Guide de transcription pour les manuscrits du Xe au XVe siècle

Ariane Pinche

Guide de transcription pour les manuscrits du X^e^ au XV^e^ siècle

Ariane Pinche, École nationale des chartes, Paris[1]

Ce guide est le fruit d'une réflexion menée au cours des années 2021 et 2022 dans la cadre du projet CREMMALab, de la confection du modèle HTR[2] Bicerin[3] pour les manuscrits littéraires du XIII^e^ au XIV^e^ siècle et du séminaire de recherche « Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le X^e^ et le XV^e^ siècle »[4].

L'utilisation croissante de la reconnaissance automatique d'écriture (HTR) pour les documents médiévaux demande de nouveaux investissements en termes de production de données. Uniformiser les pratiques de transcriptions devient donc primordial pour constituer des sets de données partageables et réutilisables pour en minimiser le coût. Aujourd'hui, les projets scientifiques cherchent à acquérir des données textuelles en masse pour entreprendre des éditions de textes longs ou pour constituer de larges corpus à interroger, rendant l'usage de l'HTR nécessaire pour traiter une telle masse de données. Toutefois, il est encore rare que les projets parviennent à partager leurs données ou leurs modèles pour réduire les coûts de production à une échelle collective. Ainsi, produire des modèles génériques[5] et des données de vérité de terrain qui soient à terme utiles à la communauté scientifique semble un nouvel objectif à atteindre. Pour ce faire, une harmonisation des pratiques est nécessaire afin de mettre en commun nos données[6] et de créer des modèles plus robustes et génériques pour réduire le temps imparti à leur création.

Nos préconisations ont pour but d'accompagner la création des données d'entraînement afin d'optimiser l'apprentissage machine des modèles d'HTR. L'établissement de « bonnes pratiques » et des normes communes assurera également la pérennité et la réutilisation des données produites. Les méthodes de transcription que nous proposons, ici, n'ont pas pour but de constituer une transcription définitive ou une édition finale dont la production demandera la mise en place d'un protocole en plusieurs étapes :

1 Ce guide n'aurait pas pu voir le jour dans l'aide de Jean-Baptiste Camps et de Frédéric Duval.

2 Handwritten Text Recognition.

3 Ariane Pinche, *CREMMA Medieval, an Old French dataset for HTR and segmentation*, 2021.

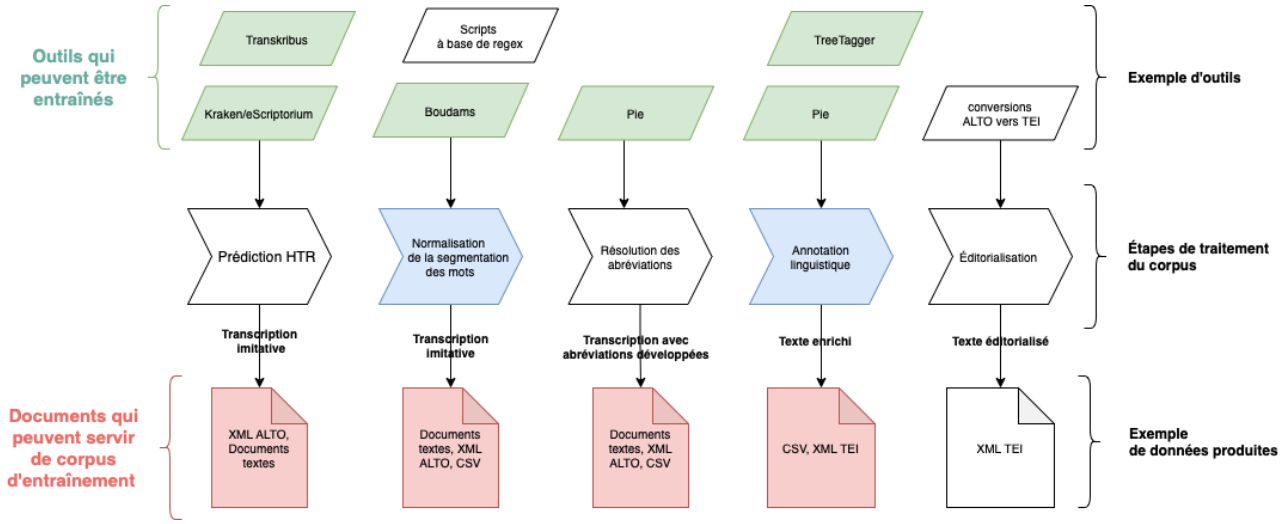
4 Ce séminaire a eu lieu entre octobre 2021 et mars 2022 à l'École nationale des chartes en présence de *Jean-Baptiste Camps, Camille Carnaille, Alix Chagué, Floriane Chiffoleau, Prunelle Deleville, Lucien Dugaz, Frédéric Duval, Simon Gabay, Lucence Ing, Vincent Jolivet, Viola Mariotti, Marco Maulu, Nicolas Perreaux, Ariane Pinche, Vera Schwarz-Ricci, Anne Rochebouet, Aurélia Rostaing, Benedetta Salvati, Peter Stokes, Sergio Torres, Dominique Stutzmann, Marguerite Vernet* que nous remercions vivement pour leur participation. L'ensemble des comptes-rendus des séances est disponible au lien suivant : https://cremmalab.hypotheses.org/seminaire-creation-de-modeles-htr. Ariane Pinche, « CREMMALAB | Constitution de corpus en ancien français pour l'HTR », 2022.

5 Ces modèles, en fonction des besoins, pourront être utilisés comme tels ou bien personnalisés.

6 Sur ce point, nous conseillons de déclarer les données dans le catalogue de *HTR-United (Alix Chagué, Thibault Clérice et Floriane Chiffoleau, HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages, 2021*, https://htr-united.github.io).

- Étape 1 : Production d'une transcription imitative avec conservation des abréviations grâce à l'HTR
- Étape 2 : Régularisation de la segmentation des mots
- Étape 3 : Développement des abréviations
- Étape 4 : Annotation linguistique du corpus
- Étape 5 : Éditorialisation du texte

Procéder par étapes distinctes permet de conserver l'ensemble des informations et d'avoir aisément accès au texte source. Chaque étape de post-traitement pour être faite manuellement ou à l'aide d'outils intermédiaires. Proposer un outil par tâche facilite leur maintenance et permet de les perfectionner individuellement les uns des autres. Enfin, les différentes sorties pourront à leur tour alimenter des corpus d'entraînement pour chacun des maillons du protocole.



Exemple de protocole de post-traitement des données issues de l'HTR

Les propositions de transcription seront aussi pragmatiques que possible, tout en essayant d'assurer une production scientifique de qualité. Certains choix ont été guidés par une part d'arbitraire dont nous espérons lever les ambiguïtés grâce à une documentation détaillée. Chacune des recommandations peut être adaptée en fonction des projets et de leurs problématiques propres[7].

I-Principes généraux de transcription

Depuis l'apparition des corpus numériques, les éditeurs cherchent à définir de nouvelles normes de transcription et à trouver un juste équilibre dans leur préconisation. L'article pionnier « *Guidelines for Transcription of the Manuscripts of the Wife of bath's Prologue* » de Peter Robin et Elizabeth Solopova (1993) relate l'une de ces premières expériences. Quoique le projet n'intégrait pas d'HTR, nos problématiques autour de la transcription des corpus sont similaires.

7 Les données pourront toujours être interopérables à condition de documenter l'ensemble des choix divergents. Toutefois, la conversion de données signifie souvent la réduction au plus petit dénominateur commun et engendre une grande perte d'informations qu'il faut essayer de limiter.

Notre tâche est de trouver un moyen de traduire la manière dont le texte est livré sur son support originel dans un système interprétable par une machine et qui favorise son apprentissage. Nos solutions seront forcément réductrices et relèveront fondamentalement d'une activité d'interprétation, car il est impossible de rendre toute la variété d'une écriture manuscrite au moyen d'un ordinateur qui possède un nombre limité de caractères[8]. Pour définir nos pratiques, nous reprendrons les deux termes définis par Dominique Stutzmann[9] de :

• *Transcription allographétique*, transcription qui vise à donner accès à toutes les formes de chaque lettre ou signe.

• *Transcription graphématique*, transcription qui préserve la suite des lettres et réduit chaque forme à son sens dans un système alphabétique.

Afin de proposer un système de transcription accessible, nous avons écarté l'idée de produire des transcriptions allographétiques, car il nous semblait impossible de parvenir à proposer des préconisations générales pour des transcriptions de ce type pour tous les documents médiévaux du X^e^ au XV^e^ siècle. En outre, si on s'appuie uniquement sur la forme de la lettre, certains allographes, certaines lettres se ressemblent tellement comme le « u » et le « n »[10], le « ſ » (s long) et le « f » qu'on pourrait les représenter par les mêmes signes, si on ne s'appuie plus sur le sens du mot, mais uniquement sur la forme du signe. Pousser l'imitation trop loin risquerait de rendre la transcription impossible à achever et inexploitable[11]. Produire des transcriptions normalisées ne nous a pas semblé être approprié non plus, parce que : d'une part, elles constituaient une perte d'information par rapport à la source et d'autre part, la résolution des abréviations étant un acte interprétatif lié à la spécificité de chacun des documents, elle relève d'un autre geste que celui de la prédiction textuelle[12]. Enfin, notre but étant de produire des modèles génériques, il nous a semblé que la résolution des abréviations aurait pu nuire à l'extension du modèle. Ainsi, pour entraîner des modèles HTR, des transcriptions graphématiques qui conservent les abréviations et la ponctuation originale nous ont semblé être les plus adaptées. Toutefois, la frontière n'étant pas toujours simple à

8 « Transcription for the computer is a fundamentally interpretative activity, composed of a series of acts of translation from one system of signs (that of the manuscript) to another (that of the computer) », Peter Robinson et Elizabeth Solopova, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », juillet 1993.

9 Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ? », BoD, 2011. Ces catégories s'appuient sur l'article fondateur de Peter Robinson et Elizabeth Solopova, Peter Robinson et Elizabeth Solopova, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », juillet 1993.

10 Anne Rochebouet, « Une "confusion" graphique fonctionnelle ? Sur la transcription du u et du n dans les textes en ancien et moyen français », Scriptorium, vol. 63/2, Persée — Portail des revues scientifiques en SHS, 2009, p. 206-219.

11 Dans un premier temps, le projet de transcription des manuscrits de the *Wife of Bath's Prologue* a envisagé de produire une transcription pour partie allographétique, notamment en distinguant les différentes formes de « r ». Alors qu'ils pensaient que cette pratique ne leur coûterait pas beaucoup plus de temps, ils se sont rendu compte qu'elle avait entraîné des erreurs dans le corpus, non seulement à cause d'une pratique hétérogène entre les différents transcripteurs, mais aussi par ce que la concentration appliquée sur la distinction des différentes formes avait entraîné des erreurs grossières dans la transcription. En outre, en essayant de distinguer les différentes formes de « s », ils sont tombés dans une impasse. Plus ils poussaient l'observation, plus ils découvraient des formes différentes et plus les frontières devenaient fragiles.

12 À ce sujet, voir le compte-rendu de la séance n° 1 du séminaire, <https://cremmalab.hypotheses.org/seminaire-creation-de-modeles-htr/compte-rendu-de-la-seance-n1>.

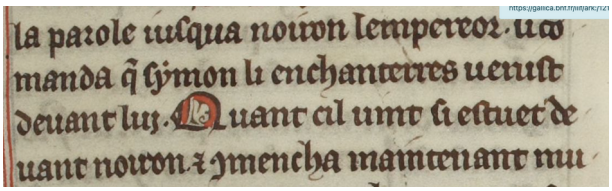
établir entre transcriptions graphématiques et allographétiques, nous chercherons à trouver des issues pragmatiques aux cas ambigus rencontrés dans les documents.

II-Les lettres

1. Généralités

Dans les transcriptions destinées à devenir des données d'entraînement pour l'HTR, la graphie originale du texte des documents doit être conservée et aucune correction d'éditeur ne doit être ajoutée. Des signes différents seront représentés par des caractères différents, mais les variantes de forme d'une même lettre seront réduites à une représentation standardisée. Ainsi, de manière générale, aucune distinction entre les différentes formes de lettres ne sera faite, les ligatures ne seront pas notées, les « ẏ » seront notés « y » pour faciliter la transcription, le point n'ayant aucune signification particulière. De la même manière, le pointage des « i » ne sera pas reproduit. Le positionnement des lettres dans la source sera rendu. Ainsi les lettres ou les signes suscrits seront ajoutés à l'aide de caractères suscrits. Les lettres ou les signes en exposants seront ajoutés à l'aide de caractères suscrits sur des espaces.

Exemple :



BnF, fr. 412

On transcrira :
« la parole iusqua noiron lempereor. li co-
manda q̃ symon li enchanterrres uenist
deuant lui. Quant cil uint si estuet de-
uant noiron ⁊ ꝯmencha maintenant mu- »

- Dans l'image ci-dessus, on observe des variations de forme entre « s » (s rond) et « ſ » (s long). Selon les normes définies plus haut l'un comme l'autre est transcrit par la lettre « s » ;

- On observe des variations de forme entre « r » et « ꝛ » (r rotunda) transcrits « r » l'un comme l'autre ;

- Le signe tironien barré « ꝛ » est représenté par un signe abréviatif tironien « ⁊ » simple, car le signe barré est considéré comme une variation de forme ;

- Le « q » surmonté d'un tilde est noté : « q̃ », l'usage des macrons (tiret droit) ayant été écarté pour éviter la multiplication des signes[13] ;

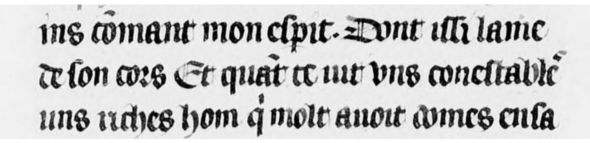
- Le « ẏ » est transcrit « y » ;

- Le pointage des « i » n'est pas pris en compte.

13 Pour la représentation des abréviations, voir *III. Les abréviations*.

2. Les distinctions des « u » et des « v », des « i » et des « j »

Dans les documents médiévaux, la distinction entre les « u » et les « v » ou les « i » et les « j » ne s'appuie pas sur un phénomène phonétique, relève d'une variation de forme.



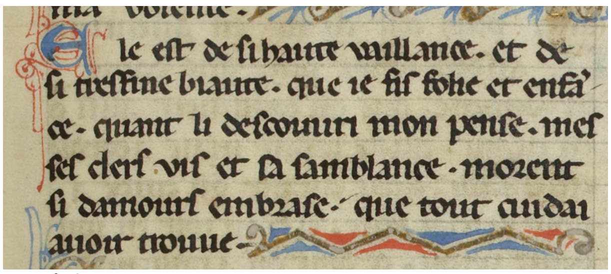
BnF, fr. 411

Dans l'illustration ci-dessus, le déterminant indéfini « un » est écrit « vns » à la ligne 2, puis « uns » à la ligne 3. Ainsi, dans une optique de constitution d'un corpus générique, nous préconisons de ne pas distinguer les « u » et les « v », ni les « i » et « j », et de systématiquement utiliser les caractères « u » et « i »[14].

2. Les majuscules

Il n'est pas toujours aisé de savoir quand mettre des majuscules dans la transcription quand celleci est imitative. En effet, l'identification des phénomènes de mise en relief de certaines lettres est complexe. Il est parfois difficile de déterminer si c'est un changement de module ou un changement de signe, ou bien si c'est un autre caractère ou une variation de ce signe. Toutefois, ces changements sont des informations textuelles importantes qui doivent être relevées quoiqu'elles ne correspondent pas à nos usages modernes (pas de majuscules en début de noms propres, en début de phrases, etc.). Dans la mesure du possible, nous engageons les transcripteurs à signaler ces lettres remarquables du moment où elles sont mises en valeur dans la source. Nous invitons à transcrire à l'aide de majuscules les lettres qui présentent une variation graphique de mise en relief quand il s'agit de :

- Lettrines ou d'initiales ornées ;

 *BnF, fr. 844*	On transcrira la première ligne : « Ele est de si haute uaillance. et de »

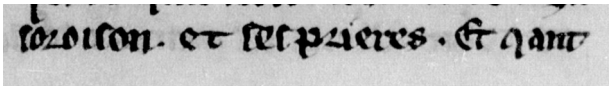
- Titres courants[15];

14 Toutefois, chaque projet peut introduire une distinction, si nécessaire, et personnaliser son propre modèle à condition de bien documenter sa pratique. Une conversion de tous les « v » en « u » et de tous les « j » en « i » *a posteriori* reste aisée, si, par la suite, on veut réintroduire le corpus dans un corpus d'entraînement générique.

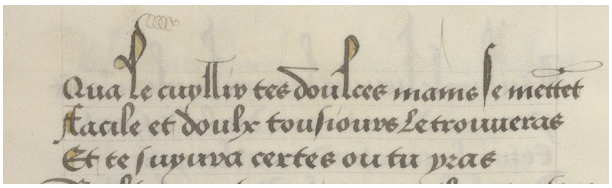
15 Attention tous les titres courants de sont pas en majuscule, conserver une transcription en minuscules quand le texte du titre ne présente de mises en valeur évidentes, exemple : https://gallica.bnf.fr/ark:/12148/btv1b52512226m/f17

 *BnF, fr. 412*	On transcrira : « MARTIN »

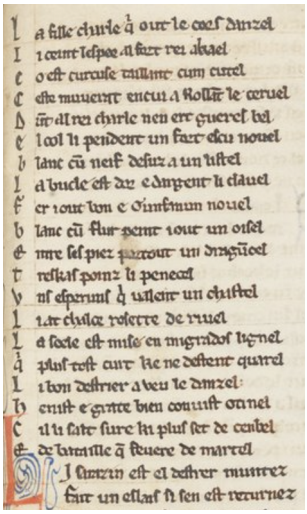
- Lettres en début de mot sémantique

 *BnF, fr.411*	On transcrira : « sorison. et ses prieres. Et quant »

- Nous excluons de cette liste les lettres cadelées.

 *Univ. of Pennsylvania, ms. codex 909*	On transcrira pour la première ligne : « Qua le cuyllir tes doulces mains se mettẽt »

Pour ce qui est des lettres mises en relief en début de vers par une espace typographique et parfois une forme de lettre différente du corps de texte, le transcripteur devra opter pour une pratique homogène adaptée à sa source sur l'ensemble de son corpus. Ainsi, soit toutes ces lettres seront signalées en majuscules, soit en minuscules.



On transcrira au choix :

« l a fille charle q̇ out le cors danzel
l i ceint lespee al fort rei akael
c o est curcuse taillant cum cutel
c este muverat encui a rollãt le cervel
d ũt al rei charle nen ert gueres bel
e l col li pendent un fort escu novel
b lanc cũ neif desuz a un listel
l a bucle est dor e dargent li clavel

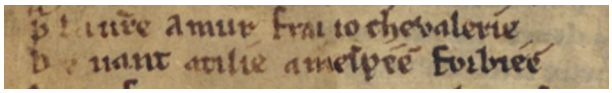
« L a fille charle q̇ out le cors danzel
L i ceint lespee al fort rei akael
C o est curcuse taillant cum cutel
C este muverat encui a rollãt le cervel
D ũt al rei charle nen ert gueres bel
E l col li pendent un fort escu novel
B lanc cũ neif desuz a un listel
L a bucle est dor e dargent li clavel

f er i out bon e gunfanun novel	F er i out bon e gunfanun novel
b lanc cũ flur peint iout un oisel	B lanc cũ flur peint iout un oisel
e ntre ses piez portout un dragũcel	E ntre ses piez portout un dragũcel
t reskas poinz li penecel	T reskas poinz li penecel
u ns esperuns q̇ valent un chastel	U ns esperuns q̇ valent un chastel
l i at chalce rosette de rivel	L i at chalce rosette de rivel
l a seele est mise en migrados lignel	L a seele est mise en migrados lignel
q̇ plus tost curt ke ne destent quarel	q̇ plus tost curt ke ne destent quarel
l i bon destrier a veu le danzel	L i bon destrier a veu le danzel
h enist e grate bien conuist otinel	H enist e grate bien conuist otinel
c il li salt sure ki plus set de cenbel	C il li salt sure ki plus set de cenbel
e de bataille q̃ fevere de martel	E de bataille q̃ fevere de martel
L i sarazín est el destrer muntez	L i sarazín est el destrer muntez
f ait un eslais si sen est returnez »	F ait un eslais si sen est returnez »

N.B. Dans le cadre d'une transcription graphématique, l'ajout de majuscules normalisées qui ne sont pas présentes dans la source est à exclure.

3. Les voyelles accentuées

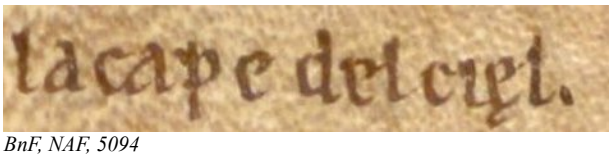
Certains manuscrits, notamment les manuscrits anglo-normands du XII^e^ siècle, peuvent présenter des voyelles accentuées. Ces accents signalent une altération phonétique ou un hiatus, mais les noter peut être la source de certaines ambiguïtés. En effet, comment distinguer un « i » pointé d'un « i » distingué quand l'accent peut désigner une palatalisation, un hiatus, une voyelle tonique ou une désambiguïsation paléographique. Ces accents étant relativement marginaux, nous invitons les transcripteurs à ne pas les noter. Toutefois, le phénomène peut être signalé de manière dérogatoire en fonction des problématiques propres à un corpus.



BnF, NAF, 5094

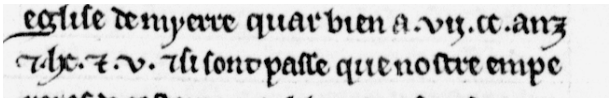
4. Les e cédillés

L'usage des « *e* » *caudata* ou cédilles est très limité en ancien français. Ils signalent l'emplacement d'une ancienne diphtongue latine. Toutefois, ce phénomène est très fréquent dans les manuscrits latins entre le X^e^ et le XII^e^ siècle. Son utilisation étant très limitée en ancien français et donc son coût de transcription moindre, afin de permettre l'interopérabilité des données en latin et en ancien français, nous conseillons de noter le « e cédillé » quand il apparaît dans une source avec un « e̦ » (formé d'un « e » et d'une cédille combinatoire [U+0327]).

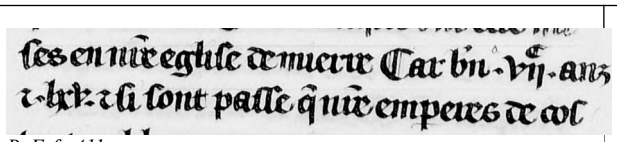
 *BnF, NAF, 5094*	On transcrira : « del cieꝛl »

5. Les chiffres

Les chiffres seront reproduits tels qu'ils sont exprimés dans les documents originaux. Quand les chiffres sont notés par des lettres, ils sont transcrits en minuscules sans distinguer les « i » des « i longs », les « u » des « v »[16]. Afin de faciliter leur repérage et leur traitement[17], nous préconisons de les encadrer systématiquement de points comme on l'observe fréquemment dans les manuscrits. On isolera bien les groupes de chiffres en fonction des coordonnants.

 *BnF, fr. 411*	On transcrira les chiffres présents sur cette image : « .vii.cc. anz ⁊ .lx. ⁊ .v. »

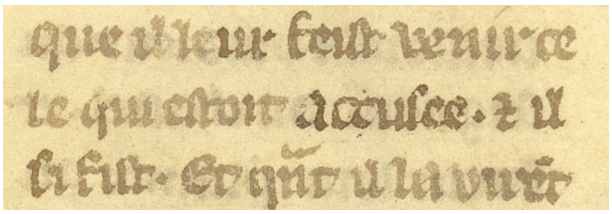
Les chiffres suscrits ou en exposant seront ajoutés à l'aide des caractères suscrits correspondants sur une espace après le dernier chiffre non suscrit et avant le point.

 *BnF, fr. 411*	On transcrira les chiffres présents sur cette image : « .vii^{c}. anz ⁊ .lxv. »

Pour le chiffre quatre-vingt, on transcrira, si le chiffre est représenté comme tel dans la source : « .iu^{xx}. », en ajoutant les « xx » suscrits sur des espaces pour les noter en exposants.

III. Les abréviations

L'ensemble des abréviations sera conservé dans les transcriptions, car leur développement est sujet à interprétation.

 *BnF, fr. 22550*	On transcrira : « que il leur feist uenir ce le qui estoit accusee. ⁊ il si fist. Et qn̂t il la virẽt »

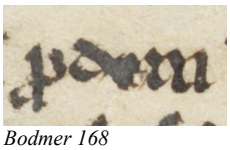
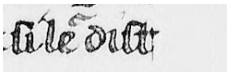
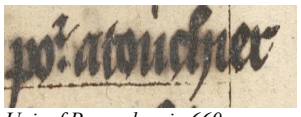
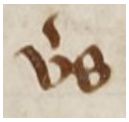
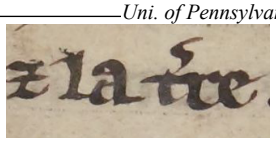
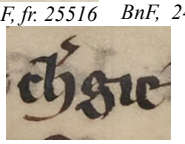
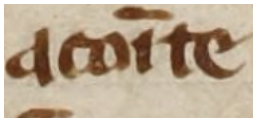
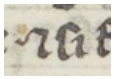
16 La restitution des v pourra être traitée *a posteriori* et restituée pour tous les chiffres.

17 Rajouter les points encadrants permettra à terme des traitements par lot ou d'ajouter des annotations pour signaler que ce sont des chiffres et ainsi les retrouver plus aisément.

Pour les signes abréviatifs, nous conseillons de choisir des signes UTF-8 rattachés au projet MUFI[18]. Ce dernier propose un très large set de caractères avec des variantes de formes. Nous avons donc réduit les possibilités de choix pour les signes abréviatifs afin d'éviter d'avoir plusieurs caractères pour des variations que nous considérons comme relevant d'un phénomène d'allographie, mais aussi afin d'éviter l'utilisation de caractères alternatifs comme « ꝑ » (Lettre minuscule arménienne Ké) pour le p barré ou le chiffre neuf suscrit pour représenter l'abréviation « us » : « ꝰ». Nous avons sélectionné un caractère par abréviation en tenant compte du ou des développements possibles de l'abréviation et de la forme du signe.

18 La *Medieval Unicode Font Initiative* (MUFI) est un projet qui vise à coordonner l'encodage et l'affichage des caractères spéciaux des textes médiévaux écrits en alphabet latin. L'objectif du MUFI est de créer un consensus sur les caractères à encoder et de présenter une proposition complète au Consortium Unicode. <https://mufi.info/m.php?p=mufi>

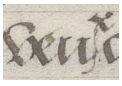
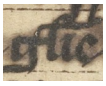
Quelques cas pratiques

 *Bodmer 168*  *BnF, fr. 412*	Nous distinguons les *p barrés courbes* des *p barrés droits* qui ne notent pas les mêmes abréviations (« pro » contre « per/par »).
 *BnF, fr. 411,*  *Uni. of Pennsylvania 660*	Pour les « r » suscrits, nous ne faisons pas de distinction de forme, ils sont tous notés « ʳ ».
 *BnF, fr. 25516*  *BnF, 24428*  *BnF, fr. 24428*	les tildes verticaux sont tous représentés par le signe « ◌̃ ».
 *BnF, fr. 22550*  *BnF, fr. 25516*	Les abréviations signalées par des tildes horizontaux ou des macrons sont rendues par un tilde horizontal suscrit : « ◌̄ ». il est recommandé de placer les tildes sur la même lettre que sur le document original. En cas de doute, il est préconisé d'adopter une méthode aussi homogène que possible.
 *BnF, fr. 412*  *BnF, fr. 22550*	Les « et » tironiens quelque soit leur forme sont tous représentés par le signe « ⁊ »

Les abréviations ont été organisées selon les catégories suivantes :

- Les tildes
- Les abréviations *par lettres suscrites.*
- Les abréviations par *signes spéciaux :* lettres barrées (d, l, p, q) parmi les plus courantes. Si jamais d'autres signes spéciaux étaient nécessaires (comme ħ ou ꝼ), chaque projet est libre d'ajouter les signes dont il aurait besoin à condition de le documenter et de s'appuyer sur les caractères proposés par la MUFI.

		Signe	**code caractère**
Les tildes			

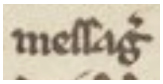
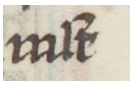
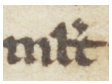
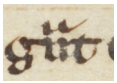
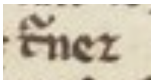
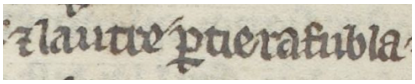
COMBINING TILDE[19] (*tilde*)		◌̃	U+0303
COMBINING VERTICAL TILDE (*tilde vertical*)		◌̾	U+033E
Les abréviations par lettres suscrites			
COMBINING LATIN SMALL LETTER A (*a suscrit*)		◌ᵃ	U+0363
COMBINING LATIN SMALL LETTER C (*c suscrit*)		◌ᶜ	U+0368
COMBINING LATIN SMALL LETTER E (*e suscrit*)		◌ᵉ	U+0364
COMBINING LATIN SMALL LETTER I (*i suscrit*)		◌ⁱ	U+0365
COMBINING LATIN SMALL LETTER M (*m suscrit*)		◌ᵐ	U+036B
COMBINING LATIN SMALL LETTER O (*o suscrit*)		◌ᵒ	U+0366
COMBINING LATIN SMALL LETTER R (*r suscrit*)		◌ᵣ	U+036C
COMBINING LATIN SMALL LETTER S (*s suscrit*)		◌ˢ	U+1DE4
COMBINING LATIN SMALL LETTER T (*t suscrit*)		◌ᵗ	U+036D
COMBINING LATIN SMALL LETTER X (*x suscrit*)		◌ˣ	U+036F
COMBINING UR ABOVE (*« ur » suscrit*)		◌᷎	U+1DD1
Les abréviations par signes spéciaux			
LATIN SMALL LETTER CON (*abréviation « com »*)		ꝯ	U+A76F
COMBINING US ABOVE (*abréviation « us » en exposant*)		◌ᵘ	U+1DD2
LATIN SMALL LETTER D WITH STROKE (*d barré*)		đ	U+0111
LATIN SMALL LETTER L WITH STROKE (*l barré*)		ł	U+0142

19 L'ensemble des noms utilisés dans cette colonne sont issus du guide de la MUFI, disponible au lien suivant :

LATIN SMALL LETTER P WITH STROKE (*p barré droit*)		ꝑ	U+A751
LATIN SMALL LETTER P WITH FLOURISH (*p barré courbe*)		ꝓ	U+A753
LATIN SMALL LETTER Q WITH DIAGONAL STROKE (*q barré*)		ꝗ	U+A759
TIRONIAN SIGN ET (*abréviation tironienne de « et »*)		⁊	U+204A
DIVISION SIGN (*abréviation de « est »*)		÷	U+00F7

La saisie des transcriptions dans l'interface *eScriptorium*[20] peut être facilitée par l'utilisation du clavier du projet CREMMALab[21] qui consigne toutes les lettres suscrites et caractères spéciaux présents dans le tableau ci-dessus.

Quelques exemples :

Terme dans la source	***Transcription***
	messag᷾
	mlt̄
	ml̓t
	normẽdie
	gn̊t
	q᷽
	t̃nez
	⁊lautre ꝑtie rafubla

20 B. Kiessling, R. Tissot, P. Stokes, [et al.], « EScriptorium: An Open Source Platform for Historical Document Analysis », 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 2, 2019, p. 19— 19.

21 https://github.com/HTR-United/cremma-medieval/blob/main/CremmaLab.json

	ů

Remarque : Les abréviations par suspension seront représentées par un point « . », quand le point (ou un autre signe équivalent) est présent dans le manuscrit. Aucun point ne sera rajouté, si aucun signe n'apparaît sur le document source. Exemple : s. (saint).

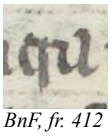
IV. La segmentation des mots et séquences

1. Séparation des mots

La segmentation des mots est un élément primordial pour la compréhension des textes. Cependant, dans les manuscrits médiévaux, elle n'est pas identique au système moderne. La distinction de la présence ou non d'une espace typographique peut parfois relever de la pure subjectivité, car cette notion n'a véritablement de sens que pour les imprimés. Toutefois, quoique nous ayons conscience que la pratique sera toujours hétérogène et comportera une part d'arbitraire, nous préconisons une pratique qui s'appuie sur le sens du texte et sépare dans la mesure du possible les mots sémantiques[22]. Une pratique imitative introduirait trop de bruit.

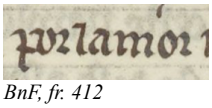
Ainsi, pour assurer la plus grande homogénéité possible, une segmentation modernisée qui sépare les mots est préconisée, même si cela n'évite pas quelques ambiguïtés.

Un cas fait exception à ce principe. Quand l'usage moderne voudrait faire apparaître une élision, alors que le manuscrit médiéval pratique une agglutination, l'agglutination sera conservée.

 *BnF, fr. 412*	On transcrira : « qil »

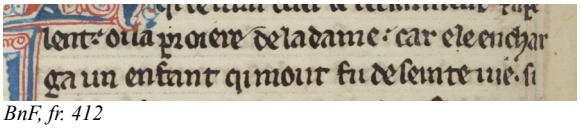
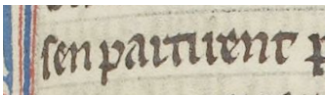
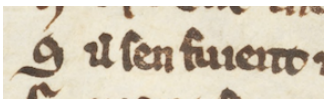
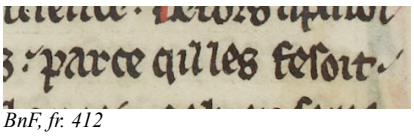
Remarques :

En cas de doute sur la segmentation des mots dans le manuscrit, nous recommandons d'adopter une segmentation conforme à celle du français contemporain.

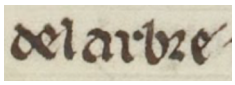
 *BnF, fr. 412*	On transcrira : « por lamor »

Certains cas peuvent être plus difficiles à trancher notamment pour des verbes comme : « enchargier », « en fuir », « en partir » ou certaines locutions. On essaiera de rester le plus proche de la source possible et en cas de doute de respecter soit l'entrée du dictionnaire de référence, soit l'usage moderne.

22 En outre une pratique sémantique facilitera le passage vers la lemmatisation

 *BnF, fr. 412*	On transcrira : « encharga »
 *BnF, fr. 412*	On transcrira : « sen partirent »
 *Bodmer 168*	On transcrira : « sil sen fuient »
 *BnF, fr. 412*	On transcrira : « parce qil »

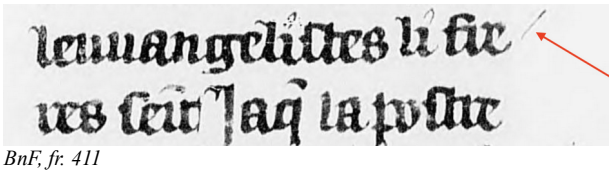
Certaines séquences peuvent avoir deux solutions possibles : « *del arbre* » ou « *de larbre* ». On s'efforcera d'imiter au mieux la source.


BnF fr. 412

2. Hyphénisation

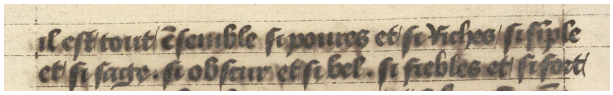
Il arrive que les copistes notent en fin de ligne les phénomènes d'hyphénisation. Quand un signe apparaît en fin de ligne, l'usage du signe « - » (hyphen, U+002D) est préconisé.

Exemple :

 *BnF, fr. 411*	On transcrira : « leuuangelistes li fre- res seĩt Iaq̃ lapostre »

3. Diastoles

Parfois, des diastoles (traits de plume vertical ou oblique) sont tracées entre deux lettres contiguës pour indiquer qu'elles appartiennent à des mots différents. Nous conseillons de les transcrire à l'aide du signe suivant : « / » (U+002F).

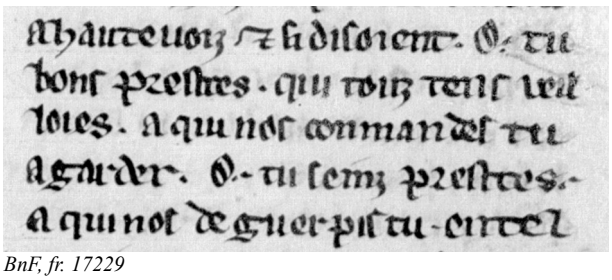

Univ. of Pennsylvania, ms. 660

V. La ponctuation

En raison de la grande variété de la ponctuation médiévale[23], nous proposons d'utiliser dans les transcriptions un système simplifié. Présumer de la fonction exacte des signes dès la phase de transcription est complexe et signaler toutes les nuances relèverait d'une annotation sémantique *a posteriori*. La variété de la ponctuation médiévale sera réduite à trois types de signes pour assurer l'homogénéité de la transcription :

- Les points simples seront transcrits par des « . »
- Les signes doubles seront transcrits par des « ; »
- Les virgules seront rendues par des « , » (les « / » étant réservés aux diastoles, *cf. supra*)

L'ensemble des signes de ponctuation seront transcrits directement après le signe qui les précède sans espace.



BnF, fr. 17229

On transcrira :
« a haute uoiz ⁊ si disoient. O; tu
bons prestres. qui touz tens ueil
loies. a qui nos conmandes tu
agarder. O; tu seinz prestres;
a qui nos deguerpis tu. en tel »

VI. Corrections, ajouts, et signes fonctionnels

Les corrections présentes dans les sources pourront être représentées en fonction des cas selon deux niveaux : par ajouts de signes dans le corps du texte ou au moyen de la mise en page. En effet, la conception des données destinées à l'entraînement de modèles HTR ne s'appuie pas uniquement sur des éléments de transcription, mais également sur la description de la mise en page des sources qui est la phase qui précède l'HTR proprement dit. Ainsi, la représentation des documents s'appuiera aussi sur l'étape de segmentation des zones et des lignes de la page afin de décrire l'emplacement du texte sur son support et de représenter, par exemple, les notes marginales ou les ajouts interlinéaires.

23 Un rapide inventaire des signes les plus courants nous a permis de distinguer huit signes différents qui ne correspondent pas à nos usages modernes. Certains signes de ponctuation ont des fonctions moindres et ne peuvent être utilisés que pour indiquer des bouts de lignes. Voici une sélection de signes de ponctuation présents dans les manuscrits médiévaux (liste établie d'après la définition de Philippe Bobichon, disponible sur codicologia, <http://codicologia.irht.cnrs.fr/theme/liste_theme/422#tr-8661>) :

- Point : ponctuation utilisée notamment pour séparer les termes d'une énumération.
- Point bas, médian : subdivisions mineures au sein de la phrase.
- Point haut : le signe le plus fort, marquant la pause la plus longue ; il est souvent suivi d'un blanc.
- Point-virgule inversé : pause moyenne, indique une montée de la voix, ou différentes pauses au milieu de la phrase.
- Point-virgule : descente de l'intonation en fin de phrase ; fins de paragraphes (ponctuation forte).
- Punctus circumflexus : point surmonté d'un accent circonflexe (sporadique). Caractéristique des manuscrits cisterciens, il apparaît aussi chez les Chartreux.
- Deux points superposés : dans les manuscrits grecs, ce procédé, devenu usuel, sera étendu à la prose pour marquer l'interlocution dans le dialogue philosophique. Dans les manuscrits hébreux bibliques, les versets sont parfois séparés par deux points superposés
- Virgula : ponctuation faible. Mince trait oblique, ancêtre de notre virgule, séparant notamment les termes d'une énumération.

1. Corrections, ajouts et renvois signalés par des éléments de mise en page

Toutes les corrections qui relèvent de la mise en page devront être signalées au moment de la segmentation de l'image grâce au nommage des zones et des lignes qui permettent de typer les différents éléments de la mise en page. Les corpus des projets de l'École nationale des chartes s'appuient sur le vocabulaire contrôlé *SegmOnto*[24] dont voici la liste de noms de zone et de ligne.

Zones	**Lignes**
DigitizationArtefactZone	DefaultLine
DropCapitalZone	DropCapitalLine
GraphicZone	HeadingLine
MainZone	InterlinearLine
DamageZone	MusicLine
MarginTextZone	
MusicZone	
NumberingZone	
QuireMarksZone	
RunningTitleZone	
SealZone	
StampZone	
TableZone	
TitlePageZone	

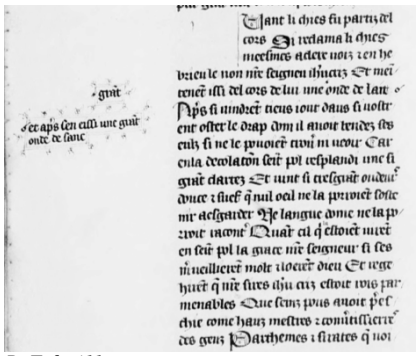
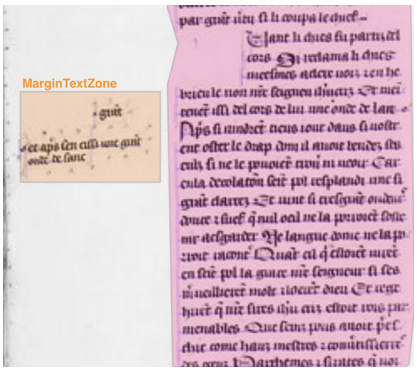
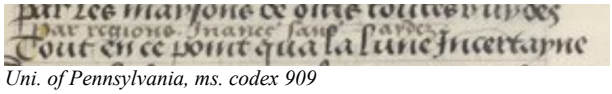
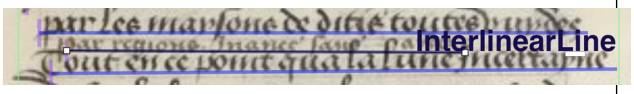
1.1 Ajouts et corrections interlinéaires ou marginaux

Les corrections, les notes ou gloses marginales seront signalées par une zone *MarginTextZone*.

Les corrections ou les gloses interlinéaires seront signalées par le type de ligne : *InterlinearLine*.

24 Dans le cadre des corpus CREMMA et CREMMALab de l'école nationale des chartes, la segmentation des documents s'appuie sur le vocabulaire contrôler du projet segmOnto. Cette initiative est le fruit d'une collaboration entre les projets eScriptorium et Kraken (Benjamin Kiessling, Daniel Stoëckl, Peter Stokes), cremmaLab (projet ENC-INRIA), l'INRIA et l'université de Genève. Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, [et al.], « SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more) », *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland, 2021. La documentation du projet est disponible à l'adresse suivante : https://segmonto.github.io.

Glose marginale

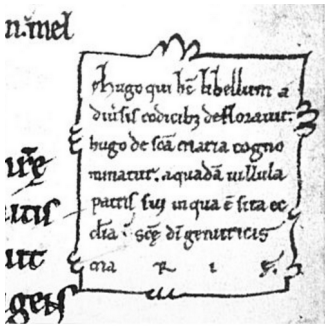
 *BnF, fr. 411*	
Ajout interlinéaire	
 *Uni. of Pennsylvania, ms. codex 909*	

1.2 Signes fonctionnels[25]

Les signes fonctionnels seront signalés s'ils sont encadrant par des *MarginTextZones* qui contiendront le texte encadré, s'ils ne le sont par des *GraphicZones* qui n'engloberont que le signe fonctionnel.

Signe	Exemple	Proposition de représentation
• Festons et accolades		zone *MarginTextZone*

25 L'ensemble des exemples cités et des illustrations qui apparaissent dans la suite du guide sont issus de *Codicologia* : http://codicologia.irht.cnrs.fr.

• Circumductions et cartouches		zone *MarginTextZone*
• Manicules		zone *GraphicZone*

2. Corrections, ajouts et renvois signalés par des signes dans la transcription

La question de la représentation corrections et des signes auxiliaires doit également être résolue. Il s'agit de proposer un système simple et généraliste qui puisse aider la plus large communauté scientifique possible[26].

2.1 Corrections

Nous traiterons ici des corrections qui relèvent d'une intervention du copiste dans le corps du texte.

Pour signaler les corrections suivantes : exponctuation, texte souligné, biffure ou rature, Le texte sera encadré par des doubles crochets droits ouvrant : « ⟦ » (U+27E6) et des doubles crochets droits fermant « ⟧ » (U+ 27E7) de la manière suivante : ⟦ texte corrigé ⟧[27].

Quand la correction consiste en un blanc laissé en attente ou en un grattage qui rend le texte illisible, nous conseillons d'utiliser les mêmes signes « ⟦» et « ⟧ » que précédement, mais cette fois sans texte à l'intérieur.

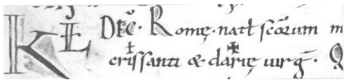
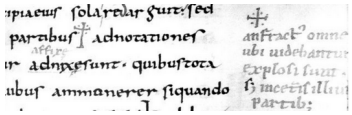
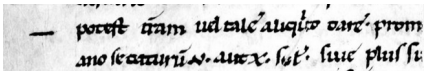
26 Définir une norme détaillée pour représenter les signes de correction ou les ajouts semble impossible, car nous aurions besoin d'une enquête codicologique exhaustive qui se concentrerait sur leurs modalités dans les manuscrits médiévaux pour établir une ontologie et des correspondances.

27 Ces signes sont ceux sélectionnés par la convention de Leiden, <https://en.wikipedia.org/wiki/Leiden_Conventions>.

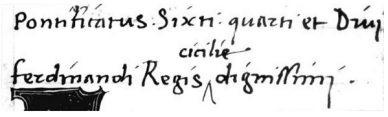
Dans le cas d'un grattage où le texte est lisible, le texte sera transcrit sans indication. La procédure sera la même en cas de transformation d'une lettre en une autre. Pour les palimpsestes, seule la dernière strate de texte sera transcrite.

2.2. Les signes de renvois

Dans les manuscrits, il n'est pas rare que les ajouts soient signalés par des signes fonctionnels qui permettent de faire le lien entre la zone de texte principale et l'ajout. Leur signalement peut se révéler utile pour le post-traitement des prédictions HTR et ainsi permettre de lier les ajouts au corps de texte. Afin de ne pas multiplier leurs représentations, des signes génériques seront utilisés en fonction du type de renvoi. Ainsi, seront représentés par des « ⁜ » **(astérisques ou *dotted crosses*, U+205C*)***, les signes de renvois suivants[28] :

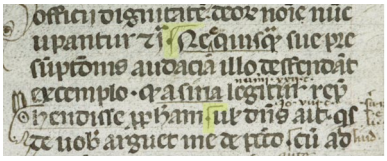
- Croisette 
- Astérisque 
- Obèle 

Les carets qui servent surtout à insérer des gloses interlinéaires seront signalés par des chevrons d'insertion « ˄ » (U+2038) pour les distinguer des autres signes de renvois qui fonctionnent selon un système d'appel et de rappel.

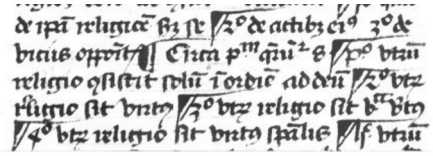


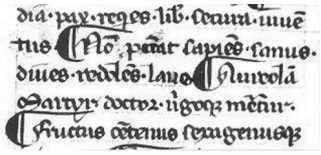
2.3. Les autres signes fonctionnels

Le texte peut être hiérarchisé au moyen de signes fonctionnels qui signalent le début d'une unité sémantique, comme un paragraphe.

- Gamma capitulaire 

28 Pour retrouver les signes correspondant à chacune des catégories, voir http://codicologia.irht.cnrs.fr.

- Crochet alinéaire 

- Pied-de-mouche 

Tous ayant la même fonction[29], ils seront représentés par un signe unique : « ¶ » (Pied-de-mouche, U+00B6).

Conclusion

Ainsi, produire des données d'entraînement demande de modéliser une représentation la matière textuelle afin de produire des données homogènes qui assureront la qualité des prédictions HTR. Ce guide ne prétend pas répondre à toutes les problématiques de transcriptions présentes dans les manuscrits médiévaux. Nous ne doutons pas que pour chaque projet de nouvelles questions émergeront. Nous espérons, toutefois, qu'il permettra aux transcripteurs de répondre aux interrogations les plus courantes et d'harmoniser les pratiques.

Ce guide ne résout pas non plus les problèmes d'interopérabilité entre des données issues de projets différents. Dans le cadre de la constitution du dépôt *cremma-medieval*[30], les transcriptions étant issues de sources diverses, la difficulté a été rencontrée et partiellement résolue. Pour régler des différences mineures et assurer l'homogénéité du jeu de caractères utilisés, un outil de contrôle et de conversion d'encodage du texte : *Choco-mufin*[31] a été mis en place. Il permet de remplacer des caractères non orthodoxes par ceux définis pour un projet donné. Toutefois restent les problèmes d'interopérabilité entre des jeux de données issues de périodes ou d'ères linguistiques différentes dont les choix de transcription seraient profondément différents. Chaque nouveau regroupement demandera la mise en place de nouvelles réflexions pour harmoniser les corpus en fonction du but visé et une simplification qui entraînera une perte d'information pour permettre l'uniformisation du jeu de données.

29 Leurs différences relèvent de la variation de forme.

30 Ariane Pinche, *op. cit.*

31 Thibault Clérice et Ariane Pinche, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, 2021.

Références bibliographiques

CHAGUÉ, Alix, CLÉRICE, Thibault et CHIFFOLEAU, Floriane, *HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages*, 2021, [En ligne : https://github.com/HTR-United/htr-united].

CLÉRICE, Thibault et PINCHE, Ariane, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, 2021, [En ligne : https://github.com/PonteIneptique/choco-mufin].

GABAY, Simon, CAMPS, Jean-Baptiste, PINCHE, Ariane, [et al.], « SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more) », *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland, 2021, [En ligne : https://hal.archives-ouvertes.fr/hal-03336528].

KIESSLING, B., TISSOT, R., STOKES, P., [et al.], « EScriptorium: An Open Source Platform for Historical Document Analysis », *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 2019, p. 19-19.

PINCHE, Ariane, *CREMMA Medieval, an Old French dataset for HTR and segmentation*, 2021, [En ligne : https://github.com/HTR-United/cremma-medieval].

PINCHE, Ariane, « CREMMALAB | Constitution de corpus en ancien français pour l'HTR », 2022, [En ligne : https://cremmalab.hypotheses.org/].

ROBINSON, Peter et SOLOPOVA, Elizabeth, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », juillet 1993, [En ligne : https://zenodo.org/record/4050360].

ROCHEBOUET, Anne, « Une « confusion » graphique fonctionnelle? Sur la transcription du u et du n dans les textes en ancien et moyen français », *Scriptorium*, vol. 63 / 2, Persée - Portail des revues scientifiques en SHS, 2009, p. 206-219.

Table des matières