

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

De la donnée avant toute chose ?

Matthias GILLE LEVENSON

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

Nancy, 10 septembre 2024

2024-09-10

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. »

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

Matthias GILLE LEVENSON

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

Nancy, 10 septembre 2024

2024-09-10

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
- Segmenter le texte et identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

4 Conclusions

De la donnée avant toute chose ?

2024-09-10

└ Plan

Plan

- Introduction
- Phase de production des données
 - Penser la production en amont
 - CATMuS, un projet de production collaboratif de données d'ATR
- Après l'HTR : tout change / rien ne change
 - Quand s'arrête la correction ?
 - Structurer les documents
 - Segmenter le texte et identifier la césure à la ligne
 - Résoudre les abréviations
 - Et l'édition critique ?
- Conclusions

1 Introduction

2 Phase de production des données

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

4 Conclusions

Introduction

2024-09-10

- Introduction
- Phase de production des données
- Après l'HTR : tout change / rien ne change
- 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^[R: fol. 279v] 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*] [ø *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om. Q* 3 pueda *AZ* | puedan *BGRQJ*

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

De la donnée avant toute chose ?
└ Introduction
└ Édition critique

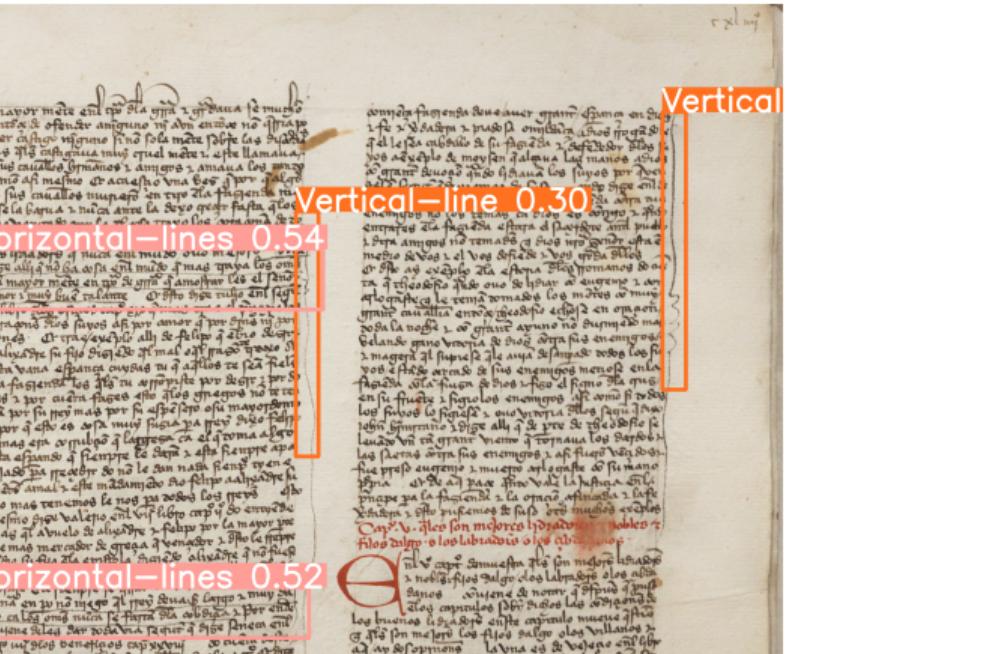
2024-09-10

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^[R: fol. 279v] 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*] [ø *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om. Q* 3 pueda *AZ* | puedan *BGRQJ*

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

└ Introduction

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

2024-09-10



Identification automatique (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*

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*
- Avec Kraken (KIESSLING 2019), pas de transcription globale du texte mais ligne par ligne pour l'instant

- 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*
- Avec Kraken (KIESSLING 2019), pas de transcription globale du texte mais ligne par ligne pour l'instant

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

2024-09-10

Penser la production en amont

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

- └ Phase de production des données
- └ Penser la production en amont
- └ Penser la production en amont

2024-09-10

- Question de pérennité : les données restent, les modèles disparaissent.

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 :

- Phase de production des données
 - Penser la production en amont
 - Penser la production en amont

2024-09-10

- 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 :

- Question de pérennité : les données restent, les modèles disparaissent.

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?

- └ Phase de production des données
 - └ Penser la production en amont
 - └ Penser la production en amont

2024-09-10

- 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?

- Question de pérennité : les données restent, les modèles disparaissent.

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?
 - Que faire des abréviations?

- └ Phase de production des données
- └ Penser la production en amont
- └ Penser la production en amont

2024-09-10

- 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?
 - Que faire des abréviations?

- Question de pérennité : les données restent, les modèles disparaissent.

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?
 - Que faire des abréviations?
 - Éviter à tout prix les modèles « boule de neige » :

```

1-23 fechura no deue paran mjero
1-24 ala color $da q qere$la los
1-25 fralcons q soy cntrados o
1-26 f Fuara a manallos.
1-27 oq<ue> torna cota umneio priua

```

Figure 1: Prediction using Transkribus *Coloso Español* model. The model uses two different ways to deal with abbreviations, line 24 (q̄) and line 27 (<ue>). Capture kindly provided by J. M. Fradecas Rueda.

Un cas de modèle produit à partir de données hétérogènes. (PINCHE et al. 2024)

De la donnée avant toute chose?

- └ Phase de production des données
- └ Penser la production en amont
- └ Penser la production en amont

2024-09-10

- 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?
 - Que faire des abréviations?
 - Éviter à tout prix les modèles « boule de neige » :

Figure 1: Prediction using Transkribus *Coloso Español* model. The model uses two different ways to deal with abbreviations, line 24 (q̄) and line 27 (<ue>). Capture kindly provided by J. M. Fradecas Rueda.

Un cas de modèle produit à partir de données hétérogènes. (PINCHE et al. 2024)

CATMuS

- Volonté autour de 2022 de réunir des producteur.ices de données (philologues) venant d'horizons distincts

- 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)

De la donnée avant toute chose?

- └ Phase de production des données
 - └ CATMuS, un projet de production collaboratif de données d'ATR
 - └ CATMuS

2024-09-10

- 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
- 9 langues
- Du IX^e au XV^e siècle, avec une prédominance du bas Moyen Âge (XIII^e-XV^e siècles)
- Corpus encore biaisé en raison de l'histoire du projet

De la donnée avant toute chose ?

Phase de production des données

CATMuS, un projet de production collaboratif de données d'ATR
Statistiques sur le corpus CATMuS

2024-09-10



- 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
- 9 langues
- Du IX^e au XV^e siècle, avec une prédominance du bas Moyen Âge (XIII^e-XV^e siècles)
- Corpus encore biaisé en raison de l'histoire du projet

Le problème des abréviations, entre usages historiens et usages philologiques

Notre point de vue est celui de philologues :

- Nous considérons cette tâche comme une tâche de TAL plutôt que de vision assistée par ordinateur (CLÉRICE et al. 2024)
- Pose des problèmes de **généralisation** ;
- Pose des problèmes d'**adaptation**.
 - Généralisation : le développement des abréviations peut poser problème dans le cadre de corpus multilingue ; les résultats sont moins bons en développant les abréviations
 - Adaptation : le développement des abréviations est dépendant du contexte linguistique et historique du document.

De la donnée avant toute chose ?

- └ Phase de production des données
- └ CATMuS, un projet de production collaboratif de données d'ATR
- └ Le problème des abréviations, entre usages historiens et usages philologiques

Notre point de vue est celui de philologues :

- Nous considérons cette tâche comme une tâche de TAL, plutôt que de vision assistée par ordinateur (CLÉRICE et al. 2024)
- Pose des problèmes de **généralisation** ;
- Pose des problèmes d'**adaptation**.

Le choix de la conservation des abréviations

- 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)

De la donnée avant toute chose ?

- └ Phase de production des données
- └ CATMuS, un projet de production collaboratif de données d'ATR
- └ Le choix de la conservation des abréviations

2024-09-10

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)

Concilier l'intérêt particulier et les besoins généraux

- Cette solution nous semble être le plus à même de concilier besoins généraux et particulier,
- La contrepartie est la nécessité de travailler en aval de l'acquisition du texte pour le normaliser

- Cette solution nous semble être le plus à même de concilier besoins généraux et particulier,
- La contrepartie est la nécessité de travailler en aval de l'acquisition du texte pour le normaliser

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
- Segmenter le texte et identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

4 Conclusions

Après l'HTR : tout change / rien ne change

2024-09-10

- Introduction
- Phase de production des données
- Après l'HTR : tout change / rien ne change
 - Quand s'arrête la correction ?
 - Structurer les documents
 - Segmenter le texte et identifier la césure à la ligne
 - Résoudre les abréviations
 - Et l'édition critique ?
- Conclusion

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/
- 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 diplomatique (qui garde les `tei:lb`)

- └ Après l'HTR : tout change / rien ne change
- └ Quand s'arrête la correction ?
- └ Quand s'arrête la correction ?

2024-09-10

- 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/
- 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 diplomatique (qui garde les `tei:lb`)

Assurer la rétroconvertibilité

```

<lb break="yes" facs="#facs_eSc_line_92b25bc7" xml:id="elem_eSc_line_92b25bc7">cū uegetablib̄ γ plātis ut pō
<lb break="?" facs="#facs_eSc_line_4fc0c0e" xml:id="elem_eSc_line_4fc0c0e"/>nūttia augm̄tia. ḡnatia γ
<lb break="?" facs="#facs_eSc_line_1e32f48c" xml:id="elem_eSc_line_1e32f48c"/>tlia q γ ip̄s arborib̄ cōpetūt.
<lb break="?" facs="#facs_eSc_line_2fddd609" xml:id="elem_eSc_line_2fddd609"/>po γ cognit̄e sfit̄e sūt uis
<lb break="yes" facs="#facs_eSc_line_b1fb0fe4" xml:id="elem_eSc_line_b1fb0fe4"/>gust̄ γ tactus. γ tlia in qb̄ q
<lb break="?" facs="#facs_eSc_line_63be87f7" xml:id="elem_eSc_line_63be87f7"/>cam̄ cū brutis. appetitie ū di
<lb break="?" facs="#facs_eSc_line_2304726d" xml:id="elem_eSc_line_2304726d"/>stīgūt. nā qdam ē appetit̄ i
<lb break="?" facs="#facs_eSc_line_9e3287cc" xml:id="elem_eSc_line_9e3287cc"/>hoie i q n̄ qcat cū brutis ut
<lb break="?" facs="#facs_eSc_line_4f2eeff0c" xml:id="elem_eSc_line_4f2eeff0c"/>appetit̄ seq̄s itllem. Qdā ū
<lb break="yes" facs="#facs_eSc_line_887a268e" xml:id="elem_eSc_line_887a268e"/>i q qcat cū eis appetit̄ seq̄s se
<lb break="?" facs="#facs_eSc_line_3c8f19df" xml:id="elem_eSc_line_3c8f19df"/>sū. Appetit̄ āt seq̄s sfm pt
<lb break="?" facs="#facs_eSc_line_cd2f421c" xml:id="elem_eSc_line_cd2f421c"/>noīari sfualitas. seq̄s itlē
<lb break="?" facs="#facs_eSc_line_5d351c82" xml:id="elem_eSc_line_5d351c82"/>ctū noīē uolūtas. f q̄mod lo-
<lb break="?" facs="#facs_eSc_line_fc03f518" xml:id="elem_eSc_line_fc03f518"/>q̄di bruta h̄ sfualitate γ appen-
<lb break="?" facs="#facs_eSc_line_3ecf893f" xml:id="elem_eSc_line_3ecf893f"/>titū sfit̄im. s̄ n̄ uolūtate h̄

```

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

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Quand s'arrête la correction ?
- └ Assurer la rétroconvertibilité



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

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. ghatia 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 ̄ q ̄ ip̄is arborib̄ cōpetūt." 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' ̄ 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 ̄ tactus. ̄ tlia in qb̄ ̄ g" 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

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Quand s'arrête la correction ?
- └ Assurer la rétroconvertibilité

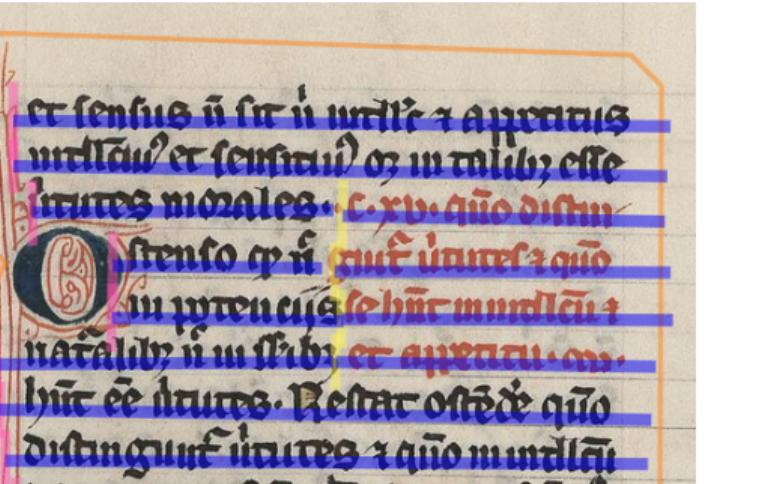
Assurer la rétroconvertibilité



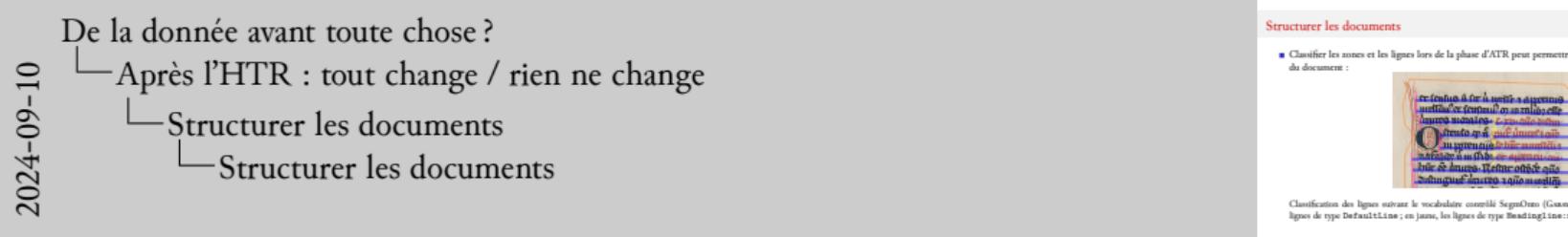
Le fichier ALTO d'origine

Structurer les documents

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

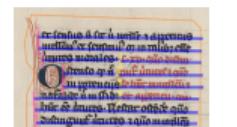


Classification des lignes suivant le vocabulaire contrôlé SegmOnto (GABAY et al. 2021). En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric



Structurer les documents

- Classifier les zones et les lignes lors de la phase d'ATR peut permettre de faciliter la structuration 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

- On utilisera la première ligne de la rubrique (dans la zone de texte principal) pour identifier une borne de division (un chapitre par exemple), et ainsi de suite pour tout le document.
- L'exemple donné ici est doublement problématique : le numéro de chapitre est erronné, ce qui peut poser problème si l'on utilise le texte pour numérotter les divisions ; en second lieu, l'ordre des lignes est évident pour l'humain mais plus difficile à identifier pour la machine, ce qui peut de même poser problème pour la structuration du texte.

Identifier la césure à la ligne

- Dans les manuscrits médiévaux la césure à la ligne n'est pas systématiquement indiquée
- La tâche d'identification de la césure est raisonnablement automatisable

*podies sensituos. Qd alli como n̄igū
omē nō es alabado n̄ es tenido por bue
no por q̄ muela bie su viāda n̄ por q̄ cie
se bie alli nō es alabado por q̄ bie agu
da mēte / o eye total mēte. Saluo ende*

*v appetitu intellectuū. litutes ḡ de q̄
loqui intendim̄ q̄ sūt q̄dam hic̄ lan
tabiles ut erit impotens n̄alibus
ul̄ m̄ ipsi s̄ibz ul̄ m̄ appetituū. v̄ituo
ul̄ m̄ appetituū intellectuū ul̄ m̄ ip̄o idē
tud m̄ om̄ibz h̄is ul̄ m̄ aliqbz horū*

*sophia. geometria metaphy
sica. r̄e. calia. virtutes. nevi si
phicē morales. sūt ille q̄ sūt
m̄ appetitu. sūne appetitus il
le sit sensitivus. sūne intelle
ctuus. cuiusmodi aut̄ sūt
iustitia. tempencia. fortitudo*

Valladolid, 251 fol. 17r ; Vatican, Borgh. 360, fol. 190r ; Cambridge, Corpus Christi College, MS 283, fol 14r ;
Valencia BH Ms 0594, fol. 23r

Dans les manuscrits médiévaux la césure à la ligne n'est pas systématiquement indiquée
La tâche d'identification de la césure est raisonnablement automatisable

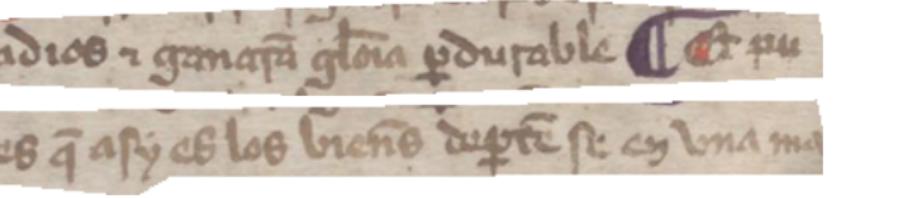
*...q̄ sūt q̄dam hic̄ lan
tabiles ut erit impotens n̄alibus
ul̄ m̄ ipsi s̄ibz ul̄ m̄ appetituū. v̄ituo
ul̄ m̄ appetituū intellectuū ul̄ m̄ ip̄o idē
tud m̄ om̄ibz h̄is ul̄ m̄ aliqbz horū*

*v appetituū intellectuū. litutes ḡ de q̄
loqui intendim̄ q̄ sūt q̄dam hic̄ lan
tabiles ut erit impotens n̄alibus
ul̄ m̄ ipsi s̄ibz ul̄ m̄ appetituū. v̄ituo
ul̄ m̄ appetituū intellectuū ul̄ m̄ ip̄o idē
tud m̄ om̄ibz h̄is ul̄ m̄ aliqbz horū*

*sophia. geometria metaphy
sica. r̄e. calia. virtutes. nevi si
phicē morales. sūt ille q̄ sūt
m̄ appetitu. sūne appetitus il
le sit sensitivus. sūne intelle
ctuus. cuiusmodi aut̄ sūt
iustitia. tempencia. fortitudo*

Valladolid, 251 fol. 17r ; Vatican, Borgh. 360, fol. 190r ; Cambridge, Corpus Christi College, MS 283, fol 14r ;
Valencia BH Ms 0594, fol. 23r

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure à la ligne
- └ Segmenter le texte et identifier la césure

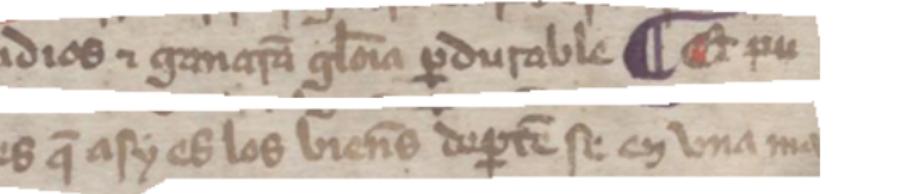
2024-09-10



Ms. 251, Valladolid, fol. 3v

Il suffit donc d'associer deux à deux les lignes (faire des bigrammes de lignes) afin d'identifier si la chaîne de caractère de fin de ligne correspond aussi à une fin de mot.

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma

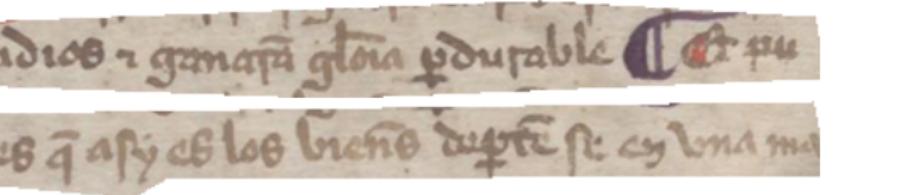
- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure à la ligne
- └ Segmenter le texte et identifier la césure

2024-09-10

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma
Ms. 251, Valladolid, fol. 3v
adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma

Il suffit donc d'associer deux à deux les lignes (faire des bigrammes de lignes) afin d'identifier si la chaîne de caractère de fin de ligne correspond aussi à une fin de mot.

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma

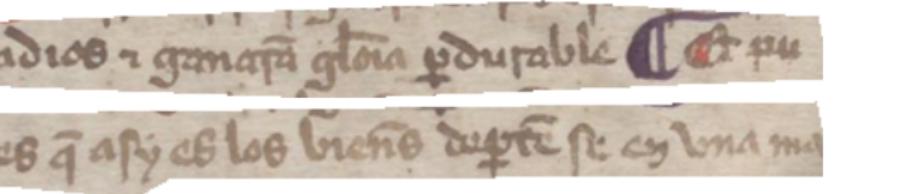


- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure à la ligne
- └ Segmenter le texte et identifier la césure

2024-09-10

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma
Ms. 251, Valladolid, fol. 3v
adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma
¶

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

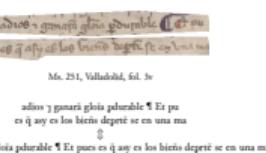
adios j ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma



adios j ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deprtē se en una ma

- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure à la ligne
- └ Segmenter le texte et identifier la césure

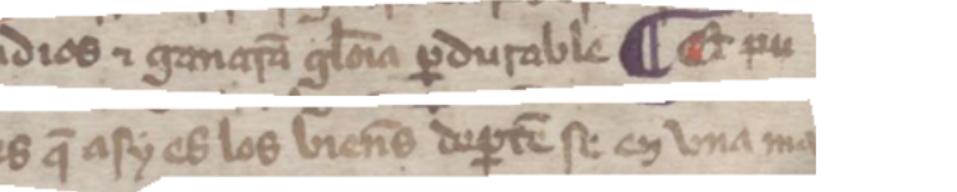
2024-09-10



Ms. 251, Valladolid, fol. 3v

adios j ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma
adios j ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deprtē se en una ma

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma



adios J ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deprtē se en una ma

Voir (CLÉRICE 2020) et https://github.com/matgille/boudams_like_tokenizer

- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure à la ligne
- └ Segmenter le texte et identifier la césure

2024-09-10

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma

Ms. 251, Valladolid, fol. 3v

adios J ganarā gloia pdurable ¶ Et pu
es q asy es los bieñs deprtē se en una ma

adios J ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deprtē se en una ma

Voir (CLÉRICE 2020) et 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 felicit e-l felicitatem
244 cit e-l citatem
245 uot e-l uoluntatem
246 <SOT>cai t e<EOT> ~caritate~
247 <SOT>.p<EOT> ~prout~
248 <SOT>ad<EOT> ~aliquid~
249 gn gener

```

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

De la donnée avant toute chose ?

- Après l'HTR : tout change / rien ne change
 - Résoudre les abréviations
 - Résoudre les abréviations

2024-09-10

- 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 felicit e-l felicitatem
244 cit e-l citatem
245 uot e-l uoluntatem
246 <SOT>cai t e<EOT> ~caritate~
247 <SOT>.p<EOT> ~prout~
248 <SOT>ad<EOT> ~aliquid~
249 gn gener

```

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

Et l'édition critique ?

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

- └ Après l'HTR : tout change / rien ne change
- └ Et l'édition critique ?
- └ Et l'édition critique ?

2024-09-10

Et l'édition critique ?

	ab	ab	ab	ab	ab	ab	ab	ab	ab
	homéotéleute								
	lex	~	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	
Bevil aqua_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	m Morales			intellectuales	et	m Morales	
Geneve Ms_Lat_92	et	medie	quedum		inter			m Morales	
Metz_Mediatheque_1234	quedam	medie		interintellectuales			et	m Morales	
Beinecke_Marston_MS_139	quadam	medie			inter	intellectumales	et	m Morales	
BNE_MSS_958	quad ^{fi}	medie			inter	intellectuales	et	m Morales	
BNF_Lat_6477	quedam	medie			inter	intellectuales	et	m Morales	
BNE_9236	quedam	medie			inter	intellectuales	et	m Morales	
BNF_Lat_1234	quedam	medie			inter	intellectumales	et	m Morales	
Valencia_BH_Ms_0594	quedam	medie		interintellectuales					
	quedam	medie			inter	intellectuales	et	m Morales	

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

2024-09-10

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Et l'édition critique ?
- └ Et l'édition critique ?

ab	ab	ab	ab	ab	ab	ab	ab		
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			
Bevil aqua_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	m Morales		intellectuales	et	m Morales		
Geneve Ms_Lat_92	et	medie	quedum	inter			m Morales		
Metz_Mediatheque_1234	quedam	medie		interintellectuales		et	m Morales		
Beinecke_Marston_MS_139	quadam	medie			inter	intellectumales	et	m Morales	
BNE_MSS_958	quad ^{fi}	medie			inter	intellectuales	et	m Morales	
BNF_Lat_6477	quedam	medie			inter	intellectuales	et	m Morales	
BNE_9236	quedam	medie			inter	intellectuales	et	m Morales	
BNF_Lat_1234	quedam	medie			inter	intellectumales	et	m Morales	
Valencia_BH_Ms_0594	quedam	medie		interintellectuales					
	quedam	medie			inter	intellectuales	et	m Morales	

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

Un texte difficilement exploitable, en sortie d'HTR désabrévié, sans correction et avec affinage minimal des modèles.
Il reste du travail sur les modèles !

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			sensitiuae	sunt	visus		gustus	auditus		et	talia	
Rome_1556			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

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Et l'édition critique ?
- └ Et l'édition critique ?

2024-09-10

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			sensitiuae	sunt	visus		gustus	auditus		et	talia	
Rome_1556			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	
BNP_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

- Il reste encore des erreurs mais le texte est exploitable.
- Un mot sur la typologie des variantes : elle s'appuie sur les annotation lexicales (lemmes) : il suffit d'une erreur de lemmatisation (dûe à une variante graphique) pour fausser la classification et créer du faux positif, d'où une tâche dont la difficulté augmente avec le nombre de témoins.

1 Introduction

2 Phase de production des données

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

4 Conclusions

De la donnée avant toute chose ?
└ Conclusions

2024-09-10

- Introduction
- Phase de production des données
- Après l'HTR : tout change / rien ne change
- Conclusions

- 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
- Se mettre en conformité (ou pas) avec les normes existantes et le documenter clairement

■ 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
- 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
- Pour la collation qui suppose un degré de précision plus élevé des outils d'ATR

2024-09-10

- 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
 - 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
 - Pour la collation qui suppose un degré de précision plus élevé des outils d'ATR

Merci!

└ Merci!

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, Malamenia VLACHOU-EFSTATHIOU, Alix CHAGUÉ, Jean-Baptiste CAMPS, Matthias Gilles 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, Malamenia VLACHOU-EFSTATHIOU, Matthias Gilles 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.

Conclusions

Références

2024-09-10

- [1] Thibault CLÉRICE. « Reducing Deep Learning Methods for Word Segmentation of Scripta Continua Text 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, Malamenia VLACHOU-EFSTATHIOU, Alix CHAGUÉ, Matthias Gilles LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, 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, Malamenia VLACHOU-EFSTATHIOU, Matthias Gilles 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.
- [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.