Océriser les imprimés du XVIe siècle en langue française

Sonia Solfrini, Simon Gabay, Maxime Humeau, Ariane Pinche, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, Geneviève Gross, Daniela Solfaroli Camillocci

Océriser les imprimés du XVI[e] siècle en langue française : le cas d'un corpus romand en caractères gothiques

Sonia Solfrini[1], **Simon Gabay**[1], **Maxime Humeau**[1], **Ariane Pinche**[2], **Pierre-Olivier Beaulnes**[1], **Aurélia Marques Oliveira**[1], **Geneviève Gross**[1] et **Daniela Solfaroli Camillocci**[1]

[1]Université de Genève
{prenom.nom}@unige.ch
[2]CIHAM–UMR 5648, C.N.R.S., Lyon
{prenom.nom}@cnrs.fr

Résumé

Depuis quelques années, la philologie computationnelle a ouvert la voie à de nouvelles approches pour l'étude des textes médiévaux et modernes. Ces approches nécessitent cependant des données en grande quantité que l'on ne peut obtenir qu'en extrayant les textes à partir des fac-similés numériques. Pour ce faire, la recherche a besoin d'outils efficaces, s'appuyant sur des guides qui garantissent une interopérabilité maximale entre les différents états d'une langue (ancien français, moyen français, etc.) et les différents types de textes (manuscrits, imprimés, etc.). Cet article se concentre sur la production imprimée du XVI[e] siècle, en langue française et en caractères gothiques, en prenant pour cas d'étude un corpus romand. Nous proposons deux modèles qui améliorent l'état de l'art actuel : l'un pour l'analyse de la mise en page et l'autre pour l'OCR. Ces modèles s'appuient sur un vocabulaire contrôlé pour la description des pages et sur un guide de transcription pour les textes en gothique.

1 Introduction

Si la plupart des imprimeurs abandonnent le gothique pour le romain et l'italique aux alentours des années 1530-1540, la typographie gothique persiste dans certaines publications du XVI[e] siècle (Pouspin, 2016). Les imprimés de la Renaissance présentent ainsi au moins trois types d'écriture, en plus d'un état relativement ancien de la langue (Catach, 1968; Vachon, 2010). Leur traitement informatique requiert donc des outils spécifiques, notamment pour l'analyse de la mise en page et la reconnaissance optique de caractères (OCR).

2 État de l'art

À partir des années 2000, le programme des Bibliothèques Virtuelles Humanistes (BVH) du CESR de Tours (Uetani et al., 2016) s'est imposé comme le projet de référence pour la diffusion de documents patrimoniaux français du XVI[e] siècle, avec notamment la base textuelle *Epistemon* [1].

L'apparition de modèles d'OCR a rapidement intéressé les spécialistes des documents historiques et des premiers modèles sont apparus, notamment pour l'allemand et le latin (Springmann et Lüdeling, 2017), accompagnés de données d'entraînement en *Frakturschrift* comme en *antiqua* (par ex. *GT4HistOCR*, cf. Springmann et al. 2018). Pour le français, le projet Gallic(orpor)a (Sagot et al., 2023) a mis à disposition des modèles d'OCR et d'HTR : le modèle Gallicorpora+ (Pinche et Gabay, 2022) pour les imprimés français en romain, et le modèle Cortado 2.0.0 (Pinche et Clérice, 2022) pour les manuscrits médiévaux, les incunables et, plus marginalement, les imprimés en gothique du XVI[e] siècle. Ce dernier modèle a connu tout récemment une révision importante, orientée vers les données médiévales, publiée sous le nom de CATMuS Medieval 1.0.0 (Pinche et al., 2023).

La production des transcriptions par l'OCR requiert aussi l'analyse de la mise en page pour en détecter les différentes lignes et zones (titre courant, pagination, corps du texte, etc.). Cette approche multimodale permet de conserver le sémantisme de l'information graphique, et ainsi de faciliter le processus de structuration des informations après la phase d'océrisation (Ramel et al., 2013). Le projet Gallic(orpor)a a mis a disposition un premier modèle [2], qui s'appuie sur un vocabulaire contrôlé pour la description standardisée des pages, SegmOnto [3] (Gabay et al., 2023b), et la ver-

1. https://www.bvh.univ-tours.fr/Epistemon/index.asp.
2. https://github.com/Gallicorpora/Segmentation-and-HTR-Models/releases.
3. https://segmonto.github.io.

sion 0.0.1 de YALTAi (Clérice, 2023), qui utilise la v5x de YOLO, pour l'entraînement du modèle. Lors de tests effectués dans le cadre du projet Gallic(orpor)a, la détection *feature-based* utilisée par YOLO a en effet démontré de meilleures performances que la détection *pixel-based* offerte par Kraken (Kiessling, 2020) pour l'analyse de la mise en page. La famille des modèles YOLO (Redmon et al., 2016) permet en effet une classification de premier niveau très efficace, qui s'adapte parfaitement à la typologie proposée par SegmOnto.

Le cas des imprimés en gothique du XVI^e^ siècle reste donc imparfaitement abordé, tant du point de vue des données que des modèles disponibles, et les outils ont rapidement évolué depuis. Notre projet vise à combler ce vide, pouvant intéresser tant les historiens que les philologues travaillant sur la Renaissance, ou encore les spécialistes de l'histoire de la Réforme.

3 Le corpus du projet SETAF

Le volet numérique du projet SETAF[4] (Solfrini et al., 2023b) vise à créer un corpus d'imprimés évangéliques romands à l'époque de la Réforme, comprenant les textes écrits par Guillaume Farel (1489-†1565) et son cercle. Le corpus primaire comprend les ouvrages publiés par les imprimeurs Pierre de Vingle et Jean Michel, actifs respectivement dans les années 1525-1535 et 1538-1545. Parmi les traits distinctifs de ces textes, nous trouvons la typographie gothique et un état de la langue relativement ancien, le moyen français, avec de nombreuses abréviations et une certaine volatilité des systèmes graphiques.

Dépôt GitHub	Textes	Pages	Lignes
SETAF-Pierre-de-Vingle	11	895	24 701
SETAF-Jean-Michel	5	404	11 778
SETAF-LesFaictzJCH	2	144	4 765
Total	18	1 443	41 244

TABLEAU 1 – Le nombre de textes, de pages et de lignes traités par l'équipe SETAF et leurs dépôts GitHub au 01-01-2024. Ces chiffres sont amenés à augmenter.

À partir des travaux de Kemp (2004) et Berthoud (1980), des bibliographies telles que GLN 15-16[5] et de la littérature sur le « groupe de Neuchâtel » (par ex. Szczech 2021), c'est-à-dire le groupe autour de Guillaume Farel, l'équipe SETAF a établi une liste d'une cinquantaine d'ouvrages. Cette liste comprend 27 textes publiés par Vingle et 33 par Michel, dont certains ne nous sont pas parvenus (quatre pour Vingle et deux pour Michel). Le tableau 1 montre le nombre de textes intégralement océrisés et corrigés jusqu'à présent (§ 4.1), classés par imprimeur, avec en plus deux éditions des *Faits de Jésus Christ et du pape* (Bodenmann, 2009) qui tiennent une place centrale au sein du projet.

4 Deux nouveaux modèles

Nous avons entraîné un modèle d'analyse de la mise en page (§ 4.2) et un modèle de reconnaissance optique de caractères gothiques (§ 4.3).

4.1 Préparation des données

Les données SETAF sont préparées grâce à l'instance genevoise[6] (Gabay et al., 2021-) d'*eScriptorium* (Kiessling et al., 2019), qui utilise le moteur d'OCR Kraken (Kiessling, 2019), et sont cataloguées sur HTR-United[7] (Chagué et Clérice, 2023), qui fournit des outils d'intégration continue (Chagué et al., 2021; Clérice et al., 2023) assurant une meilleure qualité des données et de leur documentation. Les documents respectent les normes de SegmOnto (Gabay et al., 2023c) et de notre guide de transcription (Solfrini et al., 2023a). Concernant la transcription, chaque texte est « pré-transcrit » avec un modèle d'OCR (d'abord Cortado 2.0.0, ensuite des versions « fine-tunées » de CATMuS 1.0.0), puis intégralement corrigé à la main.

4.2 Modèle 1 : analyse de la mise en page

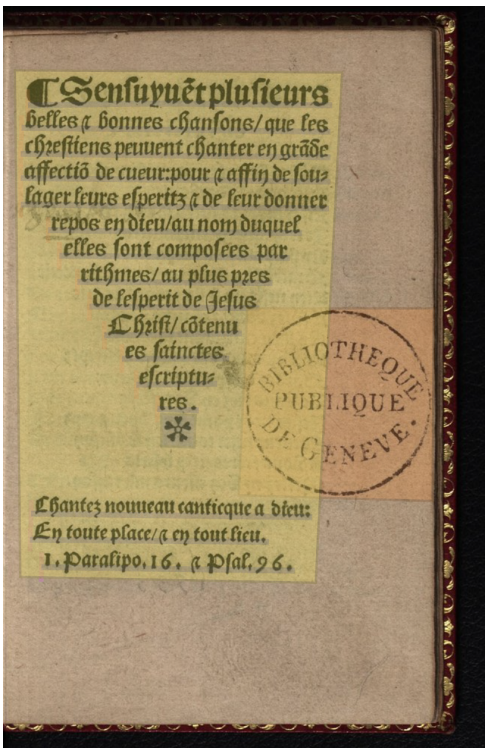
4.2.1 Règles d'annotation des images

Pour l'annotation des images, nous avons utilisé le vocabulaire contrôlé SegmOnto (Gabay et al., 2023b), en évitant au maximum l'utilisation de sous-types. Les principales zones retenues sont représentées dans la fig. 1.

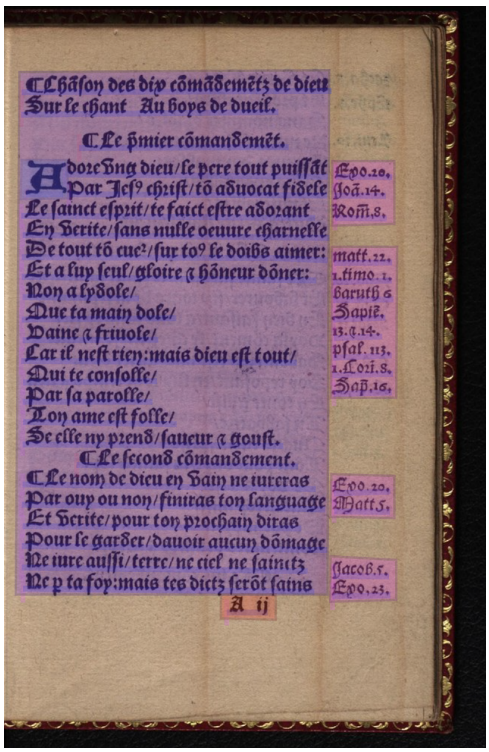
4.2.2 Jeux de données

Au-delà du corpus SETAF, grâce au choix de SegmOnto, nous avons pu rassembler d'autres jeux de données (cf. tab. 2), provenant principalement de trois projets (Gallic(orpor)a, CREMMA et FoNDUE) et tirés de documents :

4. https://www.unige.ch/setaf et https://github.com/SETAFDH.
5. https://www.ville-ge.ch/musinfo/bd/bge/gln.
6. https://www.unige.ch/lettres/humanites-numeriques/recherche/projets-de-la-chaire/fondue.
7. https://htr-united.github.io/.



(a) La page de titre est encodée avec `TitlePageZone`, les tampons de bibliothèque sont encodées avec `StampZone`, les illustrations et les décorations avec `GraphicZone`.



(b) Le corps du texte est encodé avec `MainZone`, la numérotation des cahiers avec `QuireMarksZone`, les lettrines avec `DropCapitalZone` et les notes marginales avec `MaginTextZone`.

FIGURE 1 – Principales règles de SegmOnto. Les deux pages d'exemple sont tirées de *Sensuyvent plusieurs belles et bonnes chansons*, 1533, [Pierre de Vingle], [Neuchâtel], 10.3931/e-rara-6934.

— de différentes époques ;
— de divers types (imprimé, manuscrit…) ;
— de plusieurs zones géographiques.

L'ajout de données supplémentaires, bien que différentes de celles de SETAF, devrait permettre d'améliorer l'efficacité de la détection des zones. Nous avons donc entraîné trois modèles différents pour tester l'impact de ces données hétérogènes et la capacité de SegmOnto à les homogénéiser :

— un modèle « gothique », entraîné uniquement sur les textes du XVI^{e} siècle en gothique du projet SETAF ;
— un modèle « moderne », entraîné sur les imprimés des XVI-XVIII^{e} s. ;
— un modèle « global », entraîné sur toutes les données disponibles.

Lors de l'entraînement, 10% des données sont réservées pour la validation, et trois jeux de test sont utilisés (cf. tab. 2) :

— les données de Gallic(orpor)a pour les imprimés en gothique du XVI^{e} s. ;
— les données tirées du dépôt `FONDUE-MLT-PRINT-TEST`, un jeu de données conçu pour tester des modèles avec dix pages par siècles, dans deux déclinaisons :
  — seulement les données des XVI-XVIII^{e} s. ;
  — les données des XVI-XXI^{e} s.

Il est important de noter que, tant en entraînement qu'en test, les zones sont très inégalement représentées (cf. tab. 11). Si les `MainZone` sont extrêmement fréquentes, les `QuireMarksZone` le sont par exemple beaucoup moins.

4.2.3 Métriques

Pour contrôler l'efficacité du modèle, nous utilisons cinq métriques :

— La « précision » évalue la capacité du modèle à détecter des zones, indiquant combien de zones détectées par le modèle étaient correctes par rapport à la « vérité de terrain » (*Ground Truth*, GT), c'est-à-dire les données annotées ou corrigées manuellement.
— Le « rappel » complète la précision et évalue la capacité du modèle à identifier toutes les zones présentes dans la GT.
— Le « score F1 » combine les mesures de la précision et du rappel.
— La « mAP » (*Mean Average Precision*) utilise l'intersection sur union (*Intersection over*

Projet	Type	Siècle	Pages	Zones	Jeu	Dépôt Github
Gallic(orpor)a	Manuscrit	XV^{e}	85	458	Train	HTR-MSS-15e-Siecle
Gallic(orpor)a	Incunable	XV^{e}	149	535	Train	HTR-incunable-15e-siecle
Sous-total (i)			**234**	**993**		
Gallic(orpor)a	Imprimé	XVI^{e}	80	233	Test	HTR-imprime-gothique-16e-siecle
SETAF	Imprimé	XVI^{e}	895	2 752	Train	HTR-SETAF-Pierre-de-Vingle
SETAF	Imprimé	XVI^{e}	404	1 365	Train	HTR-SETAF-Jean-Michel
SETAF	Imprimé	XVI^{e}	144	485	Train	HTR-SETAF-LesFaictzJCH
SETAF +	Imprimé	XVI^{e}	58	220	Train	HTR-Varia-Malingre-gothique
Sous-total (ii)			**1 581**	**5 055**		
SETAF +	Imprimé	XVI^{e}	202	1 062	Train	HTR-Varia-Malingre-romain
FoNDUE	Imprimé	XVI^{e}	223	688	Train	FONDUE-LA-PRINT-16
FoNDUE	Imprimé	XVI^{e}	930	2 829	Train	FONDUE-FR-PRINT-16
Gallic(orpor)a	Imprimé	XVI^{e}	180	591	Train	HTR-imprime-16e-siecle
Gallic(orpor)a	Imprimé	XVII^{e}	327	1 185	Train	HTR-imprime-17e-siecle
FoNDUE	Imprimé	XVII^{e}	69	246	Train	FONDUE-FR-PRINT-17
FoNDUE	Manuscrit	XVIII^{e}	153	460	Train	FONDUE-FR-MSS-18
Gallic(orpor)a	Imprimé	XVIII^{e}	160	624	Train	HTR-imprime-18e-siecle
Sous-total (iii)			**2 244**	**7 685**		
FoNDUE	Imprimé	XIX^{e}	48	129	Train	FONDUE-ES-PRINT-19
FoNDUE	Imprimé	XX^{e}	28	67	Train	FONDUE-IT-PRINT-20
FoNDUE	Imprimé	XX^{e}	30	72	Train	FONDUE-EN-PRINT-20
FoNDUE	Imprimé	XX^{e}	55	64	Train	FONDUE-FR-PRINT-20
-	Imprimé	XX^{e}	47	126	Train	HN2021-OCR-Poesie-Corse
Sous-total (iv)			**208**	**458**		
FoNDUE	Imprimé	XVI-XXI^{e}	60	197	Test	FONDUE-MLT-PRINT-TEST
Sous-total (v)			**60**	**197**		
Total			**4 327**	**14 388**		

TABLEAU 2 – Détail des données. Nous distinguons (i) les données médiévales, (ii) les données de la Renaissance en caractères gothiques (XVI^{e}), (iii) les données modernes en caractères romains (XVI^{e}-XVIII^{e} s.), (iv) les données contemporaines (XIX^{e}-XX^{e} s.), (v) les données de test provenant d'une sélection randomisée de Gallica et Persée.

Union, IoU), qui compare la superposition des zones détectées avec celles dans la GT. Un seuil est ensuite appliqué, au-delà duquel une prédiction est dite juste ou fausse. Plus le seuil est élevé, plus le test est difficile.

— La *mAP50* utilise un seuil de 0.50, qui donne de l'importance aux détections les plus simples ;

— La *mAP50-95* répète le test précédent en utilisant plusieurs seuils, de 0,50 à 0,95 avec un pas de 0,05. On augmente donc graduellement la difficulté du test, pour donner plus d'importance à des détections plus complexes.

4.2.4 Outils

Deux outils standards sont disponibles pour l'entraînement : Kraken (Kiessling, 2020) et YALTAi (Clérice, 2023). Des premiers essais (non-publiés) ont démontré la plus grande efficacité de YALTAi, mais son utilisation est moins commode car l'outil n'est pas intégré dans *eScriptorium*. Nous avons donc décidé d'entraîner deux modèles, le modèle entraîné avec Kraken étant distribué en supplément.

La récente publication de la v. 1.0.0 de YALTAi, qui convertit les données ALTO en données YOLO, nous permet l'utilisation du modèle YOLO v8x fournie par Ultralytics (Jocher et al., 2023) plutôt que la version v5x utilisée par le projet Gallic(orpor)a. Cette évolution devrait nous permettre d'obtenir un gain substantiel dans l'efficacité des modèles produits (Ronkin et Reshetnikov, 2023).

Lors des entraînements avec YALTAi, nous utilisons des lots de 32 images, qui sont redimensionnées en entrée à 896 pixels, avec un minimum de 150 époques.

4.2.5 Résultats

On observe (cf. tab. 3) que non seulement le modèle entraîné sur les données en gothique fonctionne raisonnablement bien sur les données de test en gothique, mais aussi sur les données des autres

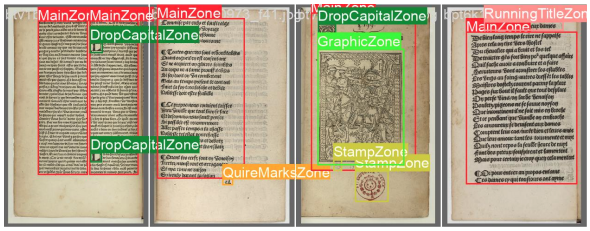
siècles, preuve de la grande stabilité de la mise en page à travers les siècles. L'ajout de données supplémentaires d'autres siècles en entraînement a un impact positif sur les données de test en gothique (cf. fig. 2), preuve de l'efficacité de SegmOnto.

Modèle	Précision	Rappel	mAP50	mAP50-95
Test sur les données gothiques				
Gothique	0.719	0.7	0.712	0.519
Moderne	0.81	**0.756**	0.777	**0.632**
Global	**0.969**	0.711	**0.789**	0.627
Test sur les données gothiques et modernes				
Gothique	0.72	0.497	0.462	0.327
Moderne	0.738	0.657	0.673	0.52
Global	**0.872**	**0.678**	**0.774**	**0.566**
Test sur toutes les données				
Gothique	0.664	0.405	0.374	0.254
Moderne	0.732	**0.535**	0.565	0.419
Global	**0.812**	0.526	**0.596**	**0.427**

TABLEAU 3 – Résultats des modèles d'analyse de la mise en page sur les trois jeux de test.

.

Il convient de prendre avec précaution ces chiffres, qui proposent une forme de macro-moyenne (i.e. moyenne de toutes les classes, peu importe leur poids respectif), et dissimulent des variations de résultats entre les zones. Si cette variation est faible pour le jeu de test en gothique (cf. tab. 10), elle est plus importante pour le jeu de test contenant toutes les données (cf. tab. 4).



(a) Vérité de terrain

(b) Prédiction

FIGURE 2 – Comparaison de la vérité de terrain et de la prédiction pour le modèle Global.

Le calcule des micro et des macro-moyennes, ainsi que du score F1 associé, permet de nuancer l'infériorité du modèle Moderne sur le modèle

Zone	Gothique	Moderne	Global
all	0.664	0.732	**0.812**
MainZone	0.904	**0.943**	0.855
QuireMarksZone	0.707	0.799	**0.855**
MarginTextZone	0.401	0.735	**0.847**
DropCapitalZone	0.669	**0.934**	0.899
TitlePageZone	0.193	0.438	**0.456**
StampZone	0.956	**1**	0.969
GraphicZone	0.804	**0.967**	0.792
NumberingZone	1	0.903	**0.919**
RunningTitleZone	1	0.803	**0.968**
DigitizationArtefactZone	0	0	**1**
TableZone	**1**	**1**	**1**
DamageZone	**1**	**1**	**1**

TABLEAU 4 – Précision par zone et par modèle pour le jeu de test avec toutes les données.

Global, même si ce dernier reste plus performant (cf. tab. 5). Dans la mesure où ce modèle comprend aussi les manuscrits (qui ne sont pas intégrés dans le test), nous choisissons le modèle Global comme meilleur modèle, à qui nous donnons le nom de « SegmOnto Capricciosa » (Humeau et al., 2024).

	P_{micro}	R_{micro}	$F1_{micro}$
Moderne	0.793	**0.578**	0.67
Global	**0.88**	0.5695	**0.691**
	P_{macro}	R_{macro}	$F1_{macro}$
Moderne	0.876	0.758	0.813
Global	**0.883**	**0.775**	**0.814**

TABLEAU 5 – Détail des résultats des modèles Moderne et Global sur le jeu de test avec toutes les données.

4.2.6 Optimisation

Modèle	Précision	Rappel	F1	mAP50	mAP50-95
N	0.782	0.532	0.632	0.531	0.376
M	0.555	**0.604**	0.579	**0.633**	0.425
X	**0.812**	0.526	**0.637**	0.596	**0.427**

TABLEAU 6 – Résultats des différentes versions (N, M, X) du modèle Global sur toutes les données de test.

Les modèles YOLO viennent en plusieurs déclinaisons : *N* pour « nano », *S* pour « small », *M* pour « medium », *L* pour « large » et *X* pour « extra-large ». Les versions les plus petites offrent une plus grande rapidité d'exécution au détriment de l'efficacité du modèle, et les version les plus grandes sont plus lentes mais plus précises dans leur prédiction. Dans l'optique d'extractions de masse, la rapidité d'exécution devient un paramètre (très) important à contrôler, et nous proposons donc trois versions de « SegmOnto Capricciosa » : « CapricciosaN »,

« CapricciosaM » et « CapricciosaX », que nous évaluons sur toutes les données de test (cf. tab. 5). Une version « CapricciosaK », entraînée avec Kraken, est aussi distribuée (cf. tab. 12).

Modèle	Paramètres (millions)	Temps d'inférence (ms)
N	3.01	2.4
M	25.86	10.5
X	68.16	24.5

TABLEAU 7 – Temps d'inférence par page.

Nous évaluons ensuite la rapidité de chacun des modèles sur notre jeu de test pour en obtenir une moyenne. On remarque une division du temps de traitement de plus de 2 pour le modèle M, et de 10 pour le modèle N par rapport à X [8].

4.3 Modèle 2 : reconnaissance optique de caractères gothiques

4.3.1 Règles de transcription

Concernant le français, il existe des guides de transcription, notamment pour les textes médiévaux (Pinche, 2022) et pour les imprimés du XVII^e siècle (Gabay et al., 2023a), dont l'objectif est la production de vérité de terrain partageant (autant que possible) les mêmes normes afin de rendre les données interopérables. Notre guide pour les imprimés du XVI^e siècle en caractères gothiques (Solfrini et al., 2023a) tente de synthétiser les grands principes de ces prédécesseurs. Il propose une transcription graphématique qui conserve la ponctuation originale et les abréviations. Comme nos textes présentent à la fois une écriture (la gothique) et un état de la langue (le moyen français) proches des textes médiévaux, nos choix philologiques se sont plutôt alignés avec ceux d'Ariane Pinche (2022). Quelques adaptations étaient cependant nécessaires et les différences principales concernent les abréviations et la ponctuation. Dans le cas de cette dernière, par exemple, nous distinguons les virgules des barres obliques (‹,› vs. ‹/›) et les traits d'union des tirets de fin de ligne (‹-› vs. ‹¬›).

4.3.2 Jeu de données

Le jeu de données comprend 18 textes du projet SETAF (Tableau 1), pour un total de 1 443 pages (41 244 lignes) qui a été réparti comme suit :

— 1 193 pages pour l'entraînement (82%) ;
— 139 pages pour la validation (10%) ;
— 111 pages pour le test (8%).

Les données des jeux d'entraînement et de validation ont été répartis de façon aléatoire, tandis que le jeu de test correspond à deux ouvrages de taille similaire (48 et 63 pages) qui ont été choisis pour être représentatifs de notre corpus : un texte imprimé par Pierre de Vingle en vers et un texte publié par Jean Michel en prose.

4.3.3 Métriques

Nous avons décidé d'utiliser l'exactitude par mot (*Word Accuracy*, WAcc) en plus de la traditionnelle exactitude par caractère (*Character Accuracy*, CAcc) : comme il y a plusieurs lettres dans un mot, il suffit en effet qu'une lettre soit fausse pour que tout le mot soit faux. Or, notre objectif étant de travailler sur les transcriptions produites par l'OCR pour d'autres tâches, telles que la normalisation et la lemmatisation, il est donc fondamental d'évaluer les résultats du modèle à l'échelle du mot – un mot mal transcrit étant logiquement mal lemmatisé ou normalisé. La WAcc n'étant pas encore proposée par Kraken, nous utilisons la librairie KaMI-lib (Terriel et Chagué, 2021-2022) pour la calculer, dans une version améliorée pour évaluer un grand nombre de pages [9].

4.3.4 Outil

En dépit de l'existence de différents moteurs d'OCR, nous avons décidé de nous concentrer sur Kraken (Kiessling, 2019), qui est actuellement le plus utilisé par la communauté francophone du fait de son intégration dans *eScriptorium*.

4.3.5 Résultats

Le Tableau 8 synthétise les résultats de nos expériences. Une *baseline* obtenue par le premier modèle utilisé, Cortado 2.0.0, offre déjà de bons résultats sans *fine-tuning*. CATMuS Medieval donne des résultats légèrement inférieurs, ce qui pourrait s'expliquer par la plus grande spécialisation de ce second modèle sur des données médiévales et son élargissement à un plus grand nombre de langues.

La quantité de données produites par le projet SETAF permet un bond qualitatif majeur (modèle « Setaf Adam ») sans qu'il ne soit nécessaire d'utiliser d'autres modèles, du point de vue de la CAcc (≈+4 pt de %) mais surtout du point de vue de la WAcc (≈+20 pt de %). Si nous tentons de « fine-tuner » Cortado et CATMuS Medieval nous obte-

8. L'évaluation a été effectué sur un GPU de type RTX 4070 Ti.

9. https://github.com/FoNDUE-HTR/kamiCLI.

Modèle de base	*Fine-tuning*	CAcc (%)	WAcc (%)
Cortado 2.0.0	-	96.31	78.89
CATMuS M.	-	95.75	77.6
Setaf Adam	-	99.46	97.25
Cortado 2.0.0	SETAF	99.72 ± 0.03	98.28 ± 0.19
CATMuS M.	SETAF	**99.73** ± 0.03	**98.35** ± 0.17

TABLEAU 8 – Notre *baseline* est obtenue avec Cortado 2.0.0 sans *fine-tuning*. La conception d'un modèle avec uniquement nos données (« Setaf Adam ») bat la *baseline*, mais les meilleurs modèles sont obtenus avec le *fine-tuning* de CATMuS Medieval sur les données SETAF, ce qui a permis de créer un nouveau modèle (« CATMuS Gothic Print »). Les résultats indiqués pour le *fine-tuning* de CATMuS Medieval et Cortado représentent la moyenne de cinq expériences réalisées pour chaque *fine-tuning* et chaque moyenne est accompagnée de son écart-type.

Pages SETAF	Lignes SETAF	CAcc (%)	WAcc (%)
0	**0**	**95.75**	**77.6**
10	280	98.98 ± 0.30	94.43 ± 1.59
25	700	99.19 ± 0.44	95.53 ± 2.30
50	1 400	99.54 ± 0.04	97.39 ± 0.29
100	2 800	99.60 ± 0.06	97.59 ± 0.45
1 332	**37 324**	**99.76**	**98.50**

TABLEAU 9 – Nous comparons les résultats de différents *fine-tuning* du modèle CATMuS Medieval à partir d'une *baseline* sans *fine-tuning* (0 pages). Le jeu de test reste toujours le même (décrit *supra*) et correspond à 111 pages. La répartition entre les jeux d'entraînement et de validation correspond toujours à 90-10 %. Les résultats indiqués, sauf dans le cas de la *baseline* et du modèle « CATMuS Gothic Print », représentent la moyenne de cinq expériences réalisées pour chaque *fine-tuning* et chaque moyenne est accompagnée de son écart-type. Le nombre de lignes est donné à titre indicatif et il est calculé sur la moyenne du nombre de lignes par page dans le corpus SETAF.

nons une légère amélioration (≈+1 pt de %) de la CAcc comme de la WAcc. La différence entre les résultats obtenus avec ces deux modèles est minime, mais ceux de CATMuS sont légèrement supérieurs et ils nous ont permis de créer un nouveau modèle, « CATMuS Gothic Print » (Solfrini et Gabay, 2024), ayant une CAcc de 99.76 % et une WAcc de 98.50%.

Cortado et CATMuS Medieval atteignaient déjà plus de 95 % pour la CAcc mais pas plus de 78 % pour la WAcc, donc l'amélioration considérable de la WAcc avec CATMuS Gothic Print nous confirme à la fois l'importance de cette mesure pour l'évaluation des modèles d'OCR et l'efficacité de notre modèle pour continuer à travailler à l'échelle du mot sur des textes en gothique.

4.4 Optimisation

Afin d'évaluer le meilleur *ratio* nombre de pages supplémentaires / gain en précision lors du *fine-tuning* d'un modèle, nous avons lancé plusieurs entraînements avec CATMuS Medieval sur une quantité de données croissante (cf. tab. 9). Le choix de *fine-tuner* CATMuS au lieu de Cortado repose également sur le multilinguisme du premier – e.g. un document de la Renaissance a de grandes probabilités de contenir des passages en latin.

On remarque (cf. tab. 9) que dix pages suffisent à nettement améliorer la CAcc (≈+3 pt de %) de CATMuS, mais surtout sa Wacc (≈+17 pt de %), ce qui est fondamental pour nombre de *downstream tasks* comme la lemmatisation. Si l'ajout de données d'entraînement supplémentaires permet une amélioration marginale de la CAcc (moins de 1 pt de % pour 100 pages de plus), l'augmentation est plus nette pour la WAcc (≈+1 pt de % pour 25 pages de plus, et ≈+3 pt de % en 100 pages de plus). Un ajout massif de données supplémentaires, avec la totalité des pages de notre corpus édité (18 textes), ne permet de gagner qu'un peu moins de 1 pt de % pour la WaCC – amélioration néanmoins importante dans un corpus qui contient des centaines de milliers de mots.

5 Conclusion

Avec ces deux modèles, l'océrisation des imprimés français du XVI^{e} siècle, notamment en gothique, est lancée dans les meilleurs conditions. Un travail important reste néanmoins à mener concernant les *downstream tasks* (e.g. la normalisation de la langue et la lemmatisation), les données issues de l'OCR étant trop « brutes » à cause des abréviations, des agglutinations, etc.

En outre, l'arrivée de la version 9 du modèle YOLO pourrait permettre d'accroître la performance des plus petits modèles d'analyse de la mise en page, et rendre plus accessible son utilisation, avec une réduction du nombre de paramètres (∼15%) (Wang et al., 2024).

Remerciements

Les calculs ont été effectués sur les clusters HPC de l'université de Genève. Merci à Thibault Clérice pour son accompagnement.

Bibliographie

Gabrielle Berthoud. 1980. Les impressions genevoises de Jean Michel (1538-1544). In Jean-Daniel Candaux et Bernard Lescaze, éditeurs, *Cinq siècles d'imprimerie genevoise*, volume 1, pages 55–88. Droz, Genève.

Reinhard Bodenmann. 2009. *Faictz de Jesus Christ et du pape*. Cahiers d'Humanisme et Renaissance. Droz.

Nina Catach. 1968. *L'Orthographe française à l'époque de la Renaissance : auteurs, imprimeurs, ateliers d'imprimerie*. Droz.

Alix Chagué et Thibault Clérice. 2023. "I'm here to fight for ground truth"': HTR-United, a solution towards a common for HTR training data. In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Alix Chagué, Thibault Clérice, et Laurent Romary. 2021. HTR-United : Mutualisons la vérité de terrain! In *DH Nord 2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France. MESHS.

Thibault Clérice. 2023. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. *Journal of Data Mining and Digital Humanities*, Historical documents and automatic text recognition.

Thibault Clérice, Alix Chagué, et Hugo Scheithauer. 2023. Workshop HTR-United: metadata, quality control and sharing process for HTR training data. In *DH 2023 - Digital Humanities Conference : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Simon Gabay, Robin Champenois, Pierre Kuenzli, Jean-Luc Falcone, et Christophe Charpilloz. 2021-. Formes Numerisées et Détection Unifiée des Écritures (FoNDUE).

Simon Gabay, Thibault Clérice, et Christian Reul. 2023a. OCR17: Ground truth and models for 17th c. french prints (and hopefully more). *Journal of Data Mining and Digital Humanities*, 2023.

Simon Gabay, Ariane Pinche, Kelly Christensen, et Jean-Baptiste Camps. 2023b. SegmOnto: a controlled vocabulary to describe and process digital facsimiles. Pré-publication / document de travail.

Simon Gabay, Ariane Pinche, Kelly Christensen, Jean-Baptiste Camps, et Nicola Carboni. 2023c. SegmOnto: A controlled vocabulary to describe the layout of pages.

Maxime Humeau, Simon Gabay, et Ariane Pinche. 2024. SegmOnto. v. 1.0.0 - Capricciosa.

Glenn Jocher, Ayush Chaurasia, et Jing Qiu. 2023. Ultralytics yolov8. v. 8.0.0.

William Kemp. 2004. La redécouverte des éditions de Pierre de Vingle imprimées à Genève et à Neuchâtel (1533-1536). In Jean-François Gilmont et William Kemp, éditeurs, *Le Livre évangélique en français avant Calvin*, pages 147–177. Brepols, Turnhout.

Benjamin Kiessling. 2019. Kraken - an Universal Text Recognizer for the Humanities. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, The Netherlands. ADHO.

Benjamin Kiessling. 2020. A modular region and text line layout analysis system. In *17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020*, pages 313–318. IEEE.

Benjamin Kiessling, Robin Tissot, Peter Stokes, et Daniel Stökl Ben Ezra. 2019. eScriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, et Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv :1910.09700*.

Ariane Pinche. 2022. Guide de transcription pour les manuscrits du Xe au XVe siècle. Pré-publication / document de travail.

Ariane Pinche et Thibault Clérice. 2022. Htr-united/cremma-medieval: Cortado 2.0.0. v. 2.0.0.

Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Marc Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Anne Boutreux, Alec Manton, et Simon Gabay. 2023. CATMuS medieval. v. 1.0.0.

Ariane Pinche et Simon Gabay. 2022. Gallicorpora+.

Marion Pouspin. 2016. *Publier la nouvelle : Les pièces gothiques, histoire d'un nouveau média (XVe-XVIe siècles)*. Éditions de la Sorbonne.

Jean-Yves Ramel, Nicolas Sidère, et Frédéric Rayar. 2013. Interactive layout analysis, content extraction, and transcription of historical printed books using Pattern Redundancy Analysis. *Literary and Linguistic Computing*, 28(2) :301–314.

Joseph Redmon, Santosh Divvala, Ross Girshick, et Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Las Vegas.

Mikhail Ronkin et Kirill Reshetnikov. 2023. Real-time yolo-family comparison for blast quality estimation in the open pit conditions. In *2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pages 254–257.

Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Javier Ortiz Suárez, Simon Gabay, Ariane Pinche, et Jean-Baptiste Camps. 2023. Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue.

Sonia Solfrini et Simon Gabay. 2024. CATMuS Gothic Print. v. 1.0.0.

Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, et Daniela Solfaroli Camillocci. 2023a. Guide de transcription pour les imprimés français du XVIe siècle

en caractères gothiques : Version A. Pré-publication / document de travail.

Sonia Solfrini, Geneviève Gross, Brigitte Roux, Nathalie Szczech, Pierre-Olivier Beaulnes, Aurélia Oliveira Marques, et Daniela Solfaroli Camillocci. 2023b. Étudier le « groupe de Neuchâtel » : de l'édition des Faits à un corpus numérique de la première Réforme romande. In *Humanistica 2023, Association francophone des humanités numériques*.

Uwe Springmann et Anke Lüdeling. 2017. OCR of historical printings with an application to building diachronic corpora : A case study using the ridges herbal corpus. *Digital Humanities Quarterly*, 11(2).

Uwe Springmann, Christian Reul, Stefanie Dipper, et Johannes Baiter. 2018. Ground truth for training OCR engines on historical documents in German fraktur and early modern Latin. *Journal for Language Technology and Computational Linguistics*, 33(1) :97–114.

Nathalie Szczech. 2021. Un groupe en polémique. Le groupe de Neuchâtel et ses pratiques concertées d'écriture dans les années 1530. In Daniela Solfaroli Camillocci, Nicolas Fornerod, Karine Crousaz, et Christian Grosse, éditeurs, *La construction internationale de la Réforme et l'espace romand à l'époque de Martin Luther*, pages 189–206. Garnier.

Lucas Terriel et Alix Chagué. 2021-2022. KaMI-lib - Kraken Model Inspector. https://github.com/KaMI-tools-project/KaMi-lib.git.

Toshinori Uetani, Guillaume Porte, Sandrine Breuil, et Mathieu Duboc. 2016. The BVH in Tours: Digital library of image, text, and data. In *TEI Conference 2016*, Vienne, Austria. TEI Consortium.

Claire Hélène Vachon. 2010. *Le changement linguistique au XVIe siècle : une étude basée sur des textes littéraires français*. Ed. de linguistique et de philologie, Strasbourg.

Chien-Yao Wang, I-Hau Yeh, et Hong-Yuan Mark Liao. 2024. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv*.

A Annexes

A.1 Analyse de la mise en page

Zone	Gothique	Moderne	Global
all	0.719	0.81	**0.969**
MainZone	0.981	**0.998**	0.996
QuireMarksZone	0.808	0.867	**0.93**
DropCapitalZone	0.759	**0.998**	0.997
StampZone	0.987	**1**	0.95
GraphicZone	0.779	**1**	0.94
RunningTitleZone	0	0	**1**

TABLEAU 10 – Précision par zone et par modèle pour le jeu de test en gothique.

Type de zone	Gothique	Moderne	Global
MainZone	1477	3792	4224
StampZone	18	201	267
QuireMarksZone	1065	2494	2507
NumberingZone	2	1208	1413
RunningTitleZone	23	1374	1407
DropCapitalZone	234	837	838
GraphicZone	115	633	634
MarginTextZone	1852	2406	2520
TitlePageZone	21	68	73
CustomZone	0	0	17
DigitizationArtefactZone	0	0	17
TableZone	6	7	7
DamageZone	9	19	21
MusicZone	0	1	1

TABLEAU 11 – Nombre de zones par type lors de l'entraînement avec le maximum de données.

Métrique	Score
Exactitude	0.998
Exactitude moyenne	0.998
IoU moyenne	0.279
IoU pondérée en fréq.	0.895

TABLEAU 12 – Résultats du Modèle CapricciosaK, entraîné avec Kraken.

A.2 Reconnaissance optique de caractères

Erreurs	Caractère correct	Caractère généré
19	No character	Space
13	.	No character
11	Space	No character
10	No character	/
5	z	No character

TABLEAU 13 – Les caractères erronés dans le test fait sur le modèle CATMuS Gothic Print. Le jeu de test correspond toujours à deux textes (48 et 63 pages, 89 074 caractères en total) du corpus SETAF.

A.3 Émissions de CO2 liées aux expériences

Les premières expériences pour le modèle d'analyse de la mise en page ont été menées en utilisant une infrastructure qui a une efficacité carbone de 0,116 kgCO_2eq/kWh [10]. Un cumul de 18 heures de calcul a été effectué entre 1 et 3 GPU de type Titan RTX (TDP de 280W). Les émissions totales sont estimées à 3.22 kgCO_2eq [11].

10. Basé sur les estimations du 11-01-2024 via Horocarbon.
11. Les estimations ont été réalisées à l'aide du *Machine Learning Impact Calculator* (Lacoste et al., 2019).