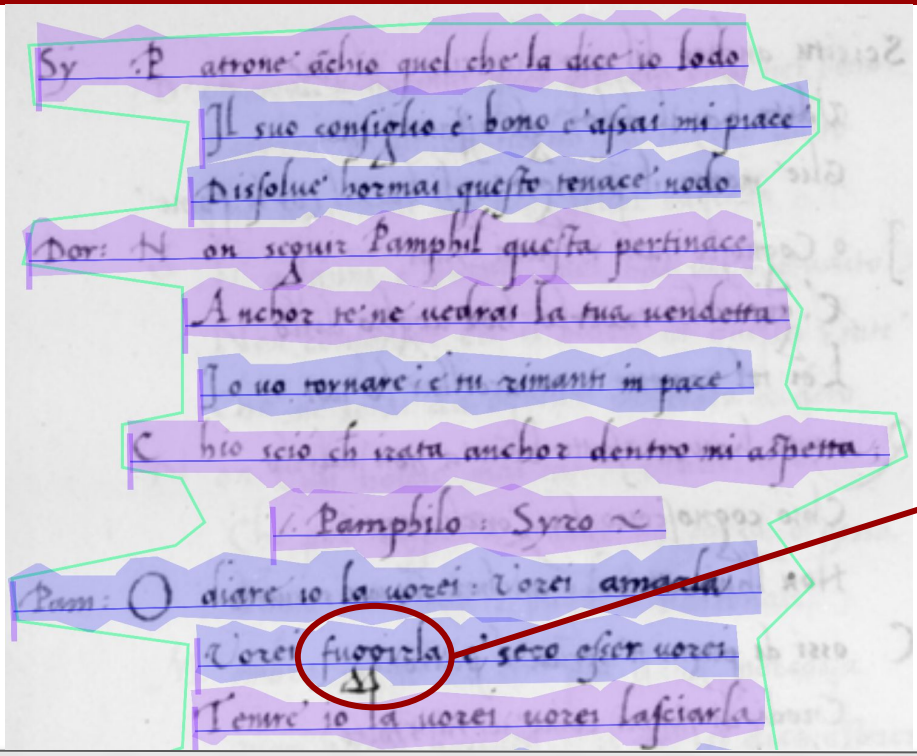


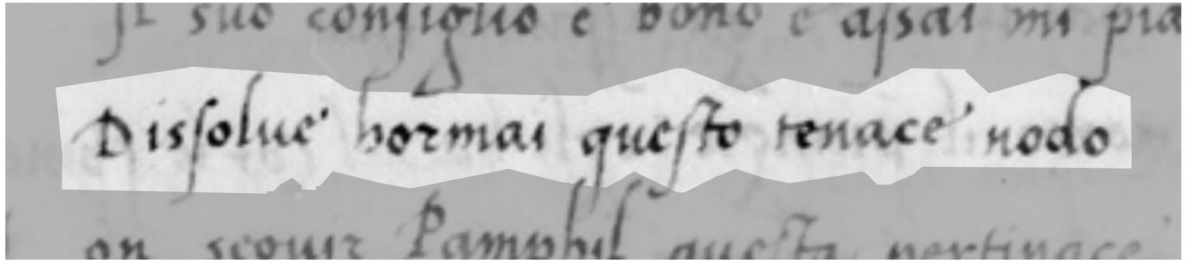
image from gallica.bnf.fr

Post-processing goals

- **detection and classification of potential errors**
  - working with non-standard varieties of language is challenging, because it is harder to distinguish between orthographic variants and HTR errors
- **self-corrections**
  - the most likely errors can be automatically substituted by their most probable corrections, according to the specific linguistic, metrical and stylistic context
- **clues and suggestions for the proofreaders**
  - proofreaders, especially when they are students or volunteers, are highly facilitated by highlights in different colors (according to the type of potential error) and lists of suggestions

Three different types of potential errors in a row

Line #4

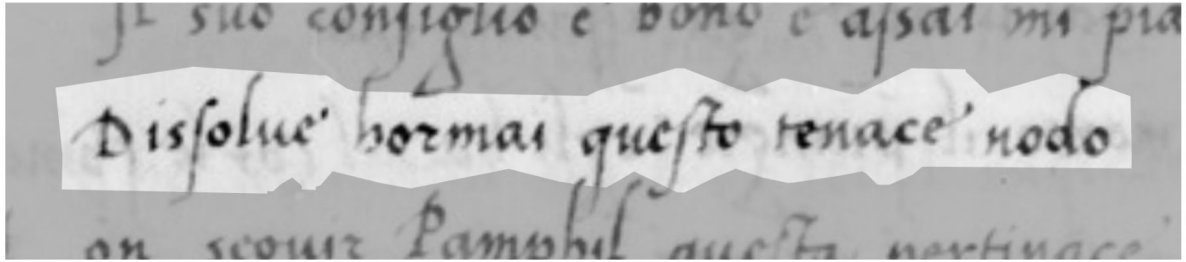


Dieſolue hormai queſt tenace nodo

by (eScriptorium) on Tue Jan 18 2022 08:36:11 GMT+0100

Three different types of potential errors in a row

Line #4



Dieſolue hormai queſt tenace nodo

by (eScriptorium) on Tue Jan 18 2022 08:36:11 GMT+0100

wrong form

false negative

wrong form in helpful context

Detection strategies

For each token

- check in the list of forms previously corrected by hand
- check in the list of forms extracted by corpora of *similar* texts
- check in the list of all available forms

- in case of elision, verify if the next word starts by a vowel
- in case of articles and pronouns, verify if they agree with the following noun phrase (noun, adjective+noun…)
- in case of poems, verify the rhymes

- classify previously unattested forms in **potential words** (i.e. sequences of characters that respect phonetic and orthographic rules) and **nonwords** (i.e. random sequences of characters)

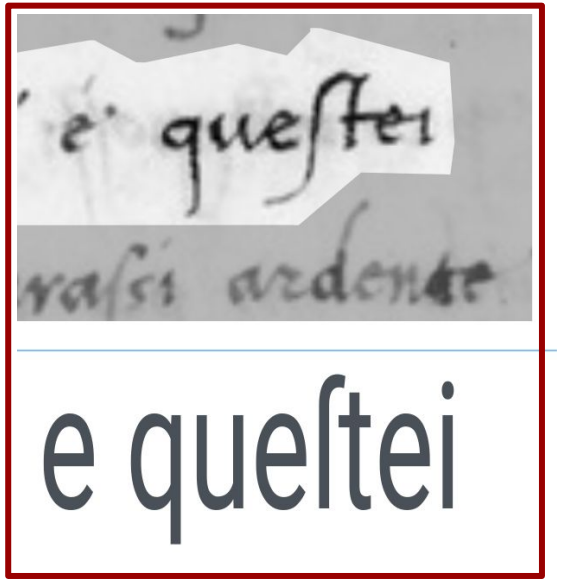
Repertories of attested historical forms

Repertories of attested historical forms

- are built on corpora of literary and documentary sources
- contain much more inflected forms than the lists used to create the most popular spell-checkers, which are based on the standard language
- may contain
  - number of occurrences (total or divided by subcorpora)
  - part of speech and morphological traits
  - diachronic information (range between first and last attestation)
  - genre information (depending on the metadata of the original corpora)
  - peculiar use by authors

Repertories of attested historical forms: an example

questei is an ancient variant of ***costei*** (=this woman) that a standard spell-checker rejects



Variant spellings

Manuscripts contain multiple spellings of the same forms, due to

- evolution of orthographic rules (e.g. principii, principij, principî, principi)
- concurrent spellings in the same manuscript

Different spellings may be

- attested (**true positives**)
- previously unattested but inferrable (**borderline false negatives**)
- previously unattested and unpredictable (**false negatives**)

Variant spellings: examples

Uses of H

- initial h before vowel
  - hoggi
  - hormai (but horamai)
- h after c
  - focho
  - secho
- h after g
  - rogho (but fogho…)
  - veggho
- h after l
  - alhora
- h after r
  - perhó
  - trarhá

Proclitics fused with the following word

- article
  - lamor
- preposition
  - detá
- pronoun
  - mha
  - lhavea

Agreement

In Italian articles and demonstrative, possessive and indefinite adjectives

- are very frequent words
- are inflected and agree with the head of the noun phrase

Whenever

- we can detect a sequence constituted by
  DET ADJ* NOUN ADJ* (e.g. *la bella chioma*, *la chioma dorata*, *la sua bella chioma dorata*)
- we know the morphological traits (i.e. gender and number) of determiners, nouns and adjectives

we can check their agreement and detect inconsistencies

Agreement: examples

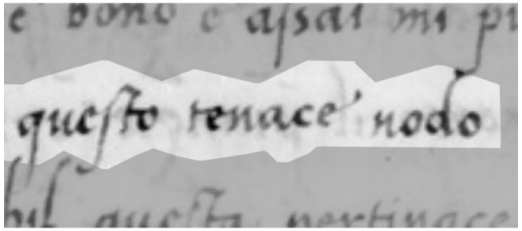
The most probable candidates for **queſt** are

- quest’ `[NP]-----[mf][sp]-  (?=[aeiou])`
- questa `[NP]-----fs-`
- queste `[NP]-----fp-`
- questi `[NP]-----mp-`
- questo `[NP]-----ms-`

but the phonetic constraint and the agreement with

- tenace `[A]-----[mf]s=`
- nodo `[N]-----ms-`

narrow the choice to **queſto**



queſt tenace nodo

wrong form in helpful context

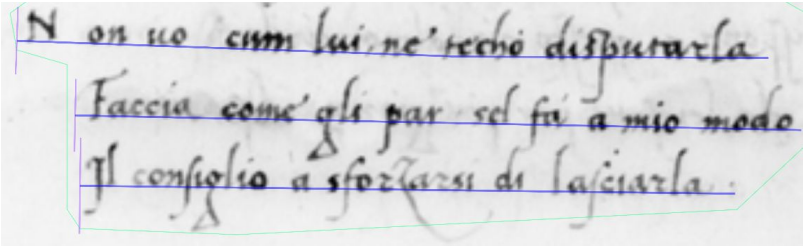
Rhyme

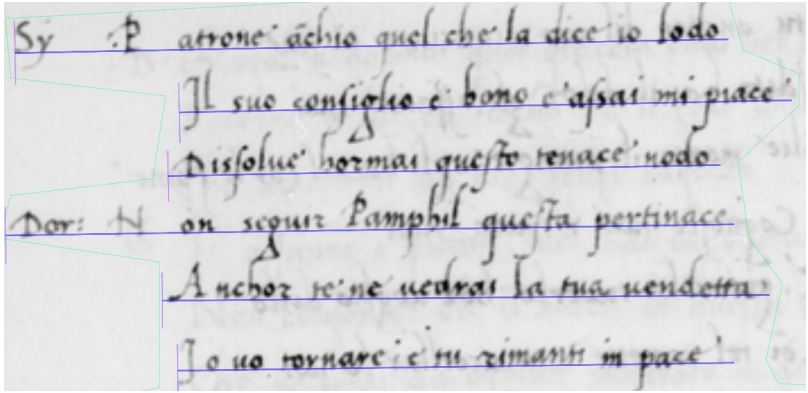
Poems can exploit the mutual information provided by rhymes, especially with fixed metrical schemes, such as

- AA
- ABA BCB CDC …
- ABAB CDCD …
- ABBA CDDC …
- ABABABCC DEDEDEFF …

Even when we cannot identify which words have the stress on the fourth- or third-last syllable, we can always check the identity of the last vowels with an accent or of the sequences of characters from the second-last vowel to the word end

Rhyme: examples





…

modo

…

ſordo →ſodo ←lodo

(manual adjustment is required)

piace

nodo

pertinae →pertinace

(self-correction is enough)

…

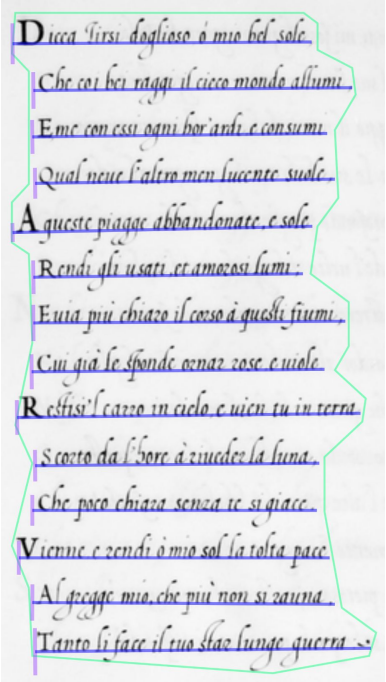
pace

Vocabulary (+morphology+orthography) restraint

- too large repertories of inflected forms are deceiving
- **theological treatises**, **notarial documents**, **Renaissance poetry** are a few examples of texts with very specific vocabularies
- focus on diachronic and diatopic variants of morphology and orthography close to your case study
- use a spell-checkers based on inflected forms extracted by corpora of texts similar to your case study

Vocabulary restraint: Leone Orsini

Leone Orsini,
Canzoniere,
c.a 1564



Dicea Tirsi doglioso ò mio bel sole,
Che coi bei raggi il cieco mondo allumi,
E me con essi ogni hor ardi, e consumi
Qual neue l'altro men lucente suole,
A queste piagge abbandonate, e sole
Rendi gli usati, et amorosi lumi;
E uia più chiaro il corso à questi fiumi,
Cui già le sponde ornar rose, e uiole.
Restisi'l carro in cielo, e uien tu in terra
Scorto da l'hore à riueder la luna,
Che poco chiara senza te si giace.
Vienne, e rendi o mio sol la tolta pace
Al gregge mio, che più non si rauna,
Tanto si face il tuo star lunge guerra :~

LEGENDA

same in Petrarch
in P. with different spelling
in P. with different inflection
not attested in P.

image from gallica.bnf.fr

Multiple texts

In several cases we transcribe multiple manuscripts of the same work or we have access to previous (sometimes normalized) editions of our manuscripts

In these cases previous transcriptions (published or validated by an accurate proof-reading) of the same or of similar manuscripts become the **collation base** for our transcription, which can be **aligned** and **merged** with the HTR output

HTR errors *vs* real variants

In order to avoid **contamination**, it is crucial to distinguish HTR errors from real variants

- agreement between the collation base (CB) and the HTR output reinforces the automated recognition
- disagreement is due
  - **to a real variant**
    - **correctly recognized by HTR** ←highlight it as a possible variant (true words very different from the CB, e.g. “biondi capelli” *vs* “crini dorati”)
    - **recognized by HTR with errors** ←highlight it as a possible variant with errors (non-words or pseudowords very different from the CB, e.g. “biordi capclli” *vs* “crini dorati”)
  - **to an artifact generated by HTR**
    - **that can be self-corrected with a high degree of confidence** ←self-correct with the word(s) in the CB but highlight it for manual check (non-words or pseudowords very close to the words in the CB, e.g. “cnini poiati” *vs* “crini dorati”
    - **that needs human intervention** ←highlight it as a possible error (true words very close to the words in the CB, e.g. “canini orati” or “crini indorati” vs “crini dorati”

K-Centres

CLARIN Knowledge Centres

CLARIN

About Language Resources Learn & Exchange Events News Contact

CLARIN and Ukraine

TA

Home / Learn & Exchange / Knowledge Centres

Knowledge Centres

CLARIN Knowledge Infrastructure

CLARIN Knowledge Centres (K-centres) are a cornerstone of the CLARIN Knowledge Infrastructure (KI), one of the main components ensuring a continuous transfer of knowledge between all players involved in the construction, operation and use of the infrastructure. The mission of the CLARIN KI is to ensure that the available knowledge and expertise does not exist as a fragmented collection of unconnected bits and pieces, but is made accessible in an organised way to both the CLARIN community and the social sciences and humanities research community more widely.

The Role of K-Centres

The focus of CLARIN is on language resources (in all modalities, from all regions and with any topical orientation) and K-centres serve researchers and educators from any discipline where language plays one of its many roles, ranging from object of study, a means of communication or expression, a means to store and extract information, object of learning or teaching activities, to training source for data-driven analytics, and many others. K-centres share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure and can be mostly found in CLARIN countries, but also exist elsewhere, and they all have a virtual presence.

Areas of K-Centre Expertise

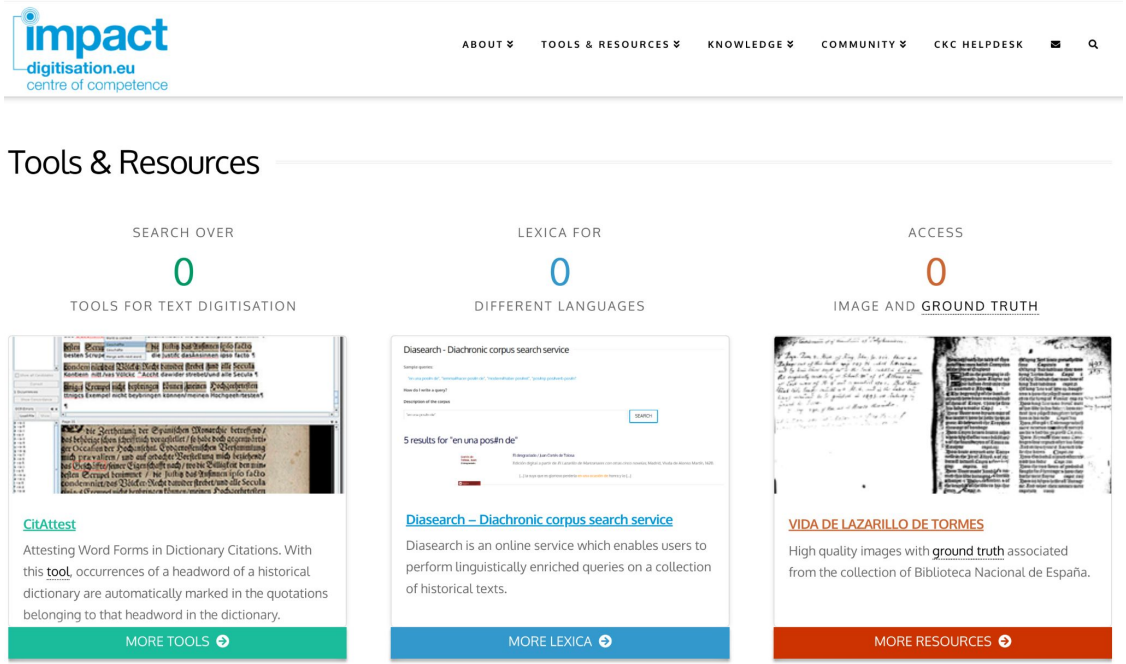
K-centres all have their own specific areas of expertise, which can belong to many different categories, such as

The Knowledge Centres of CLARIN can be contacted through their **Help Desks**

Homepage:

https://www.clarin.eu/content/knowledge-centres

IMPACT



IMPACT is focused on **digitization**

Homepage:

https://www.digitisation.eu

DiPText-KC

CLARIN K CENTRE DiPText-KC ABOUT PARTNERS PEOPLE KNOWLEDGE HELPDESK EVENTS NEWS CONTACT

DiPText-KC

CLARIN Knowledge Centre for Digital and Public Textual Scholarship

DiPText-KC offers expertise on methods, data, instruments and technologies relevant in the field of Philological and Literary Studies, History, Art History and Cultural Heritage.

Its actions aim at:

- sharing information with scholars and students about the state of the art in digital scholarly editing and text annotation through domain-specific languages;
- supporting scholars and students in the creation and publication of digital scholarly editions and resources;
- organizing training activities (for instance webinars, workshops and summer schools).

DiPText-KC is one of the Centres of CLARIN-IT, the Italian node of CLARIN (Common Language Resources and Technology Infrastructure), a digital infrastructure of pan-European interest identified by ESFRI (European Strategy Forum on Research Infrastructures) and classified as a Landmark Research Infrastucture for the Social Sciences and Humanities (ESFRI Landmarks SSH RI).

HIGHLIGHTS

The Digital and Public Textual Scholarship Knowledge Centre is focused on **digital philology**

Homepage:

https://diptext-kc.clarin-it.it

DiPText-KC

CLARIN K CENTRE DiPText-KC | ABOUT | PARTNERS | PEOPLE | KNOWLEDGE | HELPDESK | EVENTS | NEWS | CONTACT

Consortia, Associations, Centers

- Consortia
  - Unicode Consortium
  - TEI Consortium
- National and International Associations
  - ADHO
  - EADH
  - AIUCD
- DH Centres and Labs
  - Italy
    - CIRCSE
    - LabCD

Training

- Summer Schools
  - Venice Centre of Digital and Public Humanities, Department of Humanities, Ca' Foscari University of Venice
    - Summer Camp 2020
  - University of Pisa
    - Digital Tools for Humanists 2022
      (past editions: 2021 | 2020 | 2019 | 2018 | 2017)
  - Digitising, Cataloguing, Searching and Sharing the Medieval and Early-Modern Image
- Seminars / Webinars
  - VeDPH Seminars in Digital and Public Humanities – January-May 2022
    (past editions: October 2019 – May 2020 | September 2020 – December 2021)
  - Humanities Horizons – History, Hacktivism and Genetic Criticism, Solstice Seminar in DPH
  - The Public Staging of Gender in Shakespearean Theatre Discussion with Pamela Allen Brown

Digital Libraries

- Zotero Collections
  - DiPText-KC Library
  - CLARIN Library

per costruire il futur

CNR-ILC CoPhiLab @ GARR 2022

The Collaborative and Cooperative Philology Lab (CoPhiLab, CNR-ILC): data, applications, services and infrastructures (5:50:11-5:59:00)

BREAKING NEWS

- Fourth Appointment of the Workshop Cycle "Digital Philology meets Computational Linguistics"
- Third Appointment of the Workshop Cycle "Digital Philology meets Computational Linguistics"
- Concluded the First Cycle of the Permanent Seminar Series "A bridge between two worlds"
- CNR-ILC CoPhiLab @ GARR 2022
- Second Appointment of the Workshop Cycle "Digital Philology meets Computational Linguistics"

The Digital and Public Textual Scholarship Knowledge Centre keeps you informed on Consortia, Associations, Centres, Training Schools, and Digital Libraries relevant for digital philologists

https://diptext-kc.clarin-it.it

Conclusion

Conclusion

- better images and better HTR systems are crucial, but linguistic post-processing can be helpful to improve accuracy
- do not work with a document, work with a library!
- linguistic, metrical and stylistic information increase the confidence at word, phrase and line level
- many linguistic resources are available through the research infrastructure CLARIN
- CLARIN Knowledge Centres can help you to find relevant information on digitization

Shared folder

https://bit.ly/3xhBqk1

References

Boschetti, F. 2018. *Copisti digitali e filologi computazionali*, Roma.

Grossi, A.M. 2013. *Leone Orsini e il manoscritto “Italien 1535” della Bibliothèque nationale de France*, University of Toronto, PhD Thesis.

Kiessling, B., Tissot, R., Stokes, P. and Stökl Ben Ezra, D. 2019. *eScriptorium: An Open Source Platform for Historical Document Analysis*. In *International Conference on Document Analysis and Recognition Workshops (ICDARW) 2019*, pp. 19-19, doi: 10.1109/ICDARW.2019.10032.