Guide de transcription pour les imprimés français du XVIe siècle en caractères gothiques

Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia M Oliveira, Daniela Solfaroli Camillocci

Guide de transcription pour les imprimés français du XVI^{e} siècle en caractères gothiques

Version A

Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira et Daniela Solfaroli Camillocci

Institut d'Histoire de la Réformation
Université de Genève
{prénom.nom}@unige.ch

05-10-2023

Table des matières

Introduction

Ce guide de transcription a pour objectif de présenter les réflexions et les normes établies dans le cadre du projet de recherche SETAF [1], dont le corpus primaire est composé d'imprimés français en caractères gothiques, publiés à Neuchâtel et à Genève dans la première moitié du XVI^e^ siècle.

Il est important de souligner que les normes de transcription que nous proposons ici ne constituent pas les règles d'établissement du texte final, mais celles de la production de vérités de terrain, c'est-à-dire de données d'entraînement pour un moteur d'OCR (*Optical Character Recognition*). Des étapes de post-traitement sont prévues plus en aval de la chaîne de traitement pour permettre l'analyse et la publication du texte.

L'établissement de ce guide répond à deux exigences scientifiques : d'une part, la nécessité de présenter les choix philologiques faits dans le cadre de notre projet, qui doivent tenir compte des spécificités de notre corpus ; d'autre part, la volonté de partager nos pratiques et nos données afin de les rendre réutilisables par d'autres projets.

Vu que la fiabilité et l'efficacité des modèles de segmentation et d'OCR dépendent de la quantité et de la qualité des données d'entraînement, l'uniformisation des pratiques de transcription est fondamentale pour construire un ensemble toujours plus large de corpus partageant certaines normes et, par conséquent, pour élaborer des modèles toujours plus performants.

Ainsi, nous souhaitons que ce guide contribue à nourrir les réflexions d'autres équipes de recherche dans le but de créer un protocole aussi générique que possible pour les imprimés du XVI^e^ siècle en langue française.

1 Principes généraux de transcription

> *[Les différents niveaux de transcription] constituent en réalité différents maillons d'une seule et même chaîne de régularisation, chaque niveau découlant du précédent. Il est naturellement possible de sauter des étapes et de produire d'emblée une transcription régularisée – c'est d'ailleurs ce que l'on fait la plupart du temps – mais il s'agit là d'une opération délicate, […] qui fait par ailleurs perdre de manière irréversible de nombreuses informations graphiques.* (André et Rémi 2013, p. 117)

En reprenant les définitions proposées par Dominique Stutzmann (2011), deux grandes catégories de transcription peuvent être envisagées : celle qui décrit l'image, dite transcription « allographétique » , et celle qui décrit le texte, dite transcription « graphématique ». La transcription allographétique vise à conserver toute la richesse graphique et donc toutes les variantes des caractères (lettres, symboles, signes de ponctuation, etc.). En revanche, la transcription graphématique limite la variation en réduisant les différents glyphes possibles à leur sens dans un système alphabétique. Quelques points restent cependant débattus, notamment la question des abréviations.

Ariane Pinche (2022), qui a rédigé un guide de transcription pour les manuscrits gothiques du X^e^ au XV^e^ siècle, fait une distinction importante entre les transcriptions qui visent à entraîner des modèles génériques et les transcriptions pour des éditions critiques. Quant aux abréviations, elle opte pour leur conservation : « pour entraîner des modèles

1. Site du projet SETAF : https://www.unige.ch/setaf.

HTR, des transcriptions graphématiques qui conservent les abréviations et la ponctuation originale nous ont semblé être les plus adaptées »(Pinche 2022, p. 3).

Marie-Luce Demonet et l'équipe des BVH (Bibliothèques Virtuelles Humanistes), qui ont rédigé les principes éditoriaux d'*Epistemon – Corpus de textes de la Renaissance*[2], soulignent que, malgré l'hétérogénéité des documents et les évolutions techniques survenues depuis le début du programme des BVH, tous les états de transcription de leur corpus ont en commun :

- *le respect de la mise en page et du lignage ;*
- *l'absence d'adjonction d'alinéas, de guillemets, de tirets ;*
- *l'absence d'intervention sur l'usage des majuscules ;*
- *le respect des graphies (sauf ‹i›‹j›/‹u›‹v› et les abréviations pour les transcriptions les plus anciennes) et de la ponctuation ;*
- *la correction de coquilles ou d'erreurs manifestes (signalées et désormais encodées en TEI) à l'aide d'une autre édition ou de probabilités graphiques.*[3]

Dans le cadre de notre projet, nous avons opté pour une transcription graphématique qui conserve la ponctuation originale et les abréviations.

Dans les paragraphes suivants, nos choix de transcription et les cas ambigus seront expliqués en détail, à l'aide d'exemples tirés de notre corpus. Pour chaque exemple, nous indiquerons la source de façon abrégée et chaque référence sera développée dans la bibliographie, en conformité avec les conditions d'utilisation des portails numériques utilisés[4]. En outre, pour chaque caractère pris en compte, nous indiquerons son numéro Unicode et, si nécessaire, son numéro MUFI (*Medieval Unicode Font Initiative*)[5].

Nos réflexions auront comme point de départ d'autres guides, dont certains ont déjà été évoqués : le guide de transcription de Pinche (ibid.) pour les manuscrits gothiques du X^e^ au XV^e^ siècle, les principes éditoriaux d'*Epistemon – Corpus de textes de la Renaissance* des BVH et les recommandations de transcription de Gabay, Clérice et Reul (2023) pour les imprimés français du XVII^e^ siècle. Une attention particulière sera accordée au guide de Pinche (2022) ayant permis de produire un modèle d'OCR pour les manuscrits gothiques, un modèle que nous souhaitons affiner pour les imprimés français du XVI^e^ siècle en caractères gothiques.

Notre guide abordera les points suivants : la ponctuation, les lettres, les abréviations, les chiffres, la séparation des mots, la description de la mise en page et l'interopérabilité avec d'autres corpus, notamment des imprimés français du XVI^e^ siècle en caractères romains (ou en *antiqua*).

2. *Epistemon* a été développé au CESR de Tours et il s'agit d'un grand corpus d'imprimés du XVI^e^ siècle. Il est en ligne à l'adresse suivante : https://www.bvh.univ-tours.fr/Epistemon/index.asp.

3. Voir les principes éditoriaux d'*Epistemon* : https://www.bvh.univ-tours.fr/Epistemon/principeseditoriaux.asp.

4. Dans ce guide nous reproduirons beaucoup d'exemples tirés de documents numériques. Leurs droits d'usage sont précisés en ligne par les différents portails, comme e-rara ou Österreichische Nationalbibliothek (ONB). Pour ces deux exemples, les conditions d'utilisation sont disponibles aux adresses suivantes : https://www.e-rara.ch/wiki/termsOfUse?lang=en ; https://www.onb.ac.at/en/use.

5. Site de l'Unicode Consortium : https://home.unicode.org/about-unicode/ ; Site de la MUFI : https://mufi.info/q.php?p=mufi/home.

2 La ponctuation

La ponctuation française du XVI^e^ siècle a fait l'objet de nombreux travaux, comme la journée d'étude organisée par Dauvois et Dürrenmatt (2011), ou l'analyse quantitative des pratiques de ponctuation observées à partir du corpus *Epistemon* par Alexei Lavrentiev (2016). En recensant les signes de ponctuation les plus utilisés au XVI^e^ siècle, il remarque que :

> *La virgule arrive en première position (162 920 occurrences), suivie de loin par le point (68 883 occurrences) et les deux-points (21 343 occurrences). […] La barre oblique n'est employée que dans les textes les plus anciens du corpus […]. Dans l'analyse globale du corpus, on peut la fusionner avec la virgule, car dans les traités de l'époque les deux marques étaient considérées comme équivalentes.* (ibid., p. 14)

Pour des raisons de rigueur philologique et d'interopérabilité, nous avons décidé, comme l'équipe des BVH - *Epistemon*, de respecter la ponctuation originale des sources, qui ne correspond pas toujours à la ponctuation du français moderne.

En raison de la grande variété de la ponctuation médiévale, Pinche (2022) propose de transcrire les points simples par des ‹.›, tous les signes doubles par des ‹ ;›, les virgules par des ‹,› et les diastoles par des ‹/›. Notre guide concernant essentiellement la première moitié du XVI^e^ siècle, il doit tenir compte d'un système mixte, à la fois médiéval et moderne, qui nous amène à faire des choix différents. Toutefois, nos recommandations privilégiant un plus grand respect du système original (par ex. la distinction entre la virgule et la barre oblique), il reste aisément possible d'aligner à postériori nos transcriptions avec les recommandations d'A. Pinche ou les suggestions d'A. Lavrentiev par une série d'opérations simples (par ex. ‹/› → ‹,›), et donc de conserver *in fine* l'interopérabilité entre les données.

En ce qui concerne la virgule et la barre oblique, nous avons choisi de les distinguer afin de pouvoir étudier l'évolution de leur utilisation par les imprimeurs de notre corpus. Par exemple, Pierre de Vingle[6] utilise presque toujours des barres obliques, mais Jean Michel[7] introduit l'usage de la virgule de façon progressive : ses premières publications présentent peu de virgules et beaucoup de barres obliques, tandis que dans ses dernières publications cette pratique est renversée. L'introduction de la virgule se présente donc comme l'un des rares traits typographiques qui distinguent les réimpressions genevoises de Michel des publications neuchâteloises de Vingle (Berthoud 1980).

Nous avons décidé de distinguer les signes doubles parce qu'il nous semble important de garder ce type d'information graphique pour des imprimés du XVI^e^ siècle, qui témoigne de nuances sémantiques potentiellement intéressantes. Dans notre corpus, les ‹;› semblent absents mais les ‹:› sont assez répandus. En outre, nous avons remarqué la présence d'un signe qui ressemble à un point d'interrogation, servant parfois à marquer l'interrogation mais d'autres fois l'exclamation. Ce signe sera transcrit par le caractère Unicode *Question Mark* (`U+003F`) : ‹?›.

En ce qui concerne la coupure des mots en fin de ligne, c'est-à-dire l'opération qui consiste à rejeter sur la ligne suivante la fin d'un mot trop long, nous proposons d'utiliser

6. Maître-imprimeur à Lyon en 1525-1532, à Genève en 1532-1533 et à Neuchâtel en 1533-1535. Voir le R.I.E.C.H. : https://db-prod-bcul.unil.ch/riech/riech.php.

7. Maître-imprimeur à Genève de 1538 à 1544. Voir le R.I.E.C.H. : https://db-prod-bcul.unil.ch/riech/riech.php.

le caractère Unicode *Not sign* (U+00AC, ‹¬›). Ce choix est le fruit de plusieurs réflexions.

Le tiret court ou tiret quart de cadratin ou signe moins, c'est-à-dire le caractère Unicode *Hyphen-minus* (U+002D, ‹-›), est un caractère ambigu. Il peut être utilisé, selon le contexte, comme signe moins, comme séparateur de chiffres, comme tiret d'intervalle, pour la césure d'un mot en fin de ligne et comme trait d'union dans tous les autres cas (par exemple, pour les mots composés comme « belle-mère » ou pour lier les pronoms enclitiques au verbe précédent comme « dis-moi »).

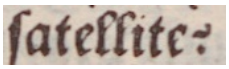
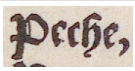
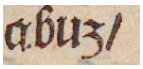
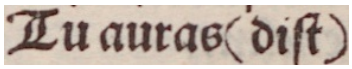
Dans notre corpus, le caractère Unicode *Double oblique hyphen* (U+2E17, ‹=›) note la coupure des mots en fin de ligne et la composition dans certains noms propres (Baddeley 1993). Dans un premier temps, nous avons pensé le transcrire par le caractère Unicode *Hyphen-minus* (U+002D, ‹-›) parce qu'il correspond à la fois au caractère utilisé par les imprimés en romain de la même époque et aux recommandations d'A. Pinche, tout en étant plus accessible sur le clavier. Ensuite, nous avons réfléchi aux prochaines étapes de la chaîne de traitement de notre corpus et nous avons réalisé qu'il était souhaitable de prévoir un signe pour indiquer la coupure des mots en fin de ligne et un autre signe pour le trait d'union dans tous les autres cas. En fait, lors de l'encodage en TEI (*Text Encoding Initiative*) et de la lemmatisation, il est nécessaire d'apporter une attention particulière à la coupure des mots en fin de ligne afin de pouvoir signaler et reconstituer le mot coupé dans son ensemble. Afin de rendre ce passage technique le plus rapide et automatique possible, il est souhaitable d'avoir un caractère spécifique pour le signe indiquant la coupure des mots en fin de ligne.

Après avoir consulté les recommandations de transcription de Gabay, Clérice et Reul (2023), qui tiennent en compte une chaîne de traitement similaire à celle prévue pour notre corpus, nous avons adopté leur suggestion et nous avons opté pour la pratique suivante : le signe indiquant la coupure des mots en fin de ligne sera transcrit par le caractère Unicode *Not sign* (U+00AC, ‹¬›) et le signe indiquant le trait d'union dans tous les autres cas sera transcrit par le caractère Unicode *Hyphen-minus* (U+002D, ‹-›).

Nous précisons qu'il reste possible d'aligner à postériori nos transcriptions avec les recommandations d'A. Pinche ou des BVH par une série d'opérations simples (par ex. ‹¬›→ ‹-›), et donc de conserver l'interopérabilité entre les données.

Voici un tableau des signes de ponctuation recensés dans notre corpus et les choix de transcription correspondants :

Tableau 1 – Les signes de ponctuation

Exemple	Source	Transcrire	Unicode
	([Malingre] 1533a), p. 5	Foy.	U+002E
‹.› n'est pas précédé d'un espace, il est suivi d'un espace. Pour les chiffres et les points, voir le paragraphe consacré aux chiffres.			
	([Malingre] 1533a), p. 5	ages:	U+003A
‹:› n'est pas précédé d'un espace, il est suivi d'un espace.			
	([Malingre] 1533 ?) p. 26	satellite?	U+003F
Le signe de l'exemple ressemble à ‹?›, mais il est utilisé parfois comme un point d'interrogation et d'autres fois comme un point d'exclamation. Il n'est pas précédé d'un espace, il est suivi d'un espace.			
	([Malingre] 1533a) p. 28	Peche,	U+002C
‹,› n'est pas précédé d'un espace, il est suivi d'un espace.			
	([Malingre] 1533a) p. 5	abuz/	U+002F
‹/› n'est pas précédé d'un espace, il est suivi d'un espace.			
	([Malingre] 1533a) p. 5	de pe¬	U+00AC
‹¬› n'est pas précédé d'un espace.			
	([Malingre] 1533a) p. 7	¶ Le docteur	U+002D
‹¶› est suivi d'un espace.			
	([Malingre] 1533 ?) p. 33	Tu auras (dist)	U+0028 U+0029
‹(› est précédé d'un espace mais il n'est pas suivi d'un espace. ‹)› n'est pas précédé d'un espace mais il est suivi d'un espace.			

3 Les lettres

Comme celle du Moyen Âge, la langue du XVI^e^ n'est pas encore standardisée, et les documents présentent des caractères aujourd'hui disparus, ce qui pose d'importants problèmes de transcription. Les ouvrages de Catach (1968), Baddeley (1993) et Vachon (2010) ont depuis plusieurs décennies permis d'améliorer notre connaissance des systèmes graphiques de l'époque, et sont donc une importante ressource pour résoudre les princi-

paux problèmes ecdotiques liés à la fabrication de notre corpus.

L'ouvrage autour duquel s'articule le projet SETAF, et par conséquent notre corpus numérique, sont les *Faictz de Jesus Christ et du pape* ([Anonyme] 2009). Il s'agit d'un traité réformé et satirique en images qui a paru sous couvert d'anonymat dans trois éditions successives, entre les années 1530 et 1560. À la fin de la postface qui accompagne la reproduction de ce traité illustré, Reinhard Bodenmann donne des indications en vue du déchiffrement du texte, nous rappelant la complexité du travail philologique que rencontre tout éditeur :

> *On notera que les lettres st ou ct correspondent souvent à un simple t (monstre = montre / faict = fait), cq à un q (vnicque = unique) ; aul et eul respectivement à au et eu (mauldit = maudit / eulx = eux) ; que es et ez correspondent généralement à un e accentué (estre = être / destruict = détruit / euesque = évêque / paouretez = paovreté = pauvreté) parfois suivi d'un s (amenez = amenés) et que le z final après une lettre autre que e équivaut à notre s (filz = fils / voz = vos / laiz = laï[c]s). Enfin il est bon de savoir que les diphtongues ai et ei sont interchangeables (plaine = pleine / feiz = fais) ; que eu peut équivaloir à u (ainsi peu = peu ou pu / pleu = plu / neust = n'eût) ; que les lettres ins et indr correspondent parfois à nos is ou ir (prins = pris / prindrent = prirent) ; et que ngn équivaut à gn ou nn (congneut = connut).* (ibid., p. 73-74)

À ces indications, Bodenmann ajoute des remarques concernant l'absence de l'apostrophe et les distinctions des ‹u›/‹v› et des ‹j›/‹i›/‹y›, deux points très importants sur lesquels nous reviendrons.

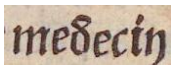
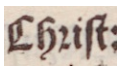
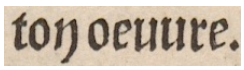
La résolution de différents problèmes présentés par Bodenmann doit de surcroît être effectuée en prenant en compte la particularité d'autres ouvrages du corpus dans lequel l'édition des *Faictz* s'inscrit – particularités qu'il convient de conserver en prévision de possibles études linguistiques. En effet, dans son épître au lecteur de *L'instruction des enfans, contenant la maniere de prononcer et escrire en francoy* ([Olivétan] 1533), Pierre-Robert Olivétan met en évidence l'introduction de l'apostrophe, des accents aigus et graves. Ce sont des éléments novateurs pour les années 1530, qui d'ailleurs ne sont pas utilisés dans les autres publications de l'imprimeur Pierre de Vingle. Même dans *L'instruction des enfans*, l'orthographe proposée par Olivétan n'est pas entièrement respectée, parce que l'imprimeur n'avait pas ces caractères. C'est d'ailleurs l'auteur lui-même qui le précise dans sa lettre adressée à Antoine Saunier qui suit l'épître *Au Lecteur* (Carbonnier-Burkard 2014).

3.1 Les variantes graphiques d'une même lettre

Comme nous l'avons déjà dit, nous avons opté pour une transcription graphématique qui conserve la ponctuation originale et les abréviations. Nous proposons donc de réduire les variantes graphiques d'une même lettre à une représentation standardisée.

Voici des exemples et nos choix de transcription :

Tableau 2 – Les variantes graphiques d'une même lettre

Exemple	Source	Transcrire	Unicode
	[Malingre] 1533a p. 5	Les	U+0073
	[Malingre] 1533a p. 5	Chrestiente	U+0073
‹s› (Latin small letter s, U+0073) et ‹ſ› (Latin small letter long s, U+017F) sont transcrits par ‹s›.			
	[Malingre] 1533a p. 5	de	U+0064
	[Malingre] 1533a p. 5	medecin	U+0064
‹ð› (Latin small letter insular d, U+A77A) et ‹δ› (Latin small letter delta, U+1E9F) sont transcrits par ‹d› (Latin small letter d, U+0064).			
	[Malingre] 1533 ? p. 6	chãter	U+0072
	[Malingre] 1533 ? p. 6	Christ	U+0072
‹r› (Latin small letter r, U+0072) et ‹ꭇ› (Latin small letter r rotunda, U+A75B) sont transcrits par ‹r›.			
	[Olivétan] 1533 p. 17	aucune	U+006E
	[Olivétan] 1533 p. 17	ton oeuure	U+006E
‹n› (Latin small letter n, U+006E) et ‹ŋ› (Latin small letter eng, U+014B) sont transcrits par ‹n›.			

3.2 Les distinctions des ‹u›/‹v› et des ‹j›/‹i›/‹y›

Comme pour les documents médiévaux, ceux du XVI^e^ siècle ne connaissent que les lettres ‹u› avec une variante pointue (‹v›) et ‹i› avec une variante à la hampe descendante (‹j›). Leur distinction relève d'une variation dite « positionnelle » ou « fonctionnelle » (Catach 1968), et ne note pas l'opposition entre sons voyelles et consonnes comme en français moderne (*vniuers* et non *univers*). Ainsi, ‹v› est utilisé à l'initiale et ‹u› à l'intérieur, la lettre ‹i› note souvent la consonne ‹j› et, enfin, ‹y› note souvent la voyelle ‹i›.

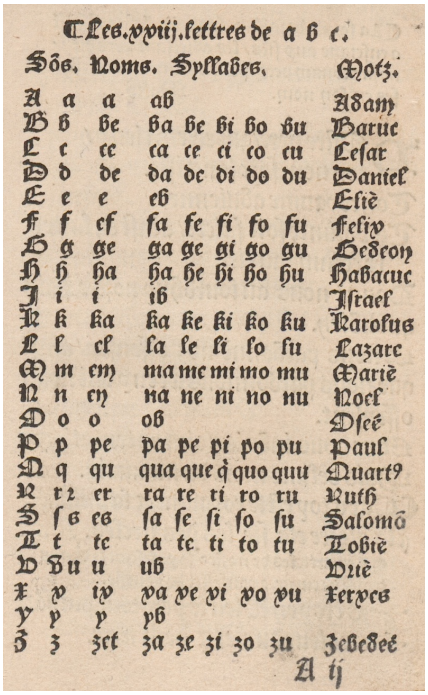
> *La répartition entre les consonnes ‹j› et ‹v› et les voyelles ‹i› et ‹u› s'est réalisée entre la fin du moyen français et la fin du XVII^e^ siècle. Le XVI^e^ siècle occupe donc un place centrale dans ce changement.* (Vachon 2010, p. 125)

Dans l'alphabet qui est proposé dans l'*Instruction des enfans* (cf. fig. 1), par exemple, ‹v› et ‹j› ne sont pas présentées comme des lettres à part entière : la variante ‹v› est placée sur la même ligne que ‹u› et une seule majuscule est indiquée pour les deux variantes ; les lettres ‹i› et ‹y› sont placées sur des lignes différentes et la variante ‹j› n'est pas présente dans cet alphabet.

Pinche (2022) opte pour transcrire systématiquement les ‹u› et ‹v› avec ‹u› et les ‹i› et ‹j› avec ‹i› :

> *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, mais relève d'une variation de forme.[…] 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 ».* (ibid., p. 5)

Figure 1 – L'alphabet dans l'*Instruction des enfans*, [Olivétan] 1533, page 13.



Suivre ce choix nous semble cohérent avec le choix de réduire les variantes graphiques d'une même lettre à une représentation standardisée : nous avons donc décidé de transcrire tous les ‹u› et ‹v› avec des ‹u›, tous les ‹i› et ‹j› avec des ‹i›. En ce qui concerne la lettre ‹y›, il s'agit d'une lettre à part entière et non d'une variation graphique de ‹i› (cf. fig. 1) : elle est donc transcrite ‹y›.

> *Il existe au XVI^e^ siècle trois différents types de ‹y› : (a) semi-consonne en position initale ou intervocalique ; (b) ‹y› grec, servant à noter un ancien upsilon ; (c) ‹y› calligraphique, employé en position initiale, finale ou comme deuxième élément de digramme.* (Vachon 2010, p. 76)

Voici des exemples et nos choix de transcription :

Tableau 3 – Les distinctions des ‹u›/‹v› et des ‹i›/‹j›/‹y›

Exemple	Source	Transcrire	Unicode
	[Olivétan] 1533 p. 13	u	U+0075
	[Olivétan] 1533 p. 13	u	U+0075
	[Olivétan] 1533 p. 13	i	U+0069
	[Olivétan] 1533 p. 13	i	U+0069
	[Olivétan] 1533 p. 13	y	U+0079

3.3 Les majuscules

Les lettres qui présentent une mise en relief (cf. fig. 2) sont transcrites à l'aide de majuscules. Aucune normalisation de l'usage concernant ces dernières n'est introduite lors de la transcription, par exemple en nous alignant sur les pratiques contemporaines (majuscules aux anthroponymes, toponymes, premières lettres d'une phrase, etc.). Outre que la standardisation actuelle n'a pas encore cours à l'époque, de nombreux cas resteraient insolubles car trop ambigus (par ex. les personnifications).

FIGURE 2 – L'alphabet en trois fontes dans l'*Instruction des enfans*, [OLIVÉTAN] 1533, page 12.

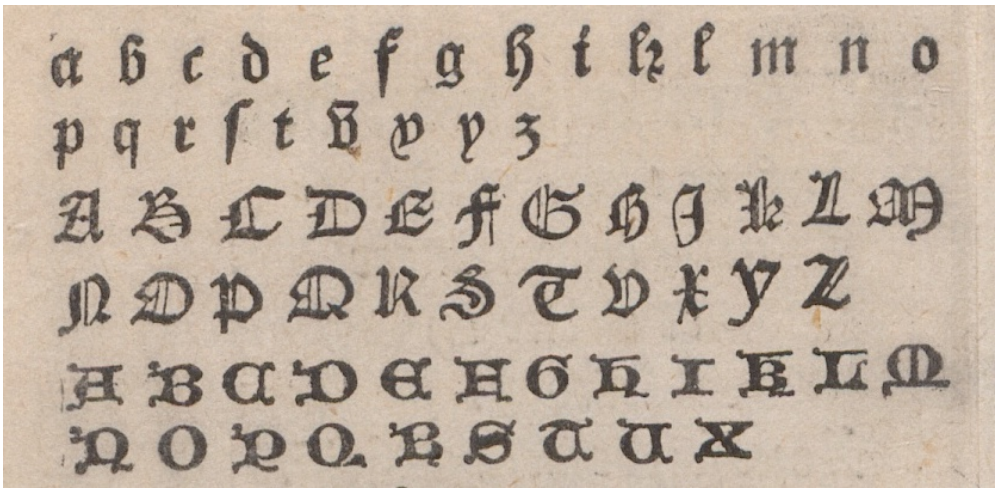
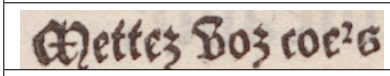
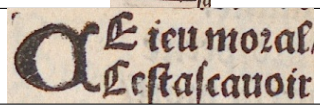
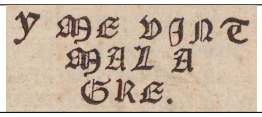


TABLEAU 4 – Les majuscules

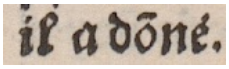
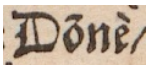
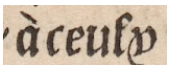
Exemple	Source	Transcrire
Lettres en début de mot		
	[MALINGRE] 1533 ? p. 6	Mettez voz coés
Lettrines ornées et non ornées		
	[MARCOURT] 1533 p. 7	E
	[MALINGRE] 1533a p. 7	CE ieu moral
Devises		
	[MALINGRE] 1533a p. 5	Y ME VINT MAL A GRE

3.4 Les voyelles accentuées

Si l'usage des accents est encore très rare à l'époque, nous en trouvons dans notre corpus, notamment dans l'*Instruction des enfans*, avec des valeurs différentes par rapport

à aujourd'hui. Ainsi, l'accent grave sur le ‹e› (‹è›) est utilisé pour noter le son [ə] (e muet), comme dans « il donnè » pour « il donne » (présent de l'indicatif). Ils sont conservés dans la transcription.

Tableau 5 – Les voyelles accentuées

Exemple	Source	Transcrire	Unicode
	[Olivétan] 1533 p. 135	il a dõné.	U+00E9
	[Olivétan] 1533 p. 134	Dõnè/	U+00E8
	[Olivétan] 1533 p. 17	à ceulx	M+00E0

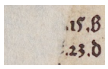
3.5 Les ligatures

En ce qui concerne les ligatures, celles qui existent encore en français (par exemple ‹œ›) sont transcrites, si présentes, mais celles qui ont disparu ne sont pas prises en compte.

3.6 Les caractères illisibles

À la place des caractères non lisibles, nous utilisons le caractère spécial suivant : ◊ (`U+25CA`). Voici un exemple :

Tableau 6 – Les caractères non lisibles

Exemple	Source	Transcrire	Unicode
	[Anonyme] [1534 ?] p. 16	◊15.b ◊23.d	U+25CA

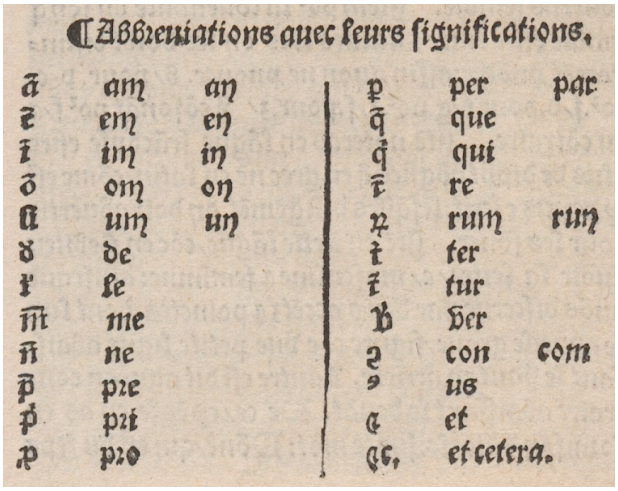
4 Les abréviations

Notre corpus se caractérise par un grand nombre d'abréviations et, à cette étape de la transcription, nous avons décidé de ne pas les développer. Comme l'a proposé A. Pinche (2022), les abréviations signalées par des tildes ou des macrons seront toujours rendues par un tilde, pour éviter la multiplication des signes, et nous avons sélectionné un caractère par abréviation, en tenant compte des développements possibles de l'abréviation et de la forme du signe. Par exemple, l'abréviation ‹ↄ› (`U+2184` = con/com) sera toujours rendue par ‹9› (`U+A76F` = con/com).

En cas de doute, nous avons consulté le dictionnaire des abréviations de Cappelli (1967) et le tableau des abréviations publié dans l'*Instruction des enfans* ([Olivétan] 1533) (cf. fig.3).

Les catégories dans lesquelles nous avons organisé les abréviations rencontrées dans notre corpus sont les suivantes :

FIGURE 3 – Le tableau des abréviations dans l'*Instruction des enfans*, [OLIVÉTAN] 1533, page 133.



- Les abréviations par lettres suscrites ;
- Les abréviations par tildes : les voyelles ;
- Les abréviations par tildes : les consonnes ;
- Les abréviations par signes spéciaux : partie 1 ;
- Les abréviations par signes spéciaux : partie 2.

TABLEAU 7 – Les abréviations par lettres suscrites

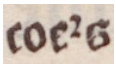
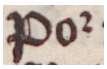
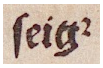
Exemple	Source	Transcrire	Unicode
	[MALINGRE] 1533 ? p. 6	coés	U+0065 U+036C
é = eur (coés = coeurs)			
	[MALINGRE] 1533 ? p. 6	pó	U+006F U+036C
ó = our/ur (pó = pour)			
	[MARCOURT] 1533 p. 6	seiǵ	U+0067 U+036C
ǵ = gneur (seiǵ = seigneur)			
	[MALINGRE] 1533 ? p. 24	q̣	U+0071 U+0365
q̣ = qui			

Tableau 8 – Les abréviations par tildes : les voyelles

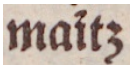
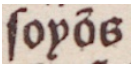
Exemple	Source	Transcrire	Unicode
	[Malingre] 1533 ? p. 6	châter	U+00E3
ã = am/an (châter = chanter)			
	[Malingre] 1533 ? p. 6	gẽsdarmes	U+1EBD
ẽ = em/en (gẽsdarmes = gensdarmes)			
	[Malingre] 1533 ? p. 26	maĩtz	U+0129
ĩ = im/in (maĩtz = maintz)			
	[Malingre] 1533 ? p. 6	soyõs	U+00F5
õ = om/on (soyõs = soyons)			
	[Malingre] 1533 ? p. 30	imũde	U+0169
ũ = um/un (imũde = imunde)			

Tableau 9 – Les abréviations par tildes : les consonnes

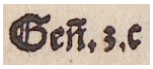
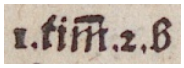
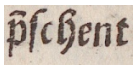
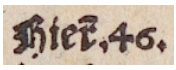
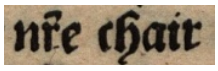
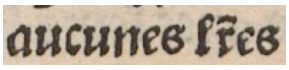
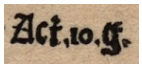
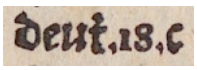
Exemple	Source	Transcrire	Unicode
	[Anonyme] [1534 ?] p. 8	Geñ.3.c	U+00F1
Geñ. = Genèse			
	[Anonyme] [1534 ?] p. 16	1.timˇ.2.b	U+006D U+0303
1.timˇ. = Ire Épître de Paul à Timothée			
	[Anonyme] [1538/1544] p. 59	pˇschent	U+0070 U+0303
pˇ = pre (pˇschent = preschent)			
	[Malingre] 1533 ? p. 6	vainqˇrons	U+0071 U+0303
qˇ = que (vainqˇrons = vainquerons)			
	[Anonyme] [1534 ?] p. 8	Hierˇ.46.	U+0072 U+0303
Hierˇ. = Jeremiah = Jérémie			
	[Malingre] 1533b p. 18	nrˇe chair	U+0072 U+0303
nrˇe = notre			
	[Malingre] 1533 ? p. 134	lrˇes	U+0072 U+0303
lrˇes = lettres			
	[Farel] 1533 p. 23	Actˇ.10.g	U+0074 U+0303
Actˇ. = Actus [apostolorum] (abbr. eccles.) = Actes des Apôtres			
	[Anonyme] [1534 ?] p. 8	deutˇ.18.c	U+0074 U+0303
Deutˇ. = Deuteronomium = Deutéronome			

Tableau 10 – Les abréviations par signes spéciaux : partie 1

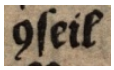
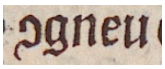
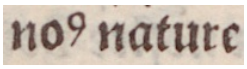
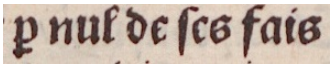
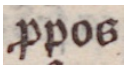
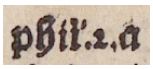
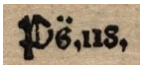
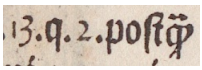
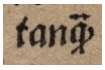
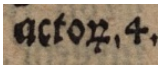
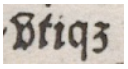
Exemple	Source	Transcrire	Unicode
	[Malingre] 1533 ? p. 6	ꝛ	U+204A
ꝛ = et			
	[Marcourt] 1533 p. 6	ꝛc.	U+204A
ꝛc. = etc.			
	[Malingre] 1533b p. 26	ꝰseil	U+A76F
ꝰ = con/com (ꝰseil = conseil)			
	[Marcourt] 1533 p. 19	ꝰgneu	U+A76F
ↄ (U+2184) = ꝰ (U+A76F) = con/com (ꝰgneu = congneu)			
	[Malingre] 1533 ? p. 13	no^{ꝰ}	U+A770
^{ꝰ} = us (no^{ꝰ} = nous = notre)			
	[Malingre] 1533 ? p. 24	ꝑ	U+A751
ꝑ = par/per			
	[Malingre] 1533 ? p. 30	ꝓpos	U+A753
ꝓ = pro (ꝓpos = propos)			
	[Malingre] 1533b p. 15	ð	U+00F0 ou M+F159 ?
ð = de (voir la figure 3)			
	[Malingre] 1533b p. 27	u͛s ? (vers)	U+0075 U+036C ?
u͛ ? = ver (voir la figure 3)			

Tableau 11 – Les abréviations par signes spéciaux : partie 2

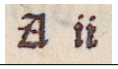
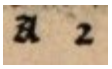
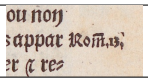
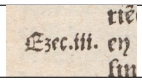
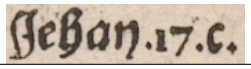
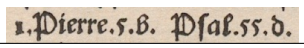
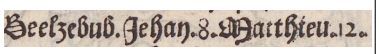
Exemple	Source	Transcrire	Unicode
	[Malingre] 1533 ? p. 5	Gł̈eux	U+0142
Gł̈eux = Glorieux			
	[Anonyme] [1534 ?] p. 8	phił.2.a	U+0142
Phił = Épître de Paul aux Philippiens			
	[Farel] 1533 p. 80	Pš.118.	U+0053 U+0308
Pš. = Psalmi = Psaumes			
	[Anonyme] [1538/1544] p. 59	postq̈	U+A757 U+0308 ou U+A759 U+0308 ou M+E68B ?
q̈ = quam (postq̈ = postquam)			
	[Malingre] 1533b p. 27	tanq̈	U+A757 U+0308 ou U+A759 U+0308 ou M+E68B ?
q̈ = quam (tanq̈ = tanquam)			
	[Malingre] 1533b p. 28	acto2̷4	U+A75D
2̷4 = -rom/-rum (acto2̷4 = actorum)			
	[Marcourt] 1534a p. 59	utiqꝫ	U+0071 U+A76B
qꝫ = que (utiqꝫ = utique)			

5 Les chiffres

Dans notre corpus, les signatures de cahiers et les références bibliques présentent parfois des chiffres romains et parfois des chiffres arabes. Nous proposons d'encadrer les chiffres romains par des points pour les distinguer des lettres et de les transcrire en minuscules, suivant ainsi la recommandation d'A. Pinche (2022). En revanche, les chiffres arabes ne seront pas encadrés par des points.

Concernant les références bibliques, des points sont présents après chaque élément de la référence dans la plupart des cas. Ces éléments forment un seul bloc sans espaces, par exemple « Iehan.17.c. ». Si des points ne sont pas présents, nous ne les rajoutons pas et nous n'ajoutons pas d'espaces, sauf s'il y a plusieurs références consécutives. Voici des exemples et nos choix de transcription :

Tableau 12 – Les chiffres

Exemple	Source	Transcrire
	[Malingre] 1533a p. 7	A .ii.
Exemple de chiffres romains dans une signature : encadrer les chiffres par des points pour les distinguer des lettres et les transcrire en minuscules.		
	[Malingre] [1538/1544] p. 11	A 2
Exemple de chiffres arabes dans une signature : ne pas encadrer les chiffres par des points.		
	[Malingre] 1533 ? p. 49	sur le Psalme.84.
Cette référence se trouve dans un titre : transcrire sans espace.		
	[Marcourt] 1533 p. 35	Rom̃.13.
Cette référence, en chiffres arabes, se trouve en marge du texte : transcrire sans espaces, telle qu'elle est dans l'exemple.		
	[Marcourt] 1534b p. 4	Ezec.iii.
Cette référence, en chiffres romains, se trouve en marge du texte : transcrire sans espaces, telle qu'elle est dans l'exemple.		
	[Olivétan] 1533 p. 39	Iehan.17.c.
Cette référence présente un troisième élément, une lettre : transcrire sans espaces, telle qu'elle est dans l'exemple.		
	[Olivétan] 1533 p. 39	1.Pierre.5.b. Psal.55.d.
Exemple de références bibliques consécutives : transcrire avec un espace entre les références. Ici l'espace était déjà présent.		
	[Anonyme] [1538/1544] p. 45	Beelzebub. Jehan.8. Matthieu.12.
Exemple de références bibliques consécutives : transcrire avec un espace entre les références. Ici l'espace n'était pas présent.		

6 La séparation des mots

Dans les imprimés du XVI^e^ siècle, la séparation des mots est différente de celle du système moderne. Les phénomènes d'agglutination sont nombreux et fréquents : certains sont clairement identifiables, par exemple l'absence de l'apostrophe ou la séquence agglutinée comprenant l'adverbe « très » suivi d'un adjectif ; d'autres, en revanche, présentent une variation très importante.

Par exemple, il est possible de trouver l'équivalent de la séquence contemporaine ‹il n'y a› en deux versions différentes : ‹il ny a› ou ‹il nya›. S'il s'agissait d'une édition critique, nous ferions un choix philologique, par exemple en fonction de la fréquence de l'une ou l'autre version. Néanmoins, dans le cadre de la conception d'un corpus, il serait trop chronophage d'effectuer cette recherche pour chaque ouvrage et un tel travail divergerait de notre principe de base, qui est celui de préparer une vérité de terrain, non directement une transcription destinée à la publication.

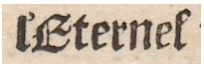
Comme le souligne A.Pinche (2022), une pratique trop imitative introduit beaucoup de bruit et ne facilite pas le passage vers la lemmatisation : elle propose donc de ne pas toujours conserver l'agglutination présente dans le texte source. Nos documents étant des imprimés écrits dans une français plus récent, ce problème nous paraît moins fréquent, et nous avons décidé de rendre toutes les agglutinations telles quelles. Les cas problématiques seront pris en compte à une autre étape de notre chaîne de traitement, lors de l'établissement d'une transcription semi-diplomatique destinée à la publication et/ou à la fouille des données.

Suivant la même logique, les agglutinations étranges, probablement dues au travail d'impression (cf. tab. 15), ne sont pas re-segmentées. Ce choix est fait en cohérence avec celui de n'intervenir sur aucun des artefacts de l'impression, que ce soit une faute (*arqre* pour *arbre*) ou non.

En ce qui concerne les phénomènes d'hyphénation, nous avons déjà expliqué le choix d'indiquer la coupure des mots en fin de ligne par le caractère *Not sign* (`U+00AC`, ‹¬›).

Les imprimeurs de notre corpus ne recourent pas à l'apostrophe, hormis quelques exceptions que nous avons mentionnées. Dans l'*Instruction des enfans*, Olivétan introduit l'apostrophe mais elle n'est pas présente dans le reste du corpus. Voici un exemple :

Tableau 13 – Un exemple d'apostrophe

Exemple	**Source**	**Transcrire**	**Unicode**
	[Olivétan] 1533 p. 17	l'Eternel	U+0027

Ci-dessous nous listons quelques exemples de phénomènes d'agglutination :

Tableau 14 – Les phénomènes d'agglutination : l'absence de l'apostrophe

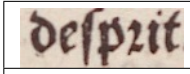
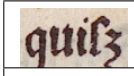
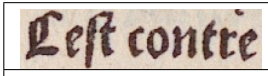
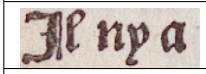
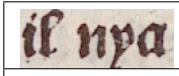
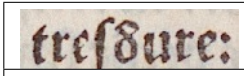
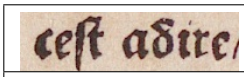
Exemple	Source	Transcrire
	[Malingre] 1533 ? p. 6	layde
Article avec ellipse + substantif ou adjectif (l'aide)		
	[Malingre] 1533 ? p. 6	desprit
Préposition avec ellipse + substantif ou adjectif (d'ésprit)		
	[Malingre] 1533a p. 24	nest
Négation avec ellipse (n'est)		
	[Malingre] 1533a p. 42	quilz
« que » avec ellipse (qu'ils)		
	[Malingre] 1533a p. 42	cest
« ce » avec ellipse (c'est)		

Tableau 15 – Les phénomènes d'agglutination liés à l'impression

Exemple	Source	Transcrire
	[Farel] 1533 p. 33	cõuoiteraspoint
	[Farel] 1533 p. 33	ꝯuoiteras point
	[Malingre] 1533a p. 26	a laChrestiente:
	[Malingre] 1533a p. 28	la Chrestiente.

Tableau 16 – Les phénomènes d'agglutination : des séquences agglutinées

Exemple	Source	Transcrire
	[Malingre] 1533a p. 22	il ny a
Forme négative de « il y a » qui dans l'exemple est écrit « il ny a »		
	[Malingre] 1533a p. 23	il nya
Forme négative de « il y a » qui dans l'exemple est écrite « il nya »		
	[Malingre] 1533a p. 29	il ya
Forme affirmative de « il y a » qui dans l'exemple est écrite « il ya »		
	[Malingre] 1533a p. 27	tresdure:
Adverbe « très » + adjectif		
	[Marcourt] 1533 p. 6	cest adire
« c'est-à-dire » qui dans l'exemple est écrit « cest adire »		
	[Marcourt] 1533 p. 22	cestadire.
« c'est-à-dire » qui dans l'exemple est écrit « cestadire »		
	[Malingre] 1533a p. 7	Cestascauoir
« c'est assavoir » qui dans l'exemple est écrit « cestascauoir »		

7 La description de la mise en page

> *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.* (Pinche 2022, p. 15)

La description de la mise en page de notre corpus numérique s'appuie sur le vocabulaire contrôlé SegmOnto (Gabay, Camps, Pinche et al. 2021), qui utilise une description à deux niveaux :

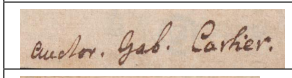
- `Zones`, pour les différents types de régions de la page, par exemple :
  - `DamageZone`
  - `DropCapitalZone`
  - `GraphicZone`
  - `MainZone`

  - `MarginTextZone`
  - `TableZone`
  - `QuireMarksZone`
  - `StampZone`
  - `TitlePageZone`
- `Lines`, pour les différents types de lignes contenues dans les zones, par exemple :
  - `DefaultLine`
  - `DropCapitalLine`
  - `HeadingLine`
  - `InterlinearLine`
  - `MusicLine`

Chaque niveau est caractérisé par des types de zone et de ligne qui ont des valeurs obligatoires et contrôlées. Chaque type de zone ou de ligne peut ensuite être caractérisé par des sous-types : il s'agit de valeurs facultatives pour lesquelles une liste ouverte est souvent proposée.

Nous avons décidé d'ajouter le sous-type *handwrittenAddition* au type *MarginTextZone* pour toutes les notes manuscrites en marge du texte : les notes des bibliothécaires, les notes de possession, les corrections, les dessins manuscrits. Comme nous nous intéressons aux textes et non aux différents exemplaires, les notes manuscrites sont signalées au niveau de la segmentation par ce type de zone, mais elles ne sont pas transcrites.

Tableau 17 – Exemples de *MarginTextZone:handwrittenAddition*

Exemple	**Source**
	[Marcourt] 1533 p. 5
	[Marcourt] 1533 p. 5

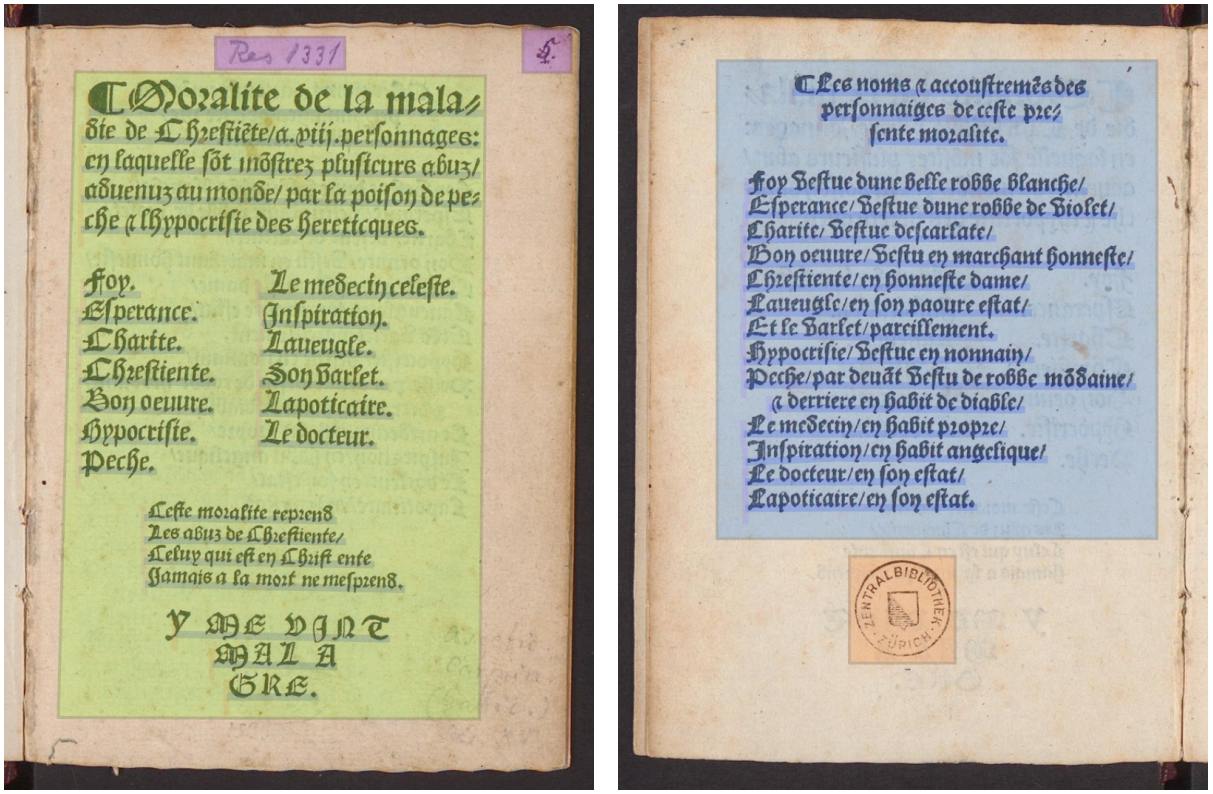
Cependant, nous avons pensé créer un sous-type pour les lignes des notes manuscrites que l'on veut garder et transcrire pour l'édition numérique des *Faictz de Jesus Christ et du pape*, qui est l'ouvrage autour duquel s'articule le projet SETAF. Comme les exigences philologiques pour une édition critique sont différentes par rapport à la construction d'un corpus générique, nous avons créé le sous-type suivant : *CustomLine:Faits*.

Un autre choix au niveau de la segmentation qui a une conséquence au niveau de la transcription concerne les symboles ornementaux ou fonctionnels. Comme certains symboles présentent trop de variantes et comme nous ne nous intéressons pas à ce type d'informations graphiques, ces symboles ne sont pas pris en compte au niveau de la transcription. Toutefois, ils sont signalés au niveau de la segmentation des régions, par la zone qui est appelée *GraphicZone*.

Tableau 18 – Exemples de *GraphicZone* : symboles ornementaux ou fonctionnels

Exemple	Source
	[Marcourt] 1533 p. 5
	[Farel] 1538 p. 26

Ci-dessous nous listons quelques exemples de types de régions et de lignes[8] :



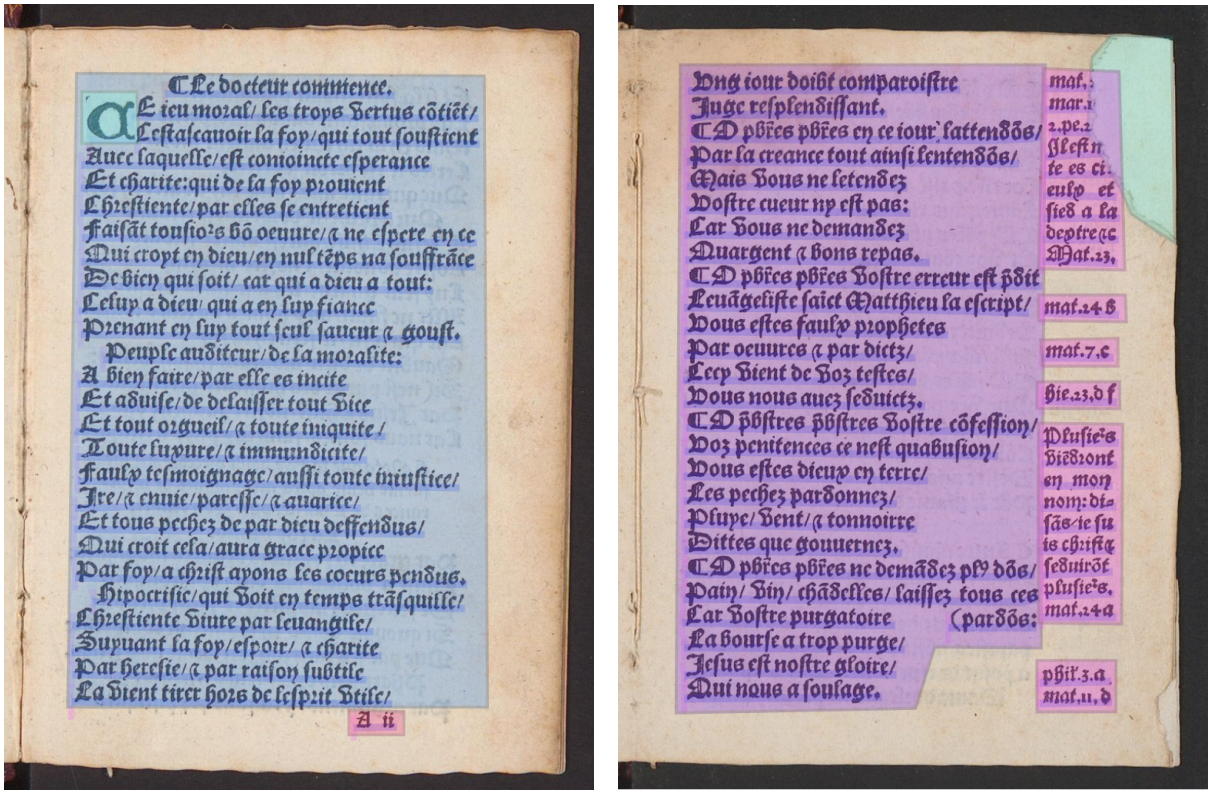
(a) [Malingre] 1533a, page 5. (b) [Malingre] 1533a, page 6.

Figure 4 – Deux exemples d'analyse de mise en page.

Dans la figure 4a, nous avons une *TitlePageZone* et deux *MarginTextZone:handwrittenAddition*. La *TitlePageZone* concerne la page de titre de l'ouvrage et les *MarginTextZone:handwrittenAddition* concernent des notes manuscrites en marge du texte.

Dans la figure 4b, nous avons une *MainZone* et une *StampZone*. La *MainZone* concerne le corps du texte, une région n'ayant pas de statut particulier, et la *StampZone* concerne les timbres des bibliothèques.

8. Nos données sont produites à l'aide de l'instance genevoise FoNDUE (https://www.unige.ch/lettres/humanites-numeriques/recherche/projets-de-la-chaire/fondue) de l'interface eScriptorium (Kiessling et al. 2019). Nous précisons que, dans les exemples qui suivent, les couleurs associées aux régions et aux lignes peuvent changer selon leur ordre dans l'*ontology* sur eScriptorium.



(a) [Malingre] 1533a, page 7. (b) [Anonyme] [1534 ?], page 13.

Figure 5 – Deux exemples d'analyse de mise en page.

Dans la figure 5a, nous avons une *MainZone*, une *DropCapitalZone* et une *QuireMarksZone*. La *MainZone* concerne le corps du texte, une région n'ayant pas de statut particulier. La *DropCapitalZone* concerne les lettrines, ornées ou non ornées. La *QuireMarksZone* concerne la signature de cahiers et la réclame. Parfois les deux sont présentes, parfois ni l'une ni l'autre, parfois l'une ou l'autre. De toute façon, elles sont placées sur la même ligne et dans la même région.

Dans la figure 5b, nous avons une *MainZone*, une *DamageZone* et plusieurs *MarginTextZone*. La *MainZone* concerne le corps du texte, une région n'ayant pas de statut particulier. La *DamageZone* concerne toute partie abîmée d'une page de l'ouvrage. La *MarginTextZone* concerne les notes en marge du texte, par exemple des commentaires ou des références bibliques.

(a) [Anonyme] [1538/1544], page 10. (b) [Anonyme] [1534 ?], page 6.

Figure 6 – Deux exemples d'analyse de mise en page.

Dans la figure 6a, nous avons une *MainZone*, une *GraphicZone* et une *QuireMarksZone*. La *MainZone* concerne le corps du texte, une région n'ayant pas de statut particulier. La *GraphicZone* concerne les images et les symboles ornementaux ou fonctionnels. La *QuireMarksZone* concerne la signature de cahiers et la réclame. Parfois les deux sont présentes, parfois ni l'une ni l'autre, parfois l'une ou l'autre. De toute façon, elles sont placées sur la même ligne et dans la même région.

Dans la figure 6b, nous avons des *HeadingLine*, une *DropCapitalLine* et plusieurs *DefaultLine*. La *DefaultLine* concerne toutes les lignes n'ayant pas de statut particulier, y compris les lignes dans des *MarginTextZone* ou des *QuireMarksZone*. La *DropCapitalLine* concerne les lettrines, de grandes lettres pouvant être rehaussées et ornées par divers procédés, qui se trouvent toujours dans une *DropCapitalZone*. La *HeadingLine* concerne différents types de lignes : les titres des ouvrages, des chapitres, des chansons, des sous-chapitres, etc. ; dans les pièces de théâtre, les noms des personnages qui prennent la parole ; certaines lignes qui sont centrées ou séparées d'autres parties du texte, comme dans le cas des didascalies ou des devises.

8 L'interopérabilité avec d'autres corpus

Jusqu'ici notre attention s'est portée sur les imprimés français du XVI^{e} siècle en caractères gothiques. Nous souhaitons cependant que ce guide contribue à créer un protocole aussi générique que possible pour tous les imprimés du XVI^{e} siècle en langue française. L'imprimé en gothique est, pour la langue française, la queue de comète d'une pratique ancienne, liée aux manuscrits médiévaux, tandis que l'imprimé en romain (ou en *antiqua*)

est le début d'une pratique moderne. La construction d'un corpus mélangeant des données provenant d'imprimés en gothique et en *antiqua* n'est pas qu'un problème typographique, mais aussi linguistique. Ainsi des oppositions apparaissent, des signes se généralisent, ce qui a des conséquences sur les normes de transcription.

Il faut aussi tenir compte du fait que les guides de transcription sont produits par des projets qui ont leurs propres objectifs de recherche. Ainsi le s long (‹ſ›) est gardé par Gabay, Clérice et Reul (2023) pour les imprimés français du XVII^e^ siècle dans le but d'étudier l'évolution de la casse des imprimeurs. Si nous voulons gagner du temps, nous devons utiliser ou affiner les modèles d'OCR produits grâce à d'autres guides de transcription et suivre *de facto* les choix d'autres équipes. Dans la mesure où il est possible de revenir en arrière, par une série d'opérations (par ex. ‹ſ› → ‹s›), cela ne pose pas de problème majeur et il est possible de conserver l'interopérabilité entre les données.

En ce qui concerne les imprimés français en caractères romains, les différences majeures entre nos recommandations de transcription et celles de Gabay, Clérice et Reul (ibid.) concernent les distinctions entre ‹s/ſ›, ‹u›/‹v› et ‹i›/‹j›. Gabay, Clérice et Reul (ibid.) conservent le s long (‹ſ›) (par ex. *meſme* n'est pas transcrit *mesme*) et ils distinguent les ‹u›/‹v› et les ‹i›/‹j› (par ex. *vniuers* n'est pas transcrit *univers* et pas non plus *uniuers*). Pour conserver l'interopérabilité entre les données, il est possible d'aligner à postériori les transcriptions par des opérations de remplacement (par ex. ‹ſ› → ‹s›, ‹v› → ‹u›, ‹j› → ‹i›).

Malgré quelques différences, nous avons essayé d'assurer autant que possible l'interopérabilité entre les données produites suivant nos recommandations et les données produites suivant les trois guides que nous avons cités dès le début : le guide de transcription de Pinche (2022) pour les manuscrits gothiques du X^e^ au XV^e^ siècle, les principes éditoriaux d'*Epistemon – Corpus de textes de la Renaissance* des BVH et les recommandations de transcription de Gabay, Clérice et Reul (2023) pour les imprimés français du XVII^e^ siècle. Enfin, une possibilité pour contourner des problèmes d'interopérabilité pourrait être le passage par la lemmatisation.

Remerciements

Nous remercions...

Bibliographie

Sources primaires

[Anonyme] ([1534 ?]). *Chansons nouvelles demonstrantz plusieurs erreurs et faulsetez : desquelles le paovre monde est remply par les ministres de Satan*. [Neuchâtel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-576. Zentralbibliothek Zürich, Zürich, Res 1327, GLN-4821.

— ([1538/1544]). *Les Faitz de Jesus Christ et du Pape, par lesquelz chascun pourra facilement congnoistre la grande difference d'entre eulx : nouvellement reveuz, corrigez, et augmentez selon la verité de la saincte Escripture, et des droictz canons, par le lecteur du sainct Palais*. Rome [Genève] : [Jean Michel]. DOI : 10.3931/e-rara-12688. Musée Historique de la Réformation, Genève, MHR K LUT 8, GLN-1406.

[Farel], [Guillaume] (1533). *La manière et fasson qu'on tient en baillant le sainct baptesme*. Neuchâtel : Pierre de Vingle. DOI : 10.3931/e-rara-5945. Bibliothèque de Genève, Genève, BGE Bd 1474, GLN-1492.
— (1538). *L'ordre et manière qu'on tient en administrant les sainctz sacremens*. [Genève] : Jean Michel. DOI : 10.3931/e-rara-5896. Bibliothèque de Genève, Genève, Bd 1473, GLN-1321.
[Malingre], [Matthieu] (1533a). *Moralite de la maladie de Chrestiente a VIII personnages*. Paris [Neuchâtel] : Pierre de Vignolle [Pierre de Vingle]. DOI : 10.3931/e-rara-563. Zentralbibliothek Zürich, Zürich, ZB Res 1331, GLN-4499.
— (1533b). *Sensuyvent plusieurs belles et bonnes chansons que les chrestiens peuvent chanter en grande affection de cueur : pour et affin de soulager leurs experitz et de leur donner repos en Dieu, au nom duquel elles sont composées par rithmes, au plus près de l'esperit de Jésus Christ, contenu es Sainctes Escriptures*. [Neuchâtel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-6934. Bibliothèque de Genève, Genève, BGE Cth 2650 (1) BGE Bd 1475 (1), GLN-4822.
— ([1538/1544]). *Moralite de la maladie de Chrestie(n)te, a. XIII. personnages*. [Genève] : [Jean Michel]. DOI : https://onb.digital/result/103BE0A8. Österreichische Nationalbibliothek, Vienne, NB 48.V.87, GLN-1411.
— (1533 ?). *Noelz nouveaulx : musiciens amateurs des Cantiques, Au nom de Dieu, chantez noelz nouveaulx, Lesquelz sont faictz sur les vieulx et antiques / Y me vint mal à gré*. [Neuchâtel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-577. Zentralbibliothek Zürich, Zürich, ZB Res 1332, GLN-4500.
[Marcourt], [Antoine] (1533). *Le livre des marchans : fort utile a toutes gens / nouvellement compose par le sire Pantapole*. Corinthe [Neuchatel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-539. Zentralbibliothek Zürich, Zürich, ZB Rés 1330, GLN-1912.
— (1534a). *Declaration de la messe, Le fruict d'icelle, la cause, et le moyen, pourquoy et comment on la doibt maintenir*. [Neuchâtel] : [Pierre de Vingle]. DOI : /10.3931/e-rara-33158. BPU Neuchâtel, Neuchâtel, A.F. C 49 B, GLN-1953.
— (1534b). *Petit traicté tres utile, et salutaire de la saincte eucharistie de nostre Seigneur Jesuchrist*. [Neuchâtel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-6082. Bibliothèque publique et universitaire, Neuchâtel, BPU A.F. C 49 D, GLN-1952.
[Olivétan], [Pierre-Robert] (1533). *L' instruction des enfans : contenant la manière de prononcer et escrire en françoys, l'oraison de Jésu Christ, les articles de la foy, les dix commandemens, la salutation angélicque : avec la déclaration d'iceux, faicte en manière de recueil des seulles sentences de l'escripture saincte, item les figures des chiphres, et leurs valeurs*. Genève : [Pierre de Vingle]. DOI : 10.3931/e-rara-6933. Bibliothèque de Genève, Genève, BGE Bd 1477, GLN-4507.

Littérature secondaire

[Anonyme] (2009). *Faictz de Jesus Christ et du pape*. Avec une postf. de Reinhard Bodenmann. Cahiers d'Humanisme et Renaissance. Genève : Droz.
André, Jacques et Jimenes Rémi (2013). « Transcription et codage des imprimés de la Renaissance. Réflexions pour un inventaire des caractères anciens ». In : *Document numérique* 16.3, p. 113-139.
Baddeley, Susan (1993). *L'Orthographe française au temps de la Réforme*. Travaux d'humanisme et Renaissance. Genève : Droz.

Bérard, Aline (2012). « Apprendre à lire et à écrire au XVIe siècle : Pierre-Robert Olivétan et L'Instruction des enfans ». Mémoire de maîtrise. Lettres, Lausanne.
Berthoud, Gabrielle (1980). « Les impressions genevoises de Jean Michel (1538-1544) ». In : *Cinq siècles d'imprimerie genevoise*. Sous la dir. de J.-D. Candaux et B. Lescaze. T. 1. Genève : Droz, p. 55-88.
Cappelli, Adriano (1967). *Dizionario di abbreviature latine ed italiane usate nelle carte e codici specialmente del medioevo = Lexicon abbreviaturarum*. 6a ed. Milano : U. Hoepli.
Carbonnier-Burkard, Marianne (2014). « Salut par la foi, salut par la lecture : les nouveaux abécédaires en français au XVIe siècle ». In : *Protestantisme et éducation dans la France moderne*. LARHRA.
Catach, Nina (1968). *L'Orthographe française à l'époque de la Renaissance : auteurs, imprimeurs, ateliers d'imprimerie*. Paris : Droz.
Dauvois, Nathalie et Jacques Dürrenmatt, éd. (2011). *La Ponctuation à la Renaissance*. Paris : Editions Classiques Garnier.
Gabay, Simon, Jean-Baptiste Camps et Thibault Clerice (2022). *Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles) : Version B*. URL : https://hal.science/hal-02571190.
Gabay, Simon, Jean-Baptiste Camps, Ariane Pinche et al. (2021). *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages, version 0.9*. https://github.com/SegmOnto.
Gabay, Simon, Thibault Clérice et Christian Reul (2023). « OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more) ». In : *Journal of Data Mining and Digital Humanities* 2023. DOI : 10.46298/jdmdh.6492.
Kiessling, Benjamin et al. (2019). « eScriptorium : An Open Source Platform for Historical Document Analysis ». In : *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. T. 2, p. 19. DOI : 10.1109/ICDARW.2019.10032.
Lavrentiev, Alexei (2016). « Ponctuation française du Moyen Âge au XVIe siècle : théories et pratiques ». In : *La ponctuation à l'aube du XXIe siècle. Perspectives historiques et usages contemporains*. Sous la dir. de S. Pétillon, F. Rinck et A. Gautier. Lambert-Lucas, p. 39-62.
Pinche, Ariane (2022). *Guide de transcription pour les manuscrits du Xe au XVe siècle*. URL : https://hal.science/hal-03697382.
Stutzmann, Dominique (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*. BoD, p. 247-277. URL : https://shs.hal.science/halshs-00596970.
Vachon, Claire Hélène (2010). *Le changement linguistique au XVIe siècle une étude basée sur des textes littéraires français*. Strasbourg : Ed. de linguistique et de philologie.

Liste des tableaux

Table des figures