

De la donnée avant toute chose ?

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

 |  |  | 

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

 |  |  | 

De la donnée avant toute chose ?

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
- Concilier l'intérêt particulier et les besoins généraux

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

- Quand s'arrête la correction ?
- Structurer les documents
- Segmenter le texte et identifier la césure
- 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

Introduction

2024-09-09

- 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*] [\emptyset *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om.* *Q* 3 pueda *AZ* | puedan *BGRQJ*

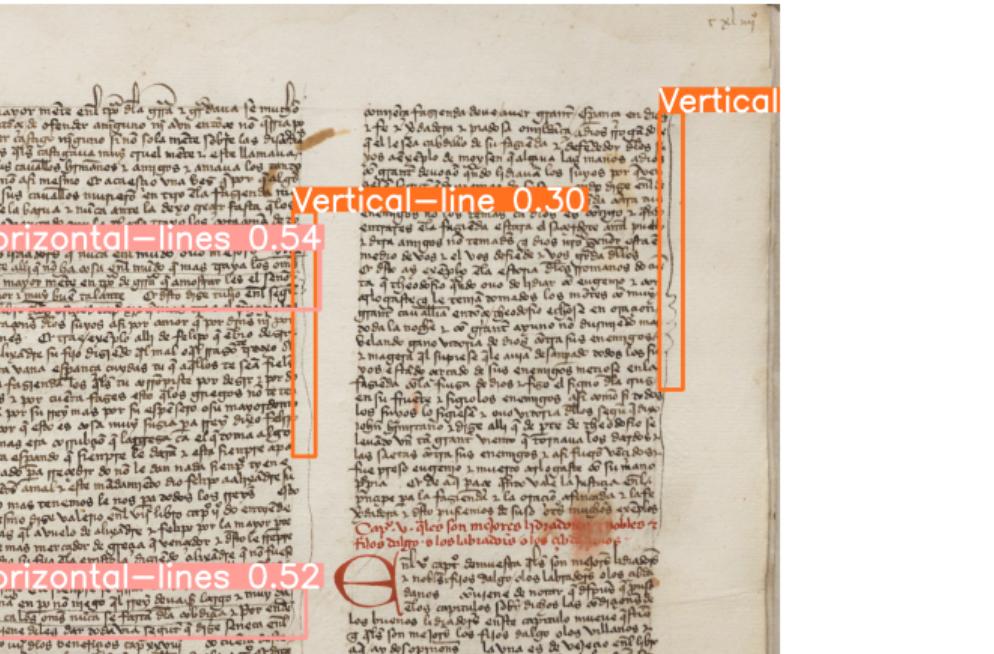
Édition critique du *Regimiento de los Príncipes*, avec collation automatisée. Les témoins A et Z sont issus d'HTR.

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*] [\emptyset *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.

Études des marques de lecture d'un manuscrit



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

└ Introduction

└ Études des marques de lecture d'un manuscrit

2024-09-09



Identification automatisée (avec YOLO v5) 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, 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, 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
- Concilier l'intérêt particulier et les besoins généraux

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

4 Conclusions

2024-09-09

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

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

De la donnée avant toute chose?
└ Phase de production des données
 └ Penser la production en amont

2024-09-09

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

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :

De la donnée avant toute chose?
└ Phase de production des données
 └ Penser la production en amont

2024-09-09

■ Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
■ Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :
 - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ?

De la donnée avant toute chose ?
└ Phase de production des données
 └ Penser la production en amont

2024-09-09

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :
 - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ?

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :
 - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ?
 - Que faire des abréviations ?

De la donnée avant toute chose ?
└ Phase de production des données
 └ Penser la production en amont

2024-09-09

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production (ou via des campagnes d'annotation-test) sur des normes d'annotation (manuel?) :
 - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ?
 - Que faire des abréviations ?

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

2024-09-09

Statistiques sur le corpus CATMuS

■ Test

De la donnée avant toute chose ?

└ Phase de production des données

 └ CATMuS, un projet de production collaboratif de données

 └ Statistiques sur le corpus CATMuS

2024-09-09

■ Test

Le problème des abréviations, entre usages historiens et usages philologiques

De la donnée avant toute chose ?

└ Phase de production des données

 └ Concilier l'intérêt particulier et les besoins généraux

 └ Le problème des abréviations, entre usages historiens et usages philologiques

2024-09-09

Le choix de la conservation des abréviations

De la donnée avant toute chose ?

└ Phase de production des données

└ Concilier l'intérêt particulier et les besoins généraux

└ Le choix de la conservation des abréviations

2024-09-09

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
- Résoudre les abréviations
- Et l'édition critique ?

4 Conclusions

2024-09-09

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
- Résoudre les abréviations
- Et l'édition critique ?

Conclusions

- La phase suivant la transcription automatisée sera généralement celle de la transformation en TEI
- Il restera des erreurs dans les données
- Faut-il intégrer ces correction 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`)

De la donnée avant toute chose ?

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

2024-09-09

- La phase suivant la transcription automatisée sera généralement celle de la transformation en TEI
- Il restera des erreurs dans les données
- Faut-il intégrer ces correction 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 uegetablib- 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 arborib- c̄petūt.
<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"/>gust⁹ γ tactus. γ tlia in qb- q
<lb break="?" facs="#facs_eSc_line_63be87f7" xml:id="elem_eSc_line_63be87f7"/>cam⁹ cū brutis. appetitie ū di
<lb break="?" facs="#facs_eSc_line_2304726d" xml:id="elem_eSc_line_2304726d"/>stīgūt. nā qdam ē appetit⁹ 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"/>appetit⁹ seq̄s itllem. Qdā ū
<lb break="yes" facs="#facs_eSc_line_887a268e" xml:id="elem_eSc_line_887a268e"/>i q qcat cū eis appetit⁹ seq̄s se
<lb break="?" facs="#facs_eSc_line_3c8f19df" xml:id="elem_eSc_line_3c8f19df"/>sū. Appetit⁹ 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 h̄t sfualitate γ appen-
<lb break="?" facs="#facs_eSc_line_3ecf893f" xml:id="elem_eSc_line_3ecf893f"/>titū sfit̄im. s- n̄ uolūtate h̄t

```

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

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

2024-09-09



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 ȝ γ 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' 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 γ tactus. γ tlia in qb̄ ȝ" 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="stīgūf. nā qdam ē appetit̄ i" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0"/>
</TextLine>

```

Le fichier ALTO en question

De la donnée avant toute chose ?

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

2024-09-09

Assurer la rétroconvertibilité ?

```

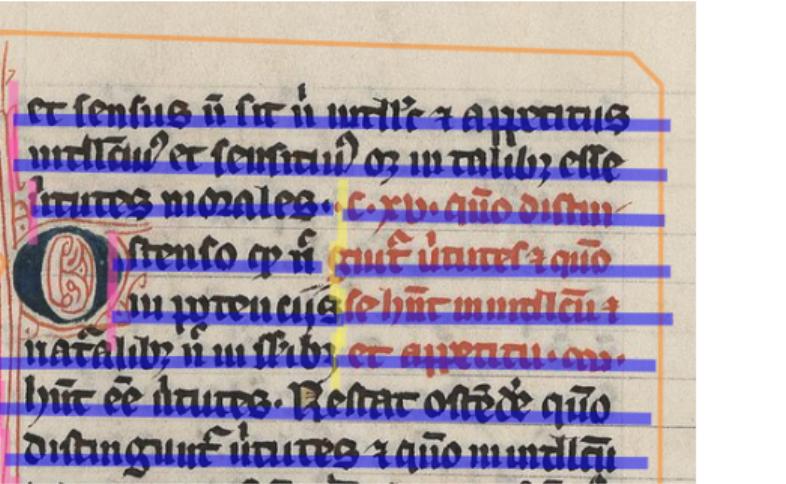
+<TextLine ID="eSc_line_4fc0c0e" TAGREFS="LT6426" BASELINE="1581 1374 2153 1366" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0">
+<Shape></Shape>
<String CONTENT="nuttia augm̄tia. ghatia γ" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0"/>
</TextLine>
-<TextLine ID="eSc_line_1e32f48c" TAGREFS="LT6426" BASELINE="1573 1434 2164 1421" HPOS="1571.0" VPOS="1375.0" WIDTH="593.0" HEIGHT="89.0">
+<Shape></Shape>
<String CONTENT="tlia ȝ γ 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' 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 γ tactus. γ tlia in qb̄ ȝ" 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="stīgūf. nā qdam ē appetit̄ i" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0"/>
</TextLine>

```

Le fichier ALTO en question

Structurer les documents

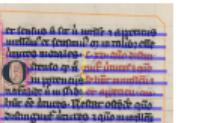
- Classifier les zones et les lignes lors de la phase d'ATR peut permettre de faciliter la structuration du document :



Classification des lignes suivant le vocabulaire contrôlé SegmOnto. En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric

- Après l'HTR : tout change / rien ne change
- Structurer les documents
- Structurer les documents

2024-09-09



Classification des lignes suivant le vocabulaire contrôlé SegmOnto. En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric

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.

- 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. Or assi como nigu
ome no es alabado ni es tenido por bue
no por q muelle bié su uiaza ni por q cre
sar bié assi no es alabado por q bee agu
da merte o oye sotil merte. Saluo ende*

*rōne ptiapant qma p se a s̄ q̄ si no
obedim̄t rōm ḡ m sensib̄ a i potenc̄
is n̄ibns vt sup̄ dicebatur no s̄int
esse v̄ntes e sō h̄nt m i tellam̄ ap
petitū m nobis an̄ d̄upser ea appen*

*.v appetitu intell̄tuū. l̄ctutes ḡ de q̄b
loqui intendim̄ q̄ s̄ut q̄dam h̄c lau
tabiles ul̄ erit in potētis h̄ali bus
ul̄ in ipsiis s̄ibz ul̄ in appetitu s̄itio
ul̄ in appetitu intell̄tuū ul̄ in ip̄o idle
tu ul̄ in om̄ibz h̄us ul̄ in aliq̄bz lorū*

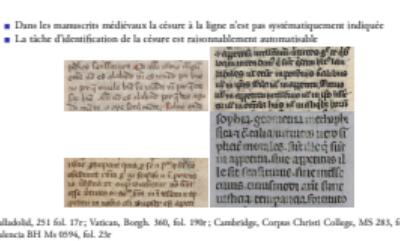
*sophia. geometria metaphi
sica. et talia uirtutes uero si
phici morales. s̄ut ille q̄ s̄ut
in appetitu. siue appetitus il
le sit sensitum. siue intellectuū
ciusmodi aut sunt
iusticia. tempancia. fortitudo*

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

De la donnée avant toute chose ?

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

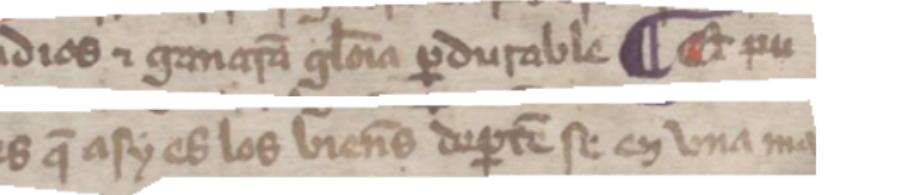
2024-09-09



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

- On montre des cas d'indications, des cas de non-indication (un texte en castillan, trois en latin ; deux manuscrits ibériques, un manuscrit anglais, et un manuscrit italien).
- Dans le corpus manuscrit castillan du 15e siècle, c'est assez peu commun, alors que l'indication sera plus fréquente en latin par exemple.
- Cette tâche ne peut être actuellement efficacement traitée par les outils d'ATR comme Kraken, étant donné que l'outil ne fonctionne que ligne par ligne sans contexte.

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
- └ Segmenter le texte et identifier la césure

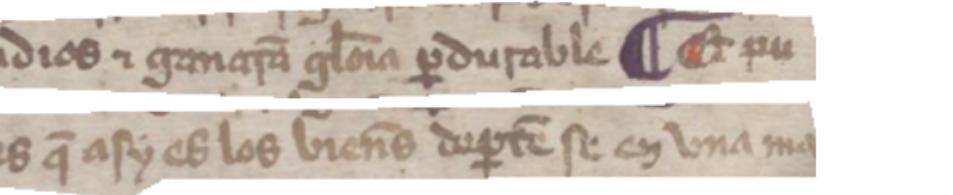
2024-09-09



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 y ganaran gloria perdurable ¶ Et pu
es que asy es los bienes departen se en una ma*

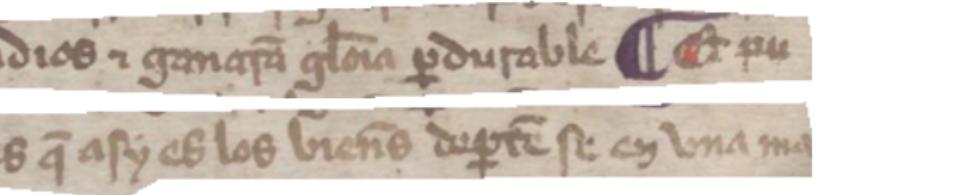
- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure
- └ Segmenter le texte et identifier la césure

2024-09-09

adios y ganaran gloria perdurable ¶ Et pu
es q asy es los bienes departen se en una ma
Ms. 251, Valladolid, fol. 3v
adios y ganaran gloria perdurable ¶ Et pu
es que asy es los bienes departen 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 y ganaran gloria perdurable ¶ Et pu
es que asy es los bienes departen se en una ma*

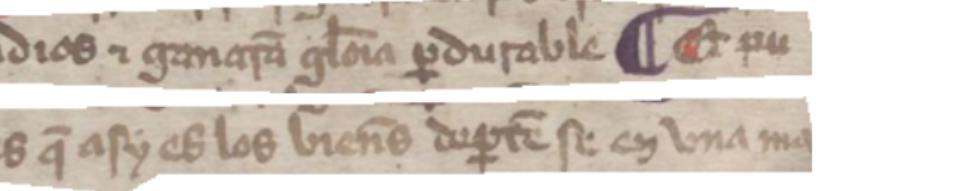


- └ Après l'HTR : tout change / rien ne change
- └ Segmenter le texte et identifier la césure
- └ Segmenter le texte et identifier la césure

2024-09-09

adios y ganaran gloria perdurable ¶ Et pu
es q asy es los bienes departen se en una ma
Ms. 251, Valladolid, fol. 3v
adios y ganaran gloria perdurable ¶ Et pu
es que asy es los bienes departen se en una ma
¶

Segmenter le texte et identifier la césure



Ms. 251, Valladolid, fol. 3v

*adios y ganaran gloria perdurable ¶ Et pu
es que asy es los bienes departen se en una ma*



adios y ganaran gloria perdurable ¶ Et pues que asy es los bienes departen se en una ma

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

2024-09-09

adios y ganaran gloria perdurable ¶ Et pu
es q a sy es los bienes departen se en una ma
Ms. 251, Valladolid, fol. 3v

adios y ganaran gloria perdurable ¶ Et pues que asy es los bienes departen se en una ma

adios y ganaran gloria perdurable ¶ Et pues que asy es los bienes departen 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.

Résoudre les abréviations

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

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

2024-09-09

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

- 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 en amont les principes d'annotation est fondamental

- Le Plan de Gestion de Données (PGD)
- 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

■ Du travail reste à mener

- Sur tout la chaîne post-ATR afin d'arriver à un texte propre
- Pour la collation qui suppose un degré d'exactitude plus élevé

Conclusions

2024-09-09

- Penser en amont les principes d'annotation est fondamental
 - Le Plan de Gestion de Données (PGD)
 - 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
- Du travail reste à mener
 - Sur tout la chaîne post-ATR afin d'arriver à un texte propre
 - Pour la collation qui suppose un degré d'exactitude plus élevé

Merci!

De la donnée avant toute chose ?

└ Conclusions

└ Merci!

Références I