Mary Shelley HTR: a Handwritten Text Recognition campaign with Transkribus

Project for the Semantic Digital Libraries course, a.a. 2023–2024

Gaia Ortona, Chiara Parravicini, Alice Picco

Introduction

The primary objective of our project is to develop a model capable of automatically transcribing the handwritten texts of Mary Shelley with a reasonably low error rate. We have utilized the online platform Transkribus[1] for training two distinct models, using approximately 70 pages (around 10000 words) from Shelley's renowned novels, *Frankenstein* and *Mathilda*.

The context

The Shelley-Godwin Archive[2] is a digital repository that provides access to the digitized manuscripts of prominent literary figures such as Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft. It serves as a platform to collect, for the first time online, the dispersed handwritten legacy of this family of writers. Through partnerships between institutions like the New York Public Library[3], the Maryland Institute for Technology in the Humanities[4], Oxford's Bodleian Library[5], the Huntington Library[6], the British Library[7], the Houghton Library[8], and the Victoria and Albert Museum[9], the archive contains over 90% of all known relevant manuscripts.

Mary Wollstonecraft, William Godwin, Mary Wollstonecraft Godwin Shelley, and Percy Bysshe Shelley collectively form "England's first family of writers." Their literary and philosophical contributions are remarkable, as they include crucial works for English literature and political philosophy. Of particular note is Mary Shelley's *Frankenstein*, *or the Modern Prometheus*, published in 1818, which profoundly influenced Western literature over the years.

---

[1] https://www.transkribus.org/
[2] http://shelleygodwinarchive.org/
[3] https://www.nypl.org/
[4] https://mith.umd.edu/
[5] https://www.bodleian.ox.ac.uk/home
[6] https://huntington.org/
[7] https://www.bl.uk/
[8] https://library.harvard.edu/libraries/houghton
[9] https://www.vam.ac.uk/

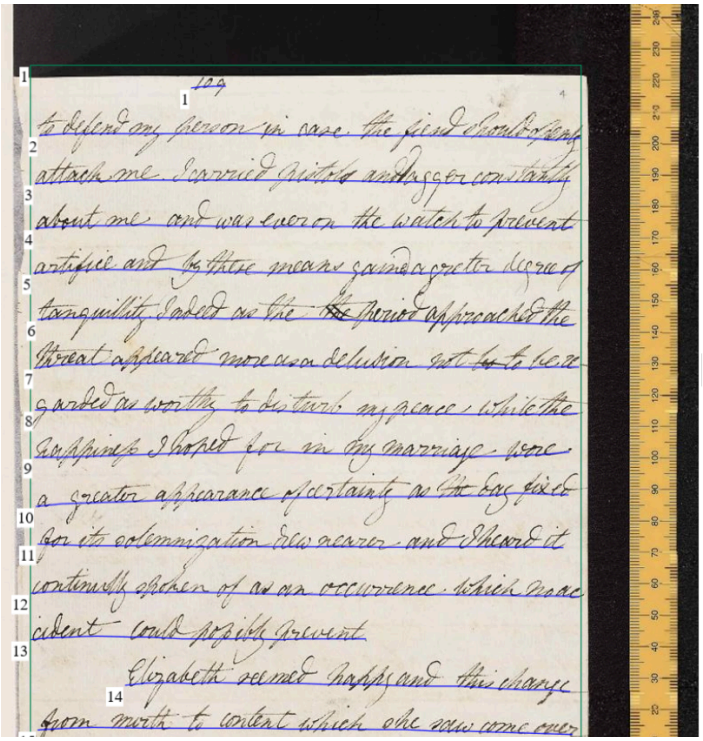
The technological infrastructure of the Shelley-Godwin Archive is built on linked data principles and adheres to emerging standards such as the Shared Canvas data model[10] and the Text Encoding Initiative (TEI)[11]. These frameworks support a participatory platform where scholars, students, and the public can engage in the curation and annotation of the archive's contents. Additionally, the archive's transcriptions and associated software applications are openly available on GitHub[12] under open licenses (Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic (CC BY-NC-SA 2.0)), fostering collaboration and innovation within the academic community.

Our project fits well in the context of the Shelley-Godwin Archive because it aims to speed up the process of adding new material to the archive. While we're initially concentrating on transcribing Mary Shelley's works, our models could easily be adjusted to transcribe texts from other authors featured in the archive. This flexibility could improve significantly to the archive's expansion and accessibility, offering a wider array of literary materials for academic study. By automating transcription, our goal is to simplify archival tasks, encouraging more scholars and enthusiasts to delve into literary history.

Workflow

As mentioned in the section above, in order to develop our models we had to select the works in the archive that were written exclusively by Mary Shelley. Therefore our overall documents set ended up consisting of 36 pages of *Frankenstein* (more specifically the chapters 5,6 and 7) and 28 pages of *Mathilda*, Shelley's second novel. Out of these 64 pages, a percentage of 10% of them was randomly selected by Transkribus as part of the validation set before the training to assess the model performance.
The number of pages was set based on what is specified in the Transkribus guidelines (it's recommended to have at least 10k words to train a HTR model[13]).
After the selection process, we leveraged the pre-trained public model available on Transkribus (the Universal Lines model) for the setting of the layout. Regarding the textual part, we relied on the transcription provided by the archive, considering it a reliable ground truth.

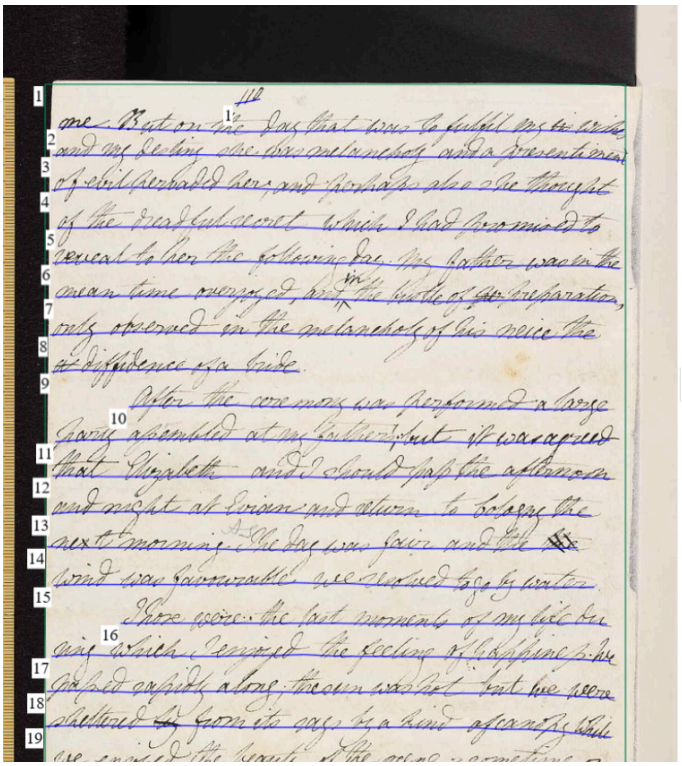
[10] https://iiif.io/api/model/shared-canvas/1.0/
[11] https://tei-c.org/
[12] https://github.com/umd-mith/sga
[13] https://help.transkribus.org/data-preparation



Region 1

1 109
2 to defend my person in case the fiend should openly
3 attack me. I carried pistols and dagger constantly
4 about me and was ever on the watch to prevent
5 artifice and by these means gain*ed* a greter degree of
6 tanquillity. Indeed as the ~~the~~ period approached the
7 threat appeared more as a delusion not ~~bo~~ to be re-
8 garded as worthy to disturb my peace while the
9 happiness I hoped for in my marriage wore
10 a greater appearance of certainty as the day fixed
11 for its solemnization drew nearer and I heard it
12 continully spoken of as an occurrence which no ac
13 cident could possibly prevent.
14 Elizabeth seemed happy and this change
15 from mirth to content which she saw come over

Figure 1 - “Frankenstein” Mary Shelley, page 109



Region 1

1 110
2 me. But on the day that was to fulfil my ~~wi~~ wishes
3 and my destiny she was melancholy and a presentiment
4 of evil pervaded her; and perhaps also she thought
5 of the dreadful secret which I had promised to
6 reveal to her the following day. My father was in the
7 mean time overjoyed, and, in the bustle of ~~pe~~ preparation,
8 only observed in the melancholy of his niece the
9 ~~a~~ diffidence of a bride.
10 After the ceremony was performed a large
11 party assembled at my father's; but it was agreed
12 that Elizabeth and I should pass the afternoon
13 and night at Evian and return to Cologny the
14 next morning. As The day was fair and the ~~the~~
15 wind ~~was~~ favourable we resolved to go by water.
16 Those were the last moments of my life du
17 ring which I enjoyed the felling of happiness. We
18 passed rapidly along; the sun was hot but we were
19 sheltered ~~by~~ from its rays by a kind of canopy while
20 we enjoyed the beauty of the scene; sometimes on

Figure 2 - “Frankenstein” Mary Shelley, page 110

As it shows in *Figure 1* and *Figure 2*, we opted for a one-region-layout and we used two tags to underline the presence in the text of deletions and additions.

Training the models with Transkribus

In the subsequent phase following transcription, our project shifted focus to model training. We devised two private models: one fine-tuned using an existing model (specifically, the English model M4), and the other without any pre-existing tuning. The latter approach allowed us to systematically assess potential disparities and inconsistencies.
The main parameters we focused on were the batch size and the number of epochs: we opted for a small batch size number (10) in order to avoid overfitting and improving the computational efficiency; regarding the number of epochs, we opted for the default setting (100) because we thought it was an appropriate number for data processing and evaluation since the dataset is not large.

Results

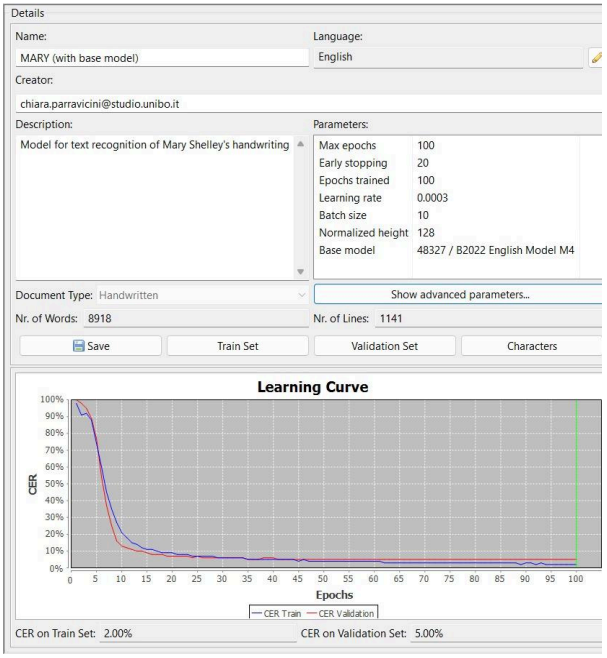


Figure 3 - HTR MaryModel (with base model)

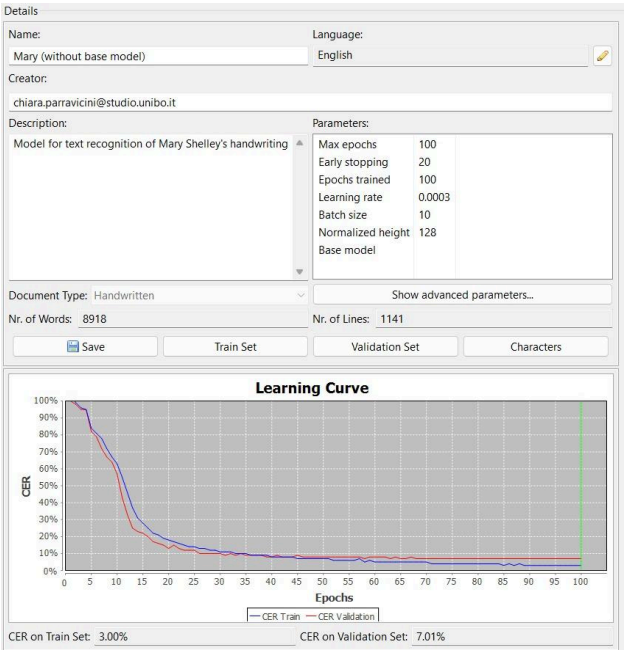


Figure 4 - HTR MaryModel (no base model)

As it shows in *Figure 3* and *Figure 4,* the graphs showcase the learning curves of both the training and validation sets, which represent how much the Character Error Rate is diminishing over the epochs. Generally speaking if CER is over 10% the transcriptions are not so accurate because a lot of manual corrections are needed. In our case, both models performed quite well, even though it's pretty clear that the one without fine-tuning has a higher CER percentage both on the training set and the validation set.

Testing the model on an unseen page of *Mathilda*

The last step of the project was to check the functionality of the model by testing it on an unseen page written by Mary Shelley. We chose to do it on a page taken from *Mathilda* that wasn't included in the initial dataset.

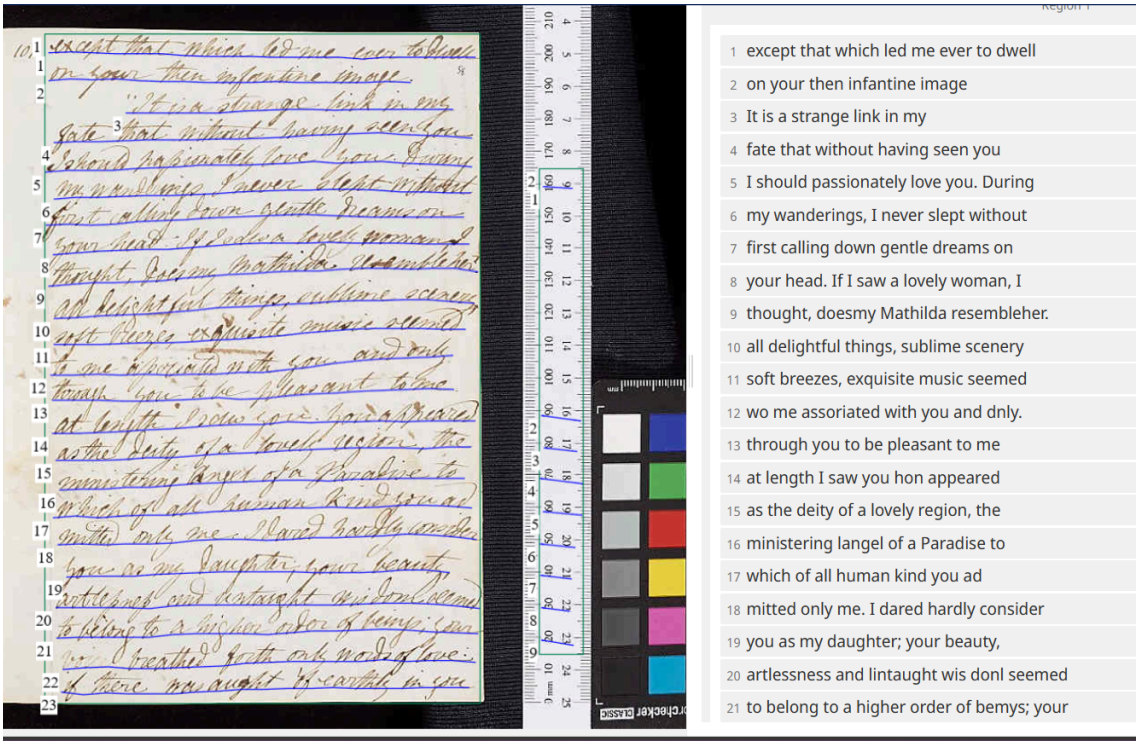
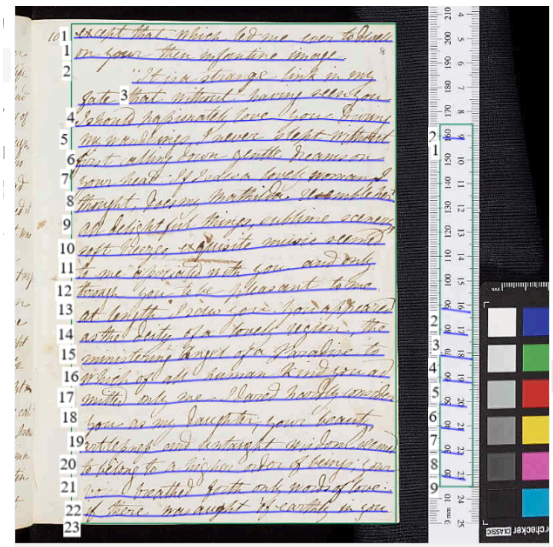


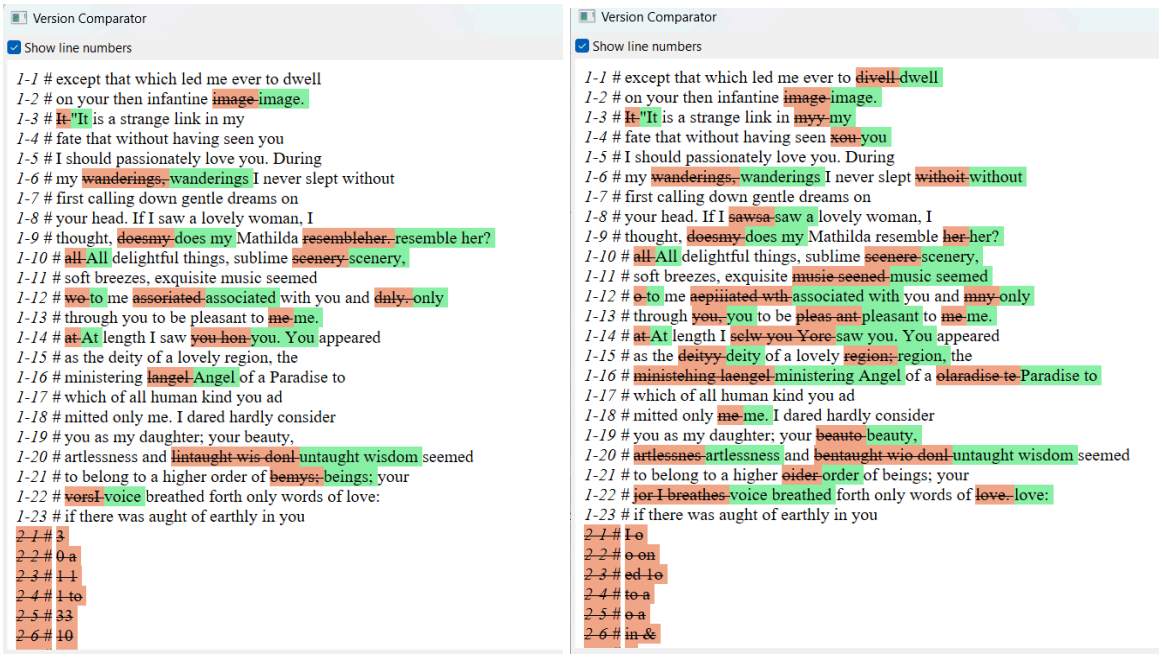
Figure 5 - Test HTR MaryModel with base model (from "Mathilda", pag. 101)



on your then infantine image
It is a strange link in myy
fate that without having seen xou
I should passionately love you. During
my wanderings, I never slept withoit
first calling down gentle dreams on
your head. If I sawsa lovely woman, I
thought, doesmy Mathilda resemble her
all delightful things, sublime scenere
soft breezes, exquisite musie seened
o me aepiiiated wth you and mny
through you, to be pleas ant to me
at length I sclw you Yorc appeared
as the deityy of a lovely region; the
ministehing laengel of a olaradise te
which of all human kind you ad
mitted only me I dared hardly consider
you as my daughter; your beauto
artlessnes and bentaught wio donl seemed
to belong to a higher oider of beings; your
jor I breathes forth only words of love.
if there was aught of earthly in you

Figure 6 - Test HTR MaryModel without base model (from "Mathilda", pag. 101)

As showcased above, the results of the test are clearly more efficient and reliable in the case of the model that was fine-tuned. This test is completely in line with what was discussed in the previous paragraph.



Version Comparator

Show line numbers

1-1 # except that which led me ever to dwell
1-2 # on your then infantine ~~image~~ image.
1-3 # ~~It~~ "It is a strange link in my
1-4 # fate that without having seen you
1-5 # I should passionately love you. During
1-6 # my ~~wanderings,~~ wanderings I never slept without
1-7 # first calling down gentle dreams on
1-8 # your head. If I saw a lovely woman, I
1-9 # thought, ~~doesmy~~ does my Mathilda ~~resembleher.~~ resemble her?
1-10 # ~~all~~ All delightful things, sublime ~~scenery~~ scenery,
1-11 # soft breezes, exquisite music seemed
1-12 # ~~wo~~ to me ~~assoriated~~ associated with you and ~~dnly.~~ only
1-13 # through you to be pleasant to ~~me~~ me.
1-14 # ~~at~~ At length I saw ~~you hon~~ you. You appeared
1-15 # as the deity of a lovely region, the
1-16 # ministering ~~langel~~ Angel of a Paradise to
1-17 # which of all human kind you ad
1-18 # mitted only me. I dared hardly consider
1-19 # you as my daughter; your beauty,
1-20 # artlessness and ~~lintaught wis donl~~ untaught wisdom seemed
1-21 # to belong to a higher order of ~~bemys;~~ beings; your
1-22 # ~~vorsI~~ voice breathed forth only words of love:
1-23 # if there was aught of earthly in you
~~*2-1 #* 3~~
~~*2-2 #* 0 a~~
~~*2-3 #* 1 1~~
~~*2-4 #* 1 to~~
~~*2-5 #* 33~~
~~*2-6 #* 10~~

Version Comparator

Show line numbers

1-1 # except that which led me ever to ~~divell~~ dwell
1-2 # on your then infantine ~~image~~ image.
1-3 # ~~It~~ "It is a strange link in ~~myy~~ my
1-4 # fate that without having seen ~~xou~~ you
1-5 # I should passionately love you. During
1-6 # my ~~wanderings,~~ wanderings I never slept ~~withoit~~ without
1-7 # first calling down gentle dreams on
1-8 # your head. If I ~~sawsa~~ saw a lovely woman, I
1-9 # thought, ~~doesmy~~ does my Mathilda resemble ~~her~~ her?
1-10 # ~~all~~ All delightful things, sublime ~~scenere~~ scenery,
1-11 # soft breezes, exquisite ~~musie seened~~ music seemed
1-12 # ~~o~~ to me ~~aepiiiated wth~~ associated with you and ~~mny~~ only
1-13 # through ~~you,~~ you to be ~~pleas ant~~ pleasant to ~~me~~ me.
1-14 # ~~at~~ At length I ~~sclw you Yorc~~ saw you. You appeared
1-15 # as the ~~deityy~~ deity of a lovely ~~region;~~ region, the
1-16 # ~~ministehing laengel~~ ministering Angel of a ~~olaradise te~~ Paradise to
1-17 # which of all human kind you ad
1-18 # mitted only ~~me~~ me. I dared hardly consider
1-19 # you as my daughter; your ~~beauto~~ beauty,
1-20 # ~~artlessnes~~ artlessness and ~~bentaught wio donl~~ untaught wisdom seemed
1-21 # to belong to a higher ~~oider~~ order of beings; your
1-22 # ~~jor I breathes~~ voice breathed forth only words of ~~love.~~ love:
1-23 # if there was aught of earthly in you
~~*2-1 #* I o~~
~~*2-2 #* o on~~
~~*2-3 #* ed Io~~
~~*2-4 #* to a~~
~~*2-5 #* o a~~
~~*2-6 #* in &~~

Figure 7 - Text Versions Comparison between the ground truth and the prediction of Mary Model with base model (on the left) and without base model (on the right)

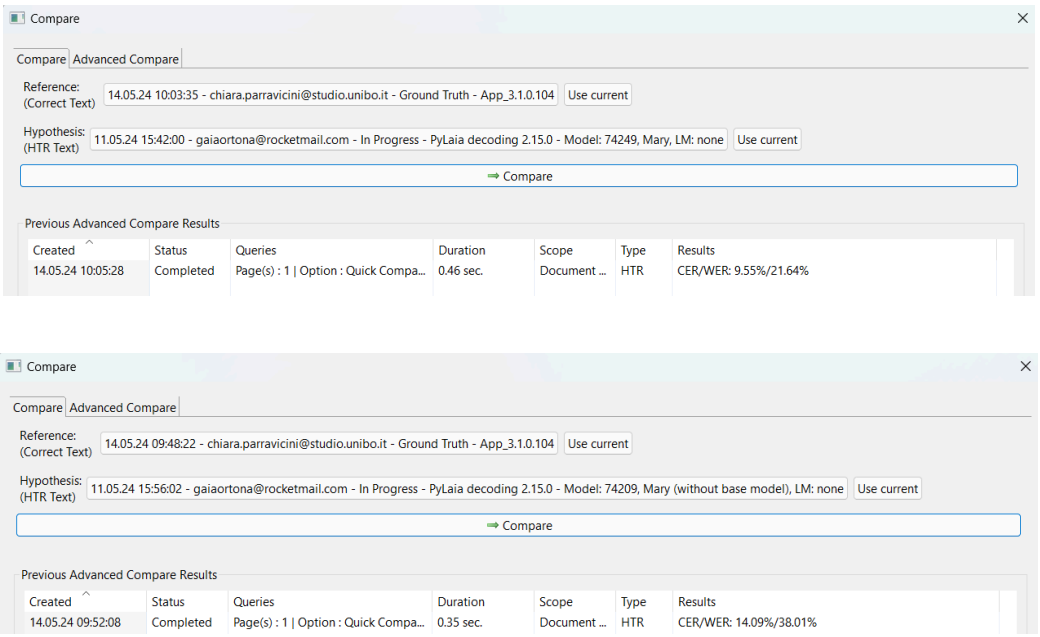


Figure 8 - CER/WER percentages of the comparison between the ground truth and the prediction of Mary Model with base model (above) and without base model (below)

One issue we wanted to focus on is the fact that our model was not well-trained for the layout recognition: in fact the last few lines in *Figure 7* refer to the numbers of a ruler that was present in the image but was not supposed to be taken into consideration in the transcription. Of course we are aware of the fact that this issue could be solved at the beginning by cropping the image, so it's something we'll consider as a possible future development of the project.

Conclusions and critics

In conclusion, we are satisfied with the results obtained, and we believe that these models could serve as a useful starting point for future automation of the transcription and digitization process for the texts in the Shelley-Godwin archive. However, we want to highlight some critical issues encountered during the project. After completing the transcription process, we trained the model on Transkribus and had to wait three days for our request to be processed by the platform's server. Due to previous unsuccessful attempts, we didn't have time to recalibrate the model and make it more efficient. In our opinion, this limitation makes Transkribus less reliable in its open-access version.

Bibliography

Mercer, A. (2016). Beyond *Frankenstein*: The Collaborative Literary Relationship of Percy Bysshe and Mary Shelley. *The Keats-Shelley Review*, *30*(1), 80–85. https://doi.org/10.1080/09524142.2016.1145937

Muñoz, T., & Viglianti, R. (2014-2015). Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive. *Journal of the Text Encoding Initiative*, Issue 8. Retrieved May 10, 2024, from journals.openedition.org/jtei/1270. https://doi.org/10.30682/aldro2201c

Sitography

The Shelley-Godwin Archive: http://shelleygodwinarchive.org/
Transkribus How To Guides: https://readcoop.eu/transkribus/resources/how-to-guides/

- How To Train and Apply Handwritten Text Recognition Models in Transkribus: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/