26/012023

SFFTRANSCRIPTION

Shakespeare First Folio Transcription

MARTINA PENSALFINI - MARTINA.PENSALFINI@STUDIO.UNIBO.IT

Lavinia Bonfatti Sabbioni – lavinia.bonfatti@studio.unibo.it
Eleonora Pasquale – eleonora.pasquale@studio.unibo.it
Martina Pensalfini – martina.pensalfini@studio.unibo.it

SFFTRANSCRIPTION *2*
ANNOTATION PIPELINE: *2*
TASK DEFINITION: 2
PILOT: 2
TRANSCRIPTION GUIDELINES: 3
LAYOUT ANALYSIS: 3
CAMPAIGN: *5*
ANNOTATION AND USE: 6
OUTCOMES AND CRITICALITIES: *6*
FURTHER WORKS: 7
SITOGRAPHY: *7*

SFFTRANSCRIPTION is a small-scale annotation campaign on three Shakespearean works present in the First Folio in 1623.The campaign has been carried out using *Transkribus*[1], a platform for the digitization, AI-powered text recognition, transcription and searching of historical documents.

Our aim while producing this project was to create a Machine Learning system able to perform a transcription task on Shakespeare's first folio of the above-mentioned tragedy: the model is trained on two other Shakespearean works – *The Twelfth Night* and *King John*, respectively a comedy and an historic play – manually annotated by the team, using a third work, *Julius Caesar* as an evaluation set.

ANNOTATION PIPELINE:

TASK DEFINITION:

Firstly, we specified what we wanted to do with our project and in this case, it was to furnish a ML model with an annotated corpus to use to train on another raw one for the evaluation phase.

Our own training corpus was composed by two dramatic works of Shakespeare – *The Twelfth Night* and *King John* – for a total of 44 pages and more than 40000 words. The First Folio is available on the Bodleian Library site[2] to be downloaded in different formats and in this case, we downloaded it as a set of images as it seemed the most appropriate for Transkribus[3].

The transcription of the two works is also available on the site; we employed it as a start for our own and merged it together with the standards (OCR-D: Ground Truth Guidelines[4]) used to better represent the text. We merged together our own little knowledge about printed text of the 16th century with the guidelines to further obtain a work that'd be loyal to its time.

We then set the work, *Julius Caesar*, as the validation set for our HTR model; similarly, to the previous plays it was available on *The Bodleian First Folio*[5] site, both as a set of images and as a PDF for the transcription, which we also adapted to our own purpose.

As noted, we adapted the transcription to our own needs, even more about the structure of the different areas of text, as we had to find a common ground to properly represent them and to annotate the text to give it a formal look.

PILOT:

We first decided to create a pilot on an individual work – *The Twelfth Night*, 20 pages and around 20000 words, to properly to segment out the layout[6]; we worked together through it to properly choose the most

[1] Transkribus Lite (url: https://transkribus.eu/lite/it - last accessed: 11/1/2023).
[2] All rights reserved to the original owner and publisher - https://www.bodleian.ox.ac.uk/home (last accessed: 11/1/2023).
[3] How To Use Transkribus in 10 Steps (url: https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/ - last accessed: 11/1/2023).
[4] Ground Truth Guidelines (url: https://ocr-d.de/en/gt-guidelines/trans/ - last accessed: 11/1/2023).
[5] All rights reserved to the original owner and publisher - (url: https://firstfolio.bodleian.ox.ac.uk - last accessed: 11/1/2023).
[6] How To Transcribe Documents with Transkribus – Introduction (url: https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/ - last accessed: 11/1/2023).

important and meaningful tags for the various parts of the text, adapt to a document of that specific époque (such as: *catchword*, *signature mark* or *header*).**Nessuna voce di sommario trovata.**

We tried to solve any problematics that may arise through creating an annotation model on this first work till we were satisfied enough to apply it to the other work for the training corpus, *King John.*

This has helped us further standardize our model to the specific guidelines chosen and adapted to our own needs and research.

TRANSCRIPTION GUIDELINES:

To carry out the annotation campaign, we relied – as said above - on the OCR-D's Ground Truth Guidelines, especially those listed in level 1. Listed below are the specific cases we encountered and the way we dealt with each one of them:

- Punctuation was left as found in the texts.
- The letter "u", representing the sound /v/, was reproduced true to the original text.
- Upper and lower cases (majuscules/minuscules) were respected.
- Abbreviations were also transcribed according to the original text, not expanded.
- S-Graphemes: long-s were transcribed as round-s.
- Ligatures – common combinations of letters to form a new character – were transcribed as two individual letters[7]. For instance: ae in Caesar.
- Hyphenation was transcribed according to the original[8].
- Distinction between I/J: "I" was employed.

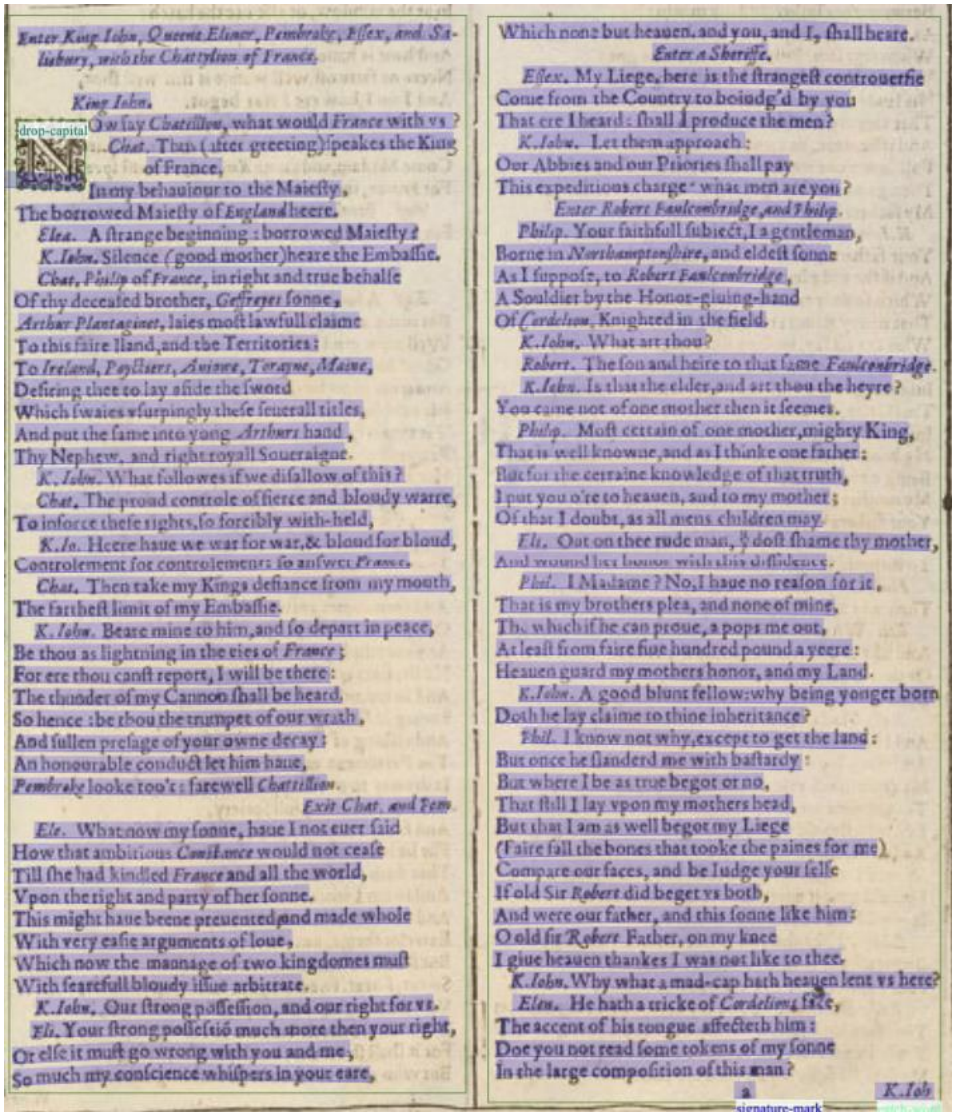
LAYOUT ANALYSIS:

We proceeded to manually segment the chosen documents into lines and text regions, careful to adjust the baselines when crooked, and to make sure two or more regions didn't merge or overlap with one another. The text was divided in various areas accordingly to the guidelines, obtaining:

- Columns of text: two per page.

[7]Ground Truth Guidelines (Ligatures- Level 1) (url: https://ocr-d.de/en/gt-guidelines/trans/tr_level_1.html#tr_level_1 - last accessed: 11/1/2023).

[8] Ground Truth Guidelines (Ligatures- Level 1) (url: https://ocr-d.de/en/gt-guidelines/trans/trSilbentrennung.html - last accessed: 11/1/2023).

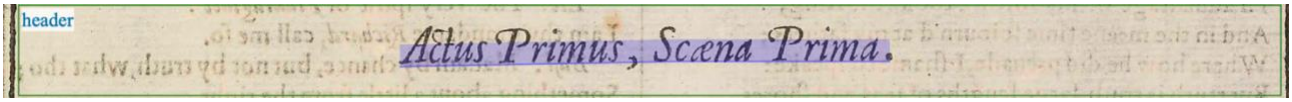


- Title (Heading).

- Page number.

- Decorations: found either at the top or bottom of the page, either in the beginning or end of the plays.

- Header.



- Drop-caps: the group found the drop capital tag trickier to use as it had to be marked by its own and consequently appears in its own line, separated from the rest of the word.

- Catch-words: placed at the foot of the page, one or two words that anticipate the first word of the following page.

- Signature marks: located below the print space, they state the sheet or the position in the book and thus they are used as a guideline for the realization of the correct sequence.

CAMPAIGN:

After the three plays were annotated using the online software Trankribus Lite, we obtained corpus counted a total of 66 pages and 43170 words.

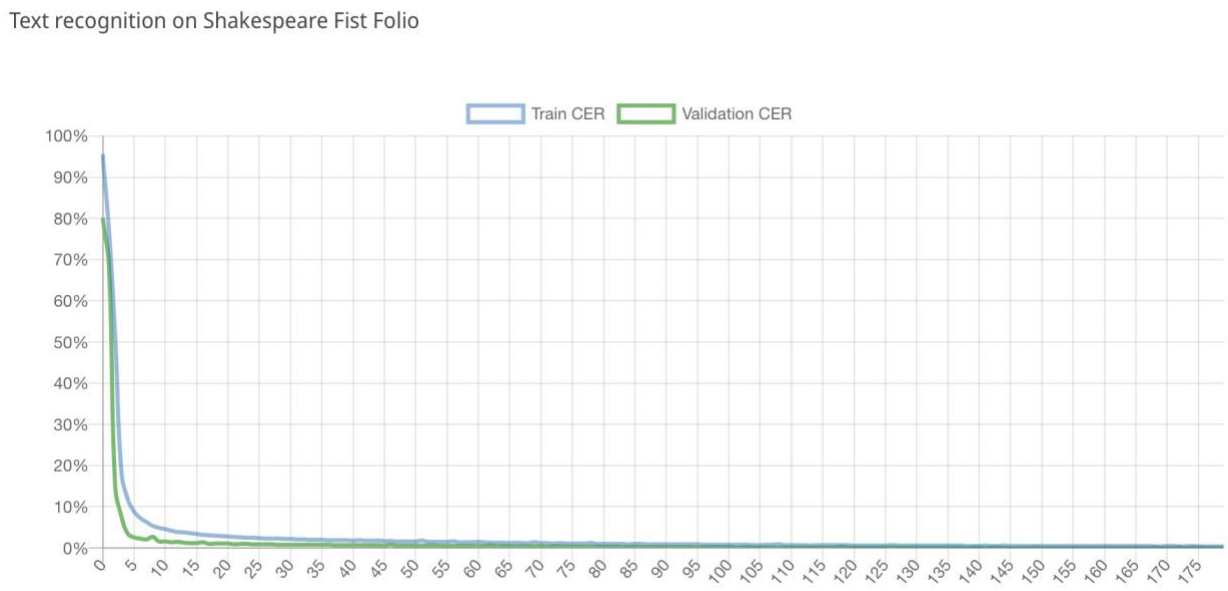
Using the data obtained we were able to train a recognition model with the already mentioned Trankribus Lite and the Pylaia HTR engine as base model. For what concerns the parameters we relied on the guidelines provided by Trankribus[9].

[9] How To Train and Apply Handwritten Text Recognition Models in Transkribus Lite (url: https://readcoop.eu/transkribus/howto/how-to-train-and-apply-handwritten-text-recognition-models-in-transkribus-lite/ - last accessed: 11/1/2023)

After having defined the name, language, and description of the model we proceeded to the definition of the main parameters: we stuck to an early stopping of 20 epochs, a maximum epoch of 300, and a learning rate of 0,0003.

Next, we selected the data we wanted to be included in our set of training, keeping in consideration just two of the three plays and setting aside Julius Caesar to for it to be used as validation set, these test pages will be used to assess the accuracy of our model.

The graph obtained is the following:



The learning curve signifies the accuracy of our model: the y-axis is defined as “Accuracy in CER” and is indicated in percentage. “CER” stands for Character Error Rate, i.e. the ratio of characters that have been transcribed incorrectly by the Text Recognition model.

The x-axis instead is defined as “Epochs”, in our model the training set was divided into 178 epochs.

The graph shows two lines: the blue one represents the progress of the training, while the red line represents the progress of evaluations on the validation set

Looking at the percentage values relating the CER for the training and the validation set we can see how, in our case, the validation set is slightly more performative than the train set showing a CER rate of 0, 00% in the first case and of 0, 05% for the second. As it is explained in the transcription guidelines, results with a CER of 10% or below can be seen as very efficient for automated transcription; considering this, the First Folio model can be considered highly effective.

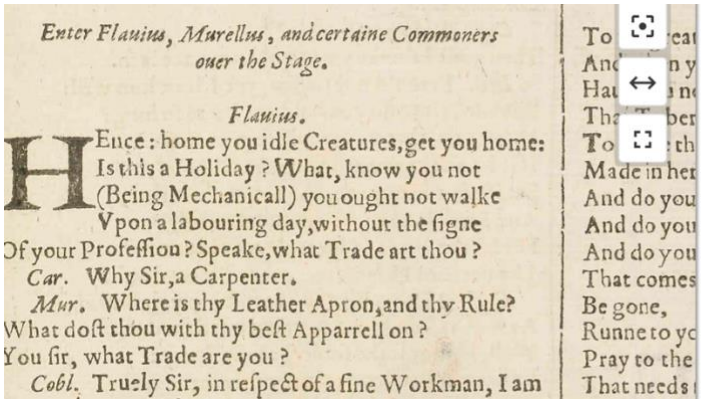
ANNOTATION AND USE:

In designing our annotation campaign, we have tried to apply the FAIR principles

OUTCOMES AND CRITICALITIES:

Given the obtained results, the amount of the starting corpus selected and the small team behind the annotation, the outcome of the campaign was considered undeniably satisfactory, although a lot can still be improved, starting from some criticalities that emerged along the process.

What we all personally all struggled with Transkribus and its strictness as a tool, for example making it difficult to represent specific elements (e.g.: drop caps) inside of the same area; alongside this Transkribus has different kind of problematics at a software level, such as its inability to automatically save progress and the fact that it is quite slow, which have disrupted our work and slowed us down.



This feedback might be further useful to create a better version of this useful and easy-to-use tool.

FURTHER WORKS:

The model we built could be extended to the entirety of the works which are hosted in The Bodleian First Folio to further and other works could be explored and transcribed to further investigate the correctness of our model and to automatically transcribe more texts.

This wasn't possible due to dimensions of the team and the need, for this first part of the work, to manually annotate which has taken a lot of time and effort to accomplice in the best way possible, still we hope in the future to further work on this.

SITOGRAPHY:

Bodleian Libraries: https://www.bodleian.ox.ac.uk/home

Ground Truth Guidelines: https://ocr-d.de/en/gt-guidelines/trans/

The Bodleian First Folio: https://firstfolio.bodleian.ox.ac.uk

Transkribus Lite: https://transkribus.eu/lite/it

Transkribus How To Guides: https://readcoop.eu/transkribus/resources/how-to-guides/

- How To Transcribe Documents with Transkribus – Introduction:

https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribusintroduction/

- How To Train and Apply Handwritten Text Recognition Models in Transkribus:

https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-intranskribus/

- How To Use Transkribus in 10 Steps:

https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/