

Transcrire (automatiquement)

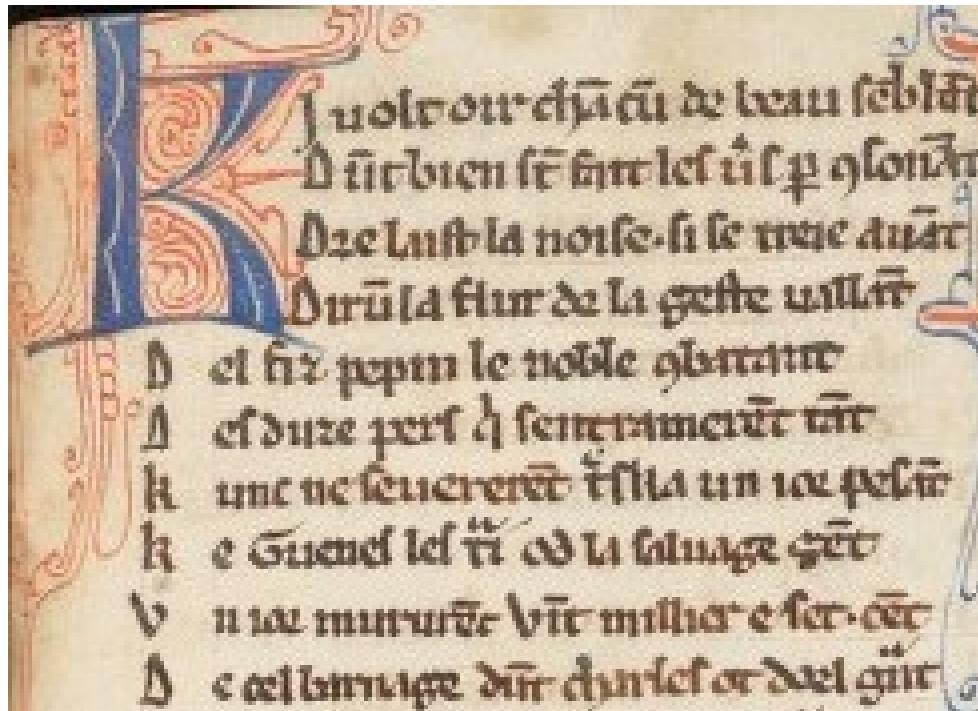
Alexandre Bartz, Simon Gabay



Une nouvelle philologie?

Editer (avant)

Ms Cologny, Fondation bodmer, Ms. 168



<https://www.e-codices.unifr.ch/en/fmb/cb-0168/211r>

Editer (avant)

Otinel, chanson de geste publiée pour la première fois, d'après les manuscrits de Rome et de Middlehill, éd. F. Guessard et H. Michelant, Paris: Jannet, 1859



UI veust oïr chançon de biau semblant,
Si face paiz, si se traie en avant,
S'orra la flor de la geste vaillant
Du fiz Pepin, le riche roi poissant,
Des .xii. pers, qui s'entramerent tant:
Tant s'entramerent, ce trovon nos lisant,
Ne se grepirent onques en lor vivant
De ci au jor que il furent morant
En Roincevaux, où furent combatant
Contre Garsile, le riche roi poissant,
Que li fel Guennes, le cuvers sodiant,
Les i vendi, ce sevent li auquant.
Cel jor méismes qu'il furent combatant,
En i morut .xxx^m. et .vii. cent
De noz barois, dont Kalles fu dblast.
Cil jugléour n'en dient tant ne quant;

<https://archive.org/details/floovantchanson00floogoog>

Editer (maintenant)

Camps, Jean-Baptiste, *La Chanson d'Otinel. Édition complète du corpus manuscrit et prolégomènes à l'édition critique*, Paris:Université de Paris-Sorbonne, 2016

I

Ki volt oïr chancun de beau semblant
dunt bien *sunt* fait les *vers par cunsonant*
ore laist la noise, si se treie avant :
dirum la flur de la geste vallant
5 del fiz Pepin le noble *cumbatant*
des duze pers qui s'entramerent tant

k'unc ne severerent tresk'a un jor pesant

ke Guenes les *traï* od la salvage gent

10 un jor mururent *vint* millier e set cent
de cel barnage, dunt Charles ot doel grant.

Editer (maintenant)

Camps, Jean-Baptiste, *La Chanson d'Otinel. Édition complète du corpus manuscrit et prolégomènes à l'édition critique*, Paris:Université de Paris-Sorbonne, 2016

I

KJ uolt oir chācū ðe beau fēblāt
Dūt bien ðt fait lef ūf p 9sonāt
Oze laſt la noise si se treie auāt
Dirū la flur ðe la geste uallāt

- s Del fiz pepin le noble 9batant
Def ðuze perf q sentramerēt tāt
kunc ne feuererēt t̄lka un 102 pesāt
ke Guenes lef t̄i oð la saluage gēt
vn 102 mururēt vñt millier e fet .cēt
10 De cel barnage ðüt charlef ot ðoel ȝnt

<https://halshs.archives-ouvertes.fr/tel-01664932>

Niveaux de transcription



Dūt bien st̄ fait les ūſ p ɔſonāt

Transcription allographétique:

Dūt bien st̄ fait les ūſ p ɔſonāt

Transcription graphématisée:

Dunt bien sunt fait les *vers* par cunsonant

Transcription modernisée:

Dont bien sont fait les vers par consonant

Terminologie

Robinson et Solopova, "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue", *The Canterbury Tales Project Occasional Papers 1*, 1993

- "Regularized": *all manuscript spellings are regularized to a particular norm*
- "Graphemic": *every manuscript spelling is preserved*
- "Graphic": *every distinct letter-type is distinguished*
- "Graphic": *every mark in the manuscript, every space, is represented in the transcription, even to the point of decomposition of letter forms into discrete marks*

Terminologie: précisions

Dominique Stutzmann, "Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin?", *Kodikologie und Paläo-graphie im digitalen Zeitalter*, 2011.

"Si encoder, c'est décrire":

- "Graphique" et "graphétique" (ou de préférence "allographétique") décrivent l'image
- "Graphémique" (ou "graphématique") décrit le texte-objet soumis aux accidents
- "Regularisé" décrit le texte-idée

Transcrire

Dūt bien ſt̄ fait leſ ūſ p ɔfonāt

est encodé:

```
Dūt bien ſt̄ fait leſ ūſ p ɔfonāt
```

On note:

- des lettres unicodes normales: t (U+0074)
- un caractère combiné: t + ḥ (U+0074 + U+0304)
- une lettre du domaine privé (MUFI): , (U+F1A6), cf. [ici](#)

Pour afficher les caractères du domaine privé, on utilise des polices spéciales, comme Junicode, qui intègre la MUFI (et doit être installée)

```
<span style="font-family:Junicode">le ūſ p ɔson</span>
```

Pour installer Junicode: <https://junicode.sourceforge.io/>

La Medieval Unicode Font Initiative (MUFI)

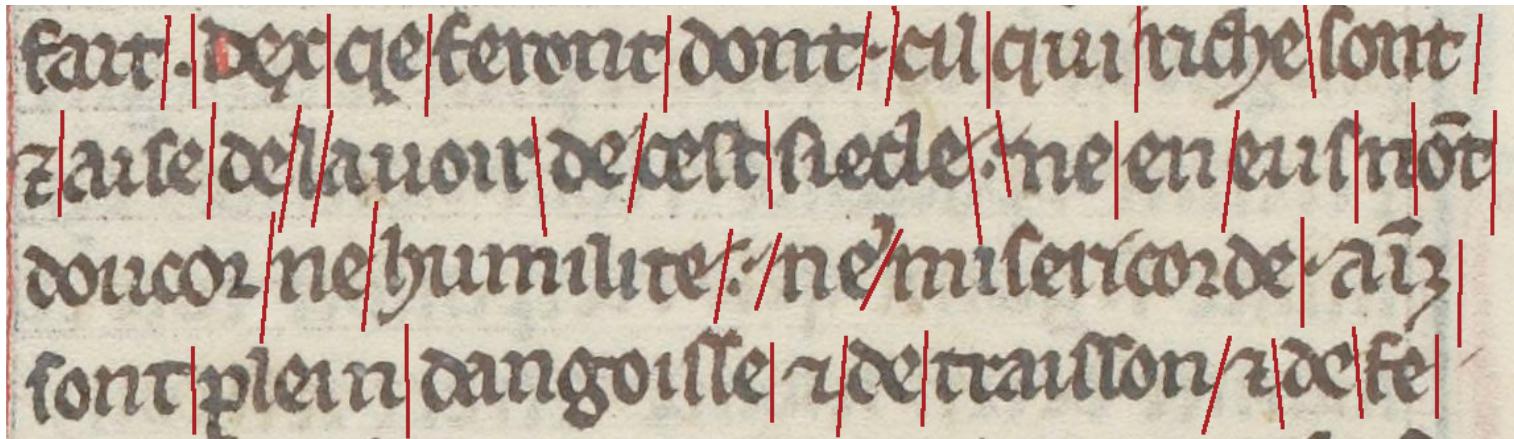
On a tous connu le problème suivant: *expérience* devrait s'afficher, mais c'est *expÚrience* qui s'affiche. C'est une problème d'unicode.

Unicode est un standard informatique qui permet des échanges de textes dans différentes langues, à un niveau mondial.

La MUFI permet d'encoder des lettres médiévales (ou plus généralement rares) absentes d'Unicode: <https://mufi.info>

Segmenter

Thibault Clérice, "Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin",
Journal of Data Mining and Digital Humanities, [Episciences.org](https://episciences.org), 2020



Doit-on conserver les agglutinations?

7 aise de lauoir de ceft siecle. ne en euf nōt

ou bien doit-on désagglutiner?

7 aise de l'auoir de ceft siecle. ne en euf nōt

De nouvelles pratiques

Les possibilités de transcription n'ont jamais été aussi vastes:

- devons-nous continuer comme avant avec du semi-diplomatique, voire une transcription interprétative?
- doit-on exploiter au maximum les possibilités offertes par le numérique?
- doit-on faire les deux?

Cela dépend évidemment de plusieurs paramètres:

- mes questions de recherche
- mes compétences personnelles
- mes moyens financiers

C'est un exemple du tournant épistémologique majeur que constitue le numérique, et non une simple informatisation des pratiques anciennes

Entraînement

Transcrire (I)

Ligne de commandes + interface dans un navigateur

truc/0001/010001.bin.png

P R E F A C E.

PREFACE

truc/0001/010002.bin.png

où il estoit tombé, apres le refus qu'on luy avoit

ou il étoit tombé, apres le refus qu'on luy avoit

truc/0001/010003.bin.png

fait des armes d'Achille. Ils ont admiré le Phi-

fait des armes d'Achille. Ins ont admiré le Phi-

truc/0001/010004.bin.png

loctete , dont tout le sujet est Ulysse, qui vient

loctete, dont tout le sujet est Ulysse, qui vient

truc/0001/010005.bin.png

pour surprendre les fleches d'Hercule. L'Oe-

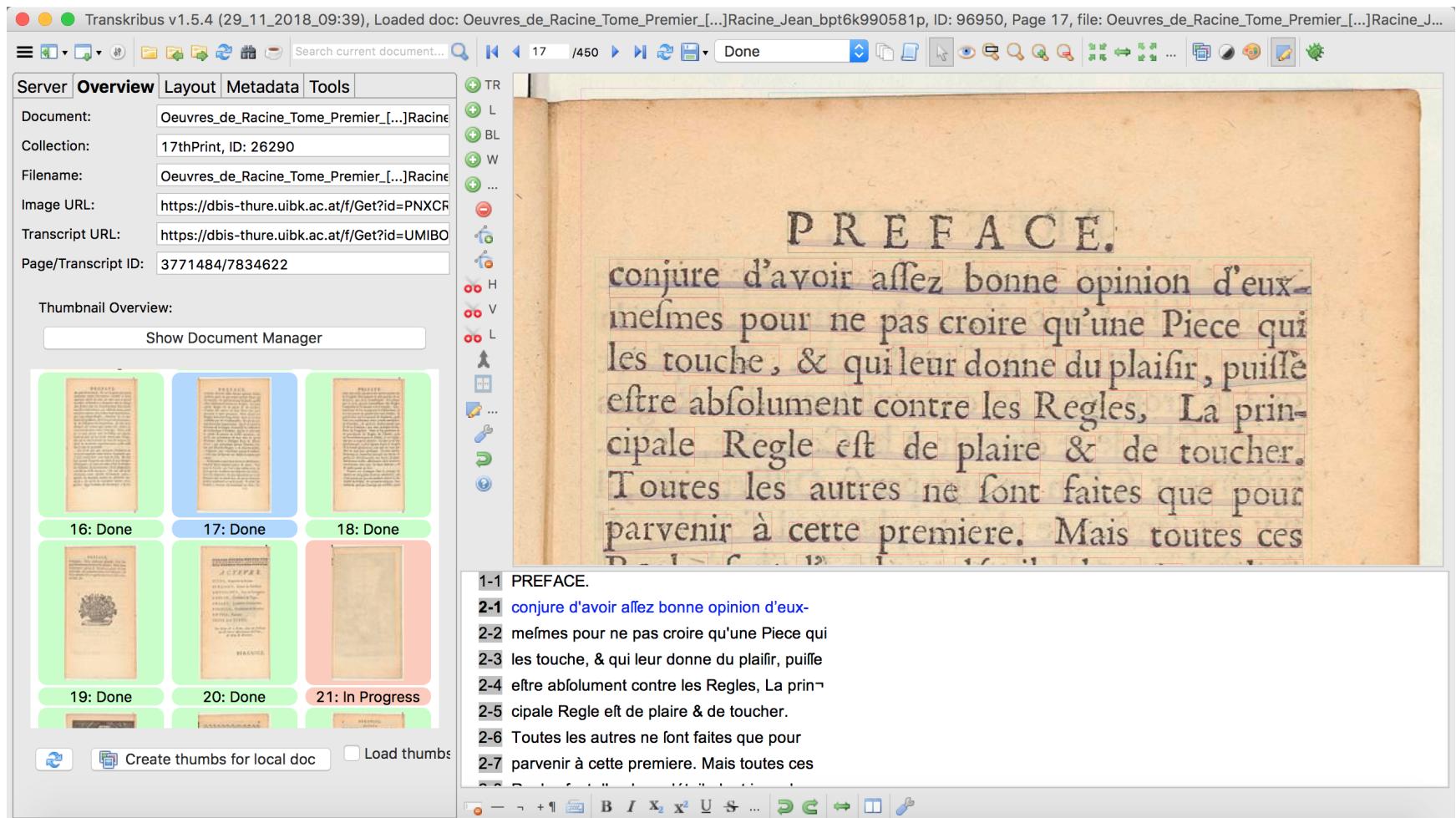
pour surprendre les fleches d'Hercule. L'Oe-

truc/0001/010006.bin.png

dipe mesme , quoy que tout plein de recon-

Transcrire (II)

Transkribus (Innsbruck)



Transcrire (III)

eScriptorium (EPHE/PSL)

eScriptorium Home About Contact

test Description Images Edit Element 1

Pause 00:00:00 Select Area Audio Record Pointer

My Documents Hello dstoekl

kraken:Josephus_abbrev_with_majuscules_best.pronn

The screenshot shows the eScriptorium application interface. At the top, there's a navigation bar with 'eScriptorium', 'Home', 'About', 'Contact', and a user dropdown 'Hello dstoekl'. Below the navigation is a toolbar with tabs for 'test', 'Description', 'Images', 'Edit' (which is selected), and 'Element 1'. To the right of the toolbar are control buttons for 'Pause', 'Select Area', 'Audio', and 'Record Pointer'. A progress bar shows '00:00:00'. On the left, a sidebar has a magnifying glass icon and a dropdown menu. The main area displays a photograph of an open medieval manuscript. The left page contains dense Latin text in two columns. The right page has some text at the top and a large, blank area below it. Overlaid on the manuscript are several digital annotations: a vertical red line with a small red square at the bottom, a horizontal red line with a small red square at the end, and a red bracket highlighting a section of text on the left page. To the right of the manuscript, there's a transcription window titled 'kraken:Josephus_abbrev_with_majuscules_best.pronn'. This window contains a large amount of Latin text in a monospaced font, which is a transcription of the manuscript's content. The transcription is in all-caps and includes some punctuation and line breaks.

Abies fusteri consumelam parvum examen promisi
et carnem mulcendum plurimi nomine
die sol plurimi. Illis autem hoc credemus
et haec quidam unde posset haec canit
milibus que pides ministrare. Dique
et regi malo audirentibus auditis non amorem praebet
sediemus operantes et si non carpe pueret
omnes exercitus collegibusque exercitum aten-
tit. Dicunt nam sedebat credidisse lebenses inuria
et maledictis iste communissimum non parvus
sequendum multa ruderent mortuas et exca-
cenus. Lebenses cognominant calvorum. Alii
quod inveniuntur perire defiderent pulchritudine
et educentur certos et marius que vocatur
camallus excedens auctoriter terminos et
certos ex calabacundum diffidalem populis
intrae congerigant est invenire loquuntur
et dicunt. Unde nobis duobus, primito
liberorum exercituum et exercituum possident
aliud datus amissione claudetur.
Exercitus celerriter. Infimis namq[ue] tene-
tur cumenrū acerbitate neclectus pre-
deret nec res ipsa curia sed his omnis
corrum gentes. Exercitus ergo pro premere
ad bellum. Non enim sine labore nobis hanc
terram concorditer sed maxime eam pro-
ficiunt obediunt. Mox etiam autem explo-
ratores quibusut terrae bona considerant
et quae ut invenientur habentur. Ame-
nis uero unumque sumat et dñi quieti
monimbius nobis adiutor et pugnandi et
sicut honoremus. Haec tunc cum dixit
se mox et multitudine ab horum exibuit
elegerat exploratores diademam nostrum
uiro, unum deinceps tribu omniumque
omne terram dominio suorum obseruantur
aegyptum acerbibus utrūcunq[ue] adiutorum
emittunt. Admetum liberos puerum
Naturamq[ue] terrae et incolarum hominum
considerant ut si de conuenienter que
dragata diebus omni loco opus expletum
tali desparatione delectos licet turpites

Création d'une vérité de terrain (*ground truth*)

Les images transcrites sont alors associées à leur transcription

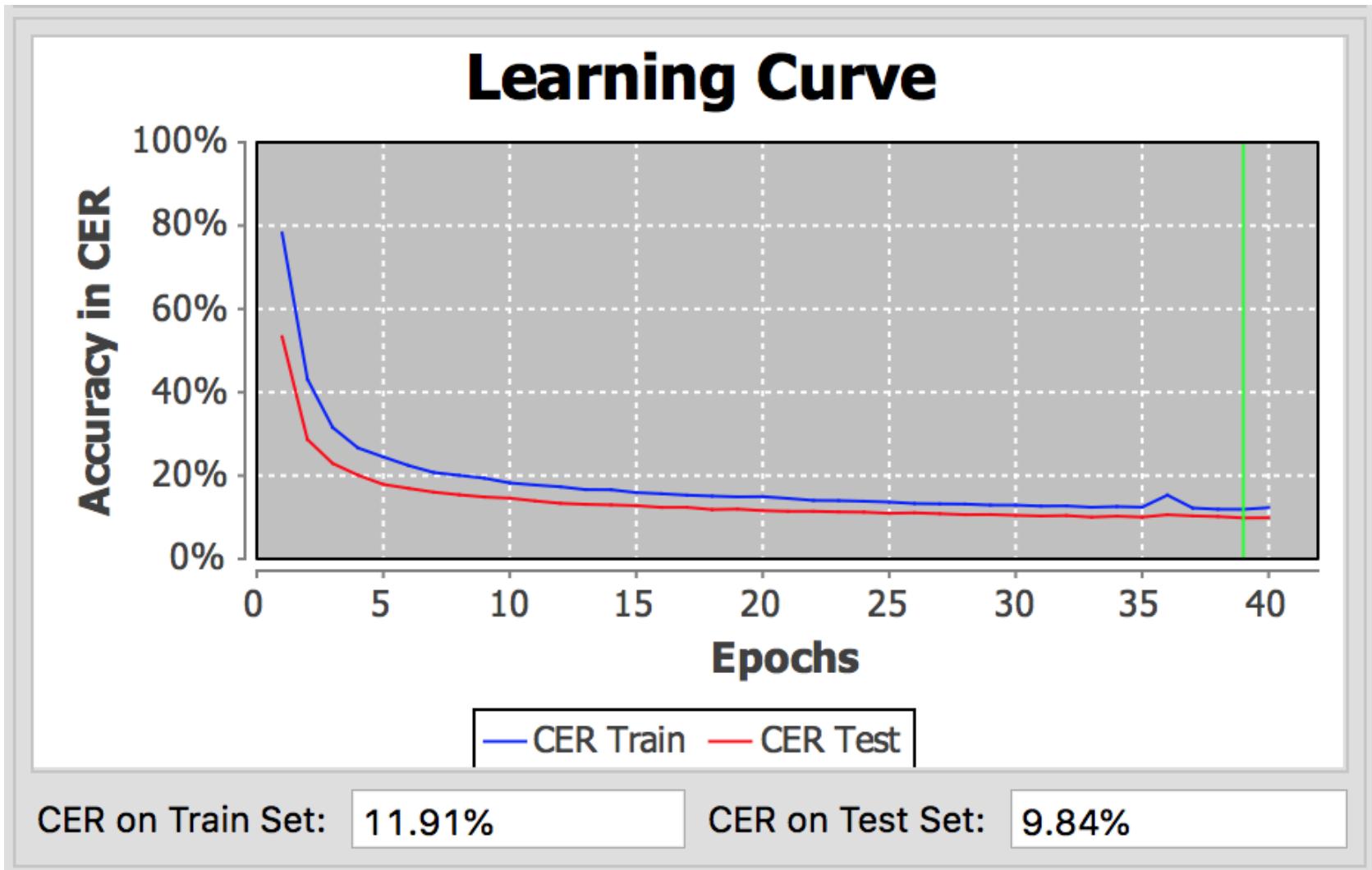


Entraînement (I)

Comme c'est du *machine learning*, on va répéter l'entraînement une multitude de fois (on parle d'*epochs*, de *stages* ...). À chaque fois un modèle est créé: celui qui performe le mieux est conservé

```
Accuracy report (17429) 0.9610 4825 188
stage 15/∞ [#####
Accuracy report (18591) 0.9621 4825 183
stage 16/∞ [#####
Accuracy report (19753) 0.9606 4825 190
stage 17/∞ [#####
Accuracy report (20915) 0.9615 4825 186
stage 18/∞ [#####
Accuracy report (22077) 0.9602 4825 192
stage 19/∞ [#####
Accuracy report (23239) 0.9617 4825 185
stage 20/∞ [#####]
```

Entraînement (II)



Scores

- On parle de CER (*Character Error Recognition*) et parfois de WER (*Word Error Recognition*).
- Distance de Levenshtein : combien d'opérations pour retrouver le résultat attendu (par exemple entre tonte et toute) ?
- Une seule lettre fausse crée un mot faux ! Le WER est donc toujours supérieur au CER !
- Ces scores peuvent être calculés sur deux jeux de données :
 - Le train set (on OCRise les images qui servent pour l'entraînement)
 - Le test set (on OCRise des images qui n'ont pas servi pour l'entraînement)

L'amélioration des scores: données artificielles

- Avec Baskerville

C'est ceux dont il est écrit au commen-

- Avec IM FELL English SC

VOUS ESTIMÉS QUELQUE CHOSE

- Avec JSL Ancient

reZ Lecteur (si je ne me trompe,) la

- Avec Chapbook

Tandis qu'autour de moy vostre Cour assemblée,

L'amélioration des scores: bruit

- Original

C'est ceux dont il est écrit au commen-

- Bruit faible

C'est ceux dont il est écrit au commen-

- Bruit fort

C'est ceux dont il est écrit au commen-

L'amélioration des scores: modification du cadre

- Cadre normal

C'est ceux dont il est écrit au commen-

- Cadre élargi

C'est ceux dont il est écrit au commen-