Bonjour à tous,

1)C'est avec grand plaisir que je vous présente aujourd'hui mon projet de fin d'études pour la validation du Master 2 en Humanités Numériques à l'école des chartes. Ce projet -comme indique le titre- porte sur le traitement numérique des manuscrits grammaticaux glosés, en mettant en lumière la tradition du "de uerbo" d'Eutychès, grammairien et disciple de Priscien de Césarée, qui enseignait le latin à Constantinople eu 6^{e} siècle. Au cours de cette présentation, je vous expliquerai en détail ma méthodologie, les outils et les ressources utilisés, ainsi que les résultats préliminaires obtenus jusqu'à présent.

2) Tout d'abord, il convient de comprendre pourquoi ce projet revêt une pertinence particulière pour un mémoire en Humanités Numériques. En effet, il existe de nombreuses difficultés inhérentes liées à la nature des manuscrits glosés, qui posent des problèmes tant aux paléographes/philologues qu'aux spécialistes en HTR et en vision par ordinateur. Plus précisément, déjà déchiffrer et éditer ces manuscrits demande un travail minutieux et chronophage, avec la difficulté supplémentaire de modéliser, analyser à grande échelle, comparer et visualiser les multiples niveaux d'information qu'ils contiennent. De plus, il n'existe actuellement pas des modèles de segmentation capable de gérer cette complexité, ni de cadre unifié pour leur traitement (même s'il existe des projets numériques isolés). En outre, pour justifier le choix d'auteur il est important de souligner que la tradition manuscrite glosée d'Eutychès est largement inédite malgré la littérature étendue.

3) Cela fait de notre sujet un véritable défi, mais aussi une bonne opportunité pour explorer les outils numériques capables à gérer et notamment à mettre en valeur la nature complexe des manuscrits glosés comme il est évident dans cette petite anatomie des éléments structuraux essentiels des manuscrits glosés.

4) En ce qui concerne notre méthodologie, ce projet repose sur trois volets distincts qui composent un flux semi-automatique allant de l'acquisition à l' analyse fe données. En même temps, chacun de ces volets dispose d'une autonomie relative afin de tirer pleinement parti des résultats obtenus à chaque étape, tout en respectant les principes de réutilisabilité, d'accessibilité et d'interopérabilité des données.

5) Notre corpus se compose de trois manuscrits glosés, chacun présentant des degrés de glose différents dont un qui contient le "commentaire marginal de Rémi d'Auxerre, ainsi qu'un glossaire. Ces manuscrits ont été tous copiés entre le neuvième et le deuxième siècle, dans des scriptoria situés dans le Nord de la France associés à l'activité carolingienne, formant ainsi, ce que nous considérons un groupe intellectuellement cohérent.

6)Passons au pipeline. La première étape consiste à l'acquisition de données, qui commence par la segmentation de nos témoins glosés. Pour ce faire, nous avons utilisé la plateforme eScriptorium et le vocabulaire contrôlé SegmOnto, spécifiquement conçu pour caractériser la mise en page de manuscrits, afin d'effectuer une annotation

manuelle de tous les registres d'information que nous souhaitons exploiter. L'objectif de cette opération était d'entraîner un modèle de segmentation sémantique performant, capable de différencier la zone principale des annotations en marge, ainsi que les lignes principales des gloses interlinéaires, comme illustré dans l'exemple. Étant donné le manque de données disponibles, nous avons enrichi notre jeu de données en adaptant le dataset de l'équipe DIVA à notre vocabulaire. Vous pouvez avoir un aperçu de la taille du jeu de données pour chaque classe dans le tableau en bas.

8) Afin d'atteindre cet objectif, nous avons procédé à l'entraînement d'un modèle en utilisant l'outil YALTAi, en suivant les directives de son développeur, Thibault Clérice. YALTAi utilise une approche de "détection d'objets" en combinant l'architecture YOLOv5 pour la détection des zones et les modalités déjà existantes de kraken pour la classification des lignes. Lors de cet entraînement, nous avons observé des performances globalement élevées pour les zones d'intérêt, à l'exception des zones marginales qui sont souvent confondues en tant qu'arrière fond et sont ignorées, comme le suggère la matrice de confusion.

9) En ce qui concerne la classification des lignes, nous avons entraîné deux modèles distincts, suite à la suggestion de M. Vidal Gorène, dans le but de réduire la complexité de la tâche mais aussi de pouvoir extraire les deux registres d'information indépendamment l'un de l'autre. Malgré certaines limitations telles que la spécificité de la mise en page et l'occurrence occasionnelle de chevauchements entre les types de lignes, cela ne diminue pas l'efficacité qualitative globale de notre approche, qui permet une réduction significative du temps nécessaire à la post-correction et à la segmentation des manuscrits par rapport à l'approche manuelle.

10) Après la segmentation, nous passons à la prochaine étape, afin de compléterla description du premier volet de notre projet : la Reconnaissance de l'Écriture Manuscrite (HTR), plus précisément la reconnaissance de la minuscule caroline. Pour ce faire, nous avons utilisé un modèle préexistant fourni par l'équipe Rescribe et Thibault Clérice, spécifiquement entraîné pour la minuscule caroline que nous avons affiné (finetuné) en fournissant la transcription manuelle de notre premier manuscrit, le Vossianus Latinus 41. Les résultats sont globalement satisfaisants pour les lignes principales, bien que la résolution et/ou le petit module des gloses rendent la reconnaissance plus difficile dans certains cas. Finalement, pour notre transcription, nous avons suivi une approche graphématique, qui reproduit les abréviations du manuscrit et préserve les variantes orthographiques de nos témoins.

12) Passons maintenant au deuxième volet de notre projet, qui concerne la Transformation et la Structuration de nos données sérialisées. Nous avons la possibilité, depuis eScriptorium, d'extraire les données directement au format XML ALTO (parce que à cause du sens de lecture non linéaire, le format txt est insiffusant), que nous transformons ensuite à l'aide d'une feuille de style XSLT en format XML-TEI, qui est le standard d'encodage utilisé pour les éditions numériques. Cette approche nous permet de préserver toutes les informations indispensables de la mise en page et de gérer

simultanément tous les registres d'information, tout en offrant une certaine flexibilité lors de la transformation.

13) Ensuite, afin d'analyser le contenu de nos manuscrits glosés, il était essentiel d'enrichir l'encodage de notre structure XML TEI avec des informations clés pour en faire une édition documentaire. Cela inclut le couple indissociable lemme-glose et/ou les annotations marginales, qui sont au cœur de notre recherche, ainsi que des descripteurs sémantiques tels que la typologie des annotations et leur taille relative, descripteurs empruntés à l'étude de Franck Cinato sur Priscien glosé. De plus, nous avons veillé à caractériser les différentes mains lorsque cela était nécessaire, comme nous le ferions dans le cas d'une édition critique traditionnelle. Pour les lemmes, il est important de mentionner que nous les avons alignés par rapport à l'édition de Keil, de façon qu'ils possèdent d'une clé unique d'indexation, qui correspond, pour tous nos témoins, à un point de référence stable.

14) Finalement, nous présentons ici certains des résultats préliminaires de notre étude, que nous considérons les plus pertinents, accompagnés d'une brève interprétation. Comment peut-on analyser ces données avec la lecture distante?

15) Tout d'abord, nous obtenons une vue d'ensemble de nos données, avec des descriptions précises concernant le nombre de lemmes, de gloses et d'annotations marginales pour chaque témoin, ainsi que les descripteurs les plus fréquents. D'emblée, nous constatons que le manuscrit de la BnF se démarque avec le plus grand nombre de lemmes uniques (absents de l'édition de Keil), de gloses, d'annotations marginales étendues (F5) et de typologies uniques.

16)Pour une exploitation pratique initiale, nous mettons en valeur le choix d'indexer nos lemmes/gloses par rapport à l'édition de Keil (pratique bien établie pour les gloses), ce qui nous permet de constituer instantanément une sorte de "base de données" à partir de laquelle nous pouvons accéder et filtrer les entrées selon les descripteurs souhaités. Cette approche facilite considérablement le processus de comparaison et de collation, qui est généralement très chronophage, en accélérant et en simplifiant les étapes requises. Voici un exemple des lemmes avec leurs gloses en notes tironiennes dans le manuscrit de Bamberg, comparés à leurs équivalents en latin dans le manuscrit de la BnF.

17) Grâce à l'annotation de plusieurs mains, ici pour le VLO14, un autre élément que nous sommes en mesure de quantifier est leur contribution au sein du manuscrit. Cela nous permet de visualiser concrètement la répartition de l'activité des glosateurs et de mettre en évidence les campagnes d'annotation non contemporaines. De plus, en étudiant les typologies privilégiées par chaque main, leur rôle devient souvent plus clair. À titre d'exemple, nous pouvons noter l'activité brève de la main "C" au début du livre (en vert) qui, après une lecture proche, semble avoir utilisé le manuscrit à des fins d'enseignement, d'où l'activité aussi restreinte.

18) Une question qui est à la fois fondamentale et très basique concerne la distinction des fonctions entre les gloses et les annotations marginales. En analysant la fréquence des différentes typologies pour chaque type d'annotation et en identifiant les 9 typologies les plus caractéristiques, nous sommes en mesure de définir plus précisément leurs fonctions respectives. En effet, même s'il existe un ensemble de typologies "génériques" telles que les synonymes, les définitions et les étymologies, cette distinction nous permet de mettre en évidence que pour les gloses, le lemme est considéré principalement comme une unité lexicale dans son contexte, tandis que pour les annotations marginales, il est perçu comme une unité méta-textuelle. Ces deux types d'annotations se complètent de manière dynamique, étant donné que souvent l'une poursuit l'autre.

19) Une autre application essentielle de la lecture distante dans le cadre d'une étude comparative consiste à mettre sur un pied d'égalité (au moins spatiale), l'activité d'annotation de tous nos témoins grâce à l'indexation, afin de mettre en évidence les points de convergence et de divergence de leur « comportement ». Dans notre cas, pour la partie couvrant le premier livre du « de uerbo », nous observons six moments distincts où les manuscrits suivent le même rythme (de croissance et de décroissance), tandis que seul le manuscrit de Bamberg se démarque en poursuivant l'annotation jusqu'à la fin.

20) Avant de conclure, nous avons mené une analyse expérimentale visant à identifier les témoins qui se distinguent par leur contenu plus "original" par rapport aux autres. Étant donné la présence d'un commentaire dans les marges du Lat7499, nous nous attendions naturellement à ce qu'un commentaire, étant essentiellement une œuvre distincte, présente une plus grande variété dans ses annotations. Nous avons pu confirmer cette hypothèse en appliquant l'indice de Shannon à nos témoins. En réponse à la question de savoir si la variabilité est directement proportionnelle à la taille de l'échantillon, ce qui est souvent notre première intuition, nous pouvons répondre par la négative.

21) La deuxième et dernière analyse expérimentale consiste à établir un indice de proximité entre nos témoins, formant ainsi un réseau ou famille d'influences. Inspirés par l'approche d'Evina Steinovà, qui avait déterminé manuellement la proximité des témoins en fonction des clusters de gloses communes pondérées par leur poids, nous adaptons la formule en utilisant l'indice de Jaccard. Nous avons adapté manuellement les metriques, définissant l'union comme le nombre de gloses "identiques" entre deux témoins et l'intersection qui représente le nombre total de lemmes partagés. Les résultats obtenus ne sont pas concluants pour le moment, nécessitant une étude plus approfondie, à l'exception de la paire la plus proche qui est raccord avec la littérature.

22) Nous souhaitons conclure avec un petit récapitulatif de la contribution de ce projet de fin d'études ainsi que quelques conclusions et observations plus générales : En première lieu, l'exploitation numérique des manuscrits glosés présente des avantages tant pour le domaine de la paléographie que pour les outils numériques existants, repoussant les limites des technologies actuelles tout en facilitant le travail des paléographes. En deuxième lieu, la lecture distante des manuscrits, bien qu'elle ne

puisse pas être considérée comme une approche exhaustive mais plutôt complémentaire, offre des informations précieuses pour valider des hypothèses et orienter la lecture proche, notamment en matière de collation et d'analyse du contenu. Finalement, il est essentiel d'accorder une attention particulière à la méthodologie de recherche, à la transparence des procédures et à la formulation de questions de recherche pertinentes afin de garantir la rigueur scientifique des résultats obtenus et la validité des interprétations.