

# SFFTRANSCRIPTION

Shakespeare First Folio Transcription

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

<b><i>SFFTRANSCRIPTION</i></b> .....	<b>2</b>
<b><i>ANNOTATION PIPELINE</i></b> : .....	<b>2</b>
<b>TASK DEFINITION</b> : .....	<b>2</b>
<b>PILOT</b> : .....	<b>2</b>
<b>TRANSCRIPTION GUIDELINES</b> :.....	<b>3</b>
<b>LAYOUT ANALYSIS</b> :.....	<b>3</b>
<b><i>CAMPAIGN</i></b> : .....	<b>5</b>
<b>ANNOTATION AND USE</b> : .....	<b>6</b>
<b><i>OUTCOMES AND CRITICALITIES</i></b> : .....	<b>6</b>
<b>FURTHER WORKS</b> : .....	<b>7</b>
<b><i>SITOGRAPHY</i></b> : .....	<b>7</b>

**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*<sup>1</sup>, a platform for the digitization, AI-powered text recognition, transcription and searching of historical documents.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<sup>2</sup> 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*<sup>3</sup>.

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<sup>4</sup>) used to better represent the text. We merged together our own little knowledge about printed text of the 16<sup>th</sup> 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*<sup>5</sup> 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<sup>6</sup>; we worked together through it to properly choose the most

---

<sup>1</sup> *Transkribus Lite* (url: <https://transkribus.eu/lite/it> - last accessed: 11/1/2023).

<sup>2</sup> All rights reserved to the original owner and publisher - <https://www.bodleian.ox.ac.uk/home> (last accessed: 11/1/2023).

<sup>3</sup> *How To Use Transkribus in 10 Steps* (url: <https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/> - last accessed: 11/1/2023).

<sup>4</sup> *Ground Truth Guidelines* (url: <https://ocr-d.de/en/gt-guidelines/trans/> - last accessed: 11/1/2023).

<sup>5</sup> All rights reserved to the original owner and publisher - (url: <https://firstfolio.bodleian.ox.ac.uk> - last accessed: 11/1/2023).

<sup>6</sup> *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<sup>7</sup>. For instance: ae in Caesar.
- Hyphenation was transcribed according to the original<sup>8</sup>.
- 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.

---

<sup>7</sup>Ground Truth Guidelines (Ligatures- Level 1) (url: [https://ocr-d.de/en/gt-guidelines/trans/tr\\_level\\_1.html#tr\\_level\\_1](https://ocr-d.de/en/gt-guidelines/trans/tr_level_1.html#tr_level_1) - last accessed: 11/1/2023).

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

*Enter King Iohn, Queen Eleanor, Pembroke, Essex, and Salisbury, with the Chastellan of France.*

*King Iohn.*  
*Drop-capital* **W**hat say Chastillon, what would France with vs?  
*Chas.* Thus (after greeting) speaks the King of France,  
 In my behauiour to the Maiesty,  
 The borrowed Maiesty of England here.  
*Elea.* A strange beginning: borrowed Maiesty?  
*K. Iohn.* Silence: (good mother) heare the Embassie.  
*Chas.* Philip of France, in right and true behalle  
 Of thy decealed brother, *Goffreyes sonne,*  
*Arthur Plantagenet,* laies most lawfull claime  
 To this faire land, and the Territories;  
 To *Ireland, Paylliers, Annes, Torayne, Maine,*  
 Desiring thee to lay aside the sword  
 Which swaies vsurpingly these severall titles,  
 And put the same into young *Arthurs* hand,  
 Thy Nephew, and right royall Soueraigne.  
*K. Iohn.* What followes if we disallow of this?  
*Chas.* The proud controule of fierce and bloody warre,  
 To enforce these rights, so forcibly with-held,  
*K. Iohn.* Heere haue we war for war, & blood for blood,  
 Contrelement for contrelements: so answer *France.*  
*Chas.* Then take my Kings defiance from my mouth,  
 The farthest limit of my Embassie.  
*K. Iohn.* Beare mine to him, and so depart in peace,  
 Be thou as lightning in the eyes of *France;*  
 For ere thou canst report, I will be there:  
 The thunder of my Cannon shall be heard,  
 So lence: be thou the trumpet of our wrath,  
 And fullen preface of your owne decay:  
 An honourable conduct let him haue,  
*Pembroke* looke tooke's: farewell *Chastillon.*

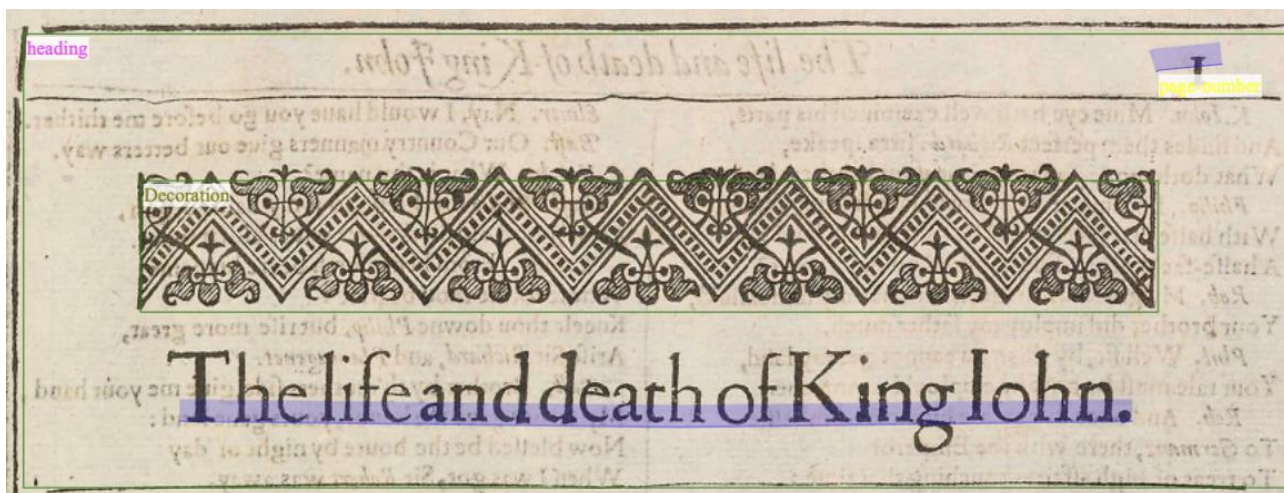
*Exit Chas. and Fem.*

*Elea.* What now my sonne, haue I not euer said  
 How that ambitious *Constance* would not cease  
 Till she had kindled *France* and all the world,  
 Vpon the right and party of her sonne,  
 This might haue bene preuented and made whole  
 With very easie arguments of loue,  
 Which now the manage of two kingdomes must  
 With fearefull bloody issue arbitrate.  
*K. Iohn.* Our strong possession, and our right for vs.  
*Eli.* Your strong possession much more then your right,  
 Or else it must go wrong with you and me,  
 So much my conscience whispers in your eare,

Which none but heauen, and you, and I, shall beare.  
*Enter a Sheriff.*  
*Essex.* My Liege, here is the strangest controuerfie  
 Come from the Country to be iudg'd by you  
 That ere I heard: shall I produce the men?  
*K. Iohn.* Let them approach:  
 Our Abbies and our Pories shall pay  
 This expeditions charge: what men are you?  
*Enter Robert Faulconbridge, and Philip.*  
*Philip.* Your faithfull subiect, I a gentleman,  
 Borne in *Northamptonshire,* and eldest sonne  
 As I suppose, to *Robert Faulconbridge,*  
 A Souldier by the Honor-gluing-hand  
 Of *Cardelou,* Knighted in the field.  
*K. Iohn.* What art thou?  
*Robert.* The son and heire to that *Isabel Faulconbridge.*  
*K. Iohn.* Is that the elder, and art thou the heyre?  
 You came not of one mother then it seemes.  
*Philip.* Most certain of one mother, mighty King,  
 That is well knowne, and as I thinke one father:  
 But for the certaine knowledge of that truth,  
 I put you o're to heauen, and to my mother;  
 Of that I doubt, as all mens children may.  
*Eli.* Out on thee rude man, I doft shame thy mother,  
 And wound her honour with this dissidence.  
*Phil.* I Madame? No, I haue no reason for it,  
 That is my brothers plea, and none of mine,  
 The which if he can proue, a pops me out,  
 At least from faire five hundred pound a yeere:  
 Heauen gurd my mothers honor, and my Land.  
*K. Iohn.* A good blunt fellow: why being younger borne  
 Doth he lay claime to thine inheritance?  
*Phil.* I know not why, except to get the land:  
 But once he slanderd me with bastardy:  
 But where I be as true begot or no,  
 That Bill I lay vpon my mothers head,  
 But that I am as well begot my Liege  
 (Faile fall the bones that tooke the paines for me)  
 Compare our faces, and be Iudge your selfe  
 If old Sir *Robert* did beget vs both,  
 And were our father, and this sonne like him:  
 O old Sir *Robert* Father, on my knee  
 I giue heauen thanks I was not like to thee.  
*K. Iohn.* Why what a mad-cap hath heauen lent vs here?  
*Eli.* He hath a tricke of *Cardelous* face,  
 The accent of his tongue affecteth him:  
 Doe you not read some tokens of my sonne  
 In the large composition of this man?

*K. Iohn.*

- Title (Heading).



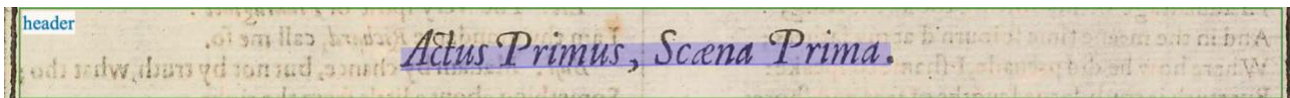
- Page number.



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



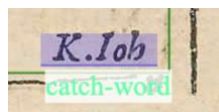
- Header.



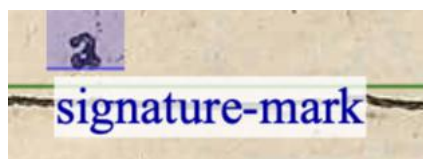
- Drop-caps: the group found the drop capital tag trickier to use as it had to be marked by its own and consequently appears in its own line, separated from the rest of the word.



- 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<sup>9</sup>.

---

<sup>9</sup> How To Train and Apply Handwritten Text Recognition Models in Trankribus 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:

Text recognition on Shakespeare First Folio



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.

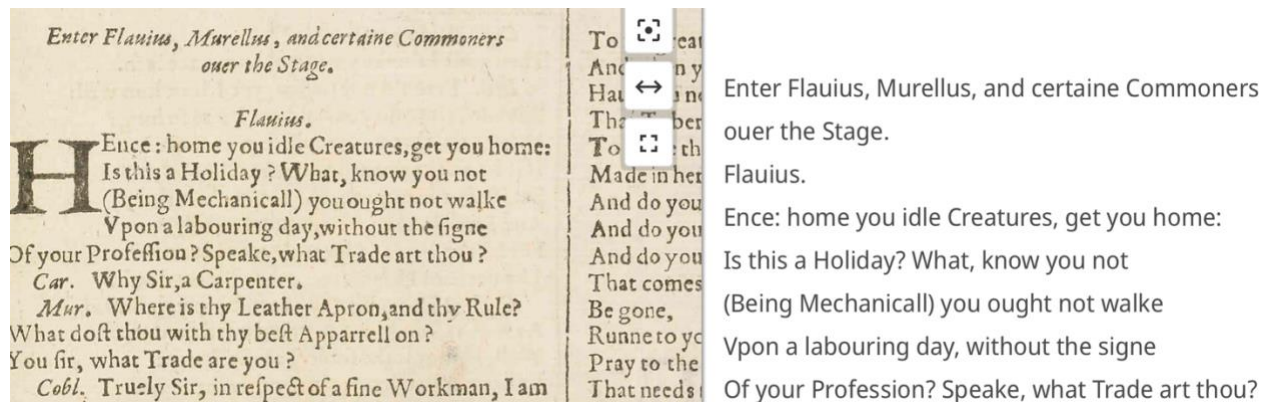
### ANNOTATION AND USE:

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 accomplish 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/>