Lire avant de faire lire

Simon Gabay, Ariane Pinche, Peter Nahon, Alix Chagué, Pauline Jacsont, Élodie Paupe, Jean-Claude Rebetez, Maxime Humeau, Christine Payot, Thibault Maillard, et al.

Humanités numériques

12 | 2025
Varia

Lire avant de faire lire. Réflexions philologiques sur la reconnaissance automatique de texte pour les manuscrits modernes français

Reading Before It Can Be Read: Philological Reflections on Automatic Text Recognition for Modern French Manuscripts

Simon Gabay, Ariane Pinche, Peter Nahon, Alix Chagué, Pauline Jacsont, Élodie Paupe, Jean-Claude Rebetez, Maxime Humeau, Christine Payot, Thibault Maillard, Yvan Jauregui, Elina Leblanc et Loraine Chappuis

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/revuehn/5300
ISSN : 2736-2337

Éditeur
Humanistica

Lire avant de faire lire. Réflexions philologiques sur la reconnaissance automatique de texte pour les manuscrits modernes français

Reading Before It Can Be Read: Philological Reflections on Automatic Text Recognition for Modern French Manuscripts

Simon Gabay, Ariane Pinche, Peter Nahon, Alix Chagué, Pauline Jacsont, Élodie Paupe, Jean-Claude Rebetez, Maxime Humeau, Christine Payot, Thibault Maillard, Yvan Jauregui, Elina Leblanc et Loraine Chappuis

Introduction

1 Une part non négligeable de la littérature et des documents d'archives rédigés après le Moyen Âge est encore conservée sous forme manuscrite. Or, contrairement à ceux de l'époque médiévale, qui ont bénéficié ces dernières années de travaux plus soutenus dans ce domaine (Pinche *et al.* 2024), les manuscrits modernes ne disposent toujours pas de ressources permettant une exploitation automatisée. Les progrès récents des outils d'extraction d'information, à commencer par ceux de reconnaissance automatique de texte (*Automatic Text Recognition*, ATR[1] ; cf. Raj et Kos 2022), rendent dès lors nécessaire le développement d'un modèle capable de récupérer efficacement le contenu textuel de ces sources.

2 La conception d'un modèle d'ATR n'est cependant pas triviale. Elle suppose notamment une compréhension minimale de l'évolution de la langue comme des formes d'écriture, afin de délimiter un périmètre raisonnable et cohérent pour les phases d'entraînement et de test. En effet, on ne peut raisonnablement espérer qu'un modèle entraîné sur des écritures coulées de la fin du XVIII[e] siècle puisse fonctionner correctement sur un texte du XVI[e] siècle rédigé en écriture financière, surtout si les données ont été transcrites

avec des normes ne prenant pas en compte les spécificités des documents les plus anciens (abréviations, modification de l'alphabet, etc.).

3 Pour résoudre ce problème, nous cherchons ici à définir un ensemble de règles visant à construire un modèle adapté aux manuscrits d'une modernité « longue », allant du Moyen Âge tardif à aujourd'hui, en nous concentrant sur le français. À cette fin, nous introduisons quelques notions importantes de philologie permettant d'établir un cadre d'analyse pour évaluer la qualité de la prédiction. Nous proposons également un nouveau modèle, destiné aux écritures françaises modernes, dont la performance surpasse celle des modèles existants, contribuant ainsi à établir un nouvel état de l'art.

État de l'art

4 La question du traitement informatique des mains modernes, qui, dans son acception la plus large, va du XV^e^ siècle jusqu'au XXI^e^ siècle pour le domaine francophone, n'est pas nouvelle, même si elle reste imparfaitement réglée. Pour la période la plus récente, on trouve notamment le projet *Lectaurep* (Chagué et Rostaing 2021), qui s'est intéressé au traitement des registres de notaires parisiens durant la période 1803-1940. Pour la partie haute de la chronologie, des premiers travaux ont, par exemple, été effectués autour du manuscrit de Richard Simon (vers 1701) conservé à la Hofbibliothek d'Aschaffenbourg (Nahon et Gabay 2023).

5 Plusieurs modèles ont déjà été publiés. Transkribus (Kahle *et al.* 2017) en propose un pour le français sans spécifier les dates et les types d'écritures couvertes[2]. Manu McFrench (Chagué et Clérice 2022), entraîné avec Kraken (Kiessling 2019), offre une alternative ouverte à celui de Transkribus, mais n'est pas plus précis quant à sa couverture paléographique – des premiers essais ayant démontré des résultats (très) faibles pour les sources de la première modernité.

6 Concernant les données d'entraînement disponibles, le catalogue HTR-United[3] (Chagué et Clérice 2023 ; Chagué, Clérice et Romary 2022) ne répertorie que peu de jeux de données pour les documents francophones écrits en cursive, surtout pour la période de l'Ancien Régime. Parmi les projets distribuant leurs données de manière ouverte[4], l'allemand (Hodel, Schoch et Dängeli 2021 ; Hodel et Schoch 2021 ; Hodel *et al.* 2021) et, dans une moindre mesure, l'italien (Cascianelli *et al.* 2022) présentent un état d'avancement assez similaire à celui du français. Une exception existe : celle du néerlandais, qui bénéficie de ressources de qualité en très grande quantité (Keijser 2020).

Rappel philologique

7 La conception d'un modèle d'ATR pour les mains « modernes » nécessite au préalable de comprendre ce que recouvre cette période, or l'établissement du *terminus post quem* est un problème important. D'une part, si la (première) modernité succède théoriquement au Moyen Âge, le passage de relais se fait à des époques et des vitesses différentes à travers la France et le continent européen, rendant difficile de définir un cadre chronologique strict. D'autre part, si l'imprimé est presque consubstantiellement lié à une modernité dont il est l'un des critères définitoires (Barbier 2006), l'histoire de

l'écriture manuscrite (Smith 2020) suit une chronologie un peu différente, du fait de particularités propres à ce médium.

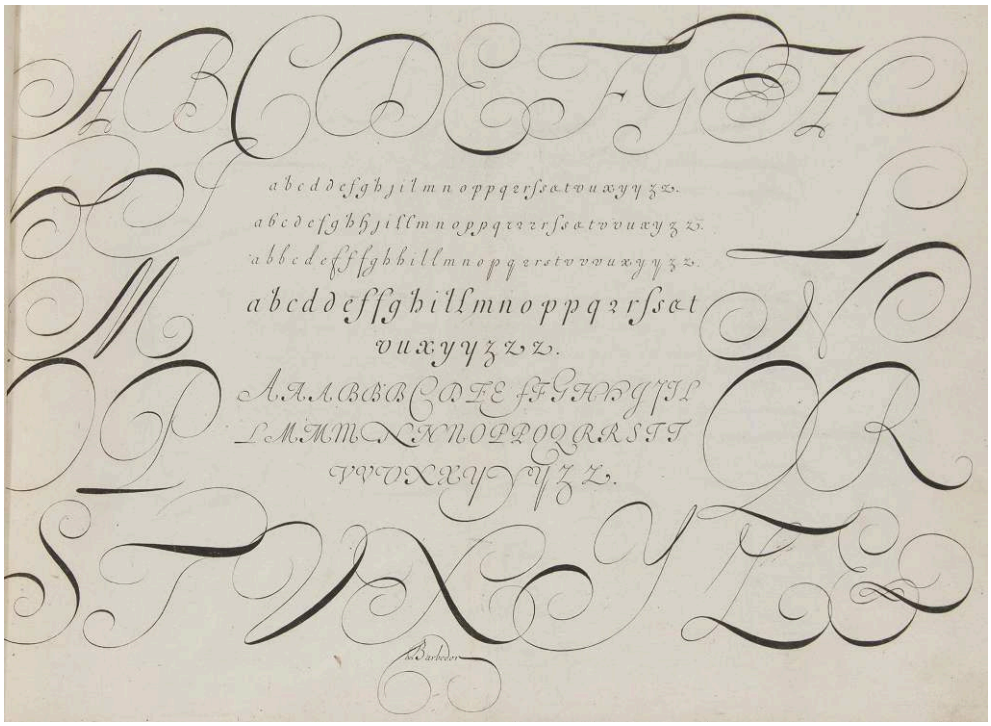
La question paléographique

8 D'un point de vue paléographique, l'écriture liée, souvent tenue pour le marqueur spécifique de la modernité, n'est pas une invention de cette période : l'écriture « mixte », qui témoigne d'une profonde refonte du ductus visant à économiser le geste avec le passage à des ligatures de séquence (Poulle 2007), apparaît en France dès le XIVe siècle. Permettant un tracé rapide qui lui assure notamment un grand succès auprès du personnel des chancelleries, elle va persister jusqu'au XVIIe siècle, non sans muter profondément au fil du temps (Poulle 1966). Si, à la fin de cette longue marche, il est encore possible de la qualifier de forme évoluée de la « gothique », cette écriture, que l'on appelle plus volontiers « financière » ou « française », a profondément changé, notamment sous l'influence du maître écrivain Guillaume Le Gangneur (1599) qui en propose une version épurée, débarrassée de ses oripeaux médiévaux.

9 À partir du XVe siècle, cette écriture financière cohabite avec d'autres, dites « cancelleresques » ou « humanistiques », importées d'Italie (Ullman 1960 ; Gasparri 1983), qui sont lentement adoptées puis formalisées sous le nom de « bâtarde » par des maîtres d'écriture comme Lucas Materot (1608). Moins liée, plus posée et imposant ainsi un tracé plus lent, la bâtarde est surtout l'apanage des lettrés, par exemple dans leur documentation personnelle (correspondance, mémoires, etc.). La financière, elle, reste plutôt l'écriture de prédilection de l'administration, qui, pour des raisons pratiques, en accentue la cursivité, parfois au détriment de la lisibilité.

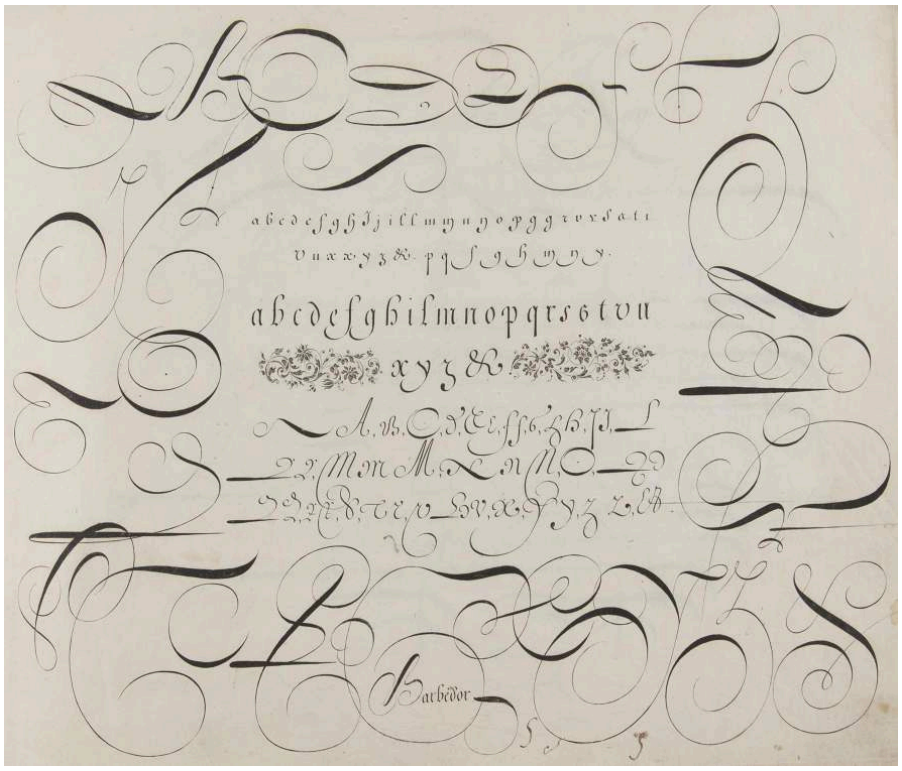
10 La créativité, voire l'exubérance des maîtres écrivains, qui jouent dans cette histoire un rôle central (Cabane 2020), va encore réduire la lisibilité des documents en ajoutant à cette cursivité parfois « échevelée » (Samaran 1967, 129) des ornementations et des ligatures volontiers extravagantes. Afin de mettre un terme à ce qui apparaît comme une dérive, le Parlement de Paris réforme la pratique par deux arrêts émis en 1632 puis en 1633, et impose deux modèles dans le royaume de France (Métayer 2001) – mais aussi, *volens nolens*, dans le reste de la francophonie. Ces deux écritures sont l'« italienne » (dite aussi « bâtarde », figure 1a), qui est une stylisation française des écritures importées d'Italie, mise au point par Étienne Le Bé, et la « française » (aussi appelée « financière » puis « ronde », figure 1b), dont le modèle est fourni par Louis Barbedor (1649).

Figure 1a. Alphabet de bâtarde (modèle de Louis Barbedor 1649)



The Newberry Library

Figure 1b. Alphabet de ronde (modèle de Louis Barbedor 1649)

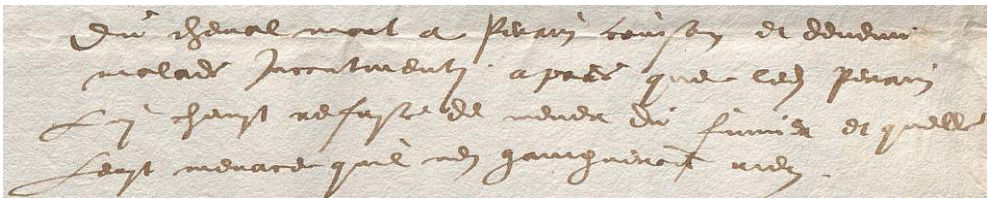


The Newberry Library

11 Si la base morphologique de ces deux écritures reste relativement stable et ne connaît plus de dérive importante, des innovations continuent de naître (Hébrard 1995) avec, par exemple, l'apparition au XVIII^e^ siècle d'un mélange entre la française et l'italienne qui s'institutionnalise sous le nom de « coulée ». Un peu plus tard, au début du XIX^e^ siècle, on voit apparaître dans la documentation une écriture dite « anglaise » (Heal 1931), signe de l'importance culturelle croissante de la moitié septentrionale de l'Europe alors que débute la révolution industrielle. En effet, l'anglaise « de France », pour ainsi dire, est le produit de diverses stylisations successives, intervenues notamment aux Pays-Bas (Smith 2020), témoignant d'une recherche stylistique intense mais aussi de la circulation des modèles à travers le continent. Parallèlement à cette circulation, on observe aussi un enracinement de certaines pratiques, comme la *kurrent* allemande (Beck 1991), qui rayonne un peu au-delà de l'espace strictement germanophone et marque les zones limitrophes en contact avec l'allemand, comme le Jura suisse (voir la figure 2b). Avec le développement de l'instruction publique (Bishop 2019), l'écriture se banalise et se diversifie du simple fait de la multiplication des scripteurs (Frondizi et Fureix 2022), notamment peu lettrés (Branca-Rosoff et Schneider 1994 ; Bergeron-Maguire 2019). Rapidement, l'anglaise s'impose dans les manuels français (Dancel 2011) et va rester le modèle dominant tout au long des XIX^e^ et XX^e^ siècles – soit plus longtemps qu'en Angleterre, qui l'a abandonnée entre-temps (Smith 2020).

12 Un modèle d'ATR pour le « français moderne » se doit donc, théoriquement, de couvrir ce vaste monde de possibles, voire plus si l'on ajoute à ce rapide panorama le problème de l'évolution de l'encre, du papier et surtout des plumes – une telle diversité pouvant se trouver dans une même série documentaire à quelques décennies de distance (voir, par exemple, les figures 2a et 2b). Étant donné l'amplitude de la variation, il convient d'être précis sur la couverture du modèle, qui ne peut (encore) traiter avec la même qualité tous les documents, et d'utiliser les catégories que nous venons de présenter pour décrire les données.

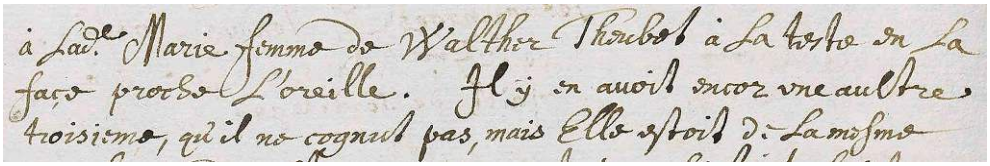
Figure 2a. Exemple d'évolution de la cursive française au cours du XVII^e^ s. dans un corpus de procès criminels : Porrentruy (Jura suisse), AAEB, B 168/15-2.3 (1608)



Archives de l'ancien évêché de Bâle · Licence CC BY

du cheual mort a Perrin coinson et deuenu
malade incontinenti apres que leꝙ Perrin
lui heust refuse de mener du fumier et quelle
leust menace quil nen gaingneroit rien.

Figure 2b. Exemple d'évolution de la cursive française au cours du XVII^e s. dans un corpus de procès criminels : Porrentruy (Jura suisse), AAEB, B 168/19-35.1 (1670)



Les deux points sur le ‹y› sont typiques de l'écriture allemande.
Archives de l'ancien évêché de Bâle · Licence CC BY

> à lad^e Marie femme de Walther Theubet à la teste en la
> façe proche l'oreille. Jl y en auoit encor vne aultre
> troisieme, qu'il ne cognut pas, mais elle estoit de la mesme

La question linguistique

13 Si l'écriture change, la langue connaît aussi des évolutions importantes qu'il convient de prendre en compte. Parmi celles qui importent le plus pour la construction d'un modèle d'ATR, on en trouve deux qui relèvent du matériel graphétique (Catach 2001 ; Cazal et Parussa 2022), à savoir les lettres ramistes et la majuscule, et une qui touche à la segmentation de la chaîne graphique.

Figure 3. À gauche, le *J* majuscule en écriture « coulée » d'après la planche de Charles Paillasson (1763). À droite, deux variantes minuscules du *J* dans la même écriture et d'après la même source

Dans la pratique, la variante minuscule de gauche est difficilement distinguable de la majuscule, dont elle est morphologiquement proche (voir, par exemple la figure 6).
Extrait de la planche « Alphabets des lettres coulées », Charles Paillasson (1763) · Wikimédia · Licence CC BY-SA 3.0

14 La majuscule a une double fonction en français contemporain : c'est un « signe-mot » et un « signe de phrase », pour reprendre la terminologie de Nina Catach (1994). Mais l'histoire de cette double fonction est complexe, tout particulièrement dans les manuscrits – quand on y trouve des majuscules (Gabay 2020). Si le XVI^e siècle voit apparaître des premières tentatives de rationalisation de la pratique (Huchon 1983), l'emploi de la majuscule de mot reste instable pendant longtemps, même dans

l'imprimé (Riffaud 2007). Concernant la majuscule de phrase, il convient de rappeler que la segmentation du discours reste longtemps rhétorique (on parle de « période ») et qu'un découpage grammatical s'impose tardivement (Siouffi 2020) – il n'est ainsi pas rare de ne trouver aucun point au XVII^e siècle (Gabay 2020). Dans la mesure où la différence entre la majuscule et la minuscule n'est souvent qu'une (infime) variation du module (figure 3) et qu'il n'est pas possible de recourir à la double fonction précédemment évoquée, il est difficile, voire impossible, de les distinguer correctement, si tant est que la différence existe…

15 Sans rentrer dans le détail des diverses propositions des grammairiens du XVI^e siècle, un problème similaire se pose avec le système alphabétique du français à l'époque moderne, qui connaît une importante révolution avec l'introduction de deux nouvelles lettres : *v* et *j* (lettres dites « ramistes »). Si, du point de vue morphologique, ces deux signes existent déjà au Moyen Âge (Cazal et Parussa 2022), ils sont à cette époque compris comme des allographes de *u* et *i* dont l'utilisation obéit à des règles graphiques (c'est-à-dire leur position au sein du mot) et non phonologiques (c'est-à-dire le son représenté). Ce n'est qu'avec Corneille, qui reprend un usage néerlando-belge introduit par Plantin, que l'Académie se décide finalement à adopter cette innovation (Catach et Golfand 1973[5]). Comme la pénétration de cette dernière se fait lentement, il n'est pas toujours aisé de savoir si la source utilise ou non *v* et *j*, y compris au XVIII^e siècle, et donc de déterminer ce qu'il faut transcrire (figures 2b et 6).

16 Enfin, la segmentation de la chaîne graphique reste un problème majeur. Dans les manuscrits, elle ne correspond pas toujours aux unités syntaxiques ou lexicales, que ce soit au XVI^e siècle (Baddeley 1998), au XVII^e siècle (Pellat 1998), au XVIII^e siècle (Seguin 1998), ou encore plus tard chez les personnes peu lettrées (Steuckardt 2014). Comme pour la capitalisation et les lettres ramistes, le facteur manuscrit joue un rôle important. D'une part, les levés de plume ne sont pas toujours très clairs et, d'autre part, il reste difficile de connaître le degré d'avancement des processus de soudure sans analyser dans le détail le système du scripteur ou de la scriptrice (*ce pendant* ou *cependant* ? Voir, par exemple, la figure 11), ce qui empêche une résolution simple de la question dès lors que l'on travaille à l'échelle de grands corpus.

La question ecdotique

17 Ces difficultés sont d'autant plus complexes à résoudre que l'on dispose d'un nombre restreint de guides pour l'édition des textes écrits pendant la période moderne, comme les *Conseils* publiés par l'École nationale des chartes (Vieillard et Guyotjeannin 2014). C'est notamment le cas pour les textes du XVII^e siècle, lesquels sont largement « modernisés » (alors qu'ils datent de l'époque moderne) sans guère se soucier du matériau original (Gabay 2014 ; Duval 2015). Aux problèmes pratiques posés par le vêtement graphique et graphétique des manuscrits s'ajoute donc celui, scientifique, d'une tradition volontiers interventionniste des éditeurs et éditrices, qui n'ont pas l'habitude, à de rares exceptions près (Sorel 2014), de remettre en question la pertinence de leurs pratiques.

Transcription

Remarques préliminaires

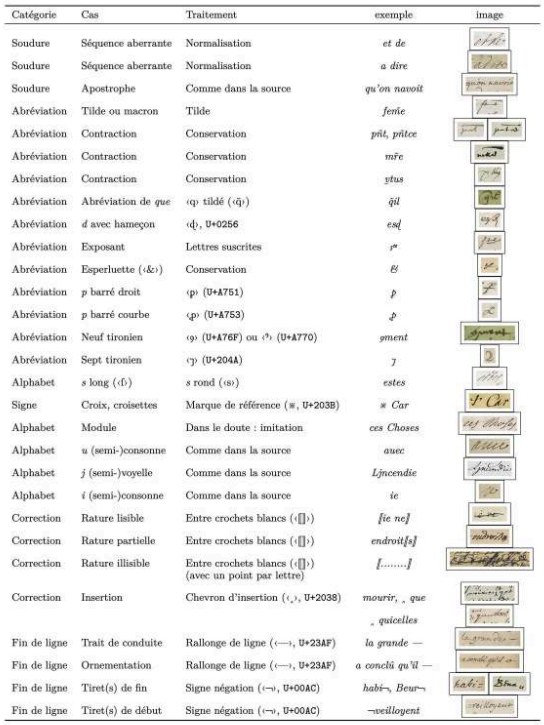
18 Plutôt que d'accumuler des transcriptions disparates, il nous a paru essentiel de présenter quelques premières propositions afin de standardiser au mieux les données qui seraient produites à l'avenir par d'autres projets. Les choix de transcription des médiévistes (Pinche 2022 ; Pinche *et al.* 2024), déjà suivis par les spécialistes des imprimés de la Renaissance (Solfrini *et al.* 2023), ont été utilisés comme lignes directrices afin de conserver, autant que possible, une interopérabilité avec leurs jeux de données et les résultats de leur récent modèle CATMuS médiéval[6] (Pinche *et al.* 2023).

19 La mise en œuvre des préconisations que nous présentons ici reste cependant sujette à caution, car les chercheurs et chercheuses ayant entamé leur projet il y a longtemps n'ont *a fortiori* pas suivi nos recommandations, qui arrivent après que le travail a été effectué. L'étude des données déjà produites laisse ainsi entrevoir des transcriptions plutôt semi-diplomatiques, la plupart du temps avec les abréviations développées. Le recours à ces données d'entraînement (dites de « vérité de terrain »), utile pour élargir la couverture paléographique et linguistique d'un modèle, peut donc créer des problèmes si les différentes pratiques tacitement en vigueur jusqu'à présent entrent en conflit entre elles ou avec les nôtres : à cause de l'hétérogénéité des principes de transcription, il pourrait s'avérer difficile de produire un modèle efficace. Il n'est ainsi pas impossible que, par pragmatisme, nous finissions dans un avenir proche par amender certaines de nos recommandations.

20 Nous souhaitons néanmoins insister sur un point majeur, qui est au cœur de notre approche et la différencie de celle des autres modernistes : nous pensons qu'il est important de distinguer la production de vérité de terrain et la préparation du texte pour l'édition. En effet, la seconde implique un toilettage du texte plus important pour faciliter la lecture humaine, alors que la première exige un plus grand respect de la source pour ne pas gêner la machine – l'abréviation ‹p› reste toujours un ‹p› et ne devient pas parfois *per* (‹pmis›→*permis*), parfois *par* (‹ptie›→*partie*). Par ailleurs, les options prises étant (très) disparates d'une tradition philologique à l'autre (Gabay 2014) et certains signes abréviatifs en français étant des diacritiques dans d'autres langues (comme le tilde en portugais), un trop grand interventionnisme ne peut *in fine* que rendre caduc tout espoir d'interopérabilité, même minimale, au-delà du français. Il s'agit de penser la création de données d'entraînement comme la première étape d'un mouvement en deux temps, le second étant celui d'une normalisation du texte pour en simplifier l'accès au lecteur ou à la lectrice[7].

21 La préparation de vérité de terrain n'a pas pour unique objectif la lecture d'une édition mais a aussi pour but l'utilisation des données, par exemple dans le cadre de recherches en linguistique de corpus (Gabay *et al.* 2022 ; Gabay et Clérice 2024). Il convient donc d'agir précautionneusement, sans pour autant s'interdire d'intervenir pour ne pas tomber dans une logique facsimilaire antiphilologique et contre-productive. Notre choix s'est donc porté sur une transcription graphématique, selon la terminologie proposée par Stutzmann (2011), qui réduit chaque forme à sa valeur dans le système alphabétique actuel et préserve la suite des lettres, sans développement des abréviations.

Figure 4. Principales règles de transcription



Catégorie	Cas	Traitement	exemple	image
Soudure	Séquence aberrante	Normalisation	*et de*	
Soudure	Séquence aberrante	Normalisation	*a dire*	
Soudure	Apostrophe	Comme dans la source	*qu'on navoit*	
Abréviation	Tilde ou macron	Tilde	*femme*	
Abréviation	Contraction	Conservation	*prit, prince*	
Abréviation	Contraction	Conservation	*mre*	
Abréviation	Contraction	Conservation	*ytus*	
Abréviation	Abréviation de *que*	‹q› tildé (‹q̃›)	*q̃il*	
Abréviation	*d* avec hameçon	‹ɖ›, U+0256	*esɖ*	
Abréviation	Exposant	Lettres suscrites	*rᵉ*	
Abréviation	Esperluette (‹&›)	Conservation	*&*	
Abréviation	*p* barré droit	‹ꝑ› (U+A751)	*ꝑ*	
Abréviation	*p* barré courbe	‹ꝓ› (U+A753)	*ꝓ*	
Abréviation	Neuf tironien	‹�unk› (U+A76F) ou ‹ꝰ› (U+A770)	*ꝭment*	
Abréviation	Sept tironien	‹ꝝ› (U+204A)	*ꝝ*	
Alphabet	*s* long (‹ſ›)	*s* rond (‹s›)	*estes*	
Signe	Croix, croisettes	Marque de référence (⁜, U+203B)	⁜ *Car*	
Alphabet	Module	Dans le doute : imitation	*ces Choses*	
Alphabet	u (semi-)consonne	Comme dans la source	*auec*	
Alphabet	*j* (semi-)voyelle	Comme dans la source	*Ljncendie*	
Alphabet	*i* (semi-)consonne	Comme dans la source	*ie*	
Correction	Rature lisible	Entre crochets blancs (‹[]›)	*[ie ne]*	
Correction	Rature partielle	Entre crochets blancs (‹[]›)	*endroit[s]*	
Correction	Rature illisible	Entre crochets blancs (‹[]›) (avec un point par lettre)	*[.....]*	
Correction	Insertion	Chevron d'insertion (‹˅›, U+2038)	*mourir, ˅ que ˅ quicelles*	
Fin de ligne	Trait de conduite	Rallonge de ligne (‹⟶›, U+23AF)	*la grande —*	
Fin de ligne	Ornementation	Rallonge de ligne (‹⟶›, U+23AF)	*a conclû qu'il —*	
Fin de ligne	Tiret(s) de fin	Signe négation (‹¬›, U+00AC)	*habi¬, Beur¬*	
Fin de ligne	Tiret(s) de début	Signe négation (‹¬›, U+00AC)	*¬veilloyent*	

TABLEAU 1 – Principales règles de transcription

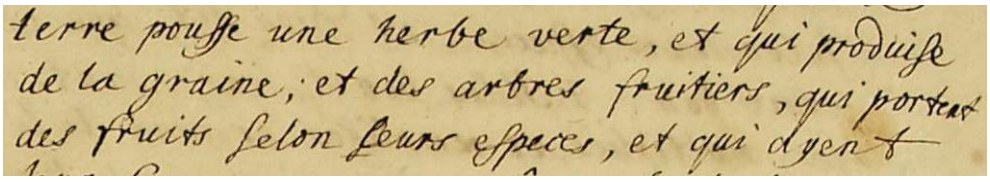
Figure produite par les auteurs

Les lettres

Les allographes

22 Les différents allographes ne sont pas notés – contrairement à ce qui peut être proposé dans les recommandations pour les imprimés (Gabay, Clérice et Reul 2023). Ainsi, la gamme des variations allant du *s* « long » (‹ſ›, normalement utilisé à l'initiale et en interne) au *s* « rond » (‹s›, plutôt utilisé en finale) est réduite au caractère que nous connaissons aujourd'hui (figure 5). Les cas similaires, comme les allographes avec un jambage plongeant (par exemple, ‹ŋ›, voir figure 2a), suivent le même principe.

Figure 5. Copie d'imprimeur, Aschaffenbourg, Hofbibliothek, Ms. 48



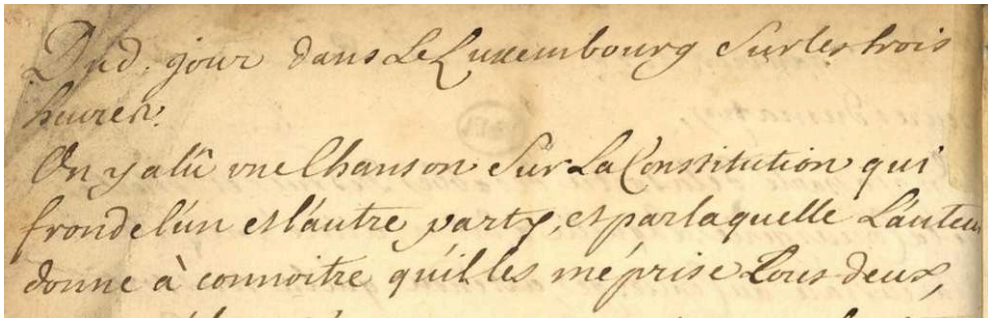
Aschaffenbourg, Hofbibliothek

> terre pousse une herbe verte, et qui produise
> de la graine; et des arbres fruitiers, qui portent
> des fruits selon leurs especes, et qui dyent

Les majuscules

23 Autant que possible, nous conservons l'usage des majuscules et des minuscules de la source, mais certaines variations de module ou des morphologies très proches de la majuscule rendent illusoire d'arriver à enseigner ces distinctions à la machine, faute d'homogénéité dans les données de vérité de terrain, mais aussi aux humains. Dans le doute, un module plus grand ou une morphologie proche de la majuscule peut être suffisant pour une transcription comme majuscule (cf. notamment le cas de ‹l› *vs* ‹L› et ‹c› *vs* ‹C› dans la figure 6).

Figure 6. Archives de la Bastille, gazetin de la police secrète, Paris, BNF, Arsenal, Ms. 10158, f. 8v



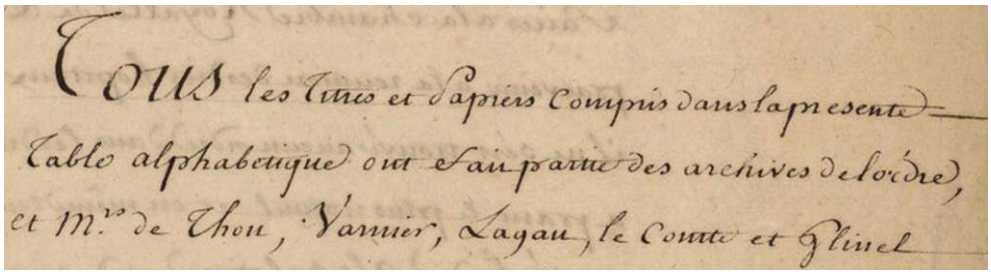
gallica.bnf.fr – BNF

> Dud. jour dans le Luxembourg sur les trois
> heures.
> On y a lû vne chanson sur la Constitution qui
> fronde l'un et l'autre party, et par laquelle l'auteur
> donne à connoitre qu'il les méprise tous deux,

La variation de module

24 Les petites majuscules, ou toutes les variations de module tendant à agrandir volontairement la lettre, sont transcrites avec des majuscules (figure 7).

Figure 7. Archive de chancellerie, Paris, BNF, Arsenal, Ms. 6118, f. 2r



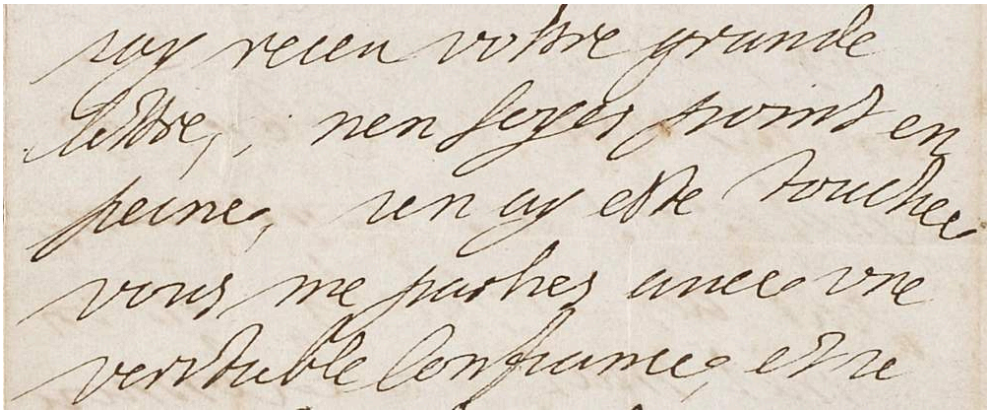
gallica.bnf.fr – BNF

> TOUS les Titres et Papiers Compris dans la presente —
> Table alphabetique ont fait partie des archives de l'ordre,
> et Mrs de Thou, Varnier, Lagau, le Comte et Glinel

Les lettres ramistes

25 Dans la mesure où l'utilisation du ‹v› et du ‹j› pour les sons consonnes et du ‹u› et du ‹i› pour les sons voyelles (ou semi-voyelles) s'impose lentement entre le XVI^e^ et le XVIII^e^ siècle, il est recommandé de conserver le système de la source (figure 8).

Figure 8. Lettre de Sévigné, Musée de Grignan, n° 1329



Musée départemental du château, Grignan · Licence Etalab

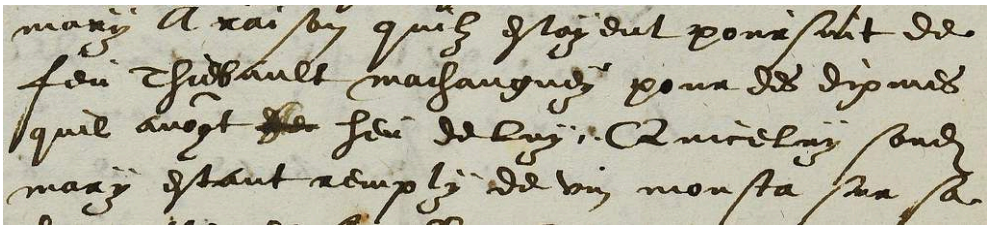
iay receu vostre grande
lettre, nen soyes point en
peine, ien ay este touchee
vous me parles auec vne
veritable confiance, et ie

L'effacement et l'enrichissement des lettres

Les signes auxiliaires

26 Les accents, trémas, apostrophes et traits d'union, etc. sont conservés tels quels. Attention, la présence de diacritiques est parfois due au contact avec une autre langue (comme dans la figure 9 avec l'allemand) : dans ces rares cas, il est possible de ne pas les conserver.

Figure 9. Procès de sorcellerie, Porrentruy, AAEB, B 168/15-10.3 (1609)



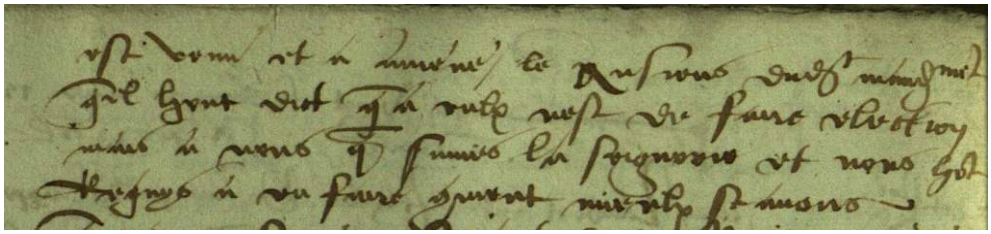
Archives de l'ancien évêché de Bâle · Licence CC BY

mary A raison quilz estoyent poursuit de
feu Thiebault machanguey pour des dixmes
quil auoyt ⟦..⟧ heu de luy, Quiceluy sonq
mary estant remply de vin monsta sur sa

Les abréviations

27 Pour les abréviations, nous suivons les recommandations de CATMuS : elles ne sont pas développées, des tentatives de développement posant *in fine* des problèmes de généralisation (Aguilar et Jolivet 2023). Le macron et le tilde sont transcrits avec un tilde pour des raisons pratiques d'accessibilité du signe sur le clavier. Pour les lettres suscrites, on utilise des caractères suscrits dans Unicode, mais il est aussi toléré d'utiliser le signe ^ (U+005E) placé avant les lettres suscrites (figure 10). Les caractères de la MUFI (*Medieval Unicode Font Initiative*[8]) sont recommandés pour couvrir les cas qui ne sont pas prévus par le standard Unicode, même s'il reste préférable de s'en tenir à ce dernier.

Figure 10. Procès-verbaux des séances du Conseil, archives de l'État de Genève, R.C. 29, 20 avril et 3 mai 1536



Archives de l'État de Genève · Licence CC BY

est venu et a amené le ansiens dud^{t} mand$^{m\tilde{e}t}$
q̃il hont dict q̃a eulx nest de faire election
mais a nous q̃ sumes la seignorie et nous hõt
requys a en faire 9ment mieulx scauons

28 ou

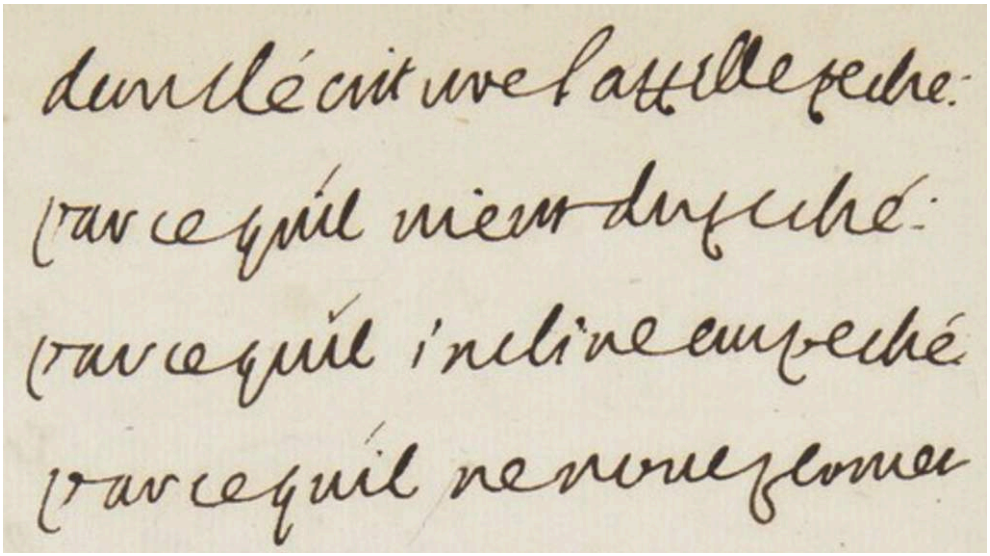
est venu et a amené le ansiens dud^t mand^mẽt
q̃il hont dict q̃a eulx nest de faire election
mais a nous q̃ sumes la seignorie et nous hõt
requys a en faire 9ment mieulx scauons

La séquence de lettres

Les levés de plume, l'espacement

29 Les levés de plume ne sont pas reproduits. Concernant la segmentation, la situation est complexe dans la mesure où l'absence de soudure peut témoigner d'un ancien état de la langue. Tant que la segmentation fait sens, elle n'est pas retouchée (par exemple *par ce que* mais *ecriture* dans la figure 11).

Figure 11. Manuscrit de Bossuet, Paris, BNF, Fr. 12820, f. 6ᵛ



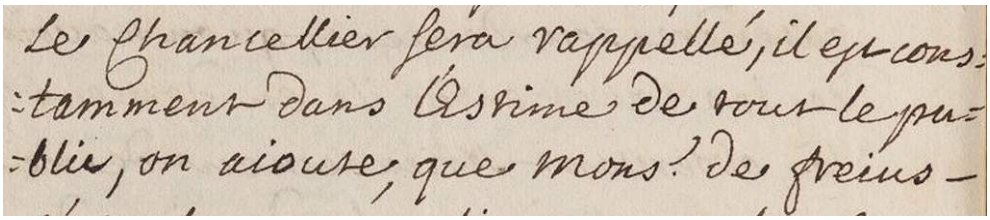
gallica.bnf.fr – BNF

dans l'ecriture sappelle peche:
par ce qu'il uient du peché:
par ce qu'il incline au peché
par ce qu'il ne nous permet

Le tiret de fin de ligne

30 Il convient de distinguer les différents tirets. Le trait d'union (‹-›, U+002D) n'est pas un tiret de fin ou de début de ligne (‹¬›, U+00AC) : cette distinction permet de simplifier le post-traitement des données, en reformant un mot dont la fin est rejetée sur la ligne d'après pour des raisons de place. De la même manière, il ne faut pas transcrire le trait de conduite (qui comble une espace vide, par exemple en fin de ligne) ou tout autre symbole équivalent par un (semi-)cadratin, mais par le caractère Unicode de la rallonge de ligne (‹⏤›, U+23AF, figure 12) – ce signe ayant pour vocation à disparaître dans une édition.

Figure 12. Archives de la Bastille, gazetin de la police secrète, Paris, BNF, Arsenal Ms. 10155, f. 6ᵛ



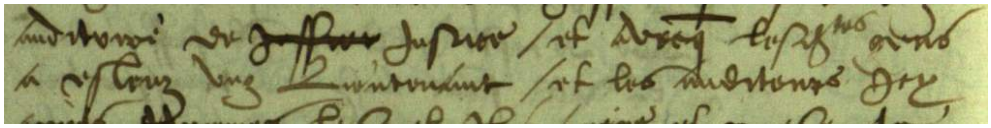
gallica.bnf.fr – BNF

le Chancellier sera rappellé, il est cons¬
¬tamment dans l'estime de tout le pu¬
¬blic, on aioute, que mons. de freius —

Ponctuation

31 La ponctuation est modernisée si besoin : la *virgula* est conservée sous la forme de virgule (le signe ‹/› étant réservé à la diastole). Les signes de ponctuation modernes (deux-points, point-virgule…) sont conservés (figure 13). On ne laisse pas d'espace avant le point-virgule, le deux-points et les points d'interrogation et d'exclamation.

Figure 13. Procès-verbaux des séances du Conseil, Archives de l'État de Genève, R.C. 29, 18 mai 1536



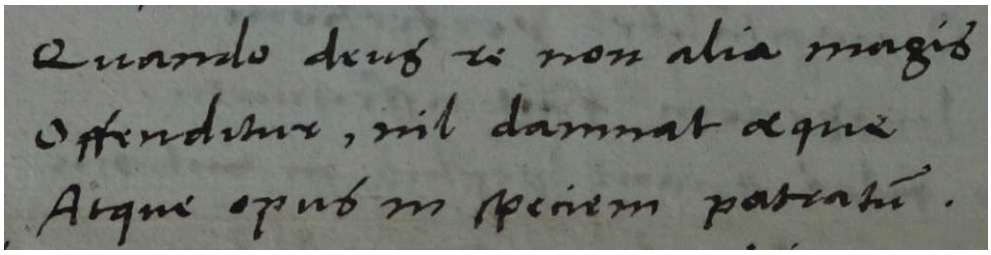
Archives de l'État de Genève · Licence CC BY

auditoire de ⟦jussier⟧ justice, et aveq̃ lesd^{tes} gens
a esleuz vng Lieutenant, et les auditeurs jcy

Les ligatures

32 Accompagnant le mouvement baroque, les ligatures se développent abondamment vers la fin du XVI^{e} siècle. Elles ne sont pas transcrites, sauf dans le cas où elles se sont maintenues aujourd'hui. C'est notamment le cas des nexus ‹œ› et ‹æ› (figure 14) qui sont donc conservés, mais ne sont pas réintroduits s'ils sont absents.

Figure 14. Archives de la ville de Strasbourg, 1 AST 212 n° 59



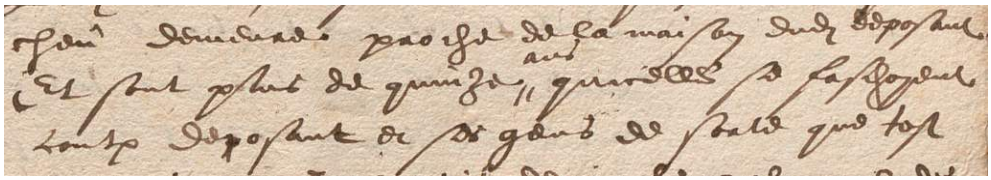
Archives de la ville de Strasbourg · Licence Etalab

Quando deus re non alia magis
Offenditur, nil damnat æque
Atque opus in speciem patratũ.

Les signes fonctionnels

33 Les signes fonctionnels sont conservés. Les cas très divers que recouvre cette catégorie (astérisque ‹*›, paragraphe ‹§›, chevron d'insertion ‹ ˆ ›, croix et croisette ‹※›, etc.) sont transcrits, avec pour recommandation de ne pas s'attacher à leur forme mais de les réduire à des types, si possible facilement accessibles sur le clavier et disponibles dans la plupart des polices standard (figure 15).

Figure 15. Procès de sorcellerie Porrentruy, AAEB, B 168/15-2.2 (1608)



heû demeure proche de la maison dud deposant
Et sont plus de quinze ˛ quicelles se faschoyent
contre deposant et ses gens de sorte que tost

Création d'un nouveau modèle

Données

34 En nous appuyant sur plusieurs projets de recherche, nous avons rassemblé une vérité de terrain en français couvrant quatre siècles, à laquelle a été ajoutée une sélection de données d'entraînement écrites dans d'autres langues (espagnol, latin, allemand et néerlandais, voir tableau 1). Nombre de documents ne sont en effet pas écrits en une seule langue et un tel mélange permet d'augmenter le nombre de mains. Parmi les textes en français, on trouve essentiellement des documents d'archives, mais aussi des manuscrits littéraires (comme celui de Zola) ou assimilés (copie d'imprimeur).

Tableau 1. Détails des données utilisées pour l'entraînement

Projet	**Langue**	**Siècle**	**Pages**	**Lignes**	**Dépôt GitHub/Zenodo**
FoNDUE†*	la	XVI^e^ s.	49	2 003	[privé]
FoNDUE†*	fr	XVII^e^ s.	63	1 668	[privé]
FoNDUE†	fr	XVIII^e^ s.	153	4 733	FONDUE-FR-MSS-18
FoNDUE†	de	XVIII^e^ s.	17	534	FONDUE-DE-MSS-18
FoNDUE†	fr	XIX^e^ s.	114	1 540	FONDUE-FR-MSS-19
FoNDUE†*	fr	XIX^e^ s.	81	2 711	[privé]
VWTM‡*	fr	XIX^e^-XX^e^ s.	190	28 611	[privé]
LECTAUREP‡	fr	XIX^e^-XX^e^ s.	104	20 304	lectaurep-mariages-et-divorces
LECTAUREP‡	fr	XIX^e^-XX^e^ s.	218	29 410	lectaurep-repertoires
Sous-total			**989**	**91 523**	
StABS†	de	XVI^e^ s.	198	8 221	zenodo.5153263

GLOBALISE†‡	nl	XVII^e-XVIII^e s.	3 263	137 414	zenodo.4159268
Conseil fédéral de suisse†		XIX^e-XX^e s.	NA	2 752	zenodo.4746341
Araucania†	es	XIX^e s.	145	3 249	HTR_Araucania_XIX
Sous-total			**3 606**	**151 636**	
Total			**4 604**	**243 159**	

Le symbole * signale les données qui ne sont pas (encore) distribuées de manière ouverte ; le symbole †, les données qui contiennent des données rédigées (lettres, rapports…) ; le symbole ‡, les données tabulaires (le format expliquant le nombre très élevé de lignes par page). La partie supérieure du tableau décrit les données dont nous maîtrisons la transcription, la partie inférieure celles pour lesquelles nous ne maîtrisons pas la transcription (la correction est beaucoup plus complexe pour des raisons linguistiques). NA signifie que le décompte est impossible (seules des lignes sont distribuées, pas les pages complètes).

35 Toutes les données n'étant pas distribuées dans le même format, nous convertissons tous les fichiers au format ALTO. Pour les transcriptions effectuées dans Transkribus, les masques sont recalculés avec Kraken pour en obtenir des similaires à ceux produits dans eScriptorium et la ligne de base est remontée de quelques pixels si besoin pour les mêmes raisons.

Entraînements

36 Nous entraînons notre modèle avec Kraken (version 5.2.1). Nous utilisons des données binaires précompilées pour simplifier le partage en minimisant la lisibilité de données personnelles (mais jamais sensibles, ni sous droit). Les hyperparamètres retenus sont les suivants : taux d'apprentissage de $1e^{-4}$, patience de 10 pour l'arrêt après plateau, normalisation Unicode NFC, taille de batch de 16. Nous utilisons l'augmentation de données et une précision mixte automatique.

37 Les données sont réparties en trois jeux. Celui pour le test (3 % du total) est sélectionné à la main. Le reste est attribué aléatoirement au jeu d'entraînement (90 %) et à celui d'évaluation (10 %). Deux types de données hors domaine sont conservés pour d'autres expériences : un jeu en français du XIX^e siècle et un autre en allemand du XVIII^e siècle (voir plus bas « Évaluation des capacités d'ajustement du modèle »).

Évaluation du modèle

38 Plusieurs expériences ont été menées pour estimer les performances du modèle obtenu à l'issue de l'entraînement. Outre une évaluation générale sur le jeu de test, quatre pages par projet ont été extraites pour une analyse plus détaillée des résultats. Trois expériences sont menées :

- avec le modèle Manu McFrench
- avec le modèle Manu McFrench ajusté (en anglais *fine tuned*) sur les données provenant du même jeu que pour le test (quatre pages en test, quatre pages en évaluation, le reste en entraînement)
- avec notre nouveau modèle Manu McFondue[9]

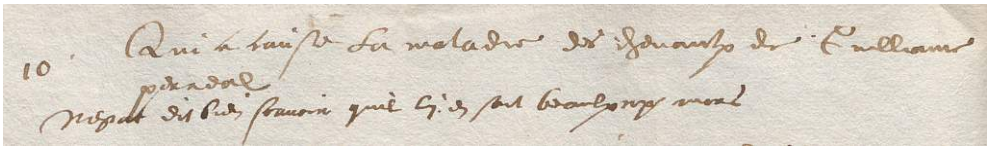
Tableau 2. Détail des résultats par jeux de données

Jeu de données	Manu McFrench (en %)	Manu McFrench ajusté (en %)	Manu McFondue (en %)
Général	66,02		89,9
AAEB	52,96	89,92	81,49
Aschaffenbourg	87,30	86,87	95,03
Arsenal	80,06	90,99	85,27
Archives privées	61,89	94,71	91,12
Contrats de mariage	92,92	NA	93,52
Actes notariés	90,43	NA	88,17
Zola	78,26	89,84	83,21
VWTM	65,17	93,63	90,12

NA signifie que les données sont déjà présentes dans Manu McFrench ou qu'aucune donnée supplémentaire n'est disponible, et donc qu'aucun *fine tuning* n'a été tenté.

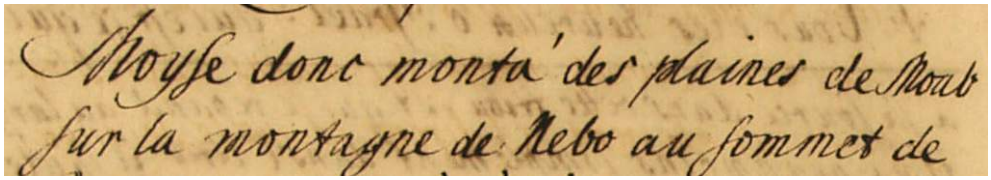
39 L'analyse des résultats (tableau 2) montre un net gain avec Manu McFondue par rapport à Manu McFrench (66,02 % *vs* 89,9 %). Les gains les plus importants sont obtenus sur la gothique tardive, dans une version peu formalisée avec de nettes influences germaniques (figure 16a). L'écriture italienne (figure 16b), relativement bien traitée par Manu McFrench (87,30 %), se trouve encore mieux traitée par Manu McFondue (95,03 %, soit + 7,73 pt de %). Pour la bâtarde coulée (figure 16c), le gain est plus faible mais non négligeable (+ 5,21 pt de %).

Figure 16a. Exemple nº 1 des données utilisées en test pour évaluer la performance du modèle : procès de sorcellerie, Porrentruy, AAEB, B 168/15-2.3 (1608)



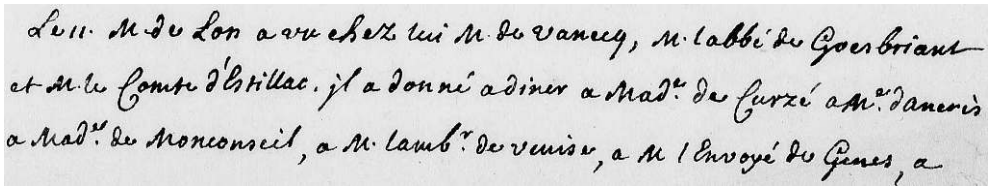
Archives de l'ancien Évêché de Bâle · Licence CC BY

Figure 16b. Exemple nº 2 des données utilisées en test pour évaluer la performance du modèle : R. Simon, *Pentateuque traduit*, copie d'imprimeur, Aschaffenbourg, Hofbibliothek, Ms. 48. (vers 1709)



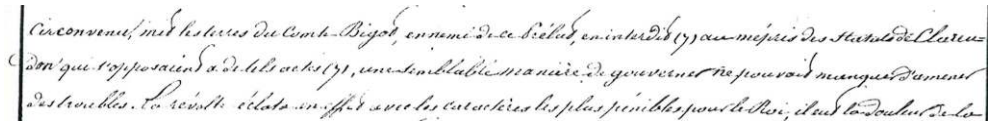
Aschaffenbourg, Hofbibliothek

Figure 16c. Exemple nº 3 des données utilisées en test pour évaluer la performance du modèle : archives de la Bastille, rapports envoyés au lieutenant de police, Paris, BNF, Arsenal, Ms. 10292 (1746)



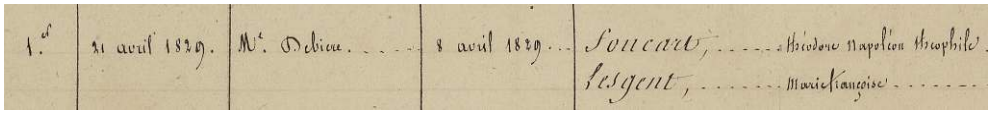
gallica.bnf.fr – BNF

Figure 16d. Exemple nº 4 des données utilisées en test pour évaluer la performance du modèle : archives privées (1818)



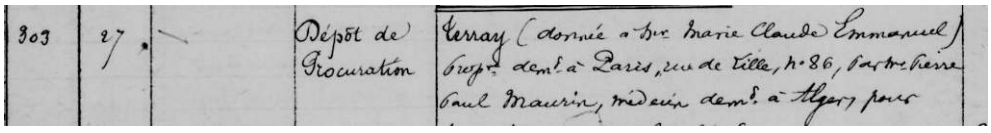
Anonyme

Figure 16e. Exemple nº 5 des données utilisées en test pour évaluer la performance du modèle : contrats de mariage, Paris, Archives nationales, CM/1 (1829)



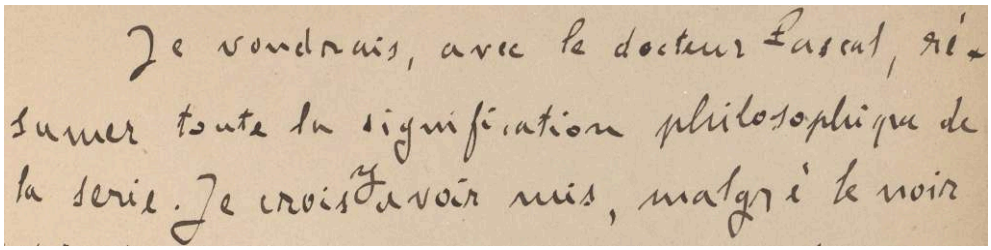
Archives nationales

Figure 16f. Exemple nº 6 des données utilisées en test pour évaluer la performance du modèle : actes notariés, Paris, Archives nationales, MC/RE/XLVIII/22 (1850)



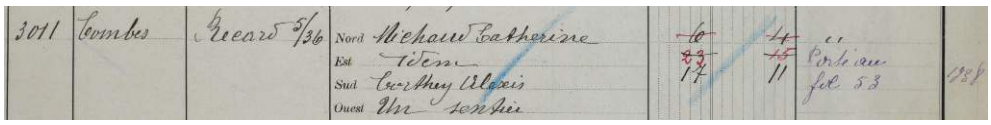
Archives nationales

Figure 16g. Exemple nº 7 des données utilisées en test pour évaluer la performance du modèle : Zola, *Le Docteur Pascal*, Cologny, Fondation Bodmer, Z-6.3* (1893)



Fondation Martin Bodmer

Figure 16h. Exemple nº 8 des données utilisées en test pour évaluer la performance du modèle : registre de l'impôt sur les biens-fonds, Le Châble, Archives communales, AC Bagnes R 72 (1894-1984)



Archives de la commune de Val de Bagnes

40 Concernant les écritures du XIXe siècle, pour lesquelles on voit bien l'apparition du trait caractéristique de la plume anglaise (sauf le manuscrit de Zola, figure 16g), les résultats sont globalement positifs. On observe un très net gain pour l'écriture anglaise dans une version assez formelle (figure 16d, + 29,23 pt de %) et un gain plus modeste pour les documents présentant un mélange de bâtarde et de ronde (figure 16e, + 0,6 pt de %). La seule contre-performance concerne une écriture majoritairement anglaise et assez peu formelle (figure 16f, — 2,26 pt de %), mais qui était déjà présente dans les données d'entraînement de Manu McFrench. Le manuscrit de Zola (figure 16g), qui utilise un lointain dérivé assez personnel de l'écriture anglaise (non liée, tracée avec une plume épaisse), montre à nouveau des résultats positifs (78,26 % *vs* 83,21 %). Enfin, les archives valaisannes (figure 16h), rédigées avec une écriture plutôt anglaise (pour les quatre premières colonnes) mais avec de nombreux ajouts postérieurs d'autres mains et beaucoup de chiffres, le gain est à nouveau très substantiel (+ 24,95 pt de %).

41 En revanche, ajuster Manu McFrench semble une meilleure stratégie que l'utilisation de Manu McFondue. La question est donc de savoir si la performance inférieure du nouveau modèle est due à une faiblesse intrinsèque (par exemple, la trop grande hétérogénéité des données), ou si l'ajustement de Manu McFrench ne donne de meilleurs résultats qu'en apparence – un ajustement de Manu McFondue étant en réalité plus efficace.

Évaluation des capacités d'ajustement du modèle

42 Afin de répondre à cette dernière interrogation, nous tentons d'ajuster Manu McFrench puis Manu McFondue :

- à un jeu de données tiré d'archives françaises du XIXe siècle [10]assez similaire à celles utilisées pour entraîner Manu McFrench et Manu McFondue, mais inconnu des deux modèles
- à un jeu de données suisses du XVIIIe siècle en allemand[11], là encore inconnu des deux modèles

43 Nous utilisons huit pages pour l'entraînement, deux pages pour la validation et cinq pages pour le test. Les résultats (tableau 3) montrent que Manu McFondue est plus efficace une fois ajusté, notamment pour des documents en langue étrangère, et que les meilleures performances de Manu McFrench présentées dans le tableau 2 sont certainement dues à l'ajustement.

Tableau 3. Expérience d'ajustement sur un jeu de données hors domaine, avec quatre puis huit pages en entraînement

Jeu de données	Manu McFrench (en %)	Manu McFondue (en %)
Français	88,48	89,40
Allemand	71,01	78,07

Conclusion

44 Le modèle Manu McFondue est encore loin de lever toutes les difficultés liées à la reconnaissance automatique des manuscrits modernes en français, mais il constitue, en dépit de ses nombreux défauts, une avancée significative dans la résolution du problème posé par ces documents. La stratégie consistant à agréger des données issues de plusieurs sources différentes, y compris en langue étrangère, semble plus efficace que la conception de modèles spécialisés par langue ou par type d'écriture. Toutefois, nos résultats montrent qu'il reste nécessaire d'augmenter considérablement la taille et la diversité du jeu d'entraînement afin d'améliorer le traitement des manuscrits en français. Par ailleurs, l'existence de textes contenant plusieurs langues dans les zones frontalières (par exemple, en Suisse) ou non (passages en latin, etc.) nous invite également à augmenter la quantité de vérité de terrain non francophone, dont la présence semble avoir un effet bénéfique sur les performances du modèle pour les documents en français.

45 Afin de mieux gérer une telle quantité de données, il nous faut continuer d'affiner notre description des mains, probablement en utilisant des méthodes de classification automatique (Stutzmann *et al.* 2020), surtout si l'enrichissement du corpus de travail passe par l'ajout de documents provenant d'autres espaces linguistiques. C'est à cette condition que nous pourrons vraiment clarifier la couverture du modèle final, dont la performance ne peut être résumée à un simple pourcentage. Une telle approche devrait également nous permettre d'optimiser l'augmentation du volume de données pour l'entraînement en ciblant prioritairement les types d'écriture les moins bien traités.

46 Il nous faut enfin consolider nos choix pour résoudre les différents problèmes de transcription rencontrés par les chercheurs et les chercheuses. Le classement que nous avons proposé ici (les lettres, leur enrichissement et leur effacement, leur séquençage) reste imparfait. S'il offre une première organisation de nos recommandations, des cas manquent et des doutes subsistent : il convient donc de l'améliorer et de le compléter. Ce faisant, il sera crucial d'aligner encore mieux nos règles de transcription avec celles proposées par le projet CATMuS, qui bénéficie d'une solide expérience dans la création de normes communes pour des données très hétérogènes[12].

BIBLIOGRAPHIE

Aguilar, Sergio Torres et Vincent Jolivet. 2023. « La reconnaissance de l'écriture pour les manuscrits documentaires du Moyen Âge ». *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.10484.

Baddeley, Susan. 1998. « Théorie et pratique de la segmentation graphique dans les textes français du premier tiers du XVI^e^ siècle ». *Langue française* 119 : 52-68. https://doi.org/10.3406/lfr.1998.6259.

Barbedor, Louis. 1649. *Les Ecritures financiere et italienne bastarde dans leur naiveté*. Paris : Nicolas Langlois. https://french.newberry.t-pen.org/www/record.html?id=https://iiif.library.utoronto.ca/presentation/v2/paleography:527/manifest.

Barbier, Frédéric. 2006. *L'Europe de Gutenberg. Le livre et l'invention de la modernité occidentale (XIII^e^-XVI^e^ siècle)*. Paris : Belin.

Bawden, Rachel, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot et Simon Gabay. 2022. « Automatic Normalisation of Early Modern French ». Dans *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, édité par Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, *et al.*, 3354-3366. European Language Resources Association. https://aclanthology.org/2022.lrec-1.358.

Beck, Friedrich. 1991. « Die „Deutsche Schrift" – Medium in fünf Jahrhunderten deutscher Geschichte ». *Archiv für Diplomatik* 37 : 453-479.

Bergeron-Maguire, Myriam. 2019. « Du Poitou en Louisiane : édition et notes à partir de la correspondance d'une peu lettrée (1802-1803) ». *Géolinguistique* 19. https://doi.org/10.4000/geolinguistique.1530.

Bishop, Marie-France. 2019. « L'enseignement de l'écriture à l'école primaire française de 1880 aux années 2000 ». Dans *L'Écriture dès le début de l'école primaire*, édité par Bernadette Kervyn, Martine Dreyfus et Catherine Brissaud, 19-39. Pessac : Presses universitaires de Bordeaux. https://doi.org/10.4000/books.pub.36757.

Branca-Rosoff, Sonia et Nathalie Schneider. 1994. *L'Écriture des citoyens. Une analyse linguistique de l'écriture des peu-lettrés pendant la période révolutionnaire*. Paris : Klincksieck.

Cabane, Célia. 2020. « Les maîtres écrivains : acteurs méconnus de la transmission des savoirs ». Dans *Écriture et transmission des savoirs de l'Antiquité à nos jours*, édité par Dominique Briquel. Paris : Éditions du Comité des travaux historiques et scientifiques. https://doi.org/10.4000/books.cths.8216.

Cascianelli, Silvia, Vittorio Pippi, Martin Maarand, Marcella Cornia, Lorenzo Baraldi, Christopher Kermorvant et Rita Cucchiara. 2022. « The LAM Dataset : a Novel Benchmark for Line-Level Handwritten Text Recognition ». Dans *26th International Conference on Pattern Recognition (ICPR)*, 1506-1513. Montréal : Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICPR56361.2022.9956189.

Catach, Nina. 1994. *La Ponctuation (histoire et système)*. Paris : Presses universitaires de France.

Catach, Nina. 2001. *Histoire de l'orthographe française*. Édition posthume réalisée par Renée Honvault. Paris : H. Champion.

Catach, Nina et Jeanne Golfand. 1973. « L'orthographe plantinienne ». *De Gulden Passer* 50 : 19-69.

Cazal, Yvonne et Gabriella Parussa. 2022. *Introduction à l'histoire de l'orthographe*. Paris : Armand Colin. https://shs.cairn.info/introduction-a-l-histoire-de-l-orthographe--9782200626518.

Chagué, Alix et Thibault Clérice. 2022. « HTR-United – Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French) ». *Zenodo*. https://zenodo.org/records/6657809.

Chagué, Alix et Thibault Clérice. 2023. « "I'm Here to Fight for Ground Truth" : HTR-United, a Solution Towards a Common for HTR Training Data ». Dans *DH2023 Book of Abstracts*, édité par Anne Baillot, Toma Tasovac, Walther Scholger et Georg Vogeler. Graz : Alliance of Digital Humanities Organizations. https://inria.hal.science/hal-04094233.

Chagué, Alix, Thibault Clérice et Laurent Romary. 2022. « HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites ». Document de travail. https://inria.hal.science/hal-04124743.

Chagué, Alix et Aurélia Rostaing. 2021. « Lectaurep : Lecture automatique des répertoires de notaires parisiens ». Communication présentée à *Les Futurs fantastiques*, Paris, 8-10 décembre. https://inria.hal.science/hal-03479303.

Dancel, Brigitte. 2011. « Apprendre à écrire, quelle histoire ! » *Carrefours de l'éducation* 2 (4) : 123-134. https://doi.org/10.3917/cdle.hs02.0123.

Duval, Frédéric. 2015. « Les éditions de textes du XVII^e^ siècle ». Dans *Manuel de la philologie de l'édition*, édité par David Trotter, 369-394. Berlin et Boston : De Gruyter. https://doi.org/10.1515/9783110302608-017.

Frondizi, Alexandre et Emmanuel Fureix. 2022. « Introduction. Vous avez dit "écritures populaires" ? » *Revue d'histoire du XIX^e^ siècle* 65 : 9-22. https://doi.org/10.4000/rh19.8513.

Gabay, Simon. 2014. « Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVII^e^ siècle ». *Vox Romanica* 73 (1) : 27-42.

Gabay, Simon. 2020. « La naissance de Marie-Blanche de Grignan. Notes sur la mise en page de la polyphonie sévignéenne ». *Fabula*. https://dx.doi.org/10.58282/colloques.13218.

Gabay, Simon et Thibault Clérice. 2024. « The Birth of French Orthography. A Computational Analysis of French Spelling Systems in Diachrony ». Dans *Proceedings of the Computational Humanities Research Conference 2024*, édité par Wouter Haverals, Marijn Koolen et Laure Thompson, 246-264. Aarhus : CEUR Workshop Proceedings. https://inria.hal.science/hal-04704549.

Gabay, Simon, Thibault Clérice et Christian Reul. 2023. « OCR17 : vérité de terrain et modèles pour les imprimés français du XVII^e^ s. (voire un peu plus) ». *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.6492.

Gabay, Simon, Philippe Gambette, Rachel Bawden et Benoît Sagot. 2022. « Ancien ou moderne ? Pistes computationnelles pour l'analyse graphématique des textes écrits au XVII^e^ siècle ». *Linx* 85. https://doi.org/10.4000/linx.9346.

Gabay, Simon, Tobias Hodel, Ronald Sluijter, Élodie Paupe, Jean-Claude Rebetez, David Rabouin, Vincent Giovannangeli, Walter Boente, Élodie Bascoul, Marion Philip *et al.* 2025. « Transcribing Western Modern Manuscripts (1500-2020) ». Dans *DH2025 Book of Abstracts*. Lisbonne : Alliance of Digital Humanities Organizations. https://hal.science/hal-05063299v1.

Gasparri, Françoise. 1983. « Enseignement et techniques de l'écriture du Moyen Âge à la fin du XVI^e siècle ». *Scrittura e civiltà* 7 : 201-222.

Heal, Ambrose. 1931. *The English Writing-Masters and Their Copy-Books, 1570-1800. A Biographical Dictionary and a Bibliography*. Cambridge : Cambridge University Press.

Hébrard, Jean. 1995. « Des écritures exemplaires : l'art du maître écrivain en France entre XVI^e et XVIII^e siècle ». *Mélanges de l'école française de Rome* 107 (2) : 473-523. https://doi.org/10.3406/mefr.1995.4394.

Hodel, Tobias et David Schoch. 2021. « Handwritten Text Recognition Test Set : Minutes of the Swiss Federal Council (1848-1903) ». *Zenodo*. https://doi.org/10.5281/zenodo.4746342.

Hodel, Tobias, David Schoch et Peter Dängeli. 2021. « Handwritten Text Recognition Ground Truth Set : StABS Ratsbücher O10, Urfehdenbuch X ». *Zenodo*. https://doi.org/10.5281/zenodo.5153263.

Hodel, Tobias, David Schoch, Christa Schneider et Jake Purcell. 2021. « General Models for Handwritten Text Recognition : Feasibility and State-of-the Art. German Kurrent as an Example ». *Journal of Open Humanities Data* 7. https://doi.org/10.5334/johd.46.

Huchon, Mireille. 1983. « Rabelais et les majuscules ». *Études rabelaisiennes* 17 : 99-113.

Kahle, Philip, Sebastian Colutto, Günter Hackl et Günter Mühlberger. 2017. « Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents ». Dans *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 04 : 19-24. Kyoto : Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/icdar.2017.307.

Keijser, Liesbeth. 2020. « 6000 Ground Truth of VOC and Notarial Deeds 3.000.000 HTR of VOC, WIC and Notarial Deeds ». *Zenodo*. https://doi.org/10.5281/zenodo.6414086.

Kiessling, Benjamin. 2019. « Kraken – An Universal Text Recognizer for the Humanities ». Dans *DH2019 Book of Abstracts*, édité par Elena Pierazzo et Fabio Ciotti. Utrecht : Alliance of Digital Humanities Organizations. https://doi.org/10.34894/Z9G2EX.

Le Gangneur, Guillaume. 1599. *La Technographie, ou Briève méthode pour parvenir à la parfaitte connoissance de l'écriture françoyse*. http://gallica.bnf.fr/ark:/12148/btv1b8609557v.

Materot, Lucas. 1608. *Les Oeuvres de Lucas Materot,... où l'on comprendra facilement la manière de bien et proprement escrire toute sorte de lettre italienne selon l'usage de ce siècle*. Avignon : J. Bramereau. http://gallica.bnf.fr/ark:/12148/btv1b86220474.

Métayer, Christine. 2001. « Normes graphiques et pratiques de l'écriture. Maîtres écrivains et écrivains publics à Paris aux XVII^e et XVIII^e siècles ». *Annales. Histoire, Sciences Sociales* 56 (4/5) : 881-901. https://doi.org/10.3406/ahess.2001.279992.

Nahon, Peter et Simon Gabay. 2023. « Modernités de Richard Simon : notes philologiques en vue d'une édition du *Pentateuque traduit, avec des Remarques* (Bibliothèque d'Aschaffenbourg, Ms. 48) ». *Dix-septième siècle* 300 (3) : 481-500. https://doi.org/10.3917/dss.233.0481.

Paillasson, Charles. 1763. « Écritures, contenant seize planches ». Dans *Recueil des planches sur la science, les arts libéraux et les arts méchaniques, avec leur explication. Seconde livraison en deux parties. Première partie. Encyclopédie ou Dictionnaire raisonné des Sciences, des Arts et des Métiers*. Paris : Briasson, David, Le Breton et Durand. https://gallica.bnf.fr/ark:/12148/bpt6k505560.

Pellat, Jean-Christophe. 1998. « Les mots graphiques dans des manuscrits et des imprimés du XVII^e siècle ». *Langue française* 119 (1) : 88-104. https://doi.org/10.3406/lfr.1998.6261.

Pinche, Ariane. 2022. « Guide de transcription pour les manuscrits du xe au xve siècle ». Document de travail. https://hal.science/hal-03697382.

Pinche, Ariane, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, *et al.* 2023. « CATMuS Medieval ». *Zenodo*. https://zenodo.org/records/10066219.

Pinche, Ariane, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, *et al.* 2024. « CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts ». Dans *DH2024 Book of Abstracts*, édité par Jajwalya Karajgikar, Andrew Janco, Jessica Otis. Washington : Alliance of Digital Humanities Organizations. https://inria.hal.science/hal-04346939.

Poulle, Emmanuel. 1966. *Paléographie des écritures cursives en France du xve au xviie siècle. Recueil de fac-similés de documents parisiens avec leur transcription, précédé d'une introduction*. Genève : Droz.

Poulle, Emmanuel. 2007. « Aux origines de l'écriture liée : les avatars de la mixte (xive-xve siècles) ». *Bibliothèque de l'École des chartes* 165 (1) : 187-200. https://doi.org/10.3406/bec.2007.463495.

Raj, Ravi et Andrzej Kos. 2022. « A Comprehensive Study of Optical Character Recognition ». Dans *29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES)*, 151-154. Wrocław : Institute of Electrical and Electronics Engineers. https://doi.org/10.23919/MIXDES55591.2022.9837974.

Riffaud, Alain. 2007. *La Ponctuation du théâtre imprimé au xviie siècle*. Genève : Droz.

Samaran, Charles. 1967. « Cursives françaises des xve, xvie et xviie siècles [Compte-rendu] ». *Journal des Savants* 3 : 129-153. https://www.persee.fr/doc/jds_0021-8103_1967_num_3_1_1153.

Seguin, Jean-Pierre. 1998. « Les incertitudes du mot graphique au xviiie siècle ». *Langue française* 119 : 105-124. https://doi.org/10.3406/lfr.1998.6262.

Siouffi, Gilles. 2020. *Une histoire de la phrase française des* Serments de Strasbourg *aux écritures numériques*. Arles : Actes Sud.

Smith, Marc. 2020. « Les modèles d'apprentissage de l'écriture en France depuis la Renaissance ». Dans *Apprendre. Archéologie de la transmission des savoirs*, édité par Patrick Pion et Nathan Schlanger, 167-179. Paris : La Découverte. https://doi.org/10.3917/dec.pion.2020.01.0167.

Solfrini, Sonia, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia M. Oliveira et Daniela Solfaroli Camillocci. 2023. « Guide de transcription pour les imprimés français du xvie siècle en caractères gothiques ». Document de travail. https://hal.science/hal-04281804.

Solfrini, Sonia, Mylène Dejouy, Aurélia Marques Oliveira et Pierre-Olivier Beaulnes. 2025. « Normaliser le moyen français : du graphématique au semi-diplomatique ». Dans *Actes des 18es Rencontres jeunes chercheurs en RI (RJCRI) et de la 27e Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (RECITAL)*, édité par Frédéric Bechet, Adrian-Gabriel Chifu, Karen Pinel-Sauvagnat, Benoît Favre, Eliot Maes et Diana Nurbakova, 239-252. Marseille : Association pour le traitement automatique des langues. https://aclanthology.org/2025.jeptalnrecital-recital.13.

Sorel, Charles. 2014. *L'Anti-roman ou l'histoire du berger Lysis*, édité par Anne-Élisabeth Spica. Paris : Honoré Champion.

Steuckardt, Agnès. 2014. « De l'écrit vers la parole. Enquête sur les correspondances peu lettrées de la Grande Guerre ». *SHS Web of Conferences : 4e Congrès mondial de linguistique française* 8 : 353-364. https://doi.org/10.1051/shsconf/20140801159.

Stutzmann, Dominique. 2011. « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ? » Dans *Kodikologie und Paläographie im digitalen Zeitalter 2/Codicology and Palaeography in the Digital Age 2*, édité par Franz Fischer, Christiane Fritze et Georg Vogeler, 247-277. Norderstedt : BoD. https://halshs.archives-ouvertes.fr/halshs-00596970.

Stutzmann, Dominique, Christopher Tensmeyer et Vincent Christlein. 2020. « Writer Identification and Script Classification : Two Tasks for a Common Understanding of Cultural Heritage ». *Manuscript cultures* 15 : 11-24. https://www.csmc.uni-hamburg.de/publications/mc/files/articles/mc15-02-stutzmann.pdf.

Ullman, Berthold Louis. 1960. *The Origin and Development of Humanistic Script*. Rome : Edizioni di Storia e Letteratura.

Vieillard, Françoise et Olivier Guyotjeannin, éd. 2014. *Conseils pour l'édition des textes médiévaux*. Nouvelle édition revue et mise à jour. Paris : École nationale des chartes et Comité des travaux historiques et scientifiques.

NOTES

1. En plus d'« ATR », il est possible de trouver différentes dénominations. En effet, à l'origine, l'OCR (pour *Optical Character Recognition*) a renvoyé à une technique opérant à l'échelle du caractère, tandis que l'expression « HTR » (pour *Handwritten Text Recognition*) a été utilisée par une partie de la communauté pour désigner des outils opérant à l'échelle de la ligne – condition *sine qua non* du traitement de la cursive. Dans le but de rassembler OCR et HTR sous une dénomination commune, une partie de la communauté scientifique a proposé de parler d'*Automatic Text Recognition* (ATR). Nous suivons cette recommandation.
2. https://app.transkribus.org/models/public/text/french-general-model/.
3. https://htr-united.github.io.
4. De grandes quantités de données de vérité de terrain ne sont malheureusement pas encore disponibles de manière ouverte et un gros travail doit être mené sur ce point. Notons que, la situation évoluant extrêmement rapidement, notre état de l'art ne présente qu'un aperçu des données disponibles au moment de l'entraînement de notre modèle (mai 2024) et non au moment de la publication.
5. Concernant *w*, il faudra attendre le quatrième tome de la neuvième édition du *Dictionnaire* (2024) pour voir l'Académie reconnaître cette lettre comme appartenant à l'alphabet du français : https://www.dictionnaire-academie.fr/article/A9W0001.
6. On consultera avec intérêt le site du projet pour plus d'information : https://catmus-guidelines.github.io.
7. Étape qui ressemblerait à celle présentée par Bawden *et al.* (2022) ou Solfrini *et al.* (2025).
8. https://mufi.info.
9. Le modèle Manu McFondue est librement disponible : https://doi.org/10.5281/zenodo.6657808.
10. Archives départementales de La Réunion, IE 5, affranchissements.
11. Franz Joseph Leonti Meyer von Schauensee, *Ausführlich Musicalisches Protocoll*, Zentral-und Hochschulbibliothek Luzern : CH-Lz, PpMsc 167 fd.

12. Des premiers résultats de cette poursuite du travail ont déjà été publiés dans Gabay *et al.* 2025.

RÉSUMÉS

Dans le domaine francophone, le manuscrit écrit après le Moyen Âge reste le dernier type de document qui n'est pas correctement traité par les outils de reconnaissance automatique de texte. Si des modèles ont déjà été publiés, leur efficacité et leur documentation restent insatisfaisantes, en grande partie à cause des difficultés que suscite l'importante évolution des documents eux-mêmes au cours des siècles, et donc la diversité des formes à traiter. Après avoir décrit le problème d'un point de vue philologique, nous proposons ici quelques réflexions préliminaires sur la transcription des documents modernes, ainsi qu'un nouveau modèle visant à améliorer les conditions de travail des chercheurs et chercheuses, en attendant de concevoir une solution pleinement satisfaisante.

In the French-speaking world, manuscripts written after the Middle Ages remain the last type of document that is not properly processed by automatic text recognition tools. Although some models have already been published, their performance and documentation are still unsatisfactory, largely because of the difficulties posed by the significant evolution that documents have undergone over the centuries, and thus by the diversity of forms to be processed. After a description of the problem from a philological point of view, we offer here some preliminary reflections on the transcription of modern documents, as well as a new model aimed at improving the working conditions of researchers, until a truly satisfactory solution can be devised.

INDEX

Keywords : data acquisition, archives, good practices, transcription, manuscript
Mots-clés : acquisition de données, archives, bonnes pratiques, transcription, manuscrit

AUTEURS

SIMON GABAY

Université de Genève, Genève, Suisse
Simon Gabay est maître-assistant en humanités numériques à l'université de Genève.
ORCID 0000-0001-9094-4475
simon.gabay@unige.ch

ARIANE PINCHE

UMR 5648 Ciham, CNRS, Lyon, France
Ariane Pinche est chargée de recherche au CNRS.

ORCID 0000-0002-7843-5050
ariane.pinche@cnrs.fr

PETER NAHON

UMR 7323 CESR, CNRS, Tours, France
Peter Nahon est chargé de recherche au CNRS.
ORCID 0000-0001-9952-811X
peter.nahon@cnrs.fr

ALIX CHAGUÉ

INRIA, Paris, France
Alix Chagué est doctorante en humanités numériques à INRIA dans l'équipe-projet Almanach à Paris, en cotutelle avec l'université de Montréal au Canada et avec l'École pratique des hautes études à Paris.
ORCID 0000-0002-0136-4434
alix.chague@inria.fr

PAULINE JACSONT

Académie suisse des sciences humaines et sociales, Berne, Suisse
Pauline Jacsont est codicologue et spécialiste d'humanités numériques au sein du projet *codices.ch*.
ORCID 0000-0002-6296-3246
pauline.jacsont@bcu.unil.ch

ÉLODIE PAUPE

Archives de l'ancien évêché de Bâle, Porrentruy, Suisse
Élodie Paupe est cheffe du projet *Crimes et châtiments* aux archives de l'ancien évêché de Bâle.
ORCID 0000-0002-6283-3652
elodie.paupe@jura.ch

JEAN-CLAUDE REBETEZ

Archives de l'ancien évêché de Bâle, Porrentruy, Suisse
Jean-Claude Rebetez est le conservateur des archives de l'ancien évêché de Bâle.
jean-claude.rebetez@aaeb.ch

MAXIME HUMEAU

Université de Lausanne, Lausanne, Suisse
Maxime Humeau est ingénieur à l'université de Lausanne au sein du projet *Grand Siècle*.
ORCID 0000-0001-6860-0916
maxime.humeau@unil.ch

CHRISTINE PAYOT

Bureau Clio, Martigny, Suisse
Christine Payot est spécialiste en humanités numériques au sein du projet *Verbier Time Machine*.
ORCID 0000-0002-2499-229X
info@bureauclio.ch

THIBAULT MAILLARD

Dicastère Culture, tourisme et sport, Val de Bagnes, Suisse
Thibault Maillard est stagiaire au sein du projet *Verbier Time Machine*.

YVAN JAUREGUI

Université de Genève, Genève, Suisse
Yvan Jauregui est assistant-doctorant en histoire moderne à l'université de Genève.
ORCID 0009-0009-1838-603X
yvan.jauregui@unige.ch

ELINA LEBLANC

Université de Genève, Genève, Suisse
Elina Leblanc est collaboratrice scientifique et développeuse à l'université de Genève au sein du projet *SETAF*.
ORCID 0009-0009-4556-8840
elina.leblanc@unige.ch

LORAINE CHAPPUIS

Université de Lausanne, Lausanne, Suisse
Loraine Chappuis est maîtresse d'enseignement et de recherche en histoire moderne à l'université de Lausanne.
ORCID 0000-0002-9598-9151
loraine.chappuis@unil.ch