LeJournal de l'UNIGE

5 mars 2026 - Marion de Vevey

Vie de l'UNIGE

Des milliers de textes de manuscrits médiévaux déchiffrés grâce à l'IA

À l'occasion du Digital Humanities Day, Simon Gabay, chercheur en humanités numériques, présentera CoMMA, un nouveau méga corpus de manuscrits médiévaux analysés grâce à l'intelligence artificielle. Ce projet ouvre de nouvelles perspectives pour l'étude de la langue, des textes et du livre au Moyen Âge.

Giovanni Boccaccio, Des cas des nobles hommes et femmes. Parchment · (I-III) + 179 folios + (IV-VI) ff. · 40.5 x 29.5 cm · France (Paris) · ca. 1410. Bibliothèque de Genève

Quand le mot «amour»est-il apparu pour la première fois à l'écrit? Sa fréquence d'emploi a-t-elle évolué au fil du temps? Pour répondre à ces questions, il faudrait pouvoir lire tous les manuscrits médiévaux, ce qui semble impossible étant donné l'ampleur des documents existants. C'est là qu'intervient CoMMA, un méga corpus de manuscrits médiévaux exploitant l'intelligence artificielle. «C'est un changement épistémologique obtenu par le croisement de plusieurs

disciplines», explique Simon Gabay, maître-assistant à la chaire des humanités numériques. Des disciplines, allant de la philologie médiévale aux sciences computationnelles, qui s'entremêlent pour poser de nouvelles questions dont les réponses étaient autrefois inaccessibles.

Un méga corpus de manuscrits

La plateforme CoMMA, publiée fin 2025, propose des dizaines de milliers de manuscrits datant de 800 à 1600. Sa force réside dans l'extraction d'information, explique le médiéviste: «Depuis quelques dizaines d'années, un important travail de numérisation a eu lieu dans différents pays. Des manuscrits ont été pris en photo, mais le problème du format image est qu'on ne peut pas y chercher de mots spécifiques.» Pour construire cette plateforme, l'extraction et la restructuration d'informations ont été effectuées grâce à une succession de modèles d'intelligence artificielle afin de reconstituer le document original dans un format numérique. On peut donc déchiffrer un manuscrit et travailler directement sur son texte. Et ce ne sont pas seulement les chercheurs, chercheuses et institutions patrimoniales qui peuvent bénéficier de CoMMA: n'importe quel-le passionné-e du Moyen Âge peut parcourir les textes sans besoin de formation préalable, grâce à l'accessibilité gratuite de la plateforme.

Évolution de la mise en page

Simon Gabay a déjà pu utiliser ce grand corpus dans le cadre d'un article sur l'évolution de la mise en page des manuscrits, pour l'instant en prépublication. «Aujourd'hui, lorsqu'on voit une page, c'est évident, nous lisons de haut en bas, de gauche à droite, il y a des paragraphes, des notes en bas, parfois un titre courant. Mais cela n'a pas toujours été le cas. Toutes ces informations sont absolument fondamentales pour étudier la manière dont se construit l'objet livre.» Or, observer cette construction n'est possible que grâce à l'intelligence artificielle. «Nous ne pouvons pas savoir comment évolue dans le temps long l'organisation de la page si nous n'avons pas regardé un grand nombre de manuscrits. C'est une chose que nous ne pouvions pas faire jusqu'à présent», s'enthousiasme le chercheur.

Grâce aux manuscrits recensés dans CoMMA, il a ainsi pu observer une phase de complexification de la mise en page au fil du Moyen Âge, avec une augmentation du nombre de zones d'écriture et d'abréviations. Ce, jusqu'au XIV^{e} siècle, date à laquelle les manuscrits se sont à nouveau simplifiés. Cette découverte est source de nouvelles questions que se pose le médiéviste: «Que s'est-il passé au XIV^{e} siècle pour que soit créée cette simplification? Est-ce que les manuscrits étaient devenus trop difficiles à lire? Ou est-ce la diminution des coûts matériels qui a amoindri le besoin de tasser autant l'information?»

Humain ou algorithme?

Des réponses peuvent donc surgir grâce à l'application de modèles d'intelligence artificielle sur ce vaste corpus. Mais ces nouvelles technologies soulèvent aussi de nouvelles interrogations: lorsqu'un mot est transcrit d'une façon inhabituelle par le logiciel, est-ce dû à une erreur du modèle IA, une faute d'orthographe ou une variante graphique, c'est-à-dire une autre manière d'écrire possiblement influencée par des habitudes? Du côté de l'humain, difficile de répondre, puisque «la notion de faute n'existe pas vraiment dans un monde sans règles», l'orthographe n'apparaissant qu'au milieu du XVIII^{e} siècle. Du côté des machines, «c'est tout l'enjeu de l'intelligence artificielle», explique le chercheur. Plus il y a de données d'entraînement, plus elle pourra reconnaître les différentes normes existantes. Une intelligence artificielle qui n'est d'ailleurs ni intelligente, ni artificielle. «Le modèle n'est pas intelligent parce qu'il ne fait que répéter ce qu'il a vu. Il n'est pas artificiel non plus car il se base sur des exemples fournis par des humains», précise le chercheur, qui ajoute que ces modèles sont le fruit de plusieurs années de travail à travers l'Europe. De grandes campagnes de numérisation en France, en Suisse et en Belgique ont été suivies par des dizaines de chercheurs et chercheuses ayant annoté des manuscrits datant de tout le Moyen Âge afin de «nourrir» les modèles. À cela s'ajoute une collaboration avec Inria Paris et le CNRS.

L'IA ne remplace pas le toucher

Malgré ce travail numérique, pas question de rester toute la journée devant son ordinateur, raconte le sourire aux lèvres le passionné de la période médiévale: «Il reste primordial d'aller voir physiquement les manuscrits. D'abord parce que, personnellement, j'ai fait ce métier pour ça. Et ensuite, parce qu'aucune reproduction n'égale l'original. Le toucher du manuscrit, son épaisseur, la manière dont les pages ont été rassemblées les unes avec les autres… tous ces détails permettent de prendre vraiment conscience de l'objet.»