

# De la donnée avant toute chose ? Retour d'expérience de l'utilisation de l'HTR dans des projets d'édition et d'étude des textes médiévaux

Journées d'Études « De la transcription manuelle participative des textes à la reconnaissance automatique de texte (ATR/HTR) : outils, théorie, pratiques. »

Matthias GILLE LEVENSON

École nationale des chartes – Centre Jean Mabillon & École Normale Supérieure de Lyon – CIHAM UMR 5648

Nancy, 10 septembre 2024



# Plan

## 1 Introduction

## 2 Phase de production des données

- Penser la production en amont
- CATMuS, un projet de production collaboratif de données d'ATR

## 3 Après l'HTR : tout change / rien ne change

- Quand s'arrête la correction ?
- Structurer les documents
- Identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

## 4 Conclusions

## 1 Introduction

## 2 Phase de production des données

## 3 Après l'HTR : tout change / rien ne change

## 4 Conclusions

# Expériences personnelles avec l'HTR

- Pour de l'édition
- Pour l'étude proprement dite du texte
- Actuellement, pour des expériences de collation multilingue

# Édition critique

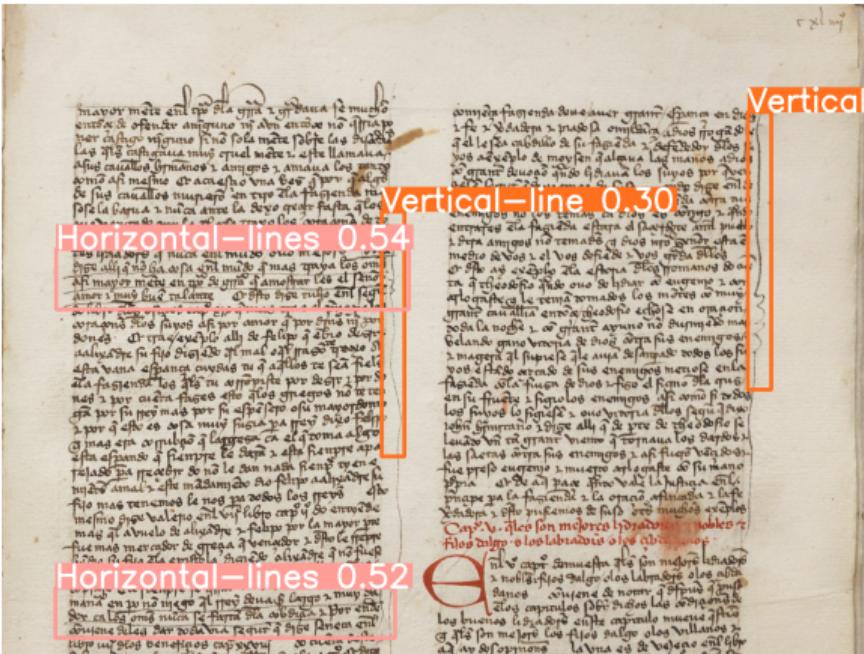
E deuen fazer otra puerta de fierro que llaman puerta de traycion con grandes fierros ayuso. E deue ser foradada la torre por que se pueda sobir e desçender con cadenas de fierro e deuese fazer ante de las puertas principales<sup>[R: fol. 279v]</sup> por que non las pueda quemar. [A, III-3-20, traduction, fol. 266r, éd. p. 647]

---

2 desçender *ABRQJZ* | defender *G* 2 deuese *ABGRJZ* | dévense *Q* 2 deuese *BAQJZ* | [dévese  
*BGRJAZ* | dévense *Q*] [ $\emptyset$  *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om. Q* 3 pueda *AZ* | puedan *BGRQJ*

Édition critique du *Regimiento de los Príncipes*, issue de collation automatisée. Les témoins A et Z sont issus d'HTR : (GILLE LEVISON 2023a) et (GILLE LEVISON 2023b)

# Études de la réception d'un manuscrit par ses marques de lecture



Identification automatisée avec YOLO v5 (REDMON 2016) de zones de texte marquées par un lecteur. Escorial Ms. K.I.5, fol. 144r

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, transcription

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité
- Avec Kraken (KIESSLING 2019), pas de transcription globale du texte mais ligne par ligne

## 1 Introduction

## 2 Phase de production des données

- Penser la production en amont
- CATMuS, un projet de production collaboratif de données d'ATR

## 3 Après l'HTR : tout change / rien ne change

## 4 Conclusions

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds

## Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :

# Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?

# Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?

# Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?

# Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

## Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

1-23 fechura no deue paran mjero

1-24 ala color \$da ñ qere\$ a los

1-25 fralcons ñ soy cntrados o

1-26 f Fuara a manallos.

1-27 oq<ue> torna contra umneio prriua

Un cas de modèle produit à partir de données hétérogènes, identifié par J.M. Fradejas (PINCHE et al. 2024)

Maria. Ense que p[re]mier en h[ab]it ier eut vierge devant le palais de los o[ro]s es ca me[me] p[re]mier  
s a p[re]mier que des me[me] que et si n[on] a p[re]mier en ferme del signe io el f[il]s en a  
mmeur du bien que au temps se pert en b[ea]tentat idem. Pour ce q[ue] p[ro]oit h[ab]it. Eala vero  
disciplina gheince s[ecundu]m se fais Dieudale vero. potestruente, la  
cune cause elle r[es]on. Im[per]iale tu te puise ioue q[ue] la f[ig]o de re[gi]o ro[bi]n[son] eut auant co  
q[ue] tellos naualloraz tal l[oc]ate ne car[re]ne. n[on] n[on] es gheu v[e]l la si hi p[re]me h[ab]it meo  
un j[an]g an h[ab]it. en t[em]ps q[ue] l[oc]ate d[omi]ne sicut ale p[re]leone p[re]erudelante  
serie turans ne r[es]o. d[omi]ne bel[is]me de la s[ecundu]m qui deute laire supplicatio. n[on] deute est enia



CATMuS  
Medieval

classe et crame[re] so de la foraleja que t[em]p[er]e d[omi]ne. d[omi]ne. d[omi]ne. d[omi]ne.  
ensamble onoste aymant de lu la serial acore. r[es]on du heuma manca es  
le En alle w[er] dat hochsta[nd] les r[es] de bone o conlas manos por fus de contre tov  
urt par bone. a Bone et en son ou p[re]to; let gremie vidor Zodrigo la m[ar]ca q[ue] d[omi]n  
scordia. TE stando en t[em]p[er]e omnia q[ue] gheine d[omi]ne omnia. q[ue] p[ro]p[ri]etate mas aqua no  
faemh[ab]de ghe h[ab]ita r[es]on. no p[er]fessio. In t[em]p[er]e d[omi]ne magis. I tercero d[omi]ne alma d[omi]n  
re dicha q[ue] ihu ad p[re]sidium. a luano el filo. m[ar]ca q[ue] excludunt amores q[ue] los principes en c  
tence r[es] de bone. r[es] de bone. fol. l[oc]ate abuas. quis ie p[re]st a my desen te comene. wa

- Volonté autour de 2022 de réunir des producteur.ices de données (philologues) venant d'horizons distincts
- Naissance du projet CATMuS pour « *Consistent Approaches for Transcribing Manuscripts* »
- Le corpus est récemment publié (CLÉRICE et al. 2024)

# Statistiques sur le corpus CATMuS

- Publié sur HuggingFace : <https://huggingface.co/datasets/CATMuS/medieval>
- Recueil de 17 dépôts différents **homogénéisés** selon les normes du projet
- Environ 175.000 lignes transcrites et intégralement corrigées (corpus *gold*)
- 245 documents différents en 9 langues
- Du IX<sup>e</sup> au XV<sup>e</sup> siècle, avec une prédominance du bas Moyen Âge (XIII<sup>e</sup>-XV<sup>e</sup> siècles)
- Corpus encore biaisé en raison de l'histoire du projet

# Statistiques sur le corpus CATMuS

- Publié sur HuggingFace : <https://huggingface.co/datasets/CATMuS/medieval>
- Recueil de 17 dépôts différents **homogénéisés** selon les normes du projet
- Environ 175.000 lignes transcrites et intégralement corrigées (corpus *gold*)
- 245 documents différents en 9 langues
- Du IX<sup>e</sup> au XV<sup>e</sup> siècle, avec une prédominance du bas Moyen Âge (XIII<sup>e</sup>-XV<sup>e</sup> siècles)
- Corpus encore biaisé en raison de l'histoire du projet

Split	Software	Training Time	Character Error Rate (%)		
			Validation	Test	Space-related
General	Kraken	2112 min $\pm$ 163	5.7 $\pm$ 0.07	4.7 $\pm$ 0.06	1.0 $\pm$ 0.02
Feature	Kraken	1464 min $\pm$ 238	6.8 $\pm$ 0.16	13.1 $\pm$ 0.24	2.7 $\pm$ 0.06
General	Pylaia	308 min $\pm$ 047	9.1 $\pm$ 0.63	8.4 $\pm$ 0.73	1.8 $\pm$ 0.11
Feature	Pylaia	295 min $\pm$ 078	11.3 $\pm$ 0.24	21.2 $\pm$ 0.92	3.8 $\pm$ 0.06

Modèles produits à partir des données de CATMuS. (CLÉRICE et al. 2024, p. 16)

# Une pomme de discorde : les abréviations

Notre point de vue est celui de philologues :

- Nous considérons la résolution des abréviations comme une tâche de TAL plutôt que de vision assistée par ordinateur (CLÉRICE et al. 2024)
- La résolution via ATR pose des problèmes de **généralisation** et d'**adaptation**.

# Le choix de la conservation des abréviations

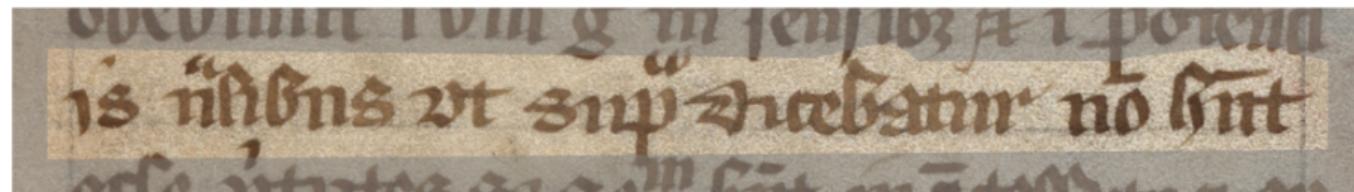
Une norme de transcription **graphématisque** (STUTZMANN 2010) :

- Réduction des allographes au graphème
- Réduction des allographes <i>/<j> et <u>/<v> à <i> et à <u>
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>

# Le choix de la conservation des abréviations

Une norme de transcription **graphématisque** (STUTZMANN 2010) :

- Réduction des allographes au graphème
- Réduction des allographes <i>/<j> et <u>/<v> à <i> et à <u>
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>



is <sup>a</sup>n̄libus ut sup̄ dicebatur nō hñt

# Concilier l'intérêt particulier et les besoins généraux

- Cette solution nous semble être le plus à même de **concilier besoins généraux et particuliers**
- La question de l'**homogénéité des données** est fondamentale : manuel d'annotation et outils de contrôle (<https://github.com/PonteIneptique/choco-mufin> et <https://github.com/HTR-United/HTRVX>)
- La contrepartie est la nécessité de **travailler en aval** de l'acquisition du texte pour normaliser le texte
- Le manuel en ligne est disponible : <https://catmus-guidelines.github.io/> (en cours de rédaction)

## 1 Introduction

## 2 Phase de production des données

## 3 Après l'HTR : tout change / rien ne change

- Quand s'arrête la correction ?
- Structurer les documents
- Identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

## 4 Conclusions

# Quand s'arrête la correction ?

- La phase suivant la transcription automatisée sera généralement celle de la transformation en TEI
- Plusieurs outils permettent de réaliser cette transformation :  
<https://github.com/Jean-Baptiste-Camps/ALTEI>,  
<https://github.com/chartes/alto2tei>,  
[https://github.com/matgille/alto\\_to\\_teii](https://github.com/matgille/alto_to_teii)/
- Il restera des erreurs dans les données
- Faut-il intégrer les corrections faites dans les données d'entraînement ?
- Si oui, cela suppose de penser en amont une modélisation en TEI qui soit **rétroconvertible**
- En d'autres termes, il faudra conserver un premier état de TEI pseudo-diplomatique (conservation des tei:lb)

# Assurer la rétroconvertibilité

```
<lb break="yes" facs="#facsc_line_92b25bc7" xml:id="elem_eSc_line_92b25bc7">cū uegetablibūt plātis ut pō  
<lb break="?" facs="#facsc_line_4fc0c0e" xml:id="elem_eSc_line_4fc0c0e">nuitia augm̄tia. ḡnatia γ  
<lb break="?" facs="#facsc_line_1e32f48c" xml:id="elem_eSc_line_1e32f48c">tlia q γ ip̄s arboribūt cōpetūt.  
<lb break="?" facs="#facsc_line_2fddd609" xml:id="elem_eSc_line_2fddd609">po γ cognit̄e sfit̄e sūt uis  
<lb break="yes" facs="#facsc_line_b1fb0fe4" xml:id="elem_eSc_line_b1fb0fe4">gust⁹ γ tactus. γ tlia in qb⁹  
<lb break="?" facs="#facsc_line_63be87f7" xml:id="elem_eSc_line_63be87f7">cam⁹ cū brutis. appetitie ū di  
<lb break="?" facs="#facsc_line_2304726d" xml:id="elem_eSc_line_2304726d">stīgūt. nā qdam ē appetit⁹ i  
<lb break="?" facs="#facsc_line_9e3287cc" xml:id="elem_eSc_line_9e3287cc">hoie i q n̄ ȳcat cū brutis ut  
<lb break="?" facs="#facsc_line_4f2eeef0c" xml:id="elem_eSc_line_4f2eeef0c">appetit⁹ seq̄s itllem. Qdā ū  
<lb break="yes" facs="#facsc_line_887a268e" xml:id="elem_eSc_line_887a268e">i q ȳcat cū eis appetit⁹ seq̄s se  
<lb break="?" facs="#facsc_line_3c8f19df" xml:id="elem_eSc_line_3c8f19df">sū. Appetit⁹ ät seq̄s sfm pt  
<lb break="?" facs="#facsc_line_cd2f421c" xml:id="elem_eSc_line_cd2f421c">noīari sfualitas. seq̄s itlē  
<lb break="?" facs="#facsc_line_5d351c82" xml:id="elem_eSc_line_5d351c82">ctū noīē uolūtas. f q̄mod lo  
<lb break="?" facs="#facsc_line_fc03f518" xml:id="elem_eSc_line_fc03f518">q̄di bruta h̄t sfualitate γ appen  
<lb break="?" facs="#facsc_line_3ecf893f" xml:id="elem_eSc_line_3ecf893f">titū sfit̄im. s̄ n̄ uolūtate h̄t
```

Le document TEI avec des identifiants présents dans le fichier ALTO originel

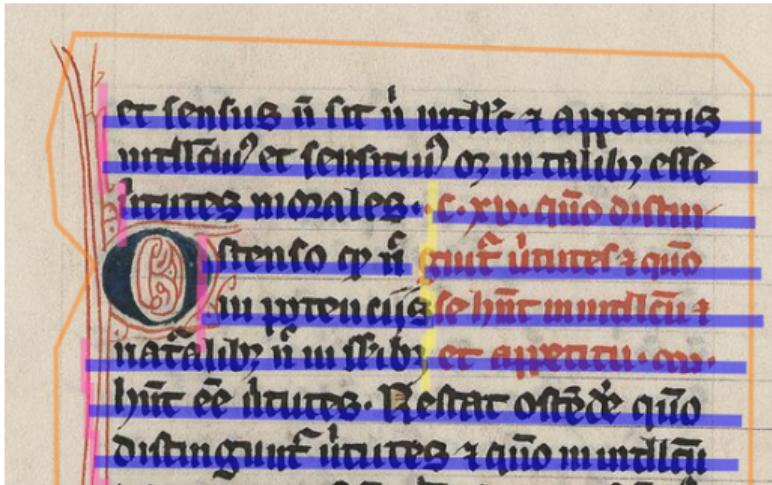
# Assurer la rétroconvertibilité

```
-<TextLine ID="eSc_line_4fc0c0e" TAGREFS="LT6426" BASELINE="1581 1374 2153 1366" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0">
+<Shape></Shape>
<String CONTENT="nuttia augm̄tia. ḡhatia γ" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0"/>
</TextLine>
-<TextLine ID="eSc_line_1e32f48c" TAGREFS="LT6426" BASELINE="1573 1434 2164 1421" HPOS="1571.0" VPOS="1375.0" WIDTH="593.0" HEIGHT="89.0">
+<Shape></Shape>
<String CONTENT="tlia ̄ q̄ ip̄is arborib̄ c̄opetūt." HPOS="1571.0" VPOS="1375.0" WIDTH="593.0" HEIGHT="89.0"/>
</TextLine>
-<TextLine ID="eSc_line_2fddd609" TAGREFS="LT6426" BASELINE="1579 1487 2155 1479" HPOS="1578.0" VPOS="1438.0" WIDTH="577.0" HEIGHT="81.0">
+<Shape></Shape>
<String CONTENT="po' v cognit̄e sfit̄e sūt uis̄" HPOS="1578.0" VPOS="1438.0" WIDTH="577.0" HEIGHT="81.0"/>
</TextLine>
-<TextLine ID="eSc_line_b1fb0fe4" TAGREFS="LT6426" BASELINE="1575 1545 2159 1534" HPOS="1573.0" VPOS="1490.0" WIDTH="586.0" HEIGHT="84.0">
+<Shape></Shape>
<String CONTENT="gust̄ t tactus. t tlia in qb̄ ̄ q̄" HPOS="1573.0" VPOS="1490.0" WIDTH="586.0" HEIGHT="84.0"/>
</TextLine>
-<TextLine ID="eSc_line_63be87f7" TAGREFS="LT6426" BASELINE="1576 1602 2164 1589" HPOS="1575.0" VPOS="1545.0" WIDTH="589.0" HEIGHT="78.0">
+<Shape></Shape>
<String CONTENT="cam̄ cū brutis. appetitie u di" HPOS="1575.0" VPOS="1545.0" WIDTH="589.0" HEIGHT="78.0"/>
</TextLine>
-<TextLine ID="eSc_line_2304726d" TAGREFS="LT6426" BASELINE="1573 1660 2150 1648" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0">
+<Shape></Shape>
<String CONTENT="st̄igūf. nā qdam ē appetit̄ i" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0"/>
</TextLine>
```

Le fichier ALTO d'origine

# Structurer les documents

- Classifier les zones et les lignes lors de la phase d'ATR peut permettre de faciliter la structuration (semi-)automatisée du document :



Classification des lignes suivant le vocabulaire contrôlé SegmOnto (GABAY et al. 2021). En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric. Vat. Borg. 360, fol 190v.

# Identifier la césure à la ligne

- Dans les manuscrits médiévaux la césure à la ligne n'est pas systématiquement indiquée

podies sensituos. **O**n assi como nigu  
ome no es alabado ni es tenido por bue  
no por q muelle bie su uata ni por q cre  
sce bie assi no es alabado por q bie agu  
da mete o eye solit mete. Saluo ende

ronē ptiapant qma p se a s<sup>m</sup> q si no  
obedimt rūm g m sensib<sup>r</sup> a i potenti  
is nūbs ut sup dñebar no hñt  
esse vñtes s<sup>a</sup> so<sup>m</sup> hñt m i tellest a ap  
petitu in nobis an dñpser est appen

.v appetitū intellētu. intutes g<sup>e</sup> de qb  
loqui intendim q sūt qdam hic lau  
tabiles ur euit in potētis nūlibus  
ul m ipis scrib ul in appetitu. Vñtu  
ul in appetitu nñtuco ul m ipo idle  
tu ul m omib<sup>s</sup> hñs ul m aliqb<sup>s</sup> hñs

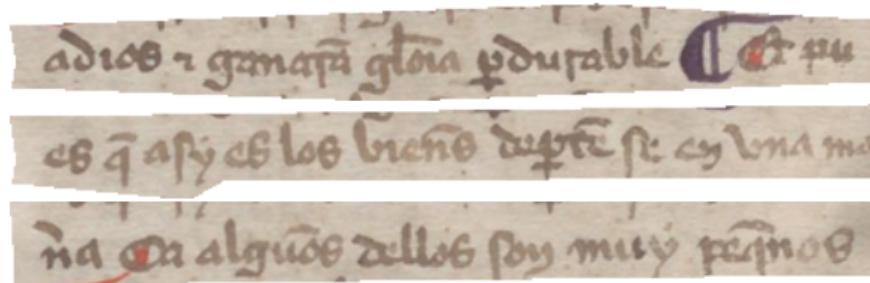
**C**on el asturiano oñodo esso alcen  
trose z fielo de manaz concis  
delrey don fernando z dela Reyna  
dona maria Su madre quele en  
biana a pedr por meqnd al papa sñ  
bre esta pason **F**inas con pedro  
que esa obispo de burgos a esa sa  
don E referencio del papa ua

S a m s<sup>r</sup> h<sup>i</sup> eti. media e int int  
tes morales intellētu. p<sup>r</sup> m  
qputa qm uñtib<sup>s</sup> moralib<sup>s</sup>. P<sup>r</sup> p<sup>r</sup>  
den n<sup>o</sup> h<sup>i</sup> hñb<sup>s</sup> leis s<sup>r</sup> n<sup>o</sup> uñtib<sup>s</sup>  
moralib<sup>s</sup>. p<sup>r</sup> e hñb<sup>s</sup> s<sup>r</sup> p<sup>r</sup> e c<sup>r</sup> h<sup>i</sup>  
entes. altitu. z uñspells. m p<sup>r</sup> dñres

sophia. geometria metaphy  
sica i c. talia uñntib<sup>s</sup> uero si  
phic<sup>r</sup> morales. s<sup>r</sup> ille q s<sup>r</sup>  
in appetitu. siue appetitus il  
le sit sensitum. siue intellē  
tuus. cuiusmodi aut sunt  
iustitia. tempencia. fortitudo

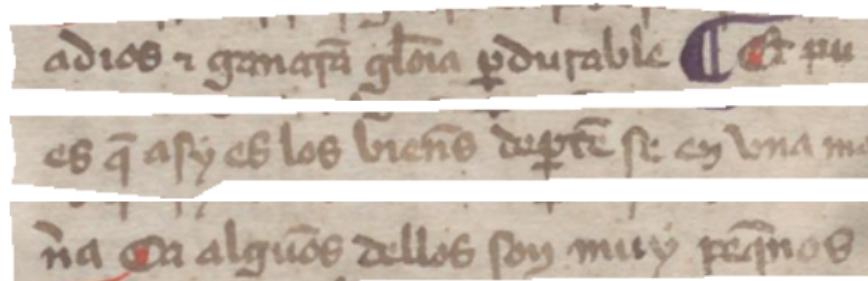
Valladolid, 251 fol. 17r ; Cambridge, Corpus Christi College, MS 283, fol 14r ; Vatican, Borg. 360, fol. 190r ;  
BNF Esp 36, fol 1v ; BNE MSS/958, fol. 14r ; Valencia BH Ms 0594, fol. 23r

# Identifier la césure



adios - ganaqā glōia p̄durable Cest p̄  
es q̄ a sy es los b̄ienos duxte se en una ma  
na en algunos de los soy muy p̄prios

# Identifier la césure



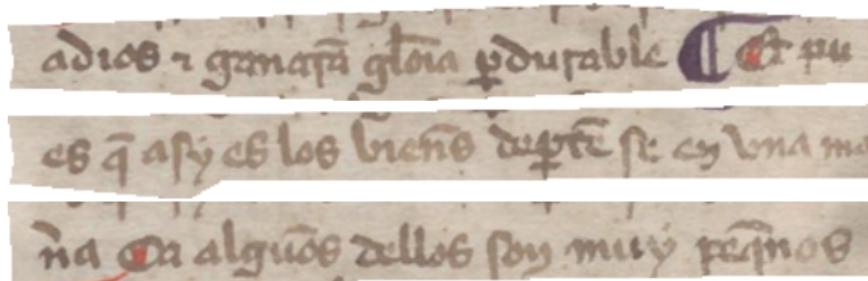
adios J ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ñá Ca alguōs dellos son muy peqnos

adios J ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ñá Ca alguōs dellos son muy peqnos



adios J ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deptē se en una mañá Ca alguōs dellos son  
muy peqnos

# Identifier la césure



adios J ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos

¶

adios J ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deptē se en una maña Ca alguōs dellos son  
muy peqnos

Voir (CLÉRICE 2020) et [https://github.com/matgille/boudams\\_like\\_tokenizer](https://github.com/matgille/boudams_like_tokenizer)

# Résoudre les abréviations

- Le plus « simple » est d'utiliser une table de conversion à base de règles
- Suppose de connaître en amont toutes les abréviations et cas possibles **en contexte**
- Un peu laborieux quand le volume de documents différents est important

243	felici <sup>t e</sup> -l	felicitatem
244	ci <sup>t e</sup> -l	citatem
245	uo <sup>t e</sup> -l	uoluntatem
246	<SOT>ca <sup>i t e</sup> <EOT>	~caritate~
247	<SOT>.p<EOT>	~prout~
248	<SOT>ad<EOT>	~aliquid~
249	gn̄	gener

Table d'abréviation. <SOT> et <EOT> viennent marquer un début et/ou une fin de mot

# Et l'édition critique ?

L'ATR appelle assez naturellement des méthodes de collation automatisée du texte

# Et l'édition critique ?

	ab homéotèleute lex	ab	ab	ab	ab	ab	ab	ab	ab	ab
		~	lex.	om./lex.	om./lex.	om./lex.	om./lex.	~	~	
Rome_1607	quaedam			mediae		inter	intellectuales	et	m Morales	
Rome_1556	quædon			mediæ		inter	intellectuales	et	m Morales	
Planck	quedam	medie				icirer	iitellctuales	et	m Morales	
Bevilaqua_1498	quadam		mediae			inter	intellectuales	et	m Morales	
Vat_Lat_590	quedam	die	me			inter	intellectuales	et	m Morales	
CCC_MSS_283	qued		mē			inter	intellectuales	et	m Morales	
Borgh_360	et	uirtutes	morales				intellectuales	et	m Morales	
Geneve Ms_Lat_92	et	medie	quedum			inter				m Morales
Metz_Mediatheque_1234	quedam	medie			inter	intellectuales		et	m Morales	
Beinecke_Marston_MS_139	quadam	medie			inter		intellectumales	et	m Morales	
BNE_MSS_958	quad <small>ē</small>	medie			inter		intellectuales	et	m Morales	
BNF_Lat_6477	quedam	medie					intellectuales	et	m Morales	
BNE_9236	quedam	medie				inter	intellectumales	et	m Morales	
BNF_Lat_1234	quedam	medie				inter	intellectuales	et	m Morales	
Valencia_BH_Ms_0594	quedam	medie			inter	intellectuales		et	m Morales	
<small>Quedam</small>										

Table d'alignement avant la phase de correction, après segmentation et résolution des abréviations. Gilles de Rome,  
*De Regimine Principum*, chapitre 1.2.2

# Et l'édition critique ?

	ab	ab	ab	ab	ab	ab	ab	ab	ab	homéotéleute	ab	ab	ab
	lex.	norm.	om./lex.	lex.	graph.	lex.	om./lex.	om.	om./lex.	om.	lex.	~	
Rome_1607	e	sensitiuae		sunt	visus			gustus	auditus		et	talia	
Rome_1556	e	sensitiæ		sunt	visus			gustus	auditus		et	talia	
Planck		sensitiue		sunt	uisus			gustus	auditus		et	talia	
Bevilacqua_1498		sensitiua		sunt	ui-sus			gustus	auditus		et	talia	
Vat_Lat_590		sensitiem		sunt	uisus			gustus	et	tactus	et	talia	
CCC_MSS_283		senfitiue		sunt	uisus	gus-tus			et	tactus	et	talia	
Borgh_360		sensituue		sunt	ui	sus		gustus		tactus	et	talia	
Geneve_Ms_Lat_92		sensitiue		sunt	uisus	auditus		gustus			et	talia	
Metz_Mediatheque_1234		sensitiue		sunt	uisus			gustus		tactus	et	talia	
Beinecke_Marston_MS_139		sensitiue		sunt	uisus			gustus	et	tactus	et	talia	
BNE_MSS_958		sensitiue		suntuisus				gustus		tactus	et	talia	
BNF_Lat_6477		sensitiue		sunt	uisus			gustus		tactus	e	talia	
BNE_9236		sensitiue		sunt	uisus			gustus		tactus	et	talia	
BNF_Lat_1234		sensitiue		sunt	uisus			gustus		tactus	et	talia	
Valencia_BH_Ms_0594		sensitiue		sunt	uisus			gustus		tactus	et	talia	

Table d'alignement après la phase de correction. Gilles de Rome, *De Regimine Principum*, chapitre 1.2.1

1 Introduction

2 Phase de production des données

3 Après l'HTR : tout change / rien ne change

4 Conclusions

- Penser en amont les principes d'annotation est fondamental
  - Identifier les intérêts et les inconvénients de chaque choix en terme de balance intérêt particulier / intérêt général
  - Lancer de courtes campagnes test pour mettre à l'épreuve les choix d'annotation
  - Identifier le type d'information nécessaire aux étapes ultérieures de traitement du texte manuscrit ou imprimé : penser l'ensemble de la production des données et pas seulement l'HTR pour l'HTR
  - Se mettre en conformité (ou pas) avec les normes existantes et le documenter clairement
- Du travail reste à mener
  - Sur tout la chaîne post-ATR afin d'arriver à un texte final plus propre
  - Et plus particulièrement pour la collation
  - Quid de l'accumulation du bruit avec la multiplication des étapes ?

# Merci!

Merci de votre attention !

- Diapos : [https://github.com/matgille/Comm\\_Nancy\\_sept\\_2024](https://github.com/matgille/Comm_Nancy_sept_2024)

# Références I

- [1] Thibault CLÉRICE. « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin ». *Journal of Data Mining & Digital Humanities* 2020 (7 avr. 2020). URL : <https://jdmdh.episciences.org/6264/pdf>.
- [2] Thibault CLÉRICE, Ariane PINCHE, Malamatenia VLACHOU-EFSTATHIOU, Alix CHAGUÉ, Jean-Baptiste CAMPS, Matthias Gille LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Patricia O'CONNOR, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Benjamin KISSLING. « CATMuS Medieval : A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond ». *Document Analysis and Recognition - ICDAR 2024*. Sous la dir. d'Elisa H. BARNEY SMITH, Marcus LIWICKI et Liangrui PENG. Cham : Springer Nature Switzerland, 2024, p. 174-194. ISBN : 978-3-031-70543-4. DOI : 10.1007/978-3-031-70543-4\_11.
- [3] Simon GABAY, Jean-Baptiste CAMPS, Ariane PINCHE et Claire JAHAN. « SegmOnto : Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More) ». *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*. 2021. ISBN : 978-3030865481. URL : <https://segmonto.github.io/>.
- [4] Matthias GILLE LEVENSON. « Le Regimiento de Los Príncipes et sa glose : étude et édition numérique de la partie sur le Gouvernement de la cité en temps de guerre (III, 3.) ». Codir. Carlos HEUSCH et Jesús R. VELASCO. École Normale Supérieure de Lyon, 2023. URL : <https://theses.hal.science/tel-04337406>.
- [5] Matthias GILLE LEVENSON. « Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR) ». *Journal of Data Mining and Digital Humanities* (2023). DOI : 10.46298/jdmdh.10416. URL : <https://zenodo.org/records/8340483>.
- [6] Odd Einar HAUGEN. « Dealing with Glyphs and Characters : Challenges in Encoding Medieval Scripts ». *Document numérique* 16.3 (2013), p. 97-111. URL : <https://www.cairn.info/revue-document-numerique-2013-3-page-97.htm>.
- [7] Benjamin KISSLING. « Kraken - an Universal Text Recognizer for the Humanities ». DH2019 : Complexity. Utrecht, 2019. URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html>.
- [8] Ariane PINCHE, Thibault CLÉRICE, Alix CHAGUÉ, Jean-Baptiste CAMPS, Malamatenia VLACHOU-EFSTATHIOU, Matthias Gille LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Patricia O'CONNOR. « CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts ». *Digital Humanities - DH2024*, ADHO. Washington, D.C., 2024.

# Références II

- [9] J. REDMON. « You Only Look Once : Unified, Real-Time Object Detection ». *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. url : [https://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016/html/Redmon\\_You\\_Only\\_Look\\_CVPR\\_2016\\_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html).
- [10] Dominique STUTZMANN. « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? » *Codicology and Palaeography in the Digital Age*. T. 2. 2010, p. 34.