Vers un modèle diachronique pour les mains modernes françaises

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

Vers un modèle diachronique pour les mains modernes françaises

Simon Gabay[1], **Ariane Pinche**[2], **Peter Nahon**[3], **Alix Chagué**[7,8,9], **Pauline Jacsont**[1], **Élodie Paupe**[4], **Jean-Claude Rebetez**[5], **Maxime Humeau**[1], **Christine Payot**[6], **Thibault Maillard**[6], **Yvan Jauregui**[1], **Elina Leblanc**[1] et **Loraine Chappuis**[1]

[1]Université de Genève
{prenom.nom}@unige.ch
[2]CIHAM–UMR 5648, C.N.R.S., Lyon
{prenom.nom}@cnrs.fr
[3]CESR–UMR 7323, C.N.R.S., Tours
{prenom.nom}@cnrs.fr
[4]Office de la culture de la République et Canton du Jura
{prenom.nom}@jura.ch
[5]Archives de l'ancien Évêché de Bâle
{prenom.nom}@aaeb.ch
[6]Archives de l'Etat du Valais
[7]Inria Paris
{prenom.nom}@inria.fr
[8]Université de Montréal
[9]École Pratique des Hautes Études, Paris

Résumé

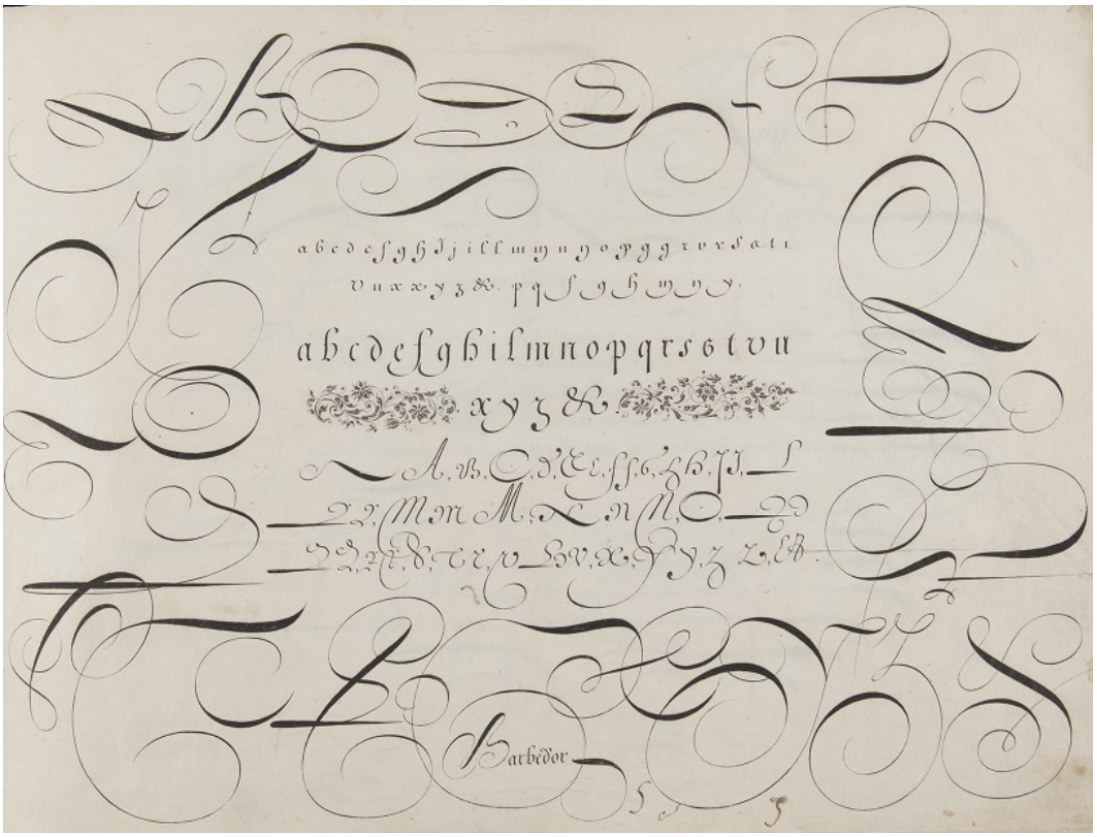
Pour le domaine francophone, les manuscrits rédigés après le Moyen Âge restent le dernier type de document qui n'est pas correctement traité par les moteurs de reconnaissance optique de caractères. Si des modèles ont déjà été publiés, leur efficacité et leur documentation restent encore insatisfaisants, en grande partie à cause des problèmes posés par l'importante évolution graphique (au sens paléographique comme linguistique) qu'a connu la langue au cours des siècles, et donc de la diversité des formes à traiter. Après une brève description du problème philologique, nous proposons donc ici quelques premières réflexions sur la transcription des documents modernes, ainsi qu'un nouveau modèle pour améliorer les conditions de travail des chercheur · se · s, le temps de concevoir une solution véritablement satisfaisante.

1 Introduction

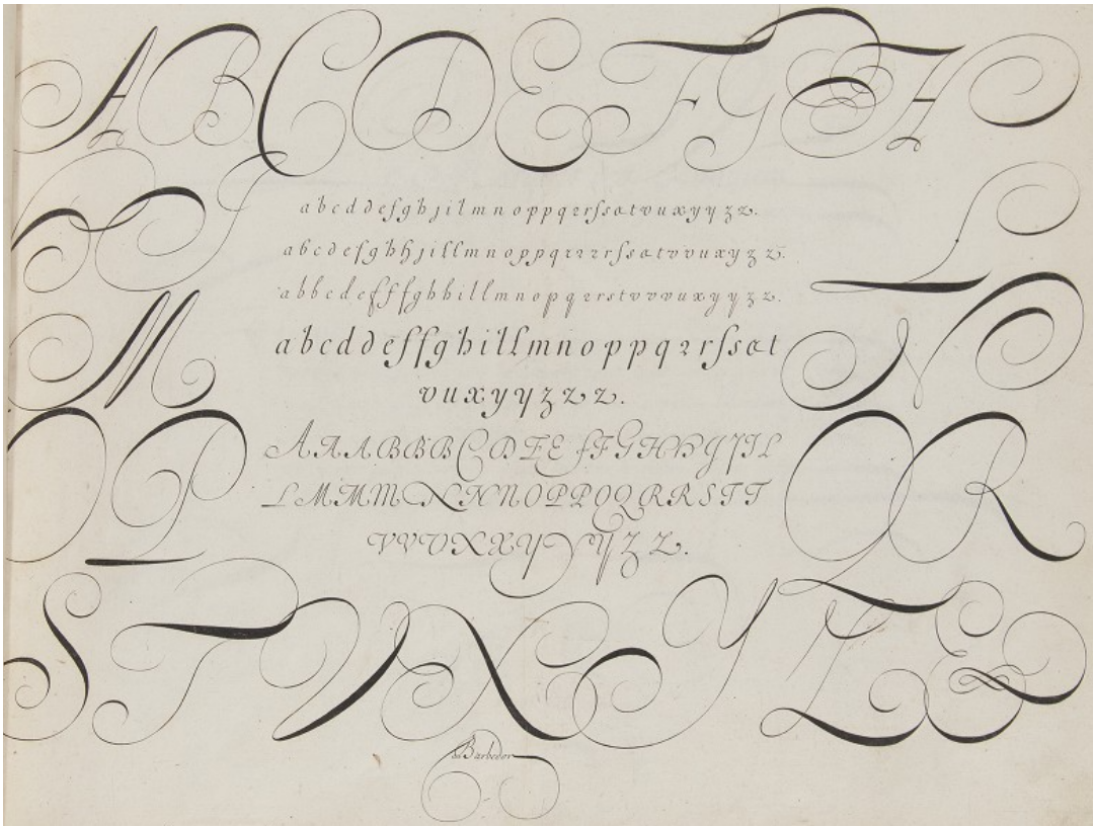
Une partie non négligeable de la littérature, mais surtout l'essentiel des documents d'archives rédigés après le Moyen Âge restent encore conservés sous forme manuscrite. Avec l'amélioration conséquente des outils d'extraction d'information, à commencer par ceux de reconnaissance optique de caractères (*Optical Character Recognition*, OCR[1], cf. Raj et Kos 2022), il devient important de développer des modèles capables de traiter numériquement ces sources manuscrites, qui ne bénéficient pas encore de ressources de qualité pour leur exploitation automatisée, contrairement à celles de l'époque médiévale (Pinche et al., 2023a).

La conception d'un tel modèle nécessite au préalable une connaissance fine de l'évolution de la langue comme des formes d'écriture, afin de délimiter un périmètre raisonnable et cohérent pour la phase d'entraînement, mais surtout pour la phase de test des modèles. En ef-

1. En plus d' « OCR », il est possible de trouver différentes dénominations. En effet, à l'origine, l'OCR a renvoyé à une technique opérant à l'échelle du caractère, tandis que l'expression « HTR » (pour *Handrwitten 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 préférons celui d'OCR, qui semble avoir absorbé les autres définitions dans la langue courante (cf. par ex. Wikipedia 2024).



(a) Alphabet de ronde.



(b) Alphabet de bâtarde.

FIGURE 1 – Modèles de Louis Barbedor (1647).

fet, on ne peut raisonnablement espérer qu'un modèle entraîné sur des écritures coulées de la fin du XVIII^e s. puisse fonctionner correctement sur un texte rédigé avec une écriture financière du XVI^e s., surtout si les données ont été transcrites avec des normes ne prenant pas en compte les spécificités des documents anciens (abréviations, modification de l'alphabet, etc.).

Nous proposons ici d'établir un cadre d'analyse pertinent pour construire et évaluer correctement un modèle pour les manuscrits d'une modernité « longue », allant du Moyen Âge tardif à aujourd'hui, en nous concentrant sur le français. Nous proposons aussi un premier modèle pour les mains françaises modernes.

2 État de l'art

La question du traitement informatique des mains modernes, dans son acception la plus large qui va du XV^e siècle jusqu'au XX^e siècle, pour le domaine francophone (cf. part. 3.1) n'est pas nouvelle, même si elle reste imparfaitement réglée. Pour la période la plus récente, on trouve le projet Lectaurep (Chagué et Rostaing, 2021) qui s'est intéressé au traitement des registres de notaires parisiens pour la période 1803-1940. Pour la partie haute de la chronologie, des premiers travaux ont par exemple été effectué autour du manuscrit de Richard Simon (c. 1701) conservé à la Hofbibliothek d'Aschaffenbourg (Nahon et Gabay, 2023).

Plusieurs modèles ont déjà été publiés. Transkribus (Kahle et al., 2017) propose un modèle pour le français sans spécifier les dates et les types d'écritures couvertes [2]. Le modèle 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 calligraphique — des premiers essais ayant démontré des résultats (très) faibles pour les sources de la première modernité.

Concernant les données d'entraînement disponibles, le catalogue HTR-United [3] (Chagué et al., 2022 ; Chagué et Clérice, 2023) ne répertorie que peu de jeux de données pour les documents francophones écrits en cursive, surtout pour la période d'Ancien Régime. Parmi les projets distribuant leurs données de manière ouverte [4], l'allemand (Hodel et al., 2021a ; Hodel et Schoch, 2021 ; Hodel et al., 2021b) ou 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).

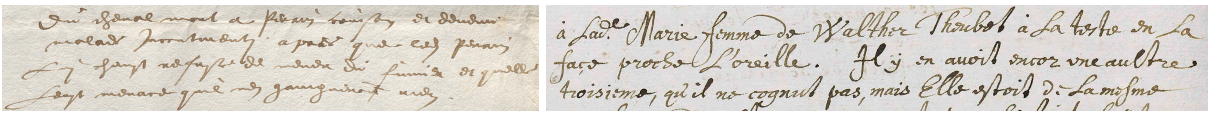
3 Rappel philologique

La conception d'un modèle d'OCR 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 la modernité, dont il est 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.

3.1 Forme calligraphique

D'un point de vue calligraphique, 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 dès le XIII^e s., notamment dans la production documentaire. Cette écriture « financière » (ausi appelée « française »), permet un tracé rapide qui lui assure un grand succès auprès du personnel administratif, et va persister jusqu'au XVII^e s., non sans muter profondément pendant cette longue période (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 a pro-

2. Modèle N°37758.
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 de ce côté.



(a) Porrentruy, AAEB, B 168/15-2.3 (1608).
du cheual mort a Perrin coinson et deuenu malade incontinenti apres que led Perrin lui heust refuse de mener du fumier et quelle leust menace quil nen gaingneroit rien.

(b) Porrentruy, AAEB, B 168/19-35.1 (1670).
à lad^e Marie femme de Walther Theubet à La teste en La façe proche L'oreille. Jl y en avoit encor vne aultre troisieme, qu'il ne cognut pas, mais Elle estoit de La mesme

FIGURE 2 – Exemple d'évolution de la cursive française au cours du XVII^e s., avant et après 1632-33. (corpus de procès criminels).

fondément changé, notamment sous l'influence du maître écrivain Guillaume Le Gangneur (1599), qui en propose une version épurée, débarrassée de ses oripeaux médiévaux.

À partir du XV^e s., 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 donc un tracé plus lent, elle est surtout l'apanage des lettrés, par exemple dans leur documentation personnelle. La financière reste donc l'écriture de prédilection de l'administration, qui pour des raisons pratiques en accentue la cursivité, parfois au détriment de la lisibilité.

La créativité, voire l'exubérance des maîtres écrivains, qui jouent dans cette histoire un rôle central (Cabane, 2020), vont encore réduire la lisibilité des documents en ajoutant à cette cursivité parfois « échevelée » (Samaran, 1967, p. 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 », cf. fig. 1b), qui est une stylisation française des écritures importées d'Italie mise au point par Étienne Le Bé, et la « française » (aussi appelée « ronde », ou encore « financière », cf. fig. 1a), dont le modèle est fournit par Louis Barbedor (1647).

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 s. 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 s., 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 activité calligraphique intense à travers toute l'Europe, mais aussi de traditions régionales qui s'enracinent en profondeur, à l'image de la *kurrentschrift* allemande (Beck, 1991), marquant certaines zones limitrophes qui sont en contact avec l'allemand, comme le Jura suisse.

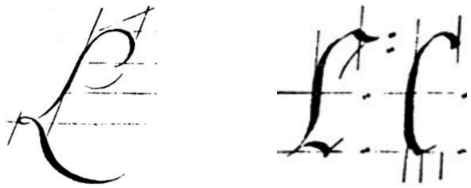
Avec le développement de l'instruction publique (Bishop, 2020), l'écriture se banalise, et se diversifie du simple fait de la multiplication des scripteurs (Frondizi et Fureix, 2022), notamment peu lettrés (Branca-Rosoff et Schneider, 1994 ; Bergeron-Maguire, 2019). Rapidement, l'anglaise s'impose dans les manuels (Dancel, 2011) et va rester le modèle dominant tout au long des XIX^e et XX^e s. – soit plus longtemps qu'en Angleterre, qui l'a abandonnée entre temps (Smith, 2020).

Un modèle d'OCR pour le « français moderne » se doit donc, théoriquement, de couvrir ce vaste monde de possible, qui connaît de profondes et multiples mutations calligraphiques à travers les siècles, voire parfois plus rapidement encore dans une même série documentaire (cf. par ex. la fig. 2). Étant donné l'amplitude de la variation, il convient d'être précis sur la couverture calligraphique du mo-

dèle, qui ne peut (encore) traiter avec la même qualité tous les types.

3.2 Vêtement graphique

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'OCR, on en trouve deux qui relèvent du matériel graphétique (Catach, 2001 ; Parussa et Cazal, 2015), à savoir les lettres ramistes et la majuscule, et une qui touche à la segmentation de la chaîne graphique.



(a) Majuscule (b) Minuscules

FIGURE 3 – Variation du ‹l› en écriture « coulée » d'après les planches de Ch. Paillasson (1763) publiées dans l'*Encyclopédie*. La variante minuscule de gauche est difficilement distinguable de la majuscule, dont elle est morphologiquement proche dans la pratique. Cf. par ex. fig. 2a et 5.

La majuscule a une double fonction en français contemporain : c'est un « signe-mot » et un « signe de phrase », pour reprendre la terminologie de N. 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^ s. 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 (Gabay, 2020). Dans la mesure où la différence entre la majuscule et la minuscule n'est souvent qu'une (infime) variation du module (cf. fig. 3) et qu'il n'est pas possible de recourir à la double fonction précédemment évoquée, il est difficile, voire impossible, d'arriver à les distinguer correctement, si tant est que la différence existe…

Sans rentrer dans le détail des diverses propositions des grammairiens du XVI^e^, 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 (Parussa et Cazal, 2015), ils sont à cette époque compris comme des allographes avec ‹u› et ‹i› des mêmes graphèmes, dont l'alternance est positionnelle, et donc graphique, et non phonologique : c'est Corneille qui, reprenant un usage néerlando-belge introduit par Plantin, convainc l'Académie d'adopter cette innovation (Catach et Golfand, 1973). La pénétration de cette dernière se faisant lentement, il n'est pas toujours aisé de savoir si la source utilise ou non les lettres ‹v|V› ou ‹j|J›, y compris au XVIII^e^ s., et donc que transcrire (cf. fig. 2b).

Enfin, la question de la segmentation de la chaîne graphique est un problème important : il est en effet rare, surtout dans les manuscrits, qu'elle se soucie des unités syntaxiques ou lexicales au XVI^e^ s. (Baddeley, 1998), au XVII^e^ s. (Pellat, 1998), au XVIII^e^ s. (Seguin, 1998), voire plus longtemps encore pour les peu lettrés (Steuckardt, 2014). Comme pour la capitalisation et les lettres ramistes, le facteur manuscrit joue un rôle important, car les levés de plume ne sont pas toujours très clairs, ni le degré d'avancement des processus de soudure, ce qui empêche une résolution simple de la question dès lors que l'on travaille à l'échelle de grands corpus.

3.3 Tradition ecdotique

Ces problèmes sont d'autant plus compliqués à résoudre que les éditeur · rice · s disposent d'un nombre faible de travaux sur les textes écrits pendant la période moderne, et particulièrement pour ceux écrits au XVII^e^ s., lesquels sont « modernisés » dans les grandes largeurs (alors qu'ils sont écrits à l'époque moderne) sans guère se soucier du matériau original (Gabay, 2014 ; Duval, 2015). Aux problèmes pratiques posés par le vêtement calligraphique et graphétique des manuscrits s'ajoute donc celui, scientifique, d'une tradition volontiers très interventionniste des chercheur · se · s, qui n'ont pas l'habitude, à de

rares exceptions près, de remettre en cause la pertinence de leurs pratiques ecdotiques.

4 Transcription

4.1 Remarques préliminaires

Plutôt que d'accumuler des transcriptions disparates, il nous a paru important de présenter quelques premières propositions afin de standardiser aux mieux les données qui seraient produites à l'avenir par d'autres projets. Les choix de transcription des médiévistes (Pinche, 2022 ; Pinche et al., 2023a), 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 (Pinche et al., 2023b).

La mise en œuvre de ces préconisations reste cependant sujette à caution, car les chercheur · se · s. ayant produit les jeux de données pour les cursives présentées *supra* (cf. part. 2) n'ont jamais explicité leurs choix dans l'établissement du texte – l'étude des données laisse entrevoir des transcriptions 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 calligraphique d'un modèle, peut donc créer des problèmes, et nous pousser à amender nos propres choix par pragmatisme dans un avenir proche.

Il est important de distinguer la production de données de vérité de terrain et la transcription à visée ecdotique. En effet, la seconde implique un toilettage du texte plus important pour faciliter la lecture par l'homme, alors que la première exige un plus grand respect de la source qui ne gêne pas la machine – au contraire, l'abréviation ‹p̱› reste toujours un ‹p̱›, et ne devient pas parfois ‹pro›, parfois ‹par›. Les options prises étant (très) disparates d'une tradition philologique à l'autre (Gabay, 2014), un trop grand interventionnisme ne peut que rendre caduque tout espoir d'interopérabilité, même minimale. Il s'agit de penser la production de vérité de terrain comme la première étape d'un mouvement en deux temps, le deuxième étant celui d'une normalisation du texte pour en simplifier l'accès au lecteur [5].

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

4.2 Les lettres

Les allographes. Les différents allographes ne sont pas notés – contrairement à ce qui peut être proposé dans les recommandations pour les imprimés (Gabay et al., 2023). Ainsi, la gamme des variations allant du *s* « long » (‹ſ›, normalement utilisé à l'initiale et en interne) au *s* « rond » (‹s›, normalement utilisé en finale) est réduite au caractère que nous connaissons aujourd'hui (cf. fig. 4). Les cas similaires, comme les allographes avec un jambage plongeant (par ex. ‹ɳ›), suivent le même principe.

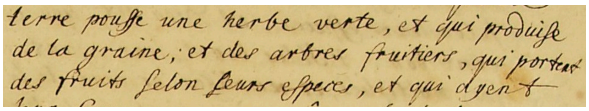


FIGURE 4 – Aschaffenbourg, Hofbibliothek, Ms. 48. terre pousse une herbe verte, et qui produise de la graine ; et des arbres fruitiers, qui portent des fruits selon leurs especes, et qui dyent

Les majuscules. Autant que possible, nous conservons l'usage des majuscules et des minuscules dans le texte, mais certaines variations de modules 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. Un module

5. Étape qui ressemblerait à celle présentée par Bawden et al. (2022), mais avec un degré d'intervention moindre.

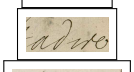
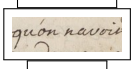
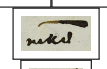
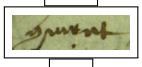
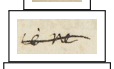
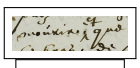
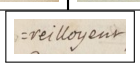
Catégorie	Cas	Traitement	exemple	image
Soudure	Agglutination illogique	Normalisation	*et de*	
Soudure	Désagglutination illogique	Normalisation	*a dire*	
Soudure	Pas d'apostrophe	Comme dans la source	*qu'on navoit*	
Abréviation	Tilde ou macron	Tilde	*femẽ*	
Abréviation	*que* abrégé	‹q› tildé (‹q̃›)	*q̃il*	
Abréviation	Contraction	Conservation	*pñt*, *pñtce*	
Abréviation	Contraction	Conservation	*mr̃e*	
Abréviation	Contraction	Conservation	*v̲tus*	
Abréviation	*di(c)t(es)*	‹d› avec hameçon (‹ɖ›, `U+0256`)	*esɖ*	
Abréviation	Exposant	Précédée du circonflexe (ˆ, `U+02C6`)	*sˆte*	
Abréviation	Esperluette (‹&›)	Conservation	*&*	
Abréviation	‹p› barré droit (*par*)	‹ꝑ› (`U+A751`)	*ꝑ*	
Abréviation	‹p› barré courbe (*pour*)	‹ꝓ› (`U+A753`)	*ꝓ*	
Abréviation	Neuf tironien	‹ꝯ› (`U+A76F`)	*ꝯment*	
Abréviation	Sept tironien	‹ɾ› (`U+204A`)	*ɾ*	
Alphabet	*s* long (‹ſ›)	*s* rond (‹s›)	*estes*	
Signe	Croix, croisettes	Marque de référence (※, `U+203B`)	*※ Car*	
Alphabet	Module	Imitation	*ces Choſes*	
Alphabet	‹u› (semi-)consonne	Comme dans la source	*auec*	
Alphabet	‹j› (semi-)voyelle	Comme dans la source	*Ljncendie*	
Alphabet	‹i› consonne	Comme dans la source	*ie*	
Correction	Rature lisible	Entre doubles crochets (‹⟦›)	*⟦ie ne⟧*	
Correction	Rature partielle	Entre doubles crochets (‹⟦›)	*endroit⟦s⟧*	
Correction	Rature illisible	Entre doubles crochets (‹⟦›) (avec un point par lettre)	*⟦.....⟧*	
Correction	Insertion	Symbole insertion (‹⸎›, `U+2380`)	*mourir, ⸎ que*	
			⸎ quicelles	
Bout de ligne	Trait de conduite	Rallonge de ligne (‹—›, `U+23AF`)	*la grande —*	
Bout de ligne	Ornementation	Rallonge de ligne (‹—›, `U+23AF`)	*a conclû qu'il —*	
Bout de ligne	Tiret(s) de fin	Signe négation (‹¬›, `U+00AC`)	*habi¬*, *Beur¬*	
Bout de ligne	Tiret(s) de début	Signe négation (‹¬›, `U+00AC`)	*¬veilloyent*	

Tableau 1 – Principales règles de transcription

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 fig. 5).

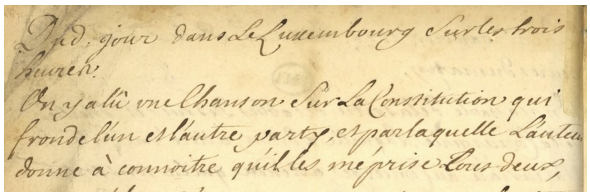


FIGURE 5 – Paris, Arsenal, Ms. 10158, f. 8v.
Dud. jour dans le Luxembourg sur les trois heures.
On y a lû vne Chanson sur la Constitution qui fronde l'un et l'autre party, et par laquelle l'auteur donne à connoitre qu'il les méprise tous deux,

La variation de module. Les petites majuscules, ou toutes les variations de module tendant à agrandir la lettre, sont transcrites avec des majuscules (cf. fig. 6).

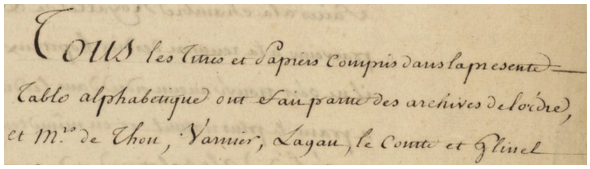


FIGURE 6 – BnF, Arsenal, Ms. 6118, f. 2r.
TOUS les Titres et Papiers compris dans la presente Table alphabétique ont partie des archives de l'ordre, et M^rs de Thou, Varnier, Lagau, le Comte et Glinel

Les lettres ramistes. Dans la mesure où l'utilisation du ‹v› et du ‹j› pour les consonnes et du ‹u› et du ‹i› pour les voyelles (ou semi-voyelles) s'impose lentement entre le XVIe et le XVIIIe, il est recommandé de conserver le système graphique du scripteur (cf. fig. 7).

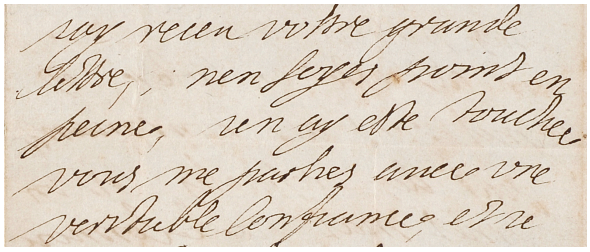


FIGURE 7 – Musée de Grignan, N°1329F4.
iay receu vostre grande lettre, nen soyes point en peine, ien ay este touchee vous me parles auec vne veritable Confiance et ie

4.3 L'effacement et l'enrichissement des lettres

Les signes auxiliaires. Les accents, trémas, apostrophes et traits d'union, etc. sont conservés tels quels, même s'il s'agit d'usages spécifiques qui ont disparu comme par exemple le tréma sur le ‹y› dans les zones de contacts avec les écritures germaniques (cf. fig. 8).

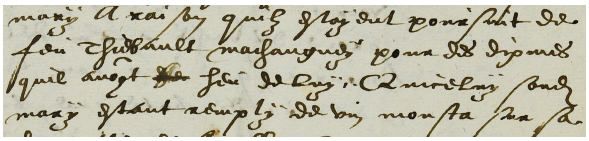


FIGURE 8 – Porrentruy, AAEB, B 168/15-10.3 (1609).
marÿ A raison quilz estoÿent poursuit de feu Thiebault machangueÿ pour des dixmes quil avoÿt ⟦..⟧ heu de luÿ, Quiceluÿ sond marÿ estant remplÿ de vin monsta sur sa

Les abréviations. Pour les abréviations, nous suivons les recommandations des médiévistes : elles ne sont pas développées, des tentatives de développement posant *in fine* des problèmes de généralisation (Torres 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. Le signe ^ (U+005E) est placé avant les lettres suscrites (cf. fig. 9). Les caractères de la MUFI[6] 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.

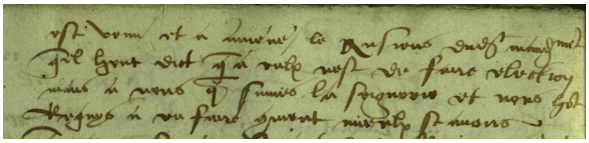


FIGURE 9 – Archives de l'Etat de Genève, R.C. 29, 20 avril et 3 mai 1536.
est venu et a amené le ansiens dud^t mand^mẽt q̃il hont dict q̃a eulx nest de faire election mais a nous q̃ sumes la seignorie et nous hõt requys a en faire ꝯment mieulx scauons.

4.4 La séquence de lettres

Les levés de plume, l'espacement Les levés de plumes 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

6. https://mufi.info.

pas retouchée (cf. par ex. *par ce que* mais *ecriture*, fig. 10).

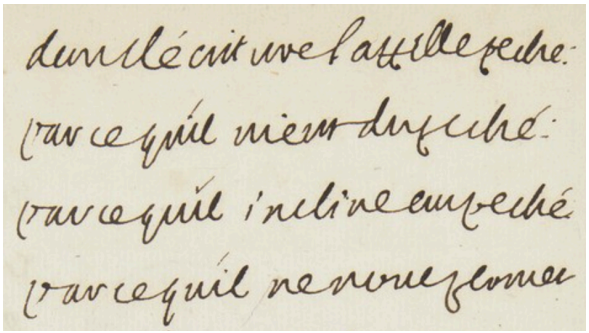


FIGURE 10 – Paris, BnF, Fr. 12820, f. 6ᵛ.
dans l'ecriture sappelle peche :
par ce qu'il uient du peché :
par ce qu'il incline au peché
par ce qu'il ne nous permet

Le tiret de fin ligne. Il convient de distinguer les différents tirets. Le trait d'union (‹-›) n'est pas un tiret de fin de ligne (‹¬›, `U+00AC`), pour simplifier le post-traitement des données. De la même manière, il faut transcrire différemment le trait de conduite (qui comble un espace vide, par exemple en fin de ligne), ou tout autre symbole équivalent par le caractère Unicode de la rallonge de ligne (‹—›, `U+23AF`).

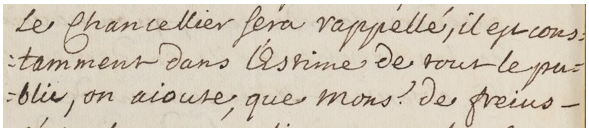


FIGURE 11 – Archives de la Bastille, Paris, BnF, Arsenal 10155 , f. 6ᵛ
le Chancellier sera rappellé, il est cons¬
¬tamment dans l'estime de tout le pu¬
¬blic, on aioute, que mons. de freius —

Ponctuation. 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.

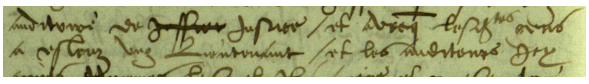


FIGURE 12 – Archives de l'Etat de Genève, R.C. 29, 18 mai 1536.
auditoire de ⟦jussier⟧ justice, et aueq̃ le s^tes gens
a esleuz vng Lieutenant, et les auditeurs jcy

Les ligatures. Accompagnant le mouvement baroque, les ligatures se développent abondamment vers la fin du XVI^e s. 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› (cf. fig. 13).

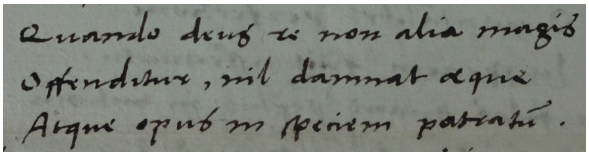


FIGURE 13 – Archives de la ville de Strasbourg, 1 AST 212 n°59.
Quando deus re non alia magis
Offenditur, nil damnat æque
Atque opus in speciem patratũ.

Les signes fonctionnels. Les signes fonctionnels sont conservés. La très large palette de cas que recouvre cette catégorie (astérisque ‹*›, paragraphe ‹§›, signe d'insertion ‹ꝰ›, 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 standards.

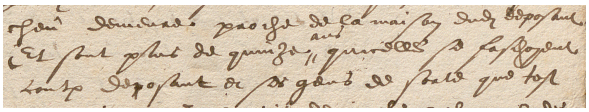


FIGURE 14 – Porrentruy, AAEB, B 168/15-2.2 (1608).
heũ demeure proche de la maison duꝺ deposant
auꝺ
Et sont plus de quinze ꝰ quicelles se faschoyent
contre deposant et ses gens de sorte que tost

5 Expérience

5.1 Données

Nous avons rassemblé un important jeu de données pour le français, couvrant quatre siècles, auquel a été ajouté un sélection des données de vérité de terrain écrits dans d'autres langues (espagnol, latin, allemand et néerlandais) distribuées par d'autres projets (cf. tab. 2). Dans les données en français, nous trouvons essentiellement des documents d'archive, mais aussi des documents littéraires (manuscrit de Zola) ou assimilés (manuscrit mis au propre d'une histoire familiale écrit en 1818).

Toutes les données n'étant pas distribuées dans le même format, nous convertissons tous les fichiers au format ALTO. Pour les données de vérité de terrain préparées avec Transkribus, les masques sont recalculés avec Kraken

Projet	Langue	Siècle	Pages	Lignes	Dépôt Github
FoNDUE†*	la	XVI^e s.	49	2 003	FONDUE-LA-MSS-16-PR
FoNDUE†*	fr	XVII^e s.	63	1 668	FONDUE-FR-MSS-17
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	FONDUE-FR-MSS-19-PR
VWTM‡*	fr	XIX^e-XX^e s.	190	28 611	FONDUE-FR-VTM-20
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	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	

TABLEAU 2 – Détails des données utilisées pour l'entraînement. Le signe * indique les données qui ne sont pas (encore) distribuées de manière ouverte. Le signe † les données qui contiennent des données rédigées (lettres, rapports…). Le signe ‡ 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 celle pour lesquelles nous ne maîtrisons pas la transcription. NA signifie que le décompte est impossible (seules des lignes sont distribuées, pas les pages complètes).

pour assurer la compatibilité des données entre elles.

5.2 Entraînements

Nous entraînons un modèles avec des données françaises et étrangères – nombre de documents ne sont en effet pas écrits en une seule langue, et un modèle multilingue est notre objectif final. 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.

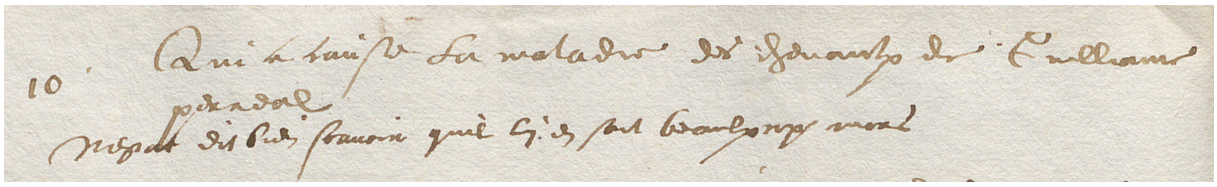
Les données sont réparties en trois jeux. Celui pour le test (3% du total) est sélectionné à la main. Le reste est réparti entre jeu de données pour l'entraînement (90%) et pour l'évaluation (10%). Deux types de données hors domaine sont conservés pour d'autres expériences : un jeu en français du XIX^e s. et un autre en allemand du XVIII^e s. (cf. part. 5.4).

5.3 Évaluation du modèle

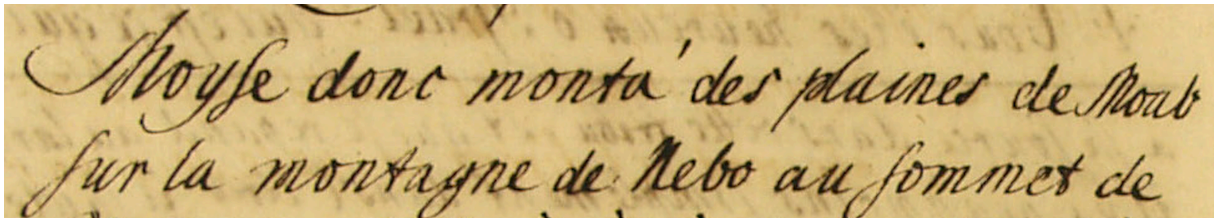
Plusieurs évaluations ont été faites pour évaluer les performances. Outre une évaluation générale du modèle, quatre pages par jeux de données ont été extraites. Trois tests sont menés :

— avec Manu Mc French ;
— avec Manu Mc French ajusté (en anglais *fine tuned*) sur les données provenant du même jeu que pour le test (4 pages en test et en évaluation, le reste en entraînement) ;
— avec notre nouveau modèle Manu McFondue.

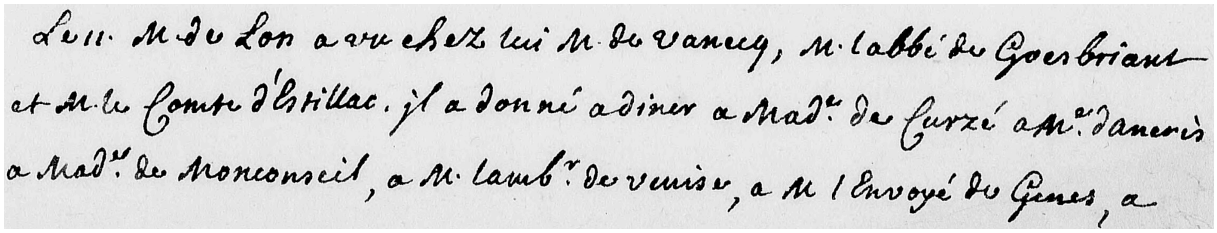
L'analyse des résultats (cf. tab. 3) montre un net gain avec Manu McFondue par rapport à Manu McFrench (66,02% vs. 89.9%). Les gains les plus importants sont obtenus sur la gothique tardive, dans une version peu formalisée avec de nettes influences germaniques (cf. fig. 15a). L'écriture italienne (cf. fig. 15b),



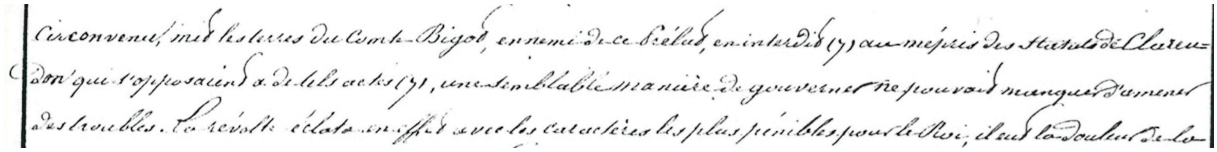
(a) Procès criminel, Porrentruy, AAEB, B 168/15-2.3 (1608)



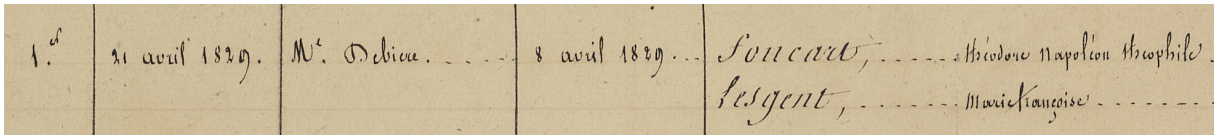
(b) R. Simon, *Pentateuque traduit*, Aschaffenoburg, Hofbibliothek, Ms. 48. (c. 1709)



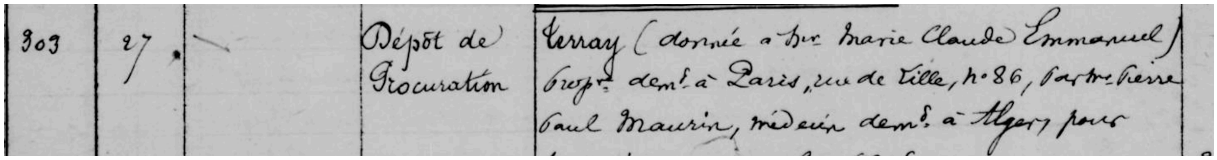
(c) Lieutenant de police, Paris, Arsenal, Ms. 10292 (1746)



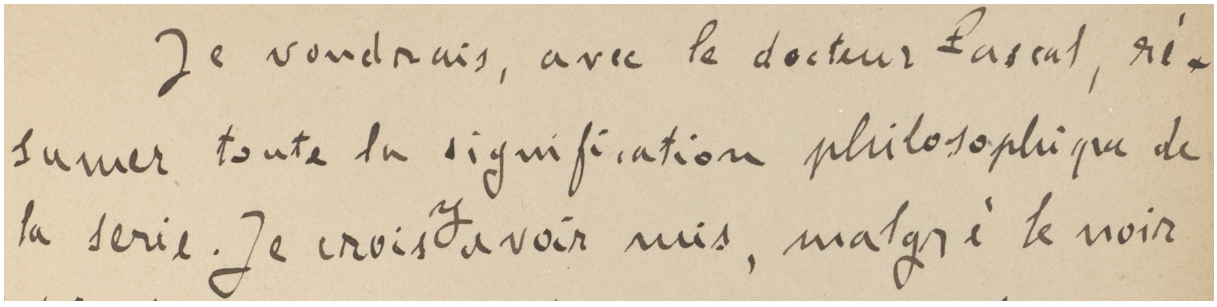
(d) Archives privées (1818)



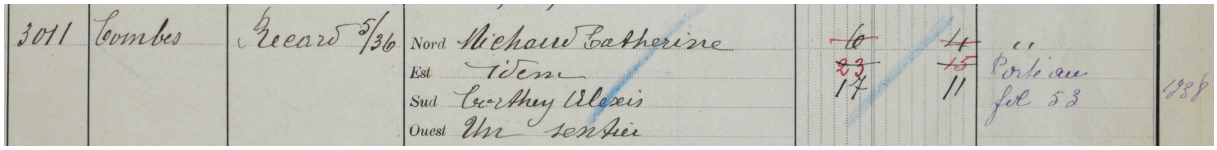
(e) Contrats de mariage, Archives nationales, CM/1 (1829)



(f) Actes notariés, Archives nationales (1850)



(g) Zola, *Le Docteur Pascal*, Cologny, Fondation Bodmer, Z-6.3* (1893)



(h) Registre de l'impôt sur les biens-fonds, Sion, Archives de l'Etat du Valais (1894-1984)

FIGURE 15 – Extrait du jeu de données de test

Dataset	ManuMc French	ManuMc French ajusté	Manu McFondue
Général	66.02		89.9
AAEB	52.96	**89.92**	81.49
Aschaffenbourg	87.30	**96.87**	95.03
Arsenal	80.06	**90.99**	85.27
Archives privées	61.89	**94.71**	91.12
Contrats de mariages	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

TABLEAU 3 – Détail des résultat par jeux de données. `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.

relativement bien traitée par Manu McFrench (87,30%) se trouve encore mieux traitée par ManuMc Fondue (95,03, soit +7,73 pt de %). Pour la bâtarde coulée (cf. fig. 15c), le gain est plus faible mais pas négligeable (+ 5,21 pt. de %).

Concernant les écritures du XIX^{e}, pour lesquelles on voit l'apparition du trait caractéristique de la plume anglaise (sauf le manuscrit de Zola), les résultats sont globalement positifs. On observe un très net gain pour l'écriture anglaise dans une version assez formelle (cf. fig. 15d, +29,23 pt de %) et un gain plus modeste pour les documents présentant un mélange de bâtarde et de ronde (cf. fig. 15e, +0,6 pt de %). La seule contre-performance concerne une écriture anglaise assez peu formelle (cf. fig. 15f, -2,26 pt de %), mais qui était déjà présente dans les données d'entraînement de ManuMc French. Le manuscrit de Zola (cf. fig. 15g), 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 (cf. fig. 15h), 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 %).

En revanche, ajuster Manu McFrench semble une stratégie plus efficace que l'utilisation de Manu McFondue sans ajustement. La question est donc de savoir si cette performance inférieure du nouveau modèle est due à une faiblesse intrinsèque (par ex. la trop grande hétérogénéité des données), ou si l'ajustement permet une amélioration en trompe l'œil – un ajustement sur Manu McFondue étant plus efficace qu'un ajustement sur Manu McFrench.

5.4 Évaluation des capacités d'ajustement du modèle

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 XIX^{e} [7] assez similaires aux données utilisées par Manu McFrench, mais inconnu des deux modèles ;

— à un jeu de données suisses du XVIII^{e} en allemand [8], là encore inconnu des deux modèles.

Dataset	ManuMc French	ManuMc Fondue
Français	88.48	**89.40**
Allemand	71.01	**78.07**

TABLEAU 4 – Expérience d'ajustement sur un jeu de données hors-domaine, avec quatre puis huit pages en entraînement.

Nous utilisons huit pages pour l'entraînement, deux pages pour la validation, et cinq pages pour le test. Les résultats (cf. tab. 4) 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 tab. 3 sont dues à l'ajustement.

6 Conclusion

Le modèle ManuMc Fondue est loin de résoudre le problème posé par l'OCRisation des manuscrits modernes, mais est une avancée significative dans la résolution de ce problème. Les prochaines tâches seront donc :

— l'ajout de données supplémentaires pour chaque siècle du français, qui fait office de « colonne vertébrale » du projet ;

— l'ajout de données supplémentaires pour les autres langues déjà couvertes (no-

7. Archives départementales de La Réunion, IE 5, affranchissements.

8. Franz Joseph Leonti Meyer von Schauensee, *Ausführlich Musicalisches Protocoll*, Zentral-und HochschulbibliothekLuzern : CH-Lz, PpMsc 167 fd.

tamment néerlandais, allemand et espagnol) ;
— la finalisation des préconisations de transcription pour le français ;
— la standardisation des règles de transcription pour un modèle multilingue de type CATMuS.

Bibliographie

Susan Baddeley. 1998. Théorie et pratique de la segmentation graphique dans les textes français du premier tiers du XVIe siècle. *119*, pages 52–68.

Louis Barbedor. 1647. *Les ecritures financiere et italienne bastarde dans leur naiveté*. N. Langlois, Paris.

Frédéric Barbier. 2006. *L'Europe de Gutenberg. Le livre et l'invention de la modernité occidentale (XIIIe-XVIe siècle)*. Belin, Paris.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, et Simon Gabay. 2022. Automatic Normalisation of Early Modern French. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. Document de travail.

Friedrich Beck. 1991. Die „deutsche schrift“ – medium in fünf jahrhunderten deutscher geschichte. *Archiv für Diplomatik*, 37 :453–479.

Myriam Bergeron-Maguire. 2019. Du poitou en louisiane : édition et notes à partir de la correspondance d'une peu lettrée (1802-1803). *Géolinguistique*, 19.

Marie-France Bishop. 2020. L'enseignement de l'écriture à l'école primaire française de 1880 aux années 2000. In *L'écriture dès le début de l'école primaire*. Presses Universitaires de Bordeaux, Pessac.

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

Célia Cabane. 2020. Les maîtres écrivains : acteurs méconnus de la transmission des savoirs. In Dominique Briquel, éditeur, *Écriture et transmission des savoirs de l'Antiquité à nos jours*, Actes des congrès nationaux des sociétés historiques et scientifiques. Éditions du Comité des travaux historiques et scientifiques.

Silvia Cascianelli, 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. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1506–1513. IEEE.

Nina Catach. 1994. *La ponctuation (Histoire et système)*. Presses Universitaires de France, Paris.

Nina Catach. 2001. *26. Graphetik / Graphétique*, pages 725–735. Max Niemeyer Verlag, Berlin, Boston.

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

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 et Aurélia Rostaing. 2021. LECTAUREP : Lecture Automatique des Répertoires de Notaires Parisiens. In *Fantastic Futures 2021 / Futures Fantastiques 2021*, Paris, France. AI4LAM and BnF and Université Paris Saclay.

Alix Chagué et Thibault Clérice. 2022. HTR-United - Manu McFrench v1 (manuscripts of modern and contemporaneous french).

Alix Chagué, Thibault Clérice, et Laurent Romary. 2022. Htr-united : un écosystème pour une approche mutualisée de la transcription automatique des écritures mansucrites. Document de travail.

Brigitte Dancel. 2011. Apprendre à écrire, quelle histoire ! *Carrefours de l'éducation*, 4 :123–134.

Frédéric Duval. 2015. Les éditions de textes du XVIIe siècle. In *Manuel de la philologie de l'édition*, pages 369–394. De Gruyter, Berlin, Boston.

Alexandre Frondizi et Emmanuel Fureix. 2022. Vous avez dit « écritures populaires » ? *Revue d'histoire du XIXe siècle*, 65 :9–22.

Simon Gabay. 2014. Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVIIe siècle. *Vox Romanica*, 73(1).

Simon Gabay. 2020. La naissance de Marie-Blanche de Grignan. Notes sur la mise en page de la polyphonie sévignéenne. In *Acta Litt&Arts*, numéro 13 in Les discours rapportés en contexte épistolaire (XVIe-XVIIIe siècles). Grenoble : Université Grenoble Alpes.

Simon Gabay, Rachel Bawden, Philippe Gambette, Jonathan Poinhos, Eleni Kogkitsidou, et Benoît Sagot. 2022. Le changement linguistique au XVIIe s. : nouvelles approches scriptométriques. In *CMLF 2022 - 8e Congrès Mondial de Linguistique Française*, volume 138 of *SHS Web of conferences*, pages 02006.1–14, Orléans, France. EDP Sciences.

Simon Gabay, Thibault Clérice, et Christian Reul. 2023. OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more). *Journal of Data Mining and Digital Humanities*, 2023.

Françoise Gasparri. 1983. Enseignement et techniques de l'écriture du Moyen Âge à la fin du XVIe siècle. *Scrittura e civiltà*, 7 :201–222.

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

Tobias Hodel et David Schoch. 2021. Handwritten Text Recognition Test Set : Minutes of the Swiss Federal Council (1848-1903).

Tobias Hodel, David Schoch, et Peter Dängeli. 2021a. Handwritten text recognition ground truth set : StABS Ratsbücher O10, Urfehdenbuch X.

Tobias Hodel, David Schoch, Christa Schneider, et Jake Purcell. 2021b. General models for handwritten text recognition : Feasibility and state-of-the art. german kurrent as an example. *Journal of Open Humanities Data*, 7.

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

Jean Hébrard. 1995. Des écritures exemplaires : l'art du maître écrivain en france entre XVIe et XVIIIe siècle. *Mélanges de l'école française de Rome*, 107(2) :473–523.

Philip Kahle, Sebastian Colutto, Günter Hackl, et Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Liesbeth Keijser. 2020. 6000 ground truth of VOC and notarial deeds 3.000.000 HTR of VOC, WIC and notarial deeds.

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

Guillaume Le Gangneur. 1599. *La Technographie, ou Brièvе méthode pour parvenir à la parfaitte connoissance de l'écriture françoyse*. s.n., s.l.

Lucas Materot. 1608. *Œuvres*. J. Bramereau, Avignon.

Christine Métayer. 2001. Normes graphiques et pratiques de l'écriture. maîtres écrivains et écrivains publics à paris aux XVIIe et XVIIIe siècles. *Annales. Histoire, Sciences Sociales*, 56(4-5) :881–901.

Peter Nahon 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.

Charles Paillasson. 1763. Écritures, contenant seize planches. In *Recueil des planches sur la science, les arts libéraux et les arts méchaniques, avec leur explication*, Encyclopédie ou Dictionnaire raisonné des Sciences, des Arts et des Métiers. Frommann, Paris. 2e livraison, 1ère partie.

Gabriella Parussa et Yvonne Cazal. 2015. *Introduction à l'histoire de l'orthographe*. Armand Colin, Paris.

Jean-Christophe Pellat. 1998. Les mots graphiques dans des manuscrits et des imprimés du XVIIe siècle. *Langue française*, 119 :88–104.

Ariane Pinche. 2022. Guide de transcription pour les manuscrits du Xe au XVe siècle. Document de travail.

Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'Connor, Wouter Haverals, Mike Kestemont, et Caroline Vandyck. 2023a. CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts. Document de travail.

Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, et Simon Gabay. 2023b. Catmus medieval.

Emmanuel Poulle. 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*. Droz, Genève.

Emmanuel Poulle. 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.

Ravi Raj et Andrzej Kos. 2022. A comprehensive study of optical character recognition. In *2022 29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES)*, pages 151–154.

Alain Riffaud. 2007. *La ponctuation du théâtre imprimé au XVIIe siècle*. Droz, Genève.

Charles Samaran. 1967. Cursives françaises des XVe, XVIe et XVIIe siècles (compte-rendu). *Journal des Savants*, 3 :129–153.

Jean-Pierre Seguin. 1998. Les incertitudes du mot graphique au XVIIIe siècle. *Langue Française*, 119 :105–124.

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

Marc Smith. 2020. Les modèles d'apprentissage de l'écriture en france depuis la renaissance. In *Apprendre*, Recherches, pages 167–179. La Découverte.

Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, et Daniela Solfaroli Camillocci. 2023. Guide de transcription pour les imprimés français du XVIe siècle en caractères gothiques : Version A. Document de travail.

Agnès Steuckardt. 2014. De l'écrit vers la parole. enquête sur les correspondances peu lettrées de

la grande guerre. In *Congrès Mondial de Linguistique Française – CMLF 2014*, pages 353 – 364, Berlin.

Dominique Stutzmann. 2011. Paléographie statistique pour décrire, identifier, dater... normaliser pour coopérer et aller plus loin ? In *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*, pages 247–277. BoD.

Sergio Octavio Torres Aguilar et Vincent Jolivet. 2023. Handwritten text recognition for documentary medieval manuscripts. *Journal of Data Mining and Digital Humanities*.

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

Wikipedia. 2024. Optical character recognition. In *Wikipedia, the free encyclopedia*. Wikimedia Foundation. Version du 25 mars 2024.