Automatic Transcription of BnF ms fr 24428 with Transkribus

Estelle Guéville https://orcid.org/0000-0003-2603-1051
Iwona Krawczyk https://orcid.org/0000-0002-9455-2732
David Joseph Wrisley https://orcid.org/0000-0002-0355-1487

Contents

In this work, we report on the training of a handwritten text recognition (HTR) model in Transkribus using corrected and normalized crowd transcribed training data from the Image du Monde challenge (8-22 January 2021). Our report is accompanied by data in plain text format, including the following files :

- BNFfr24428_GT_23Jan21.txt
- BNFfr24428_Omons_entiretranscript.txt
- BNFfr24428_GT_10Jun21.txt
- BNFfr24428_Omons2_entiretranscript.txt

As well as the following documentation:

- A Manuscript Description
- A Statement of Transcription Principles
- Comments on the Process
- Brief observations on the performance of the models
- Future Work

Manuscript Description

This report contains the automatic transcriptions of the manuscript held in the Bibliothèque nationale de France, ms. français 24428. This manuscript, dated from the 13th century, has 118 folios of parchment and measures 315×220 mm. The leather binding bears the arms of Louis-Philippe. According to archivesetmanuscrits.bnf.fr this manuscript contains the following texts:

- *L'Image du Monde* by Gautier de Metz (fol. 1r-46r)
- Summary of the previous text (fol. 47r-48v)
- *Li Volucraires*, a moralizing poem about birds by the scribe Omons (fol. 49r-52r)
- *Li Bestiaires divin* by Guillaume Le Clerc (fol. 53r-78v)

- *Li Laipidaires* (fol. 79r-88v)
- *Fables* by Marie de France (fol. 89r-114v)
- *Instruction pour la confession* (fol. 115r-118r)
- Ex-libris: *Cest livre cy est à mestre Nicholas de Lessy, et le m'a ledit mestre Nicholas presté à moy frere Jehan Cotusse, gardien des Freres Mineurs de Sens, M CCCC et XII* (fol. 118v).

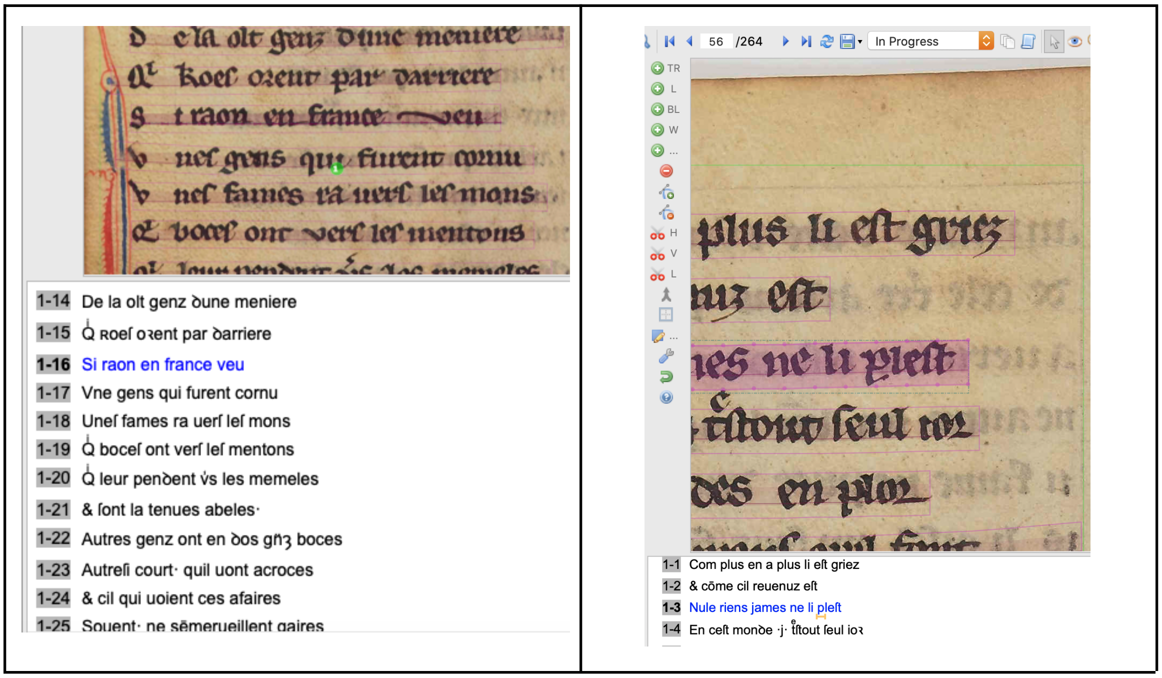
Statement of transcription principles

There is not yet consensus in the scholarly community about norms for transcription for the HTR of medieval manuscripts.

Transcribing a writing system that is not easily represented in digital text is a challenging process. We see it as a holistic one rather than a prescriptive one.

In general, creating ground truth for HTR, we used the following principles.

(1) We avoid "silent" expansion and collapsing different letter forms into modern spelling.
(2) To transcribe what we see using Unicode characters that are as simple as possible.
(3) In some examples found in a manuscript, there is an "aesthetic" quality to graphemes that we chose not to reproduce (initial letter, v, p).


1-14 De la olt genz ꝺune meniere
1-15 Qͥ ʀoeſ oꝛent par ꝺarriere
1-16 Si raon en france veu
1-17 Vne gens qui furent cornu
1-18 Uneſ fames ra uerſ leſ mons
1-19 Qͥ boceſ ont verſ leſ mentons
1-20 Qͥ leur penꝺent v́s les memeles
1-21 & ſont la tenues abeles·
1-22 Autres genz ont en ꝺos gn̄ʒ boces
1-23 Autreſi court· quil uont acroces
1-24 & cil qui uoient ces afaires
1-25 Souent· ne sēmerueillent gaires

1-1 Com plus en a plus li eſt griez
1-2 & cōme cil reuenuz eſt
1-3 Nule riens james ne li pleſt
1-4 En ceſt monꝺe ·j· tͤſtout ſeul ioꝛ

Figure 1: Examples of "aesthetic graphemes" (*v* in 1-16 at left and *p* in 1-3 at right) from BnF fr 24428, visualized in the Transkribus interface.

(4) There is also a variance in certain abbreviations (e.g., the macron) where it is difficult to encode its exact position.

(5) Since we are creating machine readable transcriptions for the purpose of computational research, we did not choose existing Unicode that will not display in regular text editors (e.g., green letters in Medieval Unicode Font Initiative - MUFI), and we avoided transcription practices that will be undone by common, downstream NLP practices (e.g., lower casing, tokenization, removing punctuation)

(6) In the case of abbreviations we either used existing Unicode characters or used combining Unicode characters.

(7) In the case of contradictory decisions, there is a need to prioritize the principles.

  (a) In the case of abbreviations such as the macron that sits between letters, we need to consider what abbreviation replaces. For the expanded word *bien* which appears as three letters *b-i-e* with a macron, we would transcribe it *biē* rather than *bīe*. Since combining characters are actually sequential characters, we chose to place a macron just before the letters that it replaces. Prior knowledge of the language has been applied with point 5 above in mind.

  (b) In the case of letter forms with some variance (w, v, h, p), we had to decide if we draw upon the richness of Unicode (see point 3 above). We established a general principle of looking at a larger sample of the letters in the manuscript before making the choice. The same is true of spacing between words. We found that it is useful to have a preliminary "scan" of the document, or even a first pass of automatic transcription, and make a general character map in order to decide on this question.

(8) When you arrive at an example of a palimpsest or corrections on or off the baseline, it is desirable to tag such a phenomenon so that it can be excluded from the model training process.

(9) If we arrive at codicological "extras", things that are not part of the main text (marginalia, catchwords), it is general practice not to transcribe these.

Comments on the Process

The two transcripts and models generated here followed slightly different procedures.

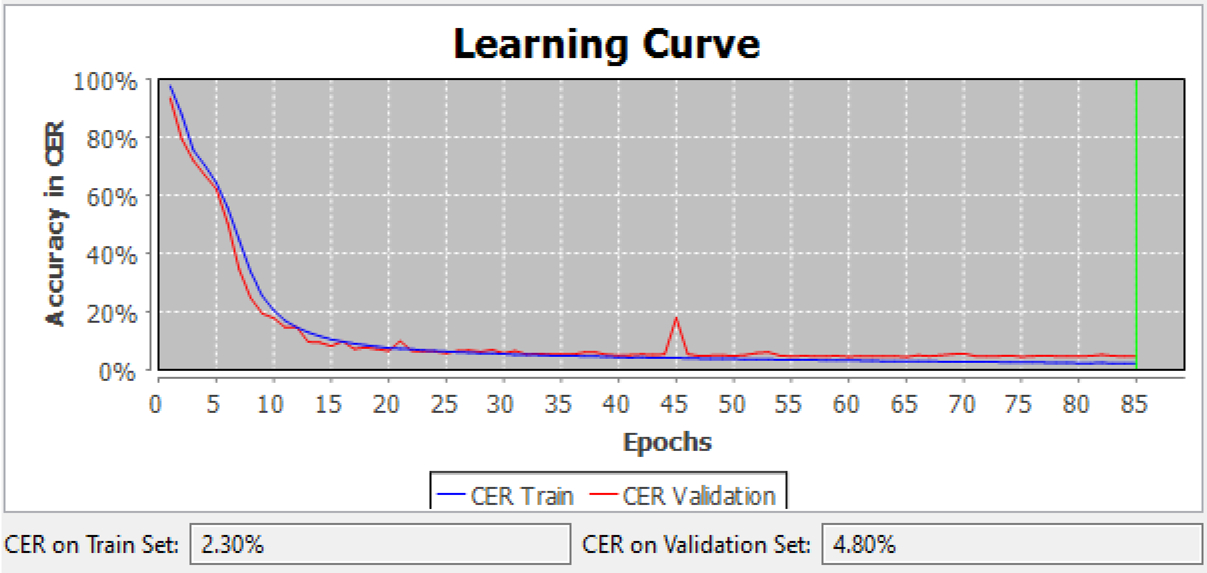
Models

Model 1:

The first model used the output from the Image du Monde Challenge 2 organised by Laura Morreale as training data. For phase II, team 5 transcribed from the beginning of the work to Book 2, Chapter 4. In the BNF français 24428, the text begins at fol. 21v, column 1, line 29 and ends at fol. 48r.

During the challenge, the team worked with some customizations of the transcription interface FromThePage recently implemented by the developers, Ben and Sara Brumfield, which allowed us to capture even more details about the scribal practices of the manuscript. This work expanded on an experiment done by Team 4 with an annotation function allowing transcribers to add the unexpanded abbreviations and for them to appear in a pop up with standard hovering behavior over the document, effectively creating--and visualizing--two layers of data in this transcription interface. The first layer adhered to the guidelines created for the first part of the challenge, which followed the typical practices for editing medieval manuscripts; on the other hand, the second layer transcribed the text capturing the original abbreviations and letter forms without their "silent" expansion and normalization. In addition to the special letter forms and abbreviations, layer 2 also preserved the spacing, punctuation and spelling of the scribe, even when the transcription team suspected that there might be an error, avoiding all modern accents. In the hours following the end of the challenge, we used the uncorrected second layer (BNFfr24428_GT_23Jan21.txt) resulting from the collectively transcribed pages of Team 5, phase II to train the model and to transcribe some sample pages (see Figures 2 and 3 below).

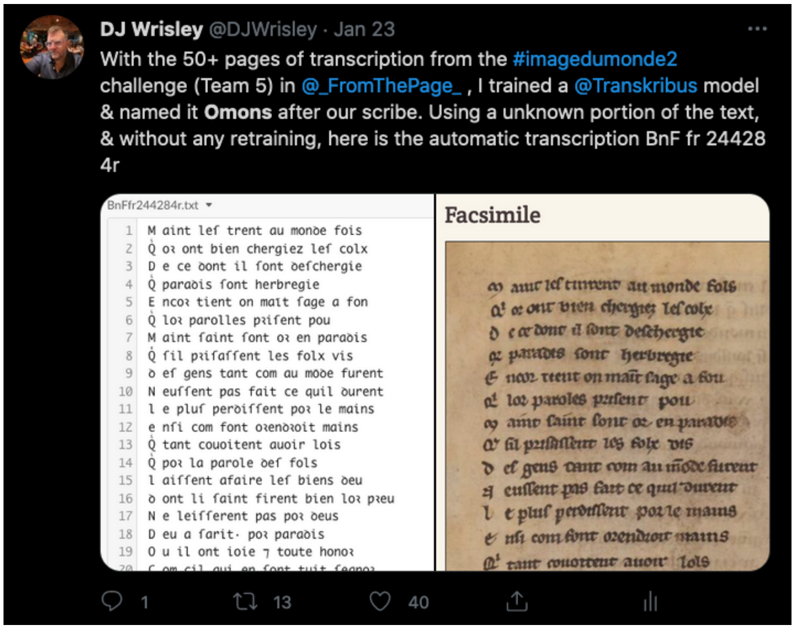
The first model was trained with Pylaia on a total of 52 pages, counting 3033 lines and 19415 words, 47 pages being the train set and 5 pages the validation set. The training of the model was configured for 250 epochs but stopped early at 85. The character error rate we obtained gets much better after 10-15 epochs (with 10% accuracy in CER) and then slowly decreases until it reaches 2.30% CER on the train set and 4.80% CER on the validation set after 85 epochs. We named the model after the supposed scribe of the manuscript, *Omons*.



CER on Train Set: 2.30%
CER on Validation Set: 4.80%

Figure 2: The learning curve for the *Omons* model in Transkribus

In the weeks after the training of the *Omons* model, a full transcript of the manuscript was achieved using it (BNFfr24428_Omons_entiretranscript.txt)



DJ Wrisley @DJWrisley · Jan 23

With the 50+ pages of transcription from the #imagedumonde2 challenge (Team 5) in @_FromThePage_ , I trained a @Transkribus model & named it **Omons** after our scribe. Using a unknown portion of the text, & without any retraining, here is the automatic transcription BnF fr 24428 4r

BnFfr244284r.txt

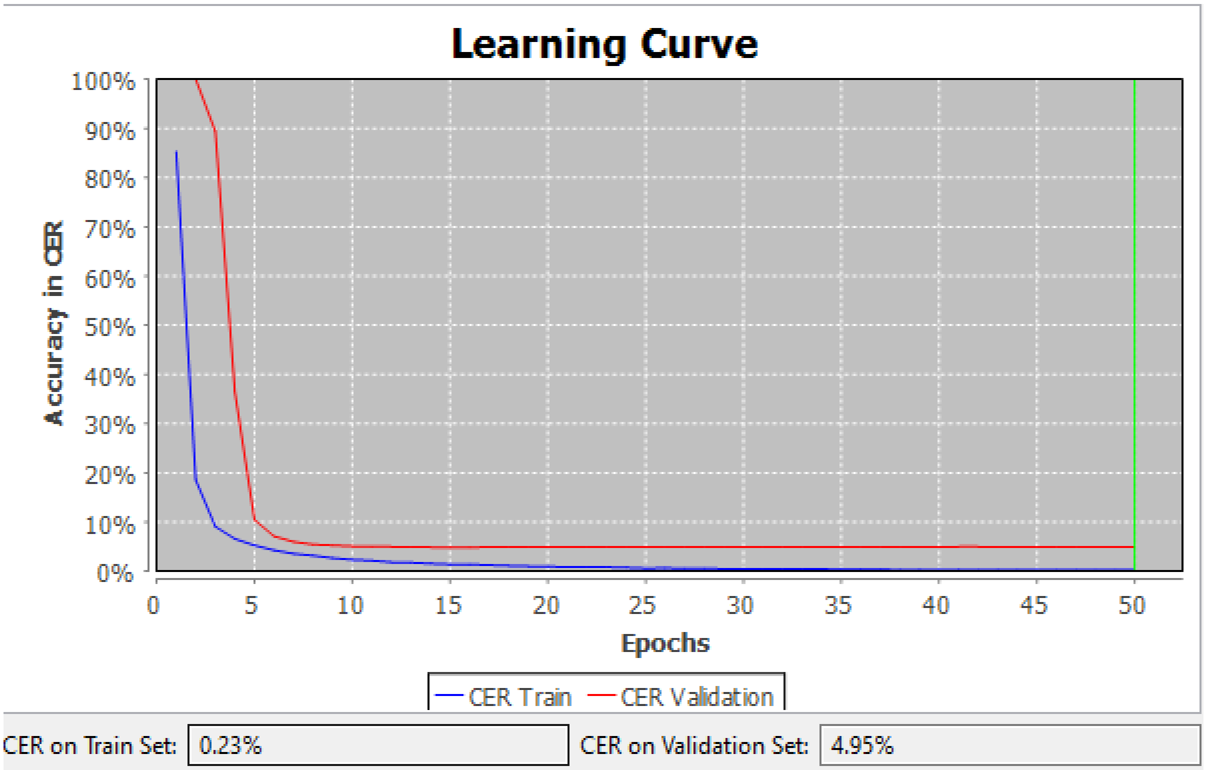
```
M aint leſ trent au monꝺe fois
Q̇ oꝛ ont bien chergiez leſ colx
D e ce ꝺont il ſont ꝺeſchergie
Q̇ paraꝺis ſont herbregie
E ncoꝛ tient on maıt ſage a ſon
Q̇ loꝛ parolles pꝛiſent pou
M aint ſaint ſont oꝛ en paraꝺis
Q̇ ſil pꝛiſaſſent les folx vis
ꝺ eſ gens tant com au mōꝺe furent
N euſſent pas fait ce quil ꝺurent
l e pluſ perꝺiſſent poꝛ le mains
e nſi com ſont oꝛenꝺꝛoit mains
Q̇ tant couoitent auoir lois
Q̇ poꝛ la parole ꝺeſ fols
l aiſſent afaire leſ biens ꝺeu
ꝺ ont li ſaint firent bien loꝛ pꝛeu
N e leiſſerent pas poꝛ ꝺeus
D eu a farit· poꝛ paraꝺis
O u il ont ioie ⁊ toute honoꝛ
C om cil qui en ſont tuit ſeanoꝛ
```

Figure 3: A tweet on 23 January 2021 by @DJWrisley illustrating the character recognition of the *Omons* model on folio 4r of BnF ms fr 24428

Model 2:

The second transcript stems from a correction of the folios transcribed during the challenge, according to the above statement of transcription principles. The following folios were corrected to ground truth status: 21v-25v, 47r-48r, 50r-50v, 52r, 53r, 53v-54r, 56r, 57v, 63v, 69v, 73r, 75r, 81r, 84v, 87v, 90r, 92v, 93v, 95v, 97r, 117r-118r. They correspond to the following pages in the IIIF document in Transkribus: 56-64, 107-109, 113-114, 117, 119, 120-121, 125, 128, 140, 152, 159, 163, 175, 182, 188, 193, 198, 200, 204, 207, 247-249. They are contained in the file BNFfr24428_GT_10Jun21.txt. The rationale for the choice of these folios was to add to the model a combination of folios with slight hand changes along with random folios within each of the individual texts mentioned above.

This model was trained with CITlab HTR+ on a total of 35 folios, 917 lines and 5520 words, 15 pages being the training set and 20 pages the validation set. Over 50 epochs, we see clearly in Figure 4 below that the performance of the model does not get significantly better after the first 5 epochs. This training run resulted in a character error rate (CER) of 0.23% of error on the train set, and 4.95% on the validation set. At this time of this report, we have not corrected and retrained the model.



CER on Train Set: 0.23%   CER on Validation Set: 4.95%

Figure 4: The learning curve for the *Omons2* model in Transkribus

Layout analysis was run on the entire remaining parts of the manuscript and were scrupulously corrected by hand.

Using *Omons2*, we auto-transcribed the remaining folios of the manuscript (1r-21r, 26r-46r, 49r-49v, 51r-51v, 54v-55v, 56v-57r, 58r-63r, 64r-69r, 69v-72v, 73v-74v, 75v-80v, 81v-84r, 85r-87r, 88r-89v, 90v-92r, 93r, 94r-95r, 96r-96v, 97v-116v), corresponding to the following pages in the IIIF document in Transkribus (15-55, 65-105, 111-112, 115-116, 122-124, 126-127, 129-139, 141-151, 152-158, 160-162, 164-174, 176-181, 183-187, 189-192, 194-197, 199, 201-203, 205-206, 208-246). Folios 46v, 48v and 52v (pages 106, 110, 118) were omitted because they are blank.

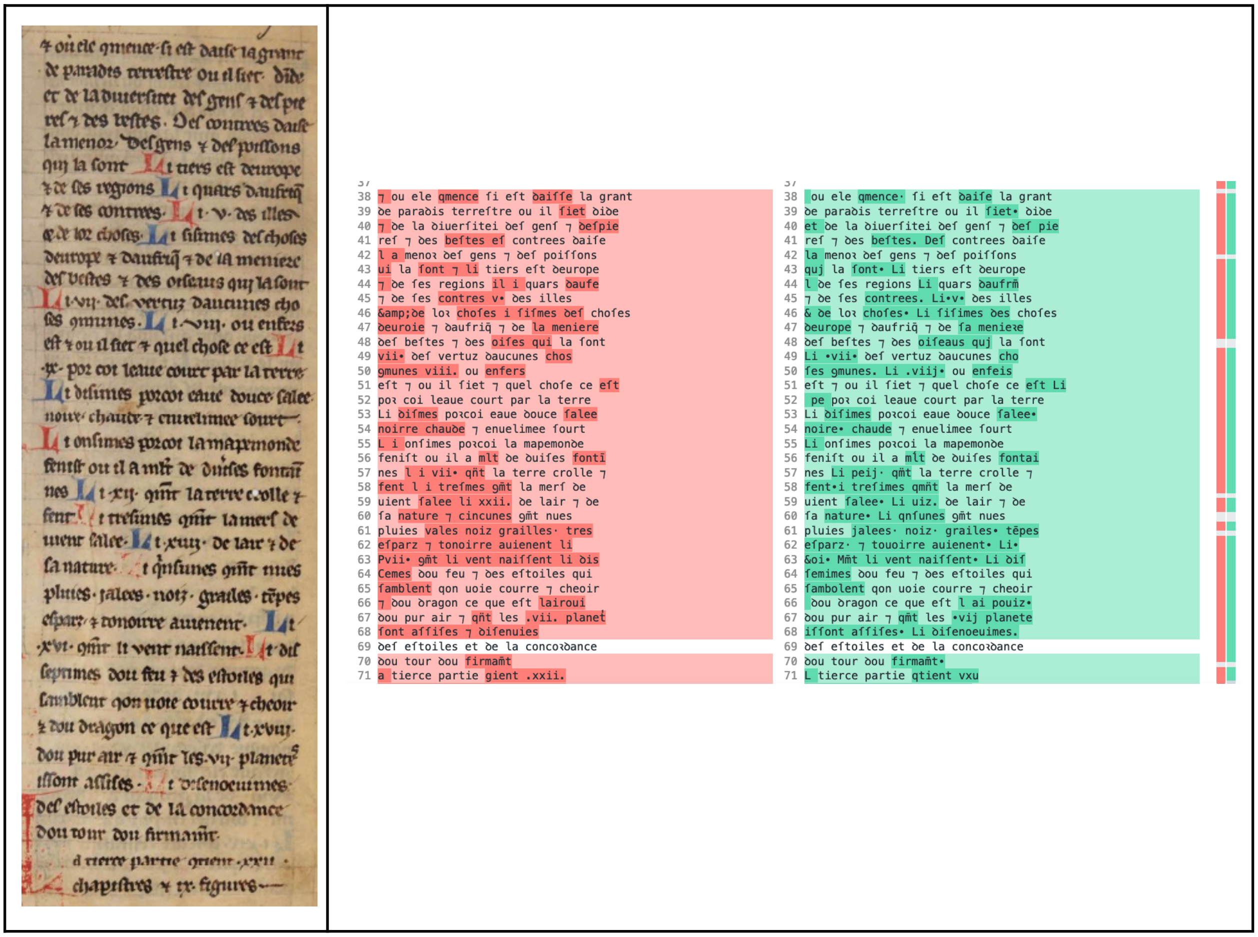
The full transcript of fol. 1r-118r (pages 15-249 in Transkribus) auto-transcribed with *Omons2* is published here as BNFfr24428_Omons2_entiretranscript.txt

Brief observations on the performance of the models

We noticed differences in the performance between the *Omons* and the *Omons2* model, namely

- Omons2 did a better job in the case of "aesthetic" spacing
- Neither model was strong in the case of Roman numerals
- Omons2 deals better with dots

- Omons2 recognized the small red and blue initials with ease
- Omons2 generally distinguishes minims better



```
⁊ ou ele qmence ſi eſt ꝺaiſſe la grant
ꝺe paraꝺis terreſtre ou il ſiet ꝺiꝺe
⁊ ꝺe la ꝺiuerſitei ꝺeſ genſ ⁊ ꝺeſpie
reſ ⁊ ꝺes beſtes eſ contrees ꝺaiſe
l a menoꝛ ꝺeſ gens ⁊ ꝺeſ poiſſons
ui la font ⁊ li tiers eſt ꝺeurope
⁊ ꝺe ſes regions il i quars ꝺaufe
⁊ ꝺe ſes contres v• ꝺes illes
&amp;ꝺe loꝛ choſes i ſiſmes ꝺeſ choſes
ꝺeuroie ⁊ ꝺaufriꝗ ⁊ ꝺe la meniere
ꝺeſ beſtes ⁊ ꝺes oiſes qui la ſont
vii• ꝺeſ vertuz ꝺaucunes chos
ꝯmunes viii. ou enfers
eſt ⁊ ou il ſiet ⁊ quel choſe ce eſt
poꝛ coi leaue court par la terre
Li ꝺiſmes poꝛcoi eaue ꝺouce ſalee
noirre chauꝺe ⁊ enuelimee ſourt
L i onſimes poꝛcoi la mapemonꝺe
feniſt ou il a mlt ꝺe ꝺuiſes fontī
nes l i vii• qn̄t la terre crolle ⁊
fent l i treſmes ꝯm̄t la merſ ꝺe
uient ſalee li xxii. ꝺe lair ⁊ ꝺe
ſa nature ⁊ cincunes ꝯm̄t nues
pluies vales noiz grailles· tres
eſparz ⁊ tonoirre auienent li
Pvii• ꝯm̄t li vent naiſſent li ꝺis
Cemes ꝺou feu ⁊ ꝺes eſtoiles qui
ſamblent qon uoie courre ⁊ cheoir
⁊ ꝺou ꝺragon ce que eſt lairoui
ꝺou pur air ⁊ qn̄t les .vii. planet
font aſſiſes ⁊ ꝺiſenuies
ꝺeſ eſtoiles et ꝺe la concoꝛꝺance
ꝺou tour ꝺou firmam̄t
a tierce partie gient .xxii.
```

```
 ou ele qmence· ſi eſt ꝺaiſe la grant
ꝺe paraꝺis terreſtre ou il ſiet• ꝺiꝺe
et ꝺe la ꝺiuerſitei ꝺeſ genſ ⁊ ꝺeſ pie
reſ ⁊ ꝺes beſtes. Deſ contrees ꝺaiſe
la menoꝛ ꝺeſ gens ⁊ ꝺeſ poiſſons
quj la ſont• Li tiers eſt ꝺeurope
l ꝺe ſes regions Li quars ꝺaufrm̄
⁊ ꝺe ſes contrees. Li•v• ꝺes illes
& ꝺe loꝛ choſes• Li ſiſimes ꝺes choſes
ꝺeurope ⁊ ꝺaufriꝗ ⁊ ꝺe ſa menieꝛe
ꝺeſ beſtes ⁊ ꝺes oiſeaus quj la ſont
Li •vii• ꝺeſ vertuz ꝺaucunes cho
ſes ꝯmunes. Li .viij• ou enfeis
eſt ⁊ ou il ſiet ⁊ quel choſe ce eſt Li
 pe poꝛ coi leaue court par la terre
Li ꝺiſimes poꝛcoi eaue ꝺouce ſalee•
noire• chaude ⁊ enuelimee ſourt
Li onſimes poꝛcoi la mapemonꝺe
feniſt ou il a mĺt ꝺe ꝺuiſes fontai
nes Li peij· qm̄t la terre crolle ⁊
fent•i treſimes qmn̄t la merſ ꝺe
uient ſalee• Li uiz. ꝺe lair ⁊ ꝺe
ſa nature• Li qnſunes ꝯm̄t nues
pluies jalees· noiz· grailes• tēpes
eſparz· ⁊ touoirre auienent• Li•
&oi• Mm̄t li vent naiſſent• Li ꝺiſ
ſemimes ꝺou feu ⁊ ꝺes eſtoiles qui
ſambolent qon uoie courre ⁊ cheoir
 ꝺou ꝺragon ce que eſt l ai pouiz•
ꝺou pur air ⁊ qm̄t les •vij planete
iſſont aſſiſes• Li ꝺiſenoeuimes.
ꝺeſ eſtoiles et ꝺe la concoꝛꝺance
ꝺou tour ꝺou firmam̄t•
L tierce partie qtient vxu
```

Figure 5: Model 1 results for fol. 1r, col. 2 (red) compared with Model 2 results for the same text (green), visualized using diffchecker.com, with the corresponding portion of the manuscript at left)

Future work

In future work, we will engage in post-correction, retraining the model after correcting selected pages of what is not understood by the model.