DECIPHERING ALDROVANDI’S COPYISTS HANDWRITING USING TRANSKRIBUS

Final project for Semantic Digital Libraries exam

Erica Andreose and Giorgia Crosilla

Index of outputs:

- BiancolinusHandwriting used to decipher Andrea Biancolino’s handwriting.
- CalzolariHandwriting used to decipher another copyist’s handwriting.

Context

One of the main manuscript funds possessed by the University of Bologna is the one of Ulisse Aldrovandi, a famous naturalist who lived in the 16th century and who is considered to be the father of natural history studies.

Prof. Monica Azzolini, one of the main experts of Aldrovandi manuscripts at the University of Bologna, expressed the need for automatic transcription in order to speed up the process for the on-going project of *Edizione Nazionale Aldrovandi*[1]. In particular, she asked for an automatic transcription of Aldrovandi's own handwriting, but since there is no existing annotation on those documents, we opted to train two models on two different hands that are frequently found in the collection.

Aldrovandi's manuscripts have been progressively uploaded on the digital library of the university *AMS Historica*[2] using IIIF, while a census of the whole collection of manuscripts is available on Manus Online. In 2001, a web portal[3] containing the transcription of some letters was published: here we can find the mere transcription in *txt*, but not the direct comparison with the original page.

Taking all these factors into account, we have decided to work on creating two models based on the hands of two different copyists, selecting those for which we had the most data available. Additionally, we have developed a prototype digital library using the "sites" feature of Transkribus.

Workflow.

The dataset of images regarding Calzolari's letters was downloaded, through a request to the IIIF Image API, directly from the IIIF manifest stored in AMS Historica, setting the image quality to maximum. On the other hand, we asked directly to the person in charge of digitization at University of Bologna for the images related to Manuscript 99, since it is still in the process of being digitized by an external company. When even this manuscript will be uploaded, then the download will be freely available using the IIIF manifest as well. In order to get good samples we did not consider pages that were slightly damaged and in which ink bleed was present. Moreover, we did not consider pages with complex layouts or in which there were lots of missing annotated words in the transcription.

In our case, the ground truth needed to provide input to the supervised machine learning-based model PYLaia HTR[4] is represented by transcriptions done by experts in the

[1] https://aldrovandiana.it/article/view/21/17 (visited in date 03/04/2024)
[2] https://historica.unibo.it/cris/fonds/fonds02020 (visited in date 03/04/2024)
[3] http://aldrovandi.dfc.unibo.it/pinakesweb/main.asp (visited in date 03/04/2024)
[4] https://web.archive.org/web/20230922124520/https://readcoop.eu/glossary/pylaia/ (visited in date 03/04/2024)

field. The transcriptions used here were already available on the web; both of them presented expanded abbreviations even though they were originally abbreviated in the manuscript.

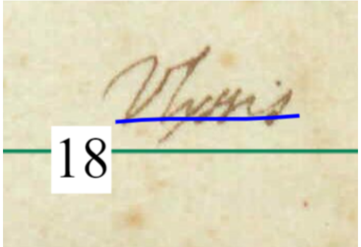
Considering all of these aspects, we defined which goals the model should achieve:

1) automatically recognize text and layout;
2) automatically recognize abbreviations and the correspondent word(s).

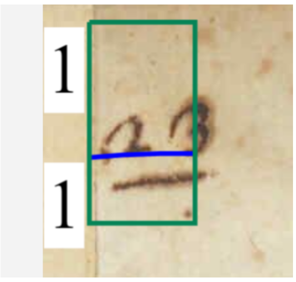
The choice to tag only the abbreviations has been taken since Transkribus states that a good performance of the model is achieved when at most two different tags are used. Our initial goal was to automatically add further semantic tags using Name Entity Recognition. However, Transkribus states that the actual version of the tool does not provide a specific model to train NER tags[5]. Moreover, a proper labeling of these semantic entities could be done only by experts in the field, while our approach could result not sufficient or incomplete.

In order to achieve an automatic recognition of text and layout, we relied on previously trained models:

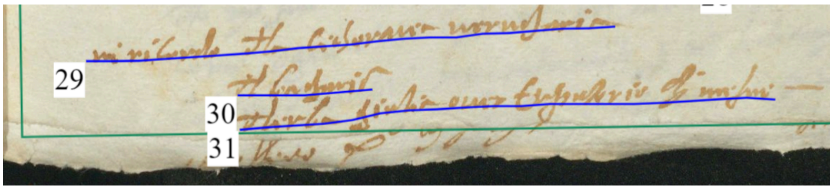
- The automatic layout recognizer "Universal Lines" was used and provided a good starting point, even though it needed some manual corrections. In particular, we noticed that the lines were not correctly recognized in the "signature" part of the letters, and the page layout did not include page numbers or cross-reference marks.



Cross reference of the next page.



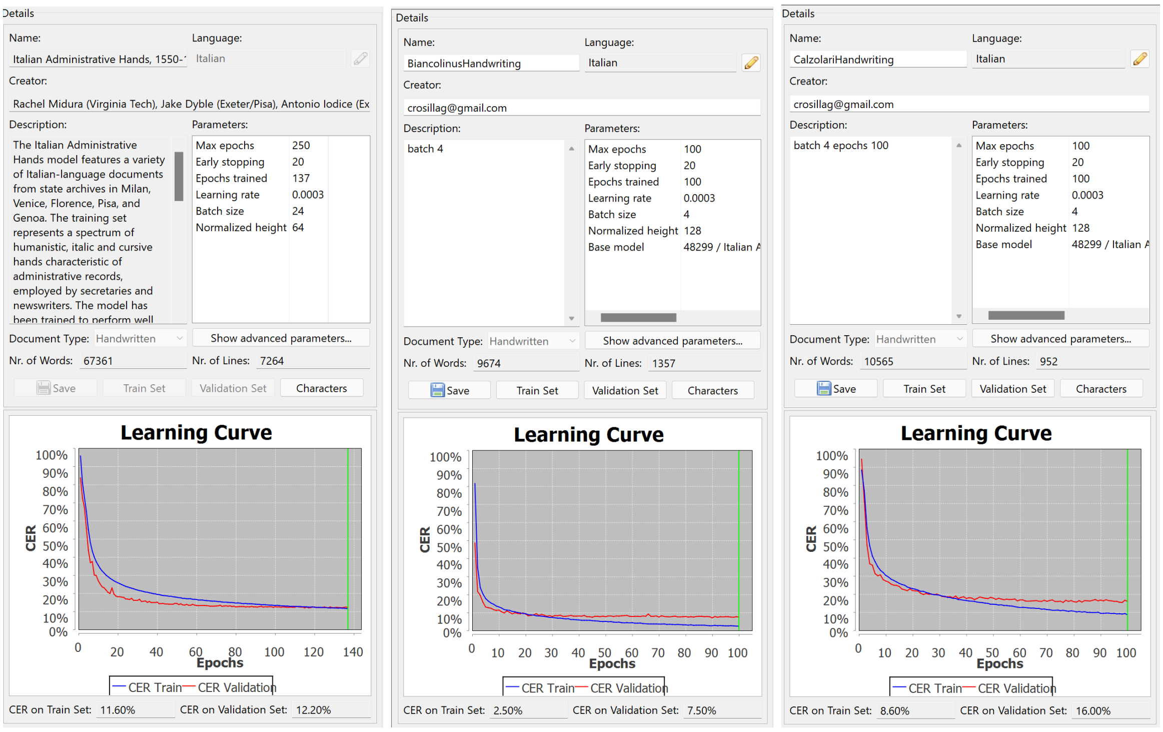
Page number (usually top left).



Signature.

[5] https://web.archive.org/web/20231004130749/https://readcoop.eu/glossary/tag-training/ (visited in date 07/04/2024).

- After having prepared our documents, we trained the models using as a base the already existing “Italian Administrative Hands 1550-1700” and tuning some parameters using the additional features of Transkribus Expert. In particular, we noticed that keeping the standard batch size used by Transkribus (24) was not effective at all, so for all the processes we chose to diminish it to 4 and saw an improvement in performance. This batch size was decided using trial and error, because we found that when using batch sizes of 8 or 12, the model's performance decreased. Even the number of epochs was reduced to 100 due to the smaller size of our dataset and because our models were already relying on a starting one.



Details

Name: Italian Administrative Hands, 1550-1

Language: Italian

Creator: Rachel Midura (Virginia Tech), Jake Dyble (Exeter/Pisa), Antonio Iodice (Ex

Description: The Italian Administrative Hands model features a variety of Italian-language documents from state archives in Milan, Venice, Florence, Pisa, and Genoa. The training set represents a spectrum of humanistic, italic and cursive hands characteristic of administrative records, employed by secretaries and newswriters. The model has been trained to perform well

Parameters:

Max epochs	250
Early stopping	20
Epochs trained	137
Learning rate	0.0003
Batch size	24
Normalized height	64

Document Type: Handwritten

Nr. of Words: 67361

Nr. of Lines: 7264

CER on Train Set: 11.60%

CER on Validation Set: 12.20%

Details

Name: BiancolinusHandwriting

Language: Italian

Creator: crosillag@gmail.com

Description: batch 4

Parameters:

Max epochs	100
Early stopping	20
Epochs trained	100
Learning rate	0.0003
Batch size	4
Normalized height	128
Base model	48299 / Italian A

Document Type: Handwritten

Nr. of Words: 9674

Nr. of Lines: 1357

CER on Train Set: 2.50%

CER on Validation Set: 7.50%

Details

Name: CalzolariHandwriting

Language: Italian

Creator: crosillag@gmail.com

Description: batch 4 epochs 100

Parameters:

Max epochs	100
Early stopping	20
Epochs trained	100
Learning rate	0.0003
Batch size	4
Normalized height	128
Base model	48299 / Italian A

Document Type: Handwritten

Nr. of Words: 10565

Nr. of Lines: 952

CER on Train Set: 8.60%

CER on Validation Set: 16.00%

Comparison between the model considered as the base model and the newly created ones.

The graphs showcase the learning curves of both the training and validation set, which represent how much the Character Error Rate value is diminishing over the epochs. Generally speaking, if CER is over 10% the transcriptions are not so accurate because they provide a model that requires lots of manual corrections[6]. For this reason, a trial and error approach has been fundamental to get to these results which are the best we could achieve given our dataset, while adjusting the parameter values.

As it can be noticed by looking at the graphs[7], in the generated models the validation curve for CER is slightly higher than the other. This may indicate overfitting, suggesting that by reducing the batch size, the model has closely memorized the training data but struggles to

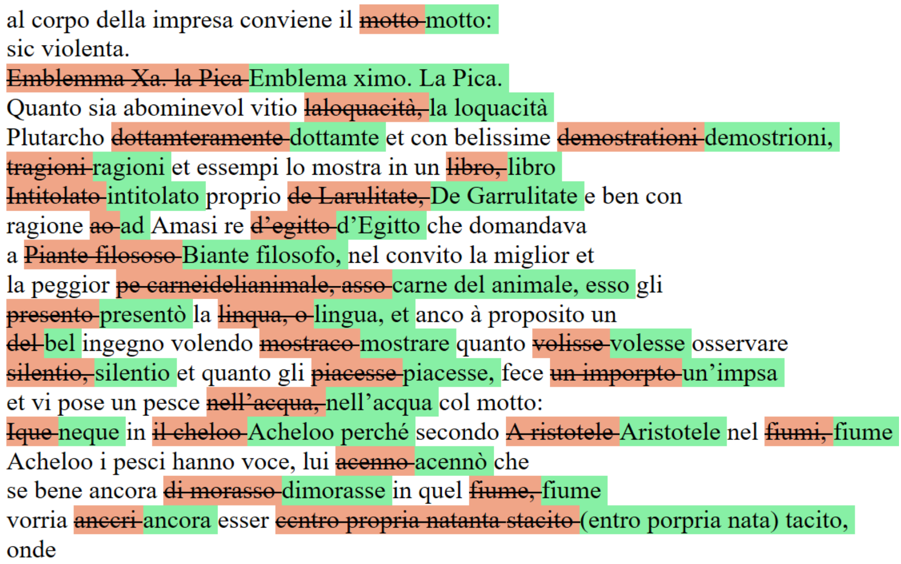
[6] Muehlberger et al. 2019.

[7] For more information regarding the graphs and their values: https://help.transkribus.org/character-error-rate-and-learning-curve (visited in date 03/04/2024)

generalize effectively to unseen data. In order to prevent this phenomenon from happening, a bigger dataset is surely needed along with tuning the correct parameters.

BiancolinusHandwriting

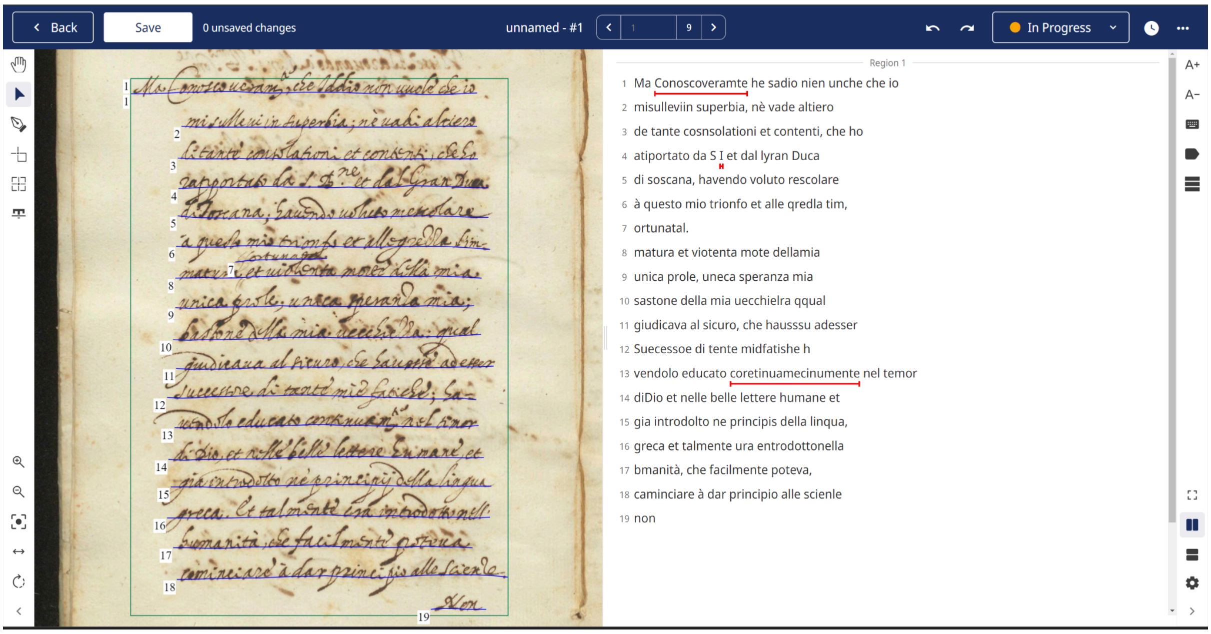
The first model was trained on the handwriting of Andrea Biancolino, one of the most prolific Aldrovandi's copyists and one of the few that has signed some documents. We used as ground truth for annotation the Manuscript 99[8], composed of 90 pages, because it was the only one that already possessed a complete transcription. The manuscript is mainly written in Italian but some parts, mainly quotes, are written in Latin, as a consequence the model created ideally transcribes both languages. Moreover, thanks to Manus Online we discovered which other manuscripts Biancolino transcribed both in Italian and Latin. Having discovered this, we can now estimate that the model could be reused for transcribing seven more manuscripts[9]. The Character Error Rate (CER) value for this model is 7.50% with a training set size of 9674 words.


al corpo della impresa conviene il ~~motto~~ motto:
sic violenta.
~~Emblemma Xa. la Pica~~ Emblema ximo. La Pica.
Quanto sia abominevol vitio ~~laloquacità,~~ la loquacità
Plutarcho ~~dottamteramente~~ dottamte et con belissime ~~demostrationi~~ demostrioni,
~~tragioni~~ ragioni et essempi lo mostra in un ~~libro,~~ libro
~~Intitolato~~ intitolato proprio ~~de Larulitate,~~ De Garrulitate e ben con
ragione ~~ao~~ ad Amasi re ~~d'egitto~~ d'Egitto che domandava
a ~~Piante filososo~~ Biante filosofo, nel convito la miglior et
la peggior ~~pe carneidelianimale, asso~~ carne del animale, esso gli
~~presento~~ presentò la ~~linqua, o~~ lingua, et anco à proposito un
~~del~~ bel ingegno volendo ~~mostraco~~ mostrare quanto ~~volisse~~ volesse osservare
~~silentio,~~ silentio et quanto gli ~~piacesse~~ piacesse, fece ~~un imporpto~~ un'impsa
et vi pose un pesce ~~nell'acqua,~~ nell'acqua col motto:
~~Ique~~ neque in ~~il cheloo~~ Acheloo perché secondo ~~A ristotele~~ Aristotele nel ~~fiumi,~~ fiume
Acheloo i pesci hanno voce, lui ~~acenno~~ acennò che
se bene ancora ~~di morasso~~ dimorasse in quel ~~fiume,~~ fiume
vorria ~~anceri~~ ancora esser ~~centro propria natanta stacito~~ (entro porpria nata) tacito,
onde

Comparison between the model's prediction and the ground truth (pag.88).

The comparison highlighted a CER of 11.26% between the ground truth transcription and the predicted one by the model.

[8] https://aldrovandiana.it/article/view/19/14 (visited in date 03/04/2024)
[9] https://manus.iccu.sbn.it/risultati-ricerca-manoscritti?nomi_id_s=1377871#1712186513176 (visited in date 03/04/2024)



Ma Conoscoveramte he sadio nien unche che io
misulleviin superbia, nè vade altiero
de tante cosnsolationi et contenti, che ho
atiportato da S I et dal lyran Duca
di soscana, havendo voluto rescolare
à questo mio trionfo et alle qredla tim,
ortunatal.
matura et viotenta mote dellamia
unica prole, uneca speranza mia
sastone della mia uecchielra qqual
giudicava al sicuro, che hausssu adesser
Suecessoe di tente midfatishe h
vendolo educato coretinuamecinumente nel temor
diDio et nelle belle lettere humane et
gia introdolto ne principis della linqua,
greca et talmente ura entrodottonella
bmanità, che facilmente poteva,
caminciare à dar principio alle scienle
non

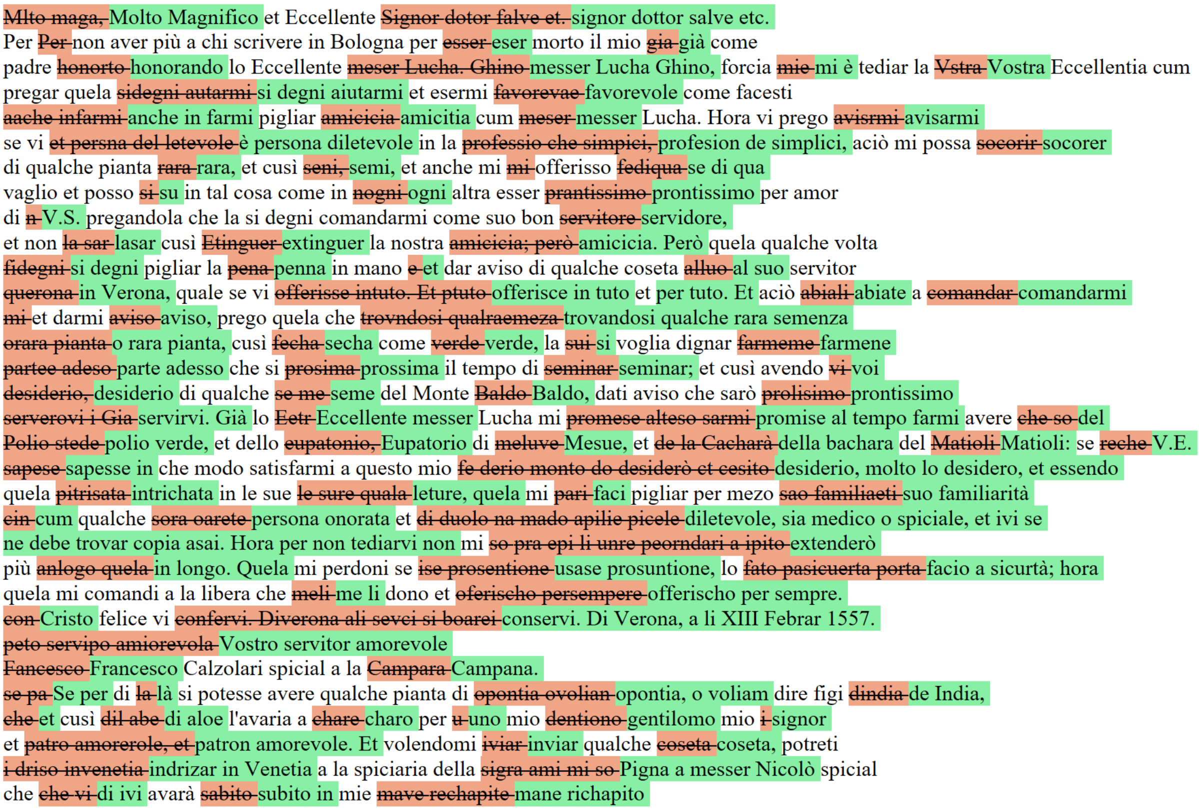
How the model behaves on an unseen text, page taken from Ms.139.

Both from the comparison and this latest test we noticed that the model is not so efficient when it comes to abbreviations, and in particular words that are written without whitespaces among them. Moreover, we noticed a few minor errors that occurred frequently: the transformation of the first letter from uppercase to lowercase; missing punctuation, and missing stress marks and apostrophe.

CalzolariHandwriting

The second model was trained on the handwriting of another anonymous copyist. We made this choice because we had a substantial number of already transcribed pages from Francesco Calzolari's letters available on “Teatro della Natura di Ulisse Aldrovandi” portal[10]. The manuscript examined is 38/2.3 present in IIIF format on *AMS Historica*[11]. We focused on the correspondence 26r-72v[12], Verona, 1554, containing Calzolari letters. However, since the name of the copyist is unknown in this case, we cannot determine to which extent this model can be further reused. The Character Error Rate (CER) value for this model is 16% with a training set size of 10565 words.

[10] http://aldrovandi.dfc.unibo.it/pinakesweb/main.asp (visited in date 03/04/2024)
[11] https://historica.unibo.it/handle/20.500.14008/79626 (visited in date 03/04/2024)
[12] https://manus.iccu.sbn.it/cnmd/0000422511 (visited in date 03/04/2024)



~~Mlto maga,~~ Molto Magnifico et Eccellente ~~Signor dotor falve et.~~ signor dottor salve etc.
Per ~~Per~~ non aver più a chi scrivere in Bologna per ~~esser~~ eser morto il mio ~~gia~~ già come
padre ~~honorto~~ honorando lo Eccellente ~~meser Lucha. Ghino~~ messer Lucha Ghino, forcia ~~mie~~ mi è tediar la ~~Vstra~~ Vostra Eccellentia cum
pregar quela ~~sidegni autarmi~~ si degni aiutarmi et esermi ~~favorevae~~ favorevole come facesti
~~aache infarmi~~ anche in farmi pigliar ~~amicicia~~ amicitia cum ~~meser~~ messer Lucha. Hora vi prego ~~avisrmi~~ avisarmi
se vi ~~et persna del letevole~~ è persona diletevole in la ~~professio che simpici,~~ profesion de simplici, aciò mi possa ~~socorir~~ socorer
di qualche pianta ~~rara~~ rara, et cusì ~~seni,~~ semi, et anche mi ~~mi~~ offerisso ~~fediqua~~ se di qua
vaglio et posso ~~si~~ su in tal cosa come in ~~nogni~~ ogni altra esser ~~prantissimo~~ prontissimo per amor
di ~~n~~ V.S. pregandola che la si degni comandarmi come suo bon ~~servitore~~ servidore,
et non ~~la sar~~ lasar cusì ~~Etinguer~~ extinguer la nostra ~~amicicia; però~~ amicicia. Però quela qualche volta
~~fidegni~~ si degni pigliar la ~~pena~~ penna in mano ~~e~~ et dar aviso di qualche coseta ~~alluo~~ al suo servitor
~~querona~~ in Verona, quale se vi ~~offerisse intuto. Et ptuto~~ offerisce in tuto et per tuto. Et aciò ~~abiali~~ abiate a ~~comandar~~ comandarmi
~~mi~~ et darmi ~~aviso~~ aviso, prego quela che ~~trovndosi qualraemeza~~ trovandosi qualche rara semenza
~~orara pianta~~ o rara pianta, cusì ~~fecha~~ secha come ~~verde~~ verde, la ~~sui~~ si voglia dignar ~~farmeme~~ farmene
~~partee adeso~~ parte adesso che si ~~prosima~~ prossima il tempo di ~~seminar~~ seminar; et cusì avendo ~~vi~~ voi
~~desiderio,~~ desiderio di qualche ~~se me~~ seme del Monte ~~Baldo~~ Baldo, dati aviso che sarò ~~prolisimo~~ prontissimo
~~serverovi i Gia~~ servirvi. Già lo ~~Eetr~~ Eccellente messer Lucha mi ~~promese alteso sarmi~~ promise al tempo farmi avere ~~che so~~ del
~~Polio stede~~ polio verde, et dello ~~eupatonio,~~ Eupatorio di ~~meluve~~ Mesue, et ~~de la Cacharà~~ della bachara del ~~Matioli~~ Matioli: se ~~reche~~ V.E.
~~sapese~~ sapesse in che modo satisfarmi a questo mio ~~fe derio monto do desiderò et cesito~~ desiderio, molto lo desidero, et essendo
quela ~~pitrisata~~ intrichata in le sue ~~le sure quala~~ leture, quela mi ~~pari~~ faci pigliar per mezo ~~sao familiaeti~~ suo familiarità
~~cin~~ cum qualche ~~sora oarete~~ persona onorata et ~~di duolo na mado apilie picele~~ diletevole, sia medico o spiciale, et ivi se
ne debe trovar copia asai. Hora per non tediarvi non mi ~~so pra epi li unre peorndari a ipito~~ extenderò
più ~~anlogo quela~~ in longo. Quela mi perdoni se ~~ise prosentione~~ usase prosuntione, lo ~~fato pasicuerta porta~~ facio a sicurtà; hora
quela mi comandi a la libera che ~~meli~~ me li dono et ~~oferischo persempere~~ offerischo per sempre.
~~con~~ Cristo felice vi ~~confervi. Diverona ali sevci si boarei~~ conservi. Di Verona, a li XIII Febrar 1557.
~~peto servipo amiorevola~~ Vostro servitor amorevole
~~Fancesco~~ Francesco Calzolari spicial a la ~~Campara~~ Campana.
~~se pa~~ Se per di ~~la~~ là si potesse avere qualche pianta di ~~opontia ovolian~~ opontia, o voliam dire figi ~~dindia~~ de India,
~~che~~ et cusì ~~dil abe~~ di aloe l'avaria a ~~chare~~ charo per ~~u~~ uno mio ~~dentiono~~ gentilomo mio ~~i~~ signor
et ~~patro amorerole, et~~ patron amorevole. Et volendomi ~~iviar~~ inviar qualche ~~coseta~~ coseta, potreti
~~i driso invenetia~~ indrizar in Venetia a la spiciaria della ~~sigra ami mi so~~ Pigna a messer Nicolò spicial
che ~~che vi~~ di ivi avarà ~~sabito~~ subito in mie ~~mave rechapite~~ mane richapito

Comparison between the model's prediction and the ground truth (p.43).

We also used Transkribus Expert to compute the comparison between the two examples and the CER was 18.10%. The main errors appear to be similar to those of Model 1: abbreviations are not correctly recognized and sometimes prepositions and words that do not possess whitespace between them are not separated by the model. In general, this model can be considered as a good starting point for a transcription, while it needs further manual refinement in order to get to the complete correct result.

Conclusion.

Both models can be considered as a good starting point and can serve as useful tools for scholars to speed up transcription tasks. The next step is to use these models to create more "ground truth" documents, which will expand the training data and further improve the models. We see this work as an initial attempt to automate the transcription of Ulisse Aldrovandi's documents, with opportunities for future development.

Bibliography.

Corrain, L. (2022). Il manoscritto 99 di Ulisse Aldrovandi. Il programma iconografico della residenza di campagna. *Aldrovandiana. Historical Studies in Natural History*, *1*(1), 35–79. https://doi.org/10.30682/aldro2201c .

Muehlberger et al. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*. https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114 .

Paolini, A. (2023). Conoscere i manoscritti aldrovandiani. Il progetto di catalogazione della Biblioteca Universitaria di Bologna. *Aldrovandiana. Historical Studies in Natural History*, *2*(2), 93–110. https://doi.org/10.30682/aldro2302f .