

Soutenance de projet de fin d'études:

---

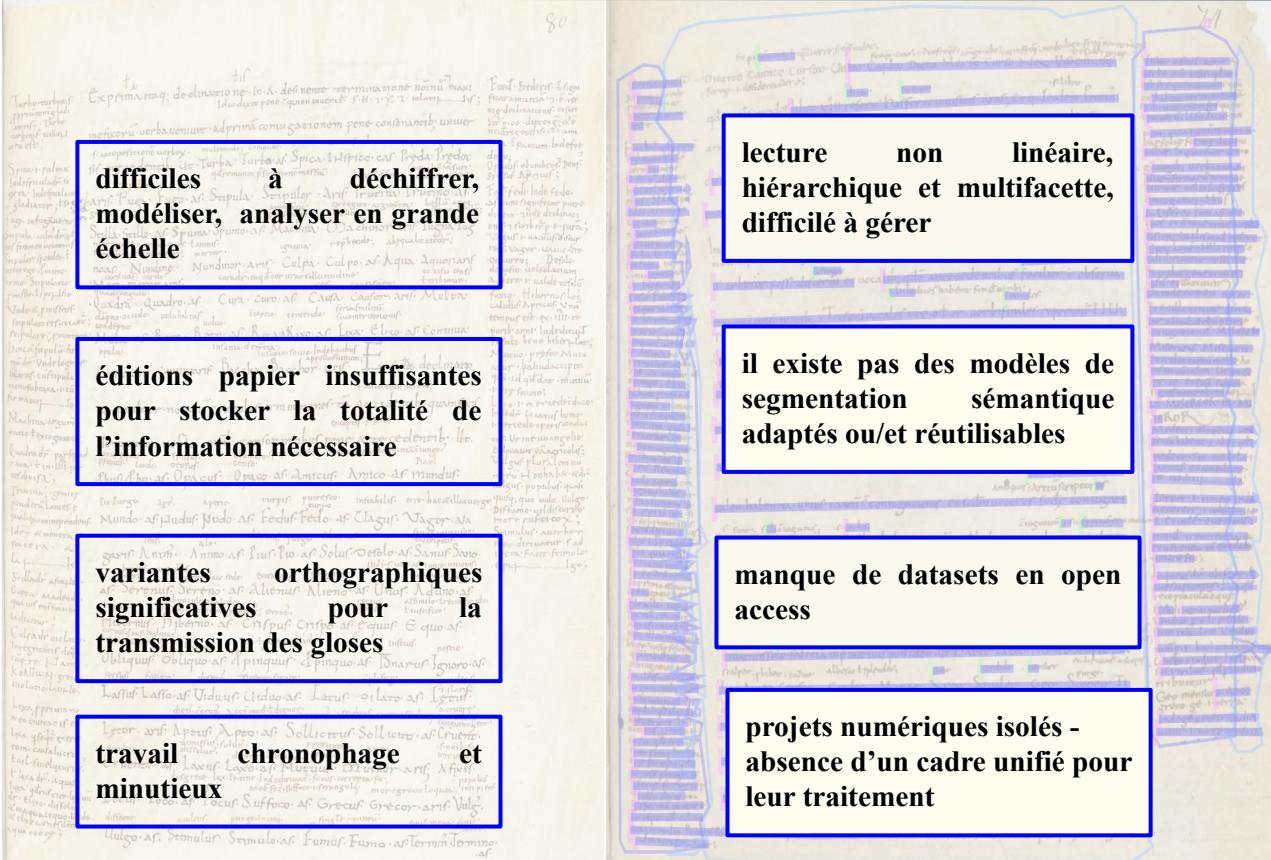
# **Éditer les manuscrits grammaticaux glosés : solutions numériques face aux défis paléographiques**

## **Le cas de la tradition manuscrite glosée d'Eutychès *grammaticus***

---

Encadrée par : Franck CINATO (HTL) et Peter STOKES (EPHE)  
3<sup>e</sup> membre du jury : Chahan VIDAL-GORENE (ENC)

# Problématiques et Motivation



**EVTYCHIS**

**ARS**

**DE VERBO**

**Littérature extensive derrière sa tradition manuscrite**

**Les gloses du « de uerbo » sont largement inédites**

**LIBER I**

**DE CONIVAGATIONIBVS VERBORVM**

**Un nouveau paradigme est nécessaire pour l'édition d'une tradition manuscrite fluide et non-linéaire**



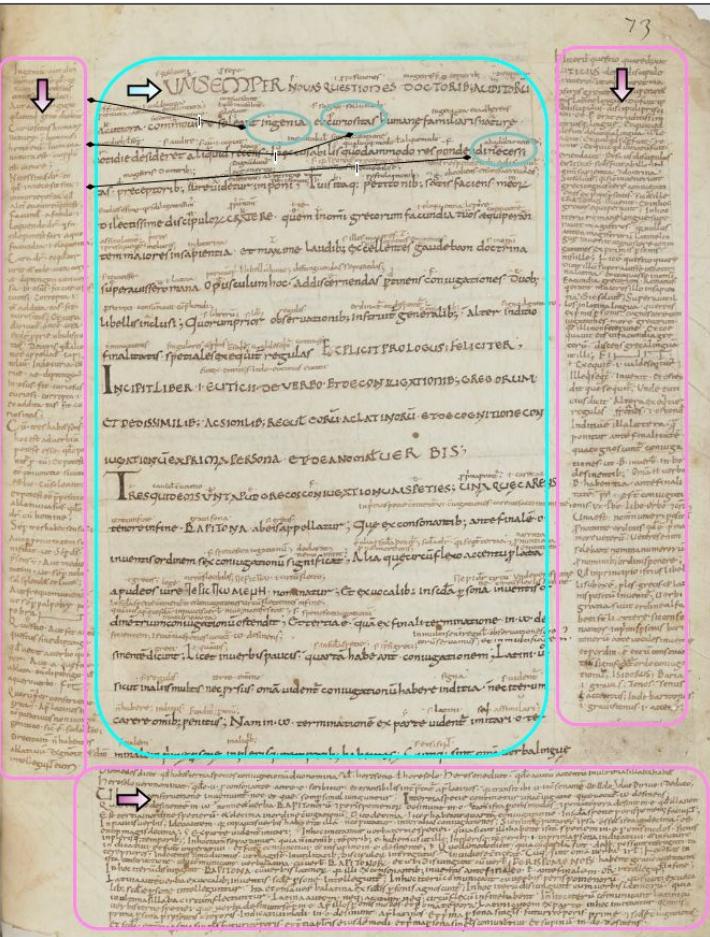
**à la fois un défi et une opportunité pour l'exploration des outils numériques adéquats et efficaces**

# Anatomie d'un manuscrit glosé

**Texte principal** au milieu, qui contient les **lemmes** (mot ou groupe de mots nécessitant explication)

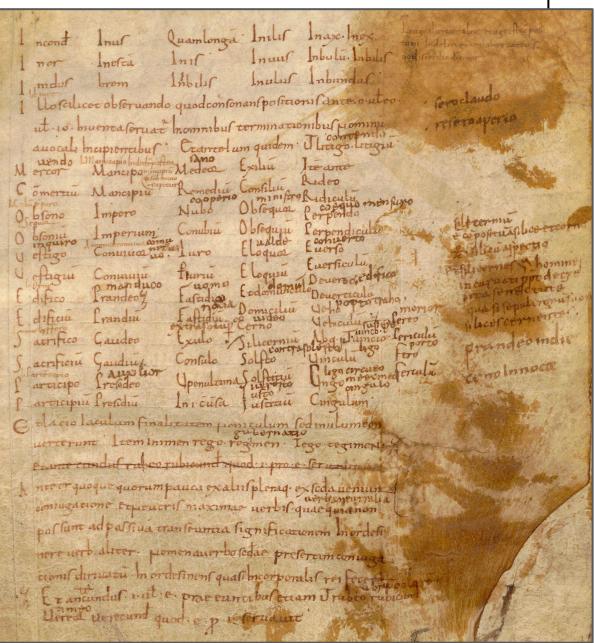
**Annotations interlinéaires** (gloses) éparses entre les lignes principales dont souvent des **notes tironiennes** et **signes de construction**

Plusieurs blocs d'**annotations marginales** (*marginalia*) - NB la lecture non linéaire des différents régistres



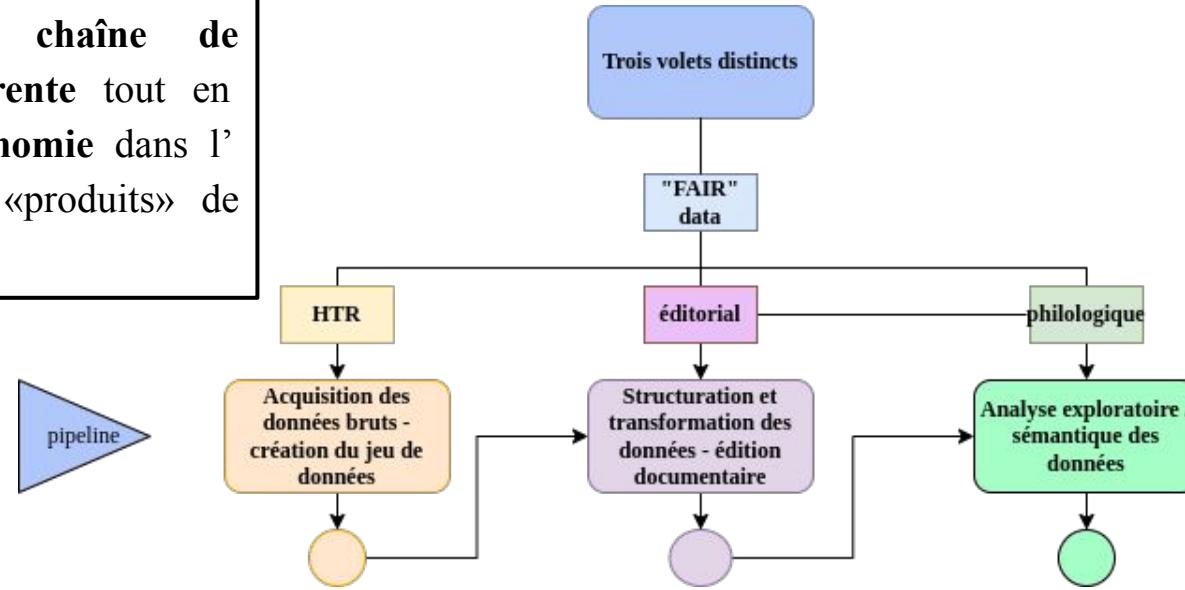
**Folioation** souvent en haut à droite, éléments de **réglure**

présence de **colonnes/tableaux** au sein du texte



# La méthodologie

Création d'une **chaîne de traitement cohérente** tout en préservant l'**autonomie** dans l'**exploitation** les «produits» de chaque volet



-ground truth (ALTO)

-modèle(s)  
segmentation

-modèle(s)  
reconnaissance

de  
éditions documentaires enrichies (XML-TEI)

-base de données de gloses

(CSV)

-analyses statistiques

-visualisations et graphiques

Construction d'un corpus cohérent pour orienter efficacement les questions de recherche et faciliter l'interprétation des résultats.

- **début du IX<sup>e</sup> - début X<sup>e</sup> s.**
- **Nord de la France**
- ***scriptoria* associés à l'  
érudition carolingienne**



- PARIS, BIBLIOTHÈQUE NATIONALE DE FRANCE, Latin 14087 ⇒ glossaire -  $\frac{1}{4}$  du IX<sup>e</sup> s. - **Corbie**
- BAMBERG, STAATSBIBLIOTEK, Msc.Class.30 ⇒  $\frac{3}{4}$  du IX<sup>e</sup> s. - **Reims**
- PARIS, BIBLIOTHÈQUE NATIONALE DE FRANCE , Latin 7499 ⇒  $\frac{1}{2}$  du X<sup>e</sup> s. - proche de **Corbie**, portant en marge “le commentaire de Rémi d’Auxerre”
- LEIDEN, UNIVERSITEITSBIBLIOTHEEK , Vossianus Latinus 8<sup>o</sup> 41 ⇒  $\frac{3}{4}$  du IX<sup>e</sup> s. - **Fleury**

[1]. Jeudy, C. (1974). *Les manuscrits de l' "Ars de uerbo" d'Eutychès et le commentaire de Rémi d'Auxerre.*

# ACQUISITION DE DONNÉES



**S** YALTAi



HTeVX



# 1a: Segmentation - Annotation

Tasks:  
**Complex Layout Analysis (CLA)**

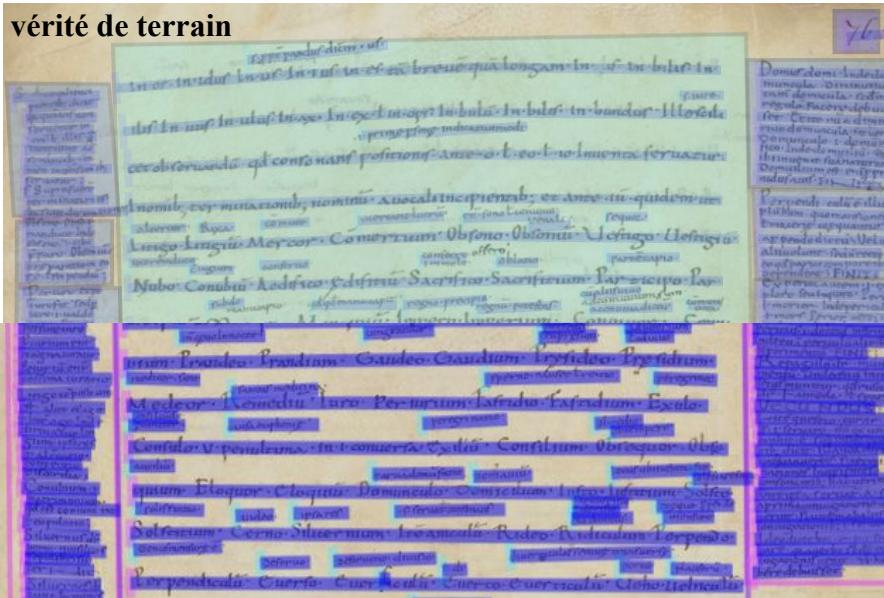
Plateforme:  
**eScriptorium**

Ontologie:  
**SegmOnto**

logiciels:  
**kraken;**  
**YALTAi**

Contrôle qualité:  
**HTRVX**

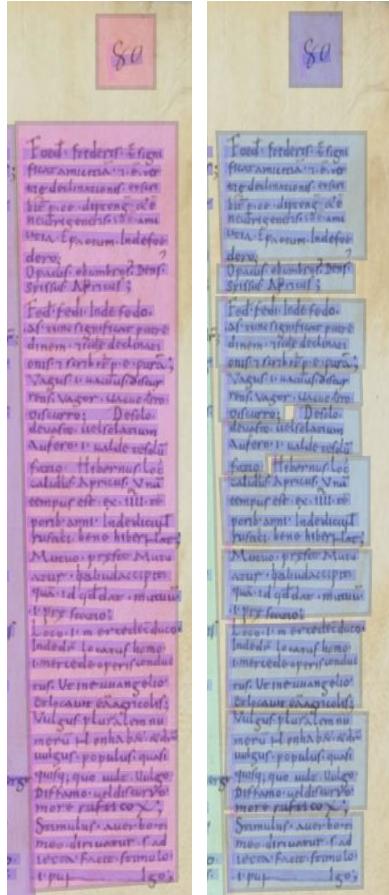
vérité de terrain



Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., & Ingold, R. (2016, October). **Diva-hisdb**: A precisely annotated large dataset of challenging medieval manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 471-476). IEEE.

Cote	Pages	MainZones	MarginTextZones	NumberingZones	DefaultLines	InterlinearLines
VLO41	63	78	51	32	1768	973
BambergMsc30	30	30	167	15	1748	1590
Lat7499	39	39	79	18	3900	2207
CSG863	40	40	87	41	1911	278
CGS18	40	40	283	34	4648	1360
CB55	48	49	176	13	3106	0
<b>Total</b>	<b>260</b>	<b>276</b>	<b>843</b>	<b>153</b>	<b>17081</b>	<b>6408</b>

TABLE 3.1 – Jeu de données pour l'entraînement des modèle de segmentation sémantique YALTAi - kraken



# 1b: Segmentation - Entraînement\* YALTAi

Tasks:  
**Complex Layout Analysis (CLA)**

Plateforme:  
**eScriptorium**

Logiciels:  
**kraken;**  
**YALTAi**

Contrôle qualité:  
**HTRVX**

Ontologie :  
**SegmOnto**

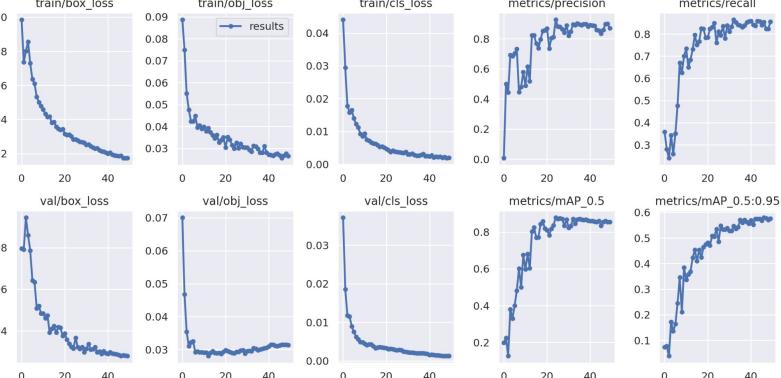
**YALTAi** : Combinaison de **YOLOv5** pour la détection d'objets (zones) avec la **classification sémantique** des baselines à l'aide de *kraken* (lignes).

P-R curve : all classes **0.855 mAP@0.5**

P-C curve : all classes **1.00 at 0.896**

R-C curve : all classes **0.92 at 0.000**

F1- C curve : all classes **0.86 at 0.353**



\*avec hyperparamètres de base sur GPU  
 \*merci à M Vidal-Gorène pour son aide



# 1c: Segmentation - Entraînement kraken

Tasks:  
**Complex Layout Analysis (CLA)**

Plateforme:  
**eScriptorium**

Logiciels:  
**kraken;**  
**YALTAi**

Contrôle qualité:  
**HTRVX**

Ontologie :  
**SegmOnto**

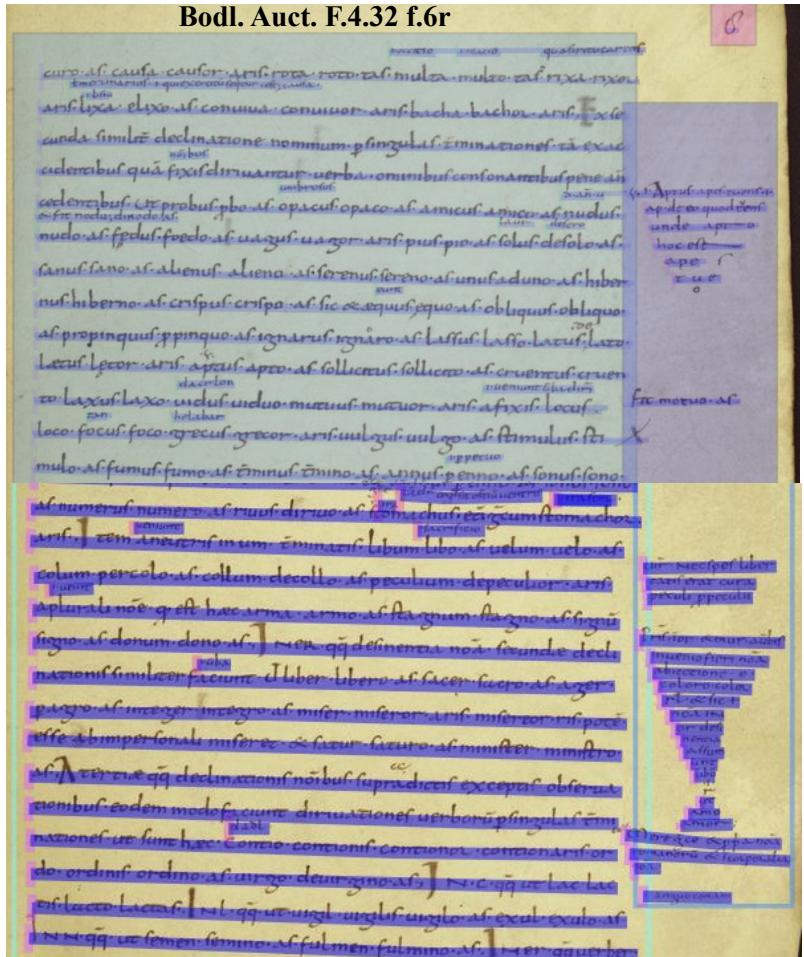
- YALTAi :  $\Rightarrow$  1 modèle pour les zones + **2 modèles kraken** pour les lignes (Default, InterLinearLine) avec possibilité d'emploi indépendant pour extraction sélective;

## Caveats !

- **zones**: Adaptée à une typologie-mise en page spécifique de manuscrits glosés
- **lignes** : Souvent un **overlap** entre classes (observation qualitative)
- besoin de **recalculer** les polygons
- Post-correction minimale 

val\_accuracy: 0.997 val\_mean\_iu: 0.375  
val\_mean\_acc: 0.997 val\_freq\_iu: 0.467  
val\_metric: 0.375

Bodl. Auct. F.4.32 f.6r



# 1d : Entraînement - Minuscule Caroline

Tasks :  
**Handwritten Text Recognition (HTR)**  
Plateforme:  
**eScriptorium**

**Fine-tuning** d'un modèle *kraken* existant de Minuscule Caroline [1] sur nos données (80/20 du VLO41) ⇒ 98,4% d'accuracy

t mercinarius. i. qui ex ercitū seq̄tur cest; causa.

1 mercinarius . i. quiex ercitū seq̄tur cest; causa.

by (eScriptorium) on Sat Jun 24 2023 19:00:24 GMT+0100

ce lā dīr eoq̄sit ostiū uentris

by (eScriptorium) on Sat Jun 24 2023 19:00:43 GMT+0100

ce lā dīr eoq̄sit ostiū uentris

**Post-correction** minimale

**Adaptation** éditoriale (eg. espacement, ponctuation)

esse ab impersonali miseret. & satur. saturo. as. minister. ministro.

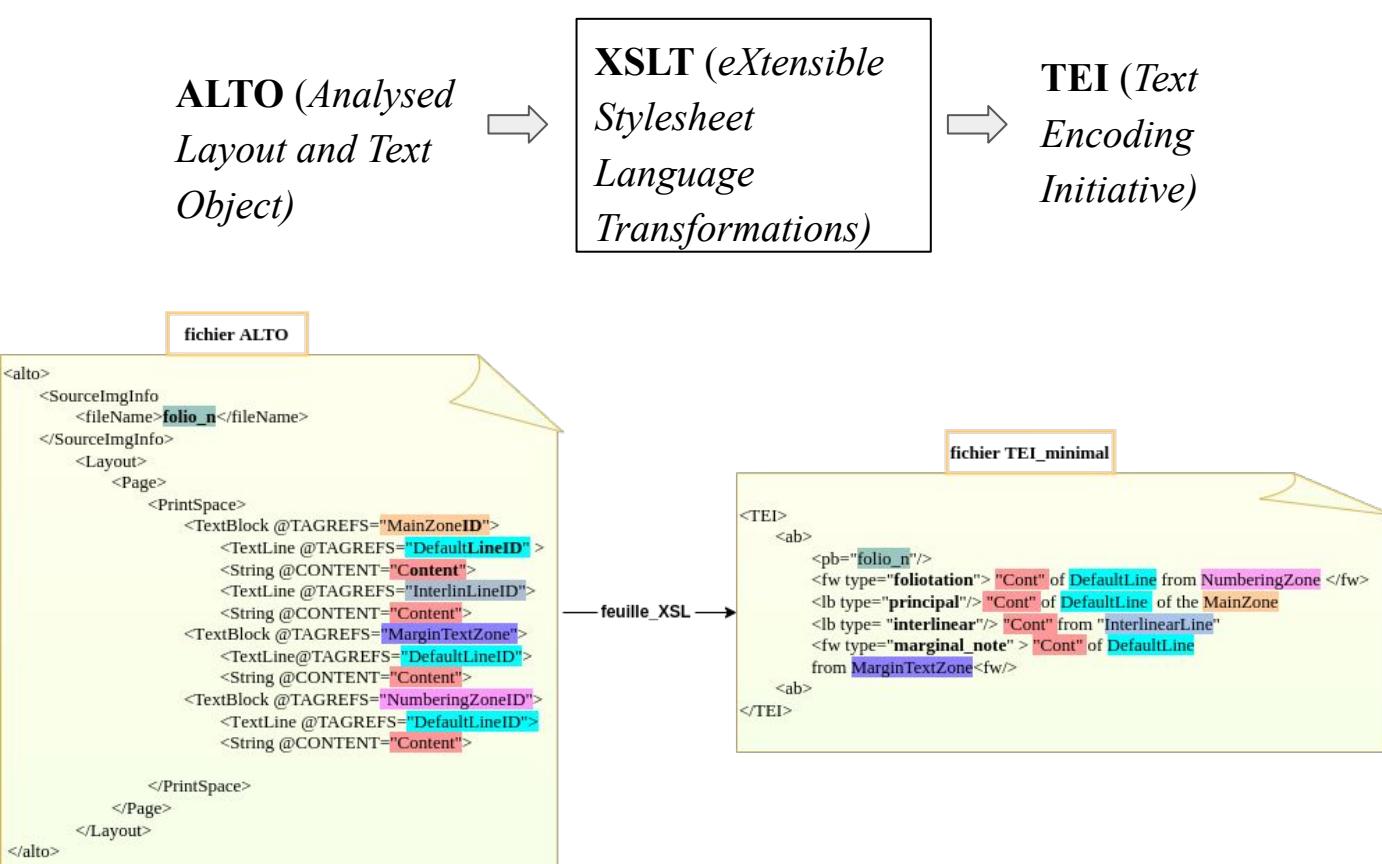
by (eScriptorium) on Sat Jun 24 2023 19:00:56 GMT+0100

esse ab impersonali miseret . & satur . saturo . as. minister . ministro.

[1]<https://github.com/rescribe/carineminuscule-groundtruth>

# TRANSFORMATION ET STRUCTURATION DES DONNÉES



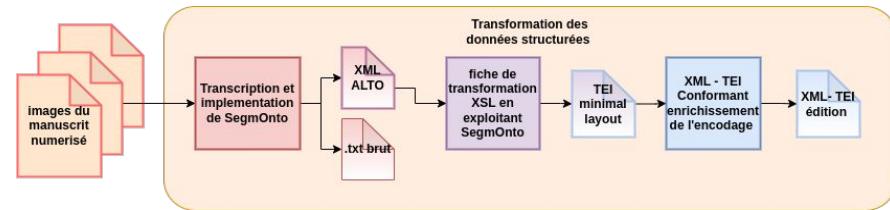


⇒ **Préservation** des informations de mise en page (segmentation sémantique) et du contenu textuel ligne par ligne.

**Optimisation de la gestion et du stockage efficaces sans perte d'informations essentielles** ✓

**Souplesse, interopérabilité et rigueur des données** ✓

## 2b: Structuration des données - ALTO to TEI



**Enrichissement** du TEI\_minimal avec des informations suivantes ⇒ édition documentaire

- 1) le **couple** indissociable **lemme-glose ou/et annotation marginale** + alignment avec l'édition de H.Keil[1] (page\_ligne\_ordre);
  - 2) descripteurs **sémantiques**[2]: **Sens** = **Typologie** et **Forme** = **Taille relative**;
  - 3) **différentes mains** des glosateurs (si plusieurs)
  - 4) *graeca*, citations des auteurs classiques, colonnes (pour VLO41)

<lb type="interlinear"/><gloss xml:id="p447\_l05\_002\_a" target="#p447\_l05\_002" ana="S22" style="F2">p sepe </gloss>  
<lb type="interlinear"/><gloss xml:id="p447\_l05\_004\_a" target="#p447\_l05\_004" ana="S22" style="F2">÷ ppositiones</gloss>  
<lb type="interlinear"/><gloss xml:id="p447\_l05\_005\_a" target="#p447\_l05\_005" ana="S22" style="F3">magistris p̄cepto<sup>r</sup>ib<sup>u</sup></gloss>  
<lb type="interlinear"/><gloss xml:id="p447\_l05\_006\_a" target="#p447\_l05\_006" ana="S22" style="F2">discipulor</gloss>

<lb type="principal"/>  
  <seg type="lemma" xml:id="p447\_l05\_001">VM </seg>  
  <seg type="lemma" xml:id="p447\_l05\_002">SEMPER </seg> NOVAS  
  <seg type="lemma" xml:id="p447\_l05\_004">QVESTIONES </seg>  
  <seg type="lemma" xml:id="p447\_l05\_005">DOCTORIB<sup>u</sup> </seg>  
  <seg type="lemma" xml:id="p447\_l05\_006">AVDITOR<sup>u</sup> </seg>

Encodage de la première ligne *cum glossis* du *de uerbo* selon le ms BnF, Latin 7499

[1] KEIL, Heinrich. *Grammatici Latini: Libros I-XII continens*. Teubner, 1855. [2] CINATO, Franck. *Priscien glosé*. Brepols, 2015 (part. adaptée)

# EXTRACTION ET ANALYSE DES DONNÉES



# 3a: Statistiques descriptives générales

Descripteurs	Lat14087	VLO41	BambergMsc30	Lat7499
folios annotés	3/3	<b>42/63</b>	30/30	39/39
lemmas	233	984	<b>2031</b>	1903
unique lemmas	6	21	33	<b>68</b>
gloses	233	972	1763	<b>1768</b>
gloses en notes tironiennes	N/A	N/A	272	N/A
typologie plus fréquente	S <sub>22</sub>	S <sub>22</sub>	S <sub>22</sub>	S <sub>22</sub>
forme plus fréquente	<b>F<sub>3</sub></b>	F <sub>2</sub>	F <sub>2</sub>	F <sub>2</sub>
marginalia	N/A	20	570	327
typologie plus fréquente	N/A	S <sub>23</sub>	<b>S<sub>63</sub></b>	S <sub>23</sub>
forme plus fréquente	N/A	F <sub>4</sub>	F <sub>2</sub>	<b>F<sub>5</sub></b>
typologies uniques	11	23	35	<b>44</b>

# 3b: « Tableaux de collation »

**Exploitation initiale** : Alignement des témoins via l'indexation des lemmes:

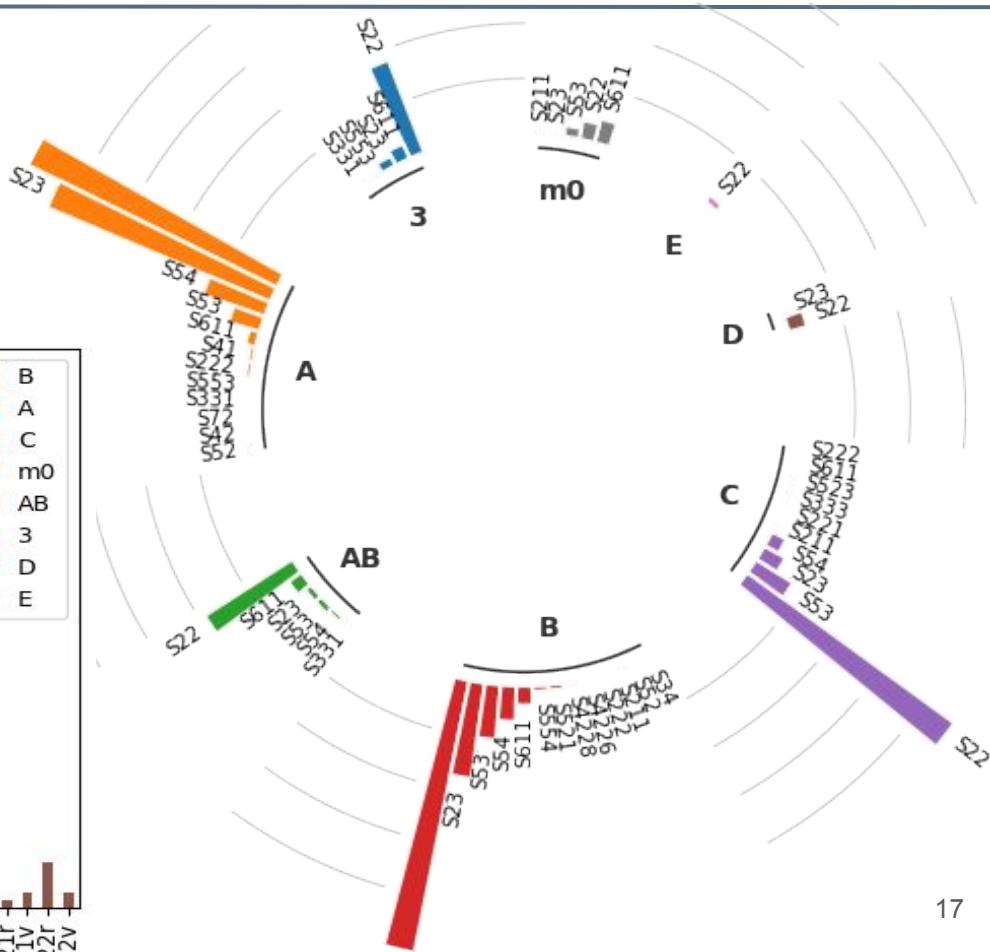
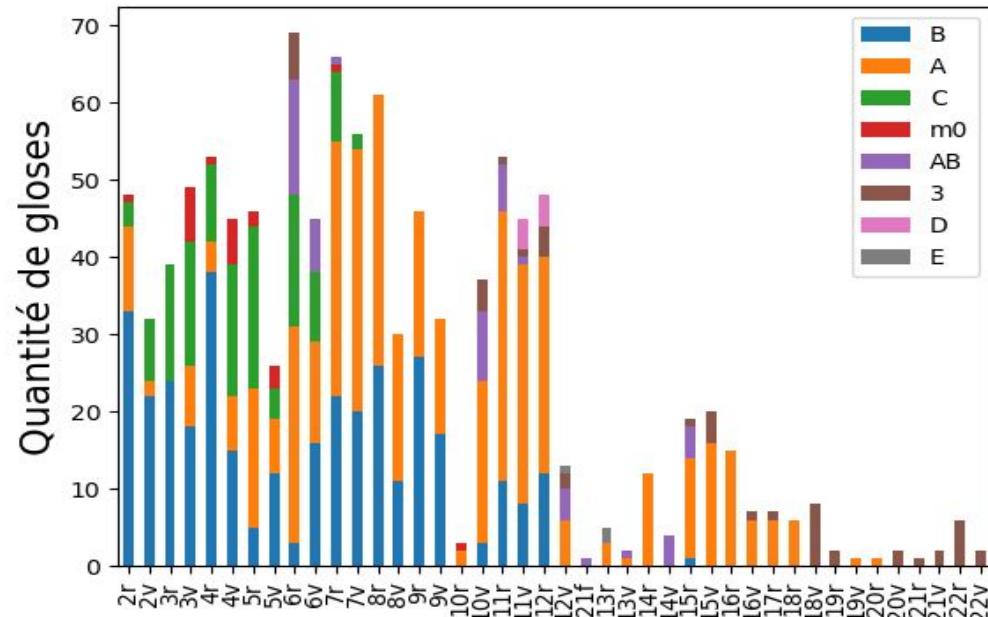
- Création d'une « base de données » rudimentaire permettant la **comparaison** et le **filtrage** des entrées:
  - selon leur **typologie**
  - leur **taille**
  - autres conditions
- **Automatisation**, gain de temps, **simplification**, et **contrôle** du processus de comparaison et de collation des témoins 

lemma_Bamberg	gloss_Bamberg	nature	lemma_Lat7499	gloss_Lat7499
DOCTORIBVS	PRAECEP <sup>toribus</sup>	shorthand	DOCTORIB?	magistris p̄ceptorib?
excellentes	EMIN <sup>entes</sup> SUBAUD <sup>i tur</sup> ILL <sup>os</sup> IDEST MA <sup>iorum</sup> TU <sup>os</sup>	shorthand	excellentes	s. illos maiores / superantes eminentes
finalitatis	TERMINA <sup>tionis</sup>	shorthand	finalitatis	̄minationis
̄minalis	FIN <sup>alis</sup>	shorthand	terminalis	finalis
prime	CONIUGA <sup>tionis</sup>	shorthand	prime	.s.čiugā
possī	CONIUNCTI <sup>us</sup> MODUS	shorthand	possim	čiunctuius modus
edo	COMED <sup>o</sup>	shorthand	Edo	manduco

### 3c: Etude individuelle (VLO41)

1) Visualisation de la répartition de l'activité non simultanée des glosateurs différents (« campagnes d'annotation ») 

## 2) Caractérisation de leur rôle dans l'annotation selon leurs typologies privilégiées ➔



### 3d: Rôle des annotations

Évaluation objective du rôle des gloses interlinéaires par rapport aux notes marginales dans notre jeu de données selon les occurrences des typologies.

Marginal Typologies	Gloss Typologies
63 Titulation, Indexation marginale	22 Synonyme
23 Définition littérale	23 Définition littérale
53 Élucide les ambiguïtés du texte	53 Élucide les ambiguïtés du texte
5 Explicative	3232 Explicite le contexte d'un pronom
22 Synonyme	54 Étymologique
54 Étymologique	611 Correction critique du texte ou Correction rédactionnelle
631 Résumé du contenu	25 Décomposition d'un mot
62 Collation et variantes	211 Traduction du grec au latin (ou l'inverse)
4181 Corrélation entre explication et citations	415 Corrélation entre un verbe et son sujet

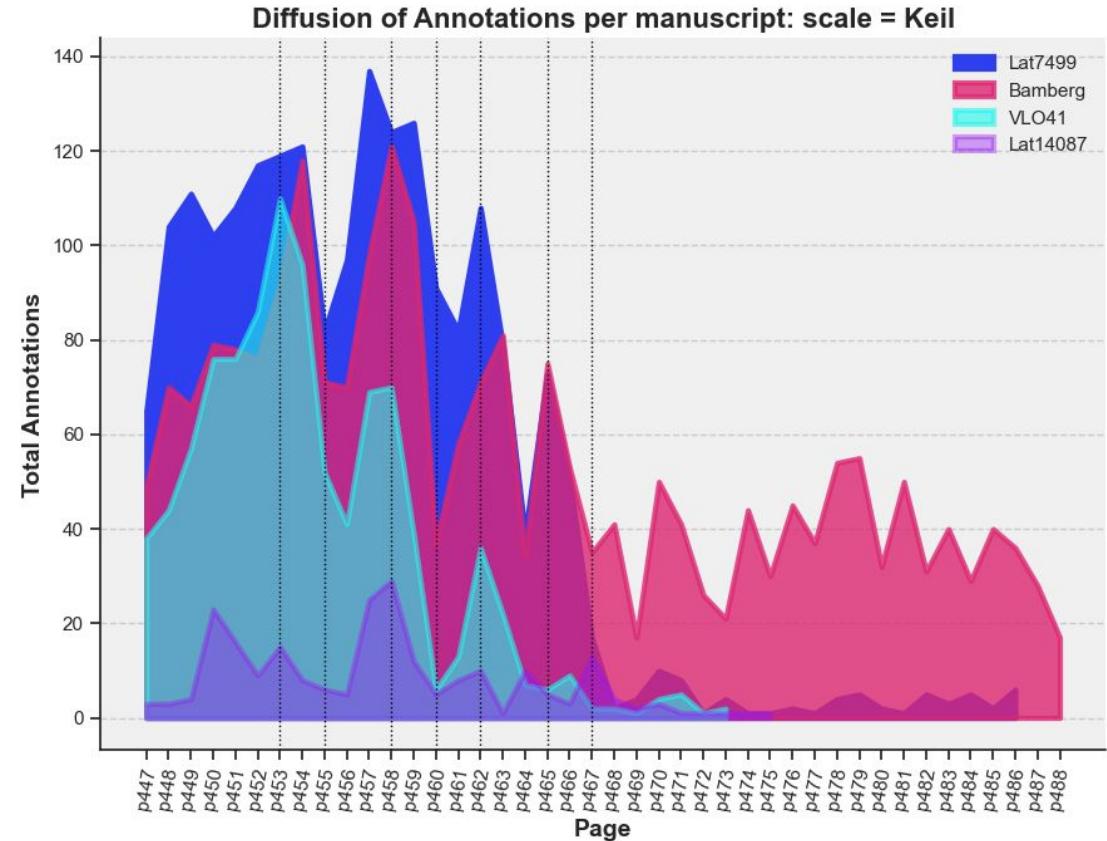
#### 1) Confirmation des hypothèses:

- Synonymes
  - Définition
  - Élucidation des ambiguïtés du texte
  - Étymologie
- ⇒ fonctions annotatives **universelles**

#### 2) Observations plus fines sur la double nature du lemme:

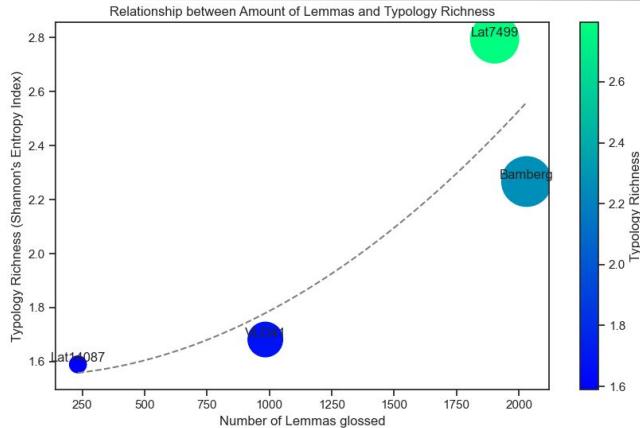
- gloses interlinéaires ⇒ **unité plutôt lexicale, considérée dans son contexte immédiat**
- marginalia ⇒ perspective plus large, englobant des aspects **sémantiques, métatextuels ou inter-textuels** du lemme.

# 3e: Mouvements d'annotation



- alignement de la quantité d'annotations des témoins et identification des convergences et des anomalies par le biais de l'indexation des lemmes:
- Six moments distincts qui déterminent le rythme de l'activité d'annotation - quatre moments de croissance et deux moments de décroissance de l'intensité.
- Quantité variable mais uniforme pour tous les témoins et finit par cesser à la fin du premier livre pour 3 sur 5 témoins;
- Seul *BambergMsc30* se démarque et poursuit l'annotation jusqu'à la fin (fond/source/famille différente ou contribution originale?).

# 3f: Variété et Originalité

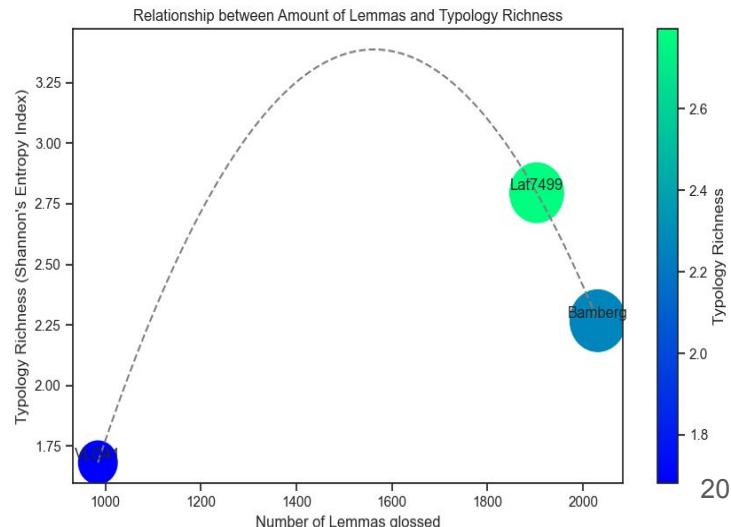


- Évaluer le «**degré d'originalité**» des témoins en étudiant **la variété des typologies** par rapport au **nombre des annotations**;
- Est-ce que la variété est **directement proportionnelle à la taille**?
- Déterminer les témoins dont les annotations apportent une plus riche variété dans le but de distinguer entre les traditions de la glose anonyme et le commentaire de RA?

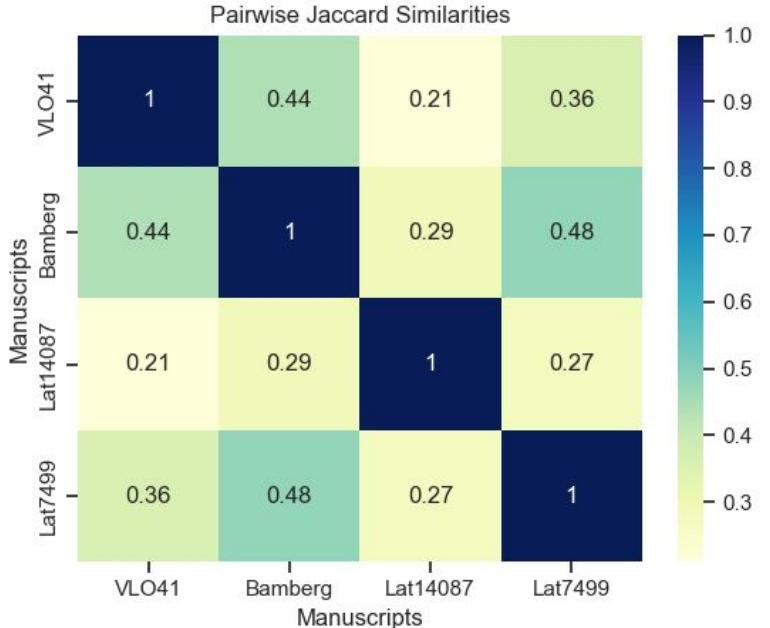
## Pourquoi l'index de Shannon?

$$H' = - \sum_{i=1}^S p_i \log_2 p_i$$

- mesure de l'hétérogénéité de la diversité d'un milieu d'étude ( $H'$ )
  - peu de données - plus simple à interpréter
  - pénalisation des probabilités très élevées eg. pour les typologies comme S22** (grâce à  $\log_2$ )
- ⇒ *Lat7499 se distingue* bien des autres en termes de variété de typologies



# 3g - Proximité relative des témoins



## Comment définir la proximité des témoins glosés?

Inspirés par l'approche de Steinovà[1], nous calculons les mesures suivantes **de manière personnalisée**, en prenant en compte les **spécificités** des données analysées et en évitant que la taille des témoins n'influence les résultats :

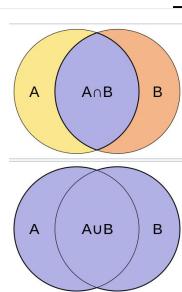
- Nombre absolu de gloses partagés (intersection)
- Nombre de gloses « identiques » partagés (union)

⇒ et nous appliquons ça à toutes les paires de témoins.

Pourquoi l'indice de Jaccard?

-Adapté à la gestion de **2 à n ensembles** de données

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



⇒ *Lat7499 et BambergMsc30 portent des gloses interlinéaires identiques selon Colette Jeudy.*

[1] Steinova, E., & Boot, P. (2021). The glosses to the first book of the *Etymologiae* of Isidore of Seville: a digital scholarly edition.

# Apport du projet - Conclusions

La contribution de ce travail se résume ainsi:

- Proposition d'un **pipeline numérique semi-automatique** offrant une **flexibilité** dans le traitement et l'exploitation des données.
- **Création d'un dataset original en libre accès** pour l'entraînement de modèles de segmentation de mise en page et de reconnaissance de caractères.
- Mise à disposition de modèles de segmentation sémantique et de détection/classification des baselines.
- **Editions documentaires enrichies** de 3 manuscrits glosés avec indexation des lemmes et descripteurs qualitatifs
- Élaboration d'un **modèle conceptuel rudimentaire** de la tradition glosée.
- **Lecture distante** des témoins avec des observations quantitatives préliminaires.

## Conclusions

- ❖ Bien que l'étude des manuscrits ne peut pas être réduite en lecture distante, celle-ci offre des **renseignements précieux** pour confirmer des hypothèses et **guider la lecture proche** (collation, édition etc.).
- ❖ Travailler avec des manuscrits glosés dans un environnement numérique **améliore à la fois la paléographie et aux outils numériques existants**, en repoussant les limites des outils actuels et en facilitant le travail des paléographes.
- ❖ Une **attention particulière** doit être accordée à la **méthodologie de recherche**, à la **transparence** des démarches et à la formulation des **questions de recherche pertinentes**. afin de garantir la **rigueur scientifique** des résultats et la validité des interprétations.

# Bibliographie

Colette, Jeudy (1974). « Les manuscrits de l'Ars de uerbo d'Eutychès et le commentaire de Rémi d'Auxerre », *Mélanges E. R. Labande Etudes de civilisation médiévale* (IXe - XIIème s.). Poitiers, 421 et 426-436.

Chagué, Alix, Clérice, Thibault et Romary, Laurent. « HTR-United : Mutualisons la vérité de terrain ! » (, 2021)

Cinato, F. (2015). *Priscien glosé* (Vol. 41, pp. 753-p). Brepols.

Gabay, Simon, Camps, Jean-Baptiste, Pinche, Ariane) et Jahan, Claire, (2021). « SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more) », dans 16th International Conference on Document Analysis and Recognition (ICDAR 2021).

Holtz, Louis (1978). « La typologie des manuscrits grammaticaux latins », *Revue d'histoire des textes*, 7-1977, p. 247-269.

- (1984). « Les manuscrits latins à gloses et à commentaires : de l'antiquité à l'époque carolingienne », dans *Il Libro e il testo*, dir. R. Raffaelli C. Questa, 1984, p. 139-167.

Irvine, M. (1994). *The Making of Textual Culture: Grammatica and Literary Theory*.

Keil, Henricus, (1857). *Grammatici latini ex recensione Henrici Keili.* - Lipsiae, BG Teubner 1857-1880, t. 5

Kiessling, Benjamin, Tissot, Robin, Stokes, Peter et Ezra, Daniel Stökl Ben, (2019) « eScriptorium : An open source platform for historical document analysis », dans 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), IEEE, t. 2, p. 19-19.

Law, Vivien A, (1986). « Late Latin grammars in the Early Middle Ages : a typological history », *Historiographia linguistica*, 13-2-3 , p. 365-380.

Lomanto, Valeria (1985), « Eutiche », dans *Enciclopedia virgiliana*

O'Sullivan, Sinéad (2010). *Glossae Aevi Carolini in libros I – II Martiani Capellae De Nuptiis Philologiae et Mercurii*, Turnhout, Brepols.

Pierazzo, Elena et Stokes, Peter A, (2011). « Putting the text back into context : a codicological approach to manuscript transcription », dans *Kodikologie und Paläographie im digitalen Zeitalter 2–Codicology and Palaeography in the Digital Age* 2, p. 397-430.

Steinovà, Evina et Boot, Peter, (2021). « The glosses to the first book of the *Etymologiae* of Isidore of Seville : a digital scholarly edition ». <https://db.innovatingknowledge.nl/edition/#right-network>

Teeuwen, Mariken (2011). “Marginal Scholarship: Rethinking the Function of Latin Gloses in Early Medieval Manuscripts,” in *Rethinking and Recontextualizing Gloses*, ed. Lendinara, Lazzari, and di Sciacca, pp. 19-37

Zetzel, James E. G. (2005) Marginal Scholarship and Textual Deviance: The “Commentum Cornuti” and the Early Scholia on Persius, *Bulletin of the Institute of Classical Studies Supplement* 84, London.

# Merci pour votre attention

**(et pour votre aide précieuse tout au long de cette année!)**

**εὐ·τυχής, ἡς, ἡς [v]**

1 *pass.* heureux, qui prospère, qui réussit, *différ.* *de ὥλθιος* (heureux par la richesse) HDT. 1, 32 ; *de εὐδαίμων* EUR. *Med.* 1229 ; *avec un inf.* εὐ. *στρατηγεῖν*, PLUT. *Ant.* 34, général heureux ; *en parl.* *d'événements, de choses, en gén.* (*sort, ESCHL. Pers.* 709 ; *action, SOPH. Tr.* 293, *etc.*) ; τὸ εὐτυχές, THC. 2, 44, *c.* εὐτυχία ;

2 *act.* qui assure le bonheur de : εὐ. *ἴκοιτο*, SOPH. *O.C.* 308, qu'il vienne pour le bonheur de, *etc.*

## Métriques YALTAi: segmentation

Precision-Recall curve : all classes **0.855 mAP@0.5**

Precision-Confidence curve : all classes **1.00 at 0.896**

Recall-Confidence Curve : all classes **0.92 at 0.000**

F1- Confidence curve : all classes **0.86 at 0.353**

La courbe de précision-rappel révèle un score moyen de précision de 0.855 avec un seuil de confiance de 0.5 (mAP@0.5), indiquant une **capacité élevée du modèle à détecter avec précision les objets recherchés tout en maintenant un bon équilibre avec le rappel**. La courbe de précision-confiance montre une précision parfaite (1.00) pour toutes les classes à un seuil de confiance de 0.896, ce qui souligne la capacité du modèle à faire des prédictions précises lorsque le seuil de confiance est élevé. La courbe de rappel-confiance montre un rappel élevé de 0.92 pour toutes les classes à un seuil de confiance très bas (0.000), ce qui signifie que **le modèle est capable de récupérer la plupart des objets positifs avec une confiance minimale requise**. Enfin, la courbe F1-confiance présente un score F1 de 0.86 à un seuil de confiance de 0.353, ce qui indique un **équilibre raisonnable (tradeoff) entre la précision et le rappel**. Dans l'ensemble, ces performances témoignent de l'efficacité du modèle dans la détection précise des objets cibles, tout en maintenant un bon équilibre entre la précision et le rappel.

## kraken

val\_accuracy: 0.997 **Cela indique que le modèle a une précision globale de 99.7%, ce qui signifie qu'il est capable de prédire avec une grande exactitude les étiquettes des objets détectés et classifiés.**

val\_mean\_acc: 0.997 **Ce score représente la précision moyenne du modèle pour la classification des différentes classes. Une valeur de 0.997 indique une excellente précision moyenne, ce qui suggère que le modèle est capable de bien distinguer les différentes catégories d'objets.**

val\_mean\_iu: 0.375 **Cela mesure la qualité de la segmentation des objets détectés par rapport à leur vérité terrain. Un score de 0.375 indique que la segmentation n'est pas très précise, car l'intersection entre les prédictions du modèle et les vérités terrain est relativement faible par rapport à leur union.**

val\_freq\_iu: 0.467 **Ce score mesure la fréquence à laquelle la segmentation prédite par le modèle correspond à la vérité terrain. Une valeur de 0.467 indique que le modèle a une fréquence de correspondance relativement modérée, ce qui signifie qu'il parvient à bien segmenter certains objets, mais pas tous.**

val\_metric: 0.375 **Un score de 0.375 indique que le modèle obtient des résultats relativement modestes en termes de précision et de segmentation.**