TRANSCRIPTION ET CODAGE DES IMPRIMÉS DE LA RENAISSANCE

Réflexions pour un inventaire des caractères anciens

Lavoisier | *Document numérique*

2013/3 - Vol. 16
pages 113 à 139

ISSN 1279-5127

Article disponible en ligne à l'adresse:

---

http://www.cairn.info/revue-document-numerique-2013-3-page-113.htm

---

Pour citer cet article :

« Transcription et codage des imprimés de la Renaissance » Réflexions pour un inventaire des caractères anciens, *Document numérique*, 2013/3 Vol. 16, p. 113-139. DOI : 10.3166/DN.16.3.113-139

Distribution électronique Cairn.info pour Lavoisier.

Transcription et codage des imprimés de la Renaissance

Réflexions pour un inventaire des caractères anciens

Jacques André[1], Rémi Jimenes[2]

1. *Inria-Rennes, rédacteur en chef honoraire de* Document numérique
Jacques.Andre35@gmail.com

2. *Centre d'études supérieures de la Renaissance*
remi.jimenes@univ-tours.fr

RÉSUMÉ. Conservant le plus grand nombre possible d'informations du document-source, une transcription de texte imprimé ancien devrait pouvoir servir de base non seulement à des analyses littéraires, mais également à des études « paléotypographiques ». Pour ce faire, il faudrait disposer d'un codage normalisé permettant d'assurer une correspondance univoque entre les caractères de la transcription numérique et ceux de la source originale. Le terme « caractère » pouvant prêter à confusion, nous introduisons un nouveau concept : celui de « typème », intermédiaire entre le caractère et le glyphe tel qu'Unicode les définit. Nous proposons d'utiliser le codage MUFI, une extension d'Unicode, augmentée des typèmes attestés dans les imprimés anciens, afin de produire une transcription dite « typémique », reproduction fidèle de la composition typographique du document original. Nous concluons sur la nécessité de réaliser l'inventaire des typèmes attestés dans les imprimés anciens, qui fera l'objet d'un Projet d'Inventaire des Caractères Anciens (PICA) actuellement à l'étude.

ABSTRACT. Preserving as many informations as possible from the original document, a transcription of ancient printed text should serve as a basis not only for literary analysis, but also for palaeotypographic studies. With this aim, we require a standardized encoding able to preserve a unequivocal link between the characters of the digital transcription and those of the original source. We define here the new concept of typem, a transitional element between the notion of character and glyph as defined by Unicode. It is proposed here to use MUFI, an extension to the Unicode standard, by adding new code points dedicated to "typems", in order to produce what we call "typemic transcriptions", reproducing all the characters of the original document. Finally, a project of a census of all the typems, named PICA (Projet d'Inventaire des Caractères Anciens), is described.

MOTS-CLÉS : typographie, MUFI, Unicode, codage, documents anciens, inventaire, caractères, typèmes, imprimés, Renaissance, PICA.

KEYWORDS: typography, MUFI, Unicode, encoding, ancient document, inventory, types, typems, printed material, Renaissance, *PICA.*

DOI:10.3166/DN.16.3.113-139 © 2013 Lavoisier

Document numérique – n° 3/2013, 113-139

1. Introduction[1]

La typographie permet de reproduire sur le papier en plusieurs dizaines de milliers d'exemplaires la trace d'un seul et même motif gravé[2]. Elle constitue de ce fait un système d'écriture *fermé*, nécessairement *normalisé* et *fini*, qui offre en tout cas bien moins de variations graphiques que les écritures manuscrites ou épigraphiques plus anciennes.

En dépit de cette relative simplicité graphique, les livres de la Renaissance se distinguent des imprimés plus récents par un certain nombre de caractéristiques propres. Remarquons d'abord la force d'inertie des pratiques manuscrites : les premiers imprimeurs tentent de reproduire le plus fidèlement possible l'aspect visuel du manuscrit, avec son florilège d'abréviations, de contractions, de chevauchements (caractères crénés) et de ligatures (figure 1). Par ailleurs, la Renaissance voit s'élaborer des principes orthographiques et stylistiques nouveaux, et ce tant pour le latin que pour les langues vernaculaires. On passe ainsi, en France, d'un « moyen français » encore médiéval à un « français classique » déjà moderne. Les débats sur le statut du vernaculaire engendrent un certain nombre d'innovations graphiques, dont les plus remarquables sont sans doute les caractères adaptés à l'orthographe phonétique préconisée par certains auteurs tel Jacques Peletier du Mans (figure 2). Toutes ces innovations linguistiques ne sont pas seulement le fait des auteurs, mais sont très étroitement liées à l'art typographique[3]. Enfin, l'introduction de nouveaux domaines de connaissance (comptabilité, algèbre, médecine, etc.) dans le champ éditorial entraîne l'apparition de nouveaux caractères.

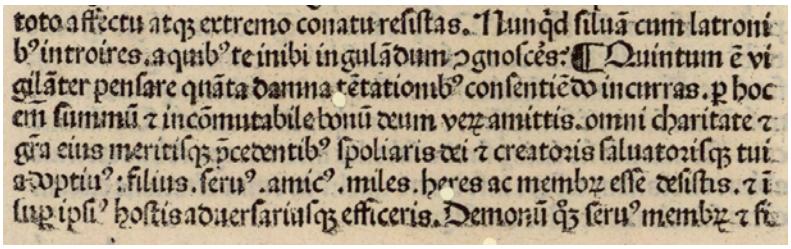

toto affectu atqz extremo conatu resistas. Nunqd siluã cum latronib⁹ introires. a quib⁹ te inibi in gulã dum ꝓgnoscés? ¶ Quintum ẽ vigilãter pensare quãta damna tẽtationib⁹ consentiẽdo incurras. ꝓ hoc eñ summũ z incõmutabile bonũ deum vez amittis. omni charitate z gr̃a eius meritisqz ꝑcedentib⁹ spoliaris dei z creatoris saluatorisqz tui adoptiu⁹: filius. seru⁹. amic⁹. miles. hẽres ac membz esse desistis. z i sup ipsi⁹ hostis aduersariusqz efficeris. Demoñ qz seru⁹ membz z fi

Figure 1. Profusion de signes imprimés hérités de la tradition manuscrite (abréviations, ligatures, crénages) : Dionysius Carthusiensis, Exhortationes novitiorum, *Deventer, 1491 (Bourges, BM)*

1. Cet article fusionne les textes largement remaniés de deux communications présentées séparément lors du colloque GIÉcA.
2. La gravure d'un seul poinçon typographique en acier permet de frapper plusieurs dizaines de matrices en cuivre, dans lesquelles pourront être fondus plusieurs centaines de milliers de caractères en plomb identiques.
3. « Il se trouve que les prémices des transformations importantes de notre orthographe au XVI[e] siècle ont d'abord apparu dans les ateliers, et avant l'intervention des auteurs » (Catach, 1968, p. XVII). Pour un exposé synthétique de ses conclusions, voir également Catach (1997).

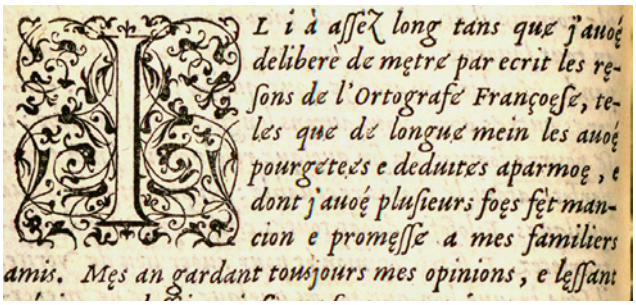


Figure 2. Un exemple d'orthographe « phonétique » : Jacques Peletier du Mans, Dialogue de l'Ortografe e Prononciacion Francoese, *Lyon, 1555. (Tours, BU)*

Inertie des habitudes manuscrites, lente élaboration d'une orthographe moderne, constitution de langages scientifiques spécialisés : nous sommes en présence de phénomènes distincts dont la conjonction aboutit à un véritable foisonnement de formes typographiques. Certaines seront pérennisées, tandis que la majorité n'aura qu'une existence éphémère. Cette situation fait de la typographie de la Renaissance un véritable laboratoire de la culture graphique occidentale. Toute la question est de savoir comment les chercheurs en histoire du livre, en littérature, en histoire des sciences ou en linguistique, peuvent s'accommoder de cette situation dans un cadre de travail numérique, en préservant autant d'informations que possible.

2. Transcrire l'information graphique

2.1. Des caractères « spéciaux » ? Plaidoyer pour une paléotypographie

Ligatures, abréviations, caractères phonétiques, signes spécifiques de ponctuation sont donc massivement présents dans les imprimés de la Renaissance. Rares sont pourtant les chercheurs qui attachent de l'importance à ces caractères que l'on dit (à tort) « spéciaux ». Le transcripteur moderne, qui perçoit ces signes comme des obstacles à la lecture et à la transcription, opte encore trop souvent pour la solution de facilité qui consiste à ignorer purement et simplement leur présence en régularisant les graphies, au mieux selon des normes explicites, au pire de manière tout à fait inconsciente.

Il est pourtant possible d'envisager un champ de recherches *paléotypographiques*, qui s'attacherait précisément à l'histoire et aux usages de ces caractères imprimés[4]. Cette histoire croisée des pratiques typographiques et des

4. Le terme *palaeotypography* a été introduit par Henry Bradshaw (1870) et récemment repris par Hendrik Vervliet (2008) pour désigner l'étude, à des fins d'identification bibliographique, des fontes employées par les imprimeurs des siècles passés. Nous élargissons ici sa définition

usages linguistiques ne semble guère avoir été explorée jusqu'à présent que par une linguiste, Nina Catach, dont les travaux doivent être considérés comme fondateurs.

Voici presque un demi-siècle, Catach publiait un livre devenu classique : *L'orthographe française à l'époque de la Renaissance* (Catach, 1968), dont le sous-titre (« Auteurs, imprimeurs, ateliers d'imprimerie ») montre assez l'attention prêtée à l'art typographique. Catach se proposait alors « d'étudier la typographie avec des yeux de linguiste » (1968, p. XVII). Il s'agissait moins pour elle de décrire l'aspect général de telle ou telle fonte, que d'analyser la diffusion de tel ou tel signe pris individuellement. En 1983, à l'occasion d'un colloque consacré à la « bibliographie matérielle », Catach définissait un véritable programme de recherche :

> *Nous en sommes actuellement aux premiers jalons de cette nouvelle discipline, qui devrait avoir sa place auprès des autres, et à mon avis au tout premier rang, dans la nouvelle bibliographie : inventaire et recensement des casses d'imprimerie, existence et importance des accents, des signes de ponctuation et des signes auxiliaires, présence de ligatures et d'abréviations, de telle ou telle capitale de signe nouveau, richesse en caractères italiques, en caractères spéciaux, etc. ; alternances d'habitudes des compositeurs d'un cahier à l'autre, systèmes graphiques différents entre l'auteur et ses secrétaires, copistes, correcteurs, etc. ; relevé des* errata *et, quand ils existent, des repentirs et des corrections sur les placards ; lettres entre auteurs et imprimeurs, déclarations et préfaces sur l'orthographe, contrats internes d'embauche, commerce d'échanges, commandes de matrices et de poinçons aux graveurs, etc. Les pistes sont innombrables et ne peuvent se désolidariser les unes des autres, ce qui en fait toute la difficulté. […] Tous ces éléments d'histoire des techniques ne sont pas indifférents aux historiens de la littérature et de la langue. Je dirais même qu'ils leur sont indispensables. […] On ne peut plus se contenter de parler de « hasards inconnaissables », d' « arbitraire total », de « ponctuation insupportable et aberrante ». […] Partons (même si ce n'est pas tout à fait vrai) de l'hypothèse que chaque signe a sa raison d'être en fonction d'un ensemble de processus, qui nous échappent encore, mais qu'il nous reste à découvrir.*
>
> (Catach, 1983)

Nina Catach est décédée en 1997. Depuis, peu de chercheurs ont repris le flambeau de cette étude croisée des pratiques ortho- et typo-graphiques. Ce champ intéresserait pourtant aussi bien la linguistique que l'histoire littéraire, l'histoire du livre ou de l'écriture, ou encore la génétique des textes. L'informatisation massive des corpus textuels devrait aujourd'hui faciliter ces recherches, à condition toutefois que les pratiques de transcription s'appuient sur des bases solides et clairement établies.

---

en désignant sous ce terme non seulement l'étude des fontes, mais aussi celle des signes imprimés en eux-mêmes.

2.2. Articuler les niveaux de transcription

Il n'est pas lieu ici de définir précisément des normes de transcription (elles doivent être adaptées à chaque objet, à chaque besoin scientifique). On peut en revanche s'attarder sur la distinction fondamentale entre différents « niveaux de transcription » d'un même texte. Robinson et Solopova (1993), éditeurs d'une version numérique des *Canterbury Tales* de Chaucer fondée sur des manuscrits, en définissaient trois[5] :

1. Une **transcription graphétique**, distinguant précisément chaque variante graphique. On y discernerait par exemple le « s » du « ſ », ou les différentes variantes de lettres initiales et finales. Stutzmann (2010) précise que ce mode de transcription suppose une réflexion préalable sur la typologie, l'ontologie des formes. Il implique de réduire les variantes graphiques à des classes explicitement désignées.

2. Une **transcription graphémique,** respectant la graphie (« *spelling* ») de chaque mot (comme les lettres quiescentes du moyen français) mais sans tenir compte des éventuelles variations graphiques de ces différentes lettres. On y distinguerait par exemple *vostre* de *votre*, mais pas *voſtre* de *vostre*. C'est également à ce niveau qu'interviendrait la régularisation des mots entièrement composés en lettres capitales.

3. Une **transcription régularisée**, qui normaliserait la graphie des mots, développerait les abréviations. Si elle conserve encore un certain intérêt pour l'histoire du lexique, de la grammaire ou de la syntaxe, cette pratique n'a déjà plus d'utilité pour l'histoire de l'écriture.

On pourrait enfin ajouter à cette typologie un quatrième niveau qui serait celui de la **transcription modernisée** (qui s'apparente, en fait, à une traduction), destinée à la diffusion d'un texte auprès du grand public.

Appliquée aux livres imprimés et à la langue française du XVI^e^ siècle, cette typologie des niveaux de transcription nécessite quelques aménagements, sur lesquels il n'est pas lieu ici de nous appesantir[6]. Il convient en revanche d'insister sur l'articulation de ces différents « niveaux de transcription ». On aurait tort de les considérer comme indépendants les uns des autres. Ils constituent en réalité différents maillons d'une seule et même chaîne de régularisation, chaque niveau

5. Conceptuellement, Robinson et Solopova prévoient l'existence d'un niveau de transcription dit « graphique », rendant compte de chaque espace du manuscrit original. Mais D. Stutzmann (2010) remarque à juste titre qu'il s'agit d'une « illusion », et que seule une véritable image peut rendre compte de toutes les particularités graphiques d'un texte.

6. Un commentaire cependant. On a indiqué que c'est au niveau de la transcription graphémique qu'intervenait le rétablissement des bas-de-casse pour les mots entièrement composés en (petites) capitales. Cette opération ne va pas sans poser quelques problèmes : si l'on souhaite rendre compte des diphtongues et digrammes, pour les textes du XVI^e^ siècle, il convient de porter une grande attention aux espaces interlettrées. De ce point de vue, « OE V V R E » et « O E V V R E » constituent deux entités distinctes : au niveau graphémique, la première devrait être transcrite « œuvre », et la seconde « oeuvre ».

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 nécessite l'existence préalable de normes de transcription très détaillées, et qui fait par ailleurs perdre de manière irréversible de nombreuses informations graphiques. Dans une perspective paléotypographique héritière des travaux de Catach, la transcription graphétique serait la seule véritablement utile[7].

2.3. Codage ou balisage : les niveaux d'enregistrement de l'information graphique

Mais l'existence même d'une « transcription graphétique » requiert les moyens techniques de reproduire les caractères des textes originaux. Or les technologies numériques accessibles aux chercheurs ne permettent pas encore d'assurer une correspondance parfaite du texte numérique avec la composition du document original. Le problème est double. Il se situe d'une part au niveau du codage typographique proprement dit : même le plus complet des codages, Unicode (voir *infra*, § 3.1), ignore l'existence d'un grand nombre de signes typographiques attestés à la Renaissance. D'autre part, même lorsque les caractères concernés sont intégrés à un codage typographique, rares sont les polices qui permettent de les afficher sur écran ou de les imprimer.

Ne disposant ni des codes, ni des fontes adéquates, les universitaires répondent au cas par cas aux difficultés qu'ils rencontrent, sans toujours se préoccuper d'harmoniser leurs pratiques. Une solution communément adoptée consiste à produire une transcription en « quasi-facsimilé » (pour les principes, voir Bowers, 1949, p. 135-179), en choisissant des caractères ressemblant aux originaux dans diverses fontes numériques, selon une conception exclusivement graphique du signe (on remplacera par exemple l'abréviation latine *–us*, « ⁹ » par le chiffre 9 en exposant). Certains projets bibliographiques vont jusqu'à dessiner des polices comprenant des caractères spécifiques, sans toutefois se préoccuper de rationaliser leur codage[8]. De tels « bricolages » ont pu rendre de nombreux services et n'avaient

7. Il nous faut ici préciser que par « transcription » nous désignons toujours un texte brut, issu d'un relevé obtenu manuellement ou automatiquement (par OCR), avant tout enrichissement ultérieur. La transcription telle que nous l'entendons n'intègre donc aucune mise en forme (taille, graisse ou couleur des caractères, choix d'une fonte spécifique, etc.), même si la mise en forme contribue à conférer son sens au texte (nous conservons en revanche à ce niveau les coupures de ligne). Nous supposons que c'est au niveau du texte enrichi (mise en forme sous traitement de texte, balisage TEI, etc.) que l'on rendra compte de l'aspect du document original (choix des fontes, italique, graisse, corps, etc.). Toutefois, certaines structures spéciales (telles que les tableaux, les formules mathématiques, les arborescences textuelles, etc.) devront faire appel à un balisage structuré (tel que MathML) mais, quel que soit le niveau de son utilisation, ces structures verront leurs éléments de base codés en termes de typèmes (sur cette notion, voir *infra*, § 4). Ainsi les accolades, composées par morceaux, utiliseront les codes U+23A7 (ARC SUPÉRIEUR D'ACCOLADE GAUCHE) et suivants.

8. Voir, par exemple, Pedraza Garcia *et al*., 1998, qui propose une fonte pour le catalogage des imprimés anciens, et Bettens, 2006, qui offre une fonte, BaifB, simulant l'écriture de Baïf.

pas de conséquences néfastes lorsque la publication sur papier était la finalité des travaux de recherche[9]. Mais à l'heure de l'informatisation des corpus textuels, ces solutions d'appoint, qui ne garantissent ni la pérennité ni l'interopérabilité des données, doivent être considérées comme insuffisantes.

Les problèmes strictement typographiques peuvent cependant être contournés par le recours à des modes alternatifs d'enregistrement de l'information. Le format XML permet de produire une transcription typographiquement simple, tout en signalant les particularités graphiques du document original par un balisage spécifique. La TEI offre ainsi aux paléographes la possibilité de signaler efficacement les variantes allographétiques des manuscrits[10]. La tentation est grande d'adapter cette pratique à la transcription de documents imprimés anciens, mais il nous faudra sans doute y résister.

Enregistrer les spécificités graphiques d'un document par le moyen d'une couche de balisage n'est pas une opération anodine : elle sépare le contenu sémantique du texte d'une part, et sa représentation d'autre part. Elle suppose l'existence d'une transcription simplifiée, enrichie *a posteriori* par le signalement de certains caractères considérés comme « spéciaux ». Une telle opération change le statut même du signe : *décrit* plutôt que *transcrit*, il n'est plus un élément constitutif du texte proprement dit, au sein duquel il se voit remplacé par un caractère ou groupe de caractères plus simple. On substitue alors au *signe* une simple *information graphique*, enregistrée sous forme de métadonnée.

La perspective paléotypographique que nous défendons ici repose sur le postulat (discutable mais méthodologiquement nécessaire en l'état actuel de nos connaissances ; voir *supra*, § 2.1) qu'il n'existe pas de caractères « spéciaux » dans le système typographique, ni d'allographe au sein d'une même fonte. C'est précisément parce que la typographie constitue un système *fermé*, et à ce titre bien différent du modèle manuscrit, que de tels postulats demeurent raisonnables. Les solutions techniques que nous préconisons dans la suite de cet article ne sont donc pas adaptables à la transcription de documents manuscrits, qui nécessitent des traitements différents. Distribués séparément dans la casse du compositeur, le s-long (ſ) et le s-rond (s) ne peuvent être considérés comme deux variantes graphiques d'un même caractère, mais doivent être interprétés comme deux signes bien distincts ; il en va de même pour les lettres initiales ou finales, les ligatures, les capitales plus ou moins calligraphiques, etc. Dans ce cadre méthodologique, tous les caractères d'un texte doivent être traités informatiquement de manière identique et l'on ne peut se contenter de signaler la présence de caractères réputés « spéciaux ».

Cette question du niveau d'enregistrement des informations graphiques (codage ou balisage) n'est pas d'ordre strictement conceptuel ; elle a des conséquences

9. On en trouvera un bon exemple dans Legros (2010), qui emploie la typographie pour rendre compte des allographes de l'écriture manuscrite de Montaigne (voir notamment p. 39-50, « Conventions typographiques »).
10. Voir, par exemple, dans ce volume, les solutions proposées par D. Stutzmann (notamment section 3.2 sur la définition d'entités).

techniques immédiates et concrètes : en rendant l'information tributaire d'un format spécifique (par exemple le XML-TEI), et non plus d'un codage typographique à vocation universelle, on restreint d'emblée les possibilités de transfert des données d'une technologie à l'autre. C'est donc selon nous au niveau du codage typographique même qu'il importe de reproduire le caractère.

3. Les codages Unicode et MUFI

3.1. Unicode

Le codage Unicode a vu officiellement le jour en octobre 1991 (version 1.0)[11]. Géré par un consortium privé (contrairement à l'Iso), il évolue sans cesse ; la version 6.3 a été publiée le 15 novembre 2013. Unicode est un codage de transmission de caractères entre ordinateurs et périphériques. Il ne s'agit donc ni d'un logiciel, ni d'un outil d'édition ou d'impression. Son principe est simple : on associe à chaque caractère un numéro (en anglais *code point*) qui lui est propre (par exemple pour *R* le numéro 0052) et un nom (pour *R* le nom *Lettre majuscule latine R*)[12]. Unicode a une vocation universelle : tous les caractères sont, ou seront, présents dans Unicode, sans restriction géographique (on trouve les caractères de toutes les langues européennes, aussi bien qu'orientales ou africaines), ni chronologique (par exemple les oghams celtes ou les caractères ougaritiques). Au cours des révisions successives, le nombre de caractères Unicode augmente. Tout le monde peut proposer de nouveaux caractères, un consortium statuant en dernier recours sur leur intégration.

Unicode associe à chaque caractère une série de propriétés : sens d'écriture, ordre alphabétique pour les tris, etc. Certains caractères peuvent être définis comme la combinaison de caractères plus élémentaires ; par exemple un « è » peut être défini comme composé d'un « e » surmonté d'un accent grave « ` ». On parle alors de caractères composites.

Unicode distingue très nettement le caractère, objet abstrait, plutôt linguistique, de ses représentations graphiques concrètes appelées « glyphes ». Il reconnaît donc la majuscule latine *R*, mais considère que les traces imprimées ou affichées sur écran de *R* en Palatino ou en Times, en corps 8 ou 12, en italique ou en gras, voire en position supérieure, ne sont que des glyphes différents du même caractère *R* codé U+0052. Cette distinction entre caractère et glyphe semble assez raisonnable

11. La documentation sur Unicode se trouve centralisée et mise à jour en continu sur le site internet du consortium (Unicode, 1991) ; voir aussi, pour une approche générale, (Andries, 2008 ; Unicode, 2013). Sur les rapports d'Unicode avec la typographie, voir André et Hudrisier (2002). Pour une synthèse générale sur les principes et les formats de codage, voir Haralambous (2004).

12. Ces noms sont en anglais. Mais il en existe une version française normative définie par ISO 10646:2003, accessible dans (Unicode2013a, Andries 2013). Ce sont ces formes françaises que nous utilisons ici, mais nous conservons l'anglais lorsque les noms français ne sont pas encore normés.

puisque les caractéristiques graphiques des caractères (taille, graisse, couleur, forme,...) peuvent être manipulées après codage, à un niveau supérieur (celui du texte « enrichi », voir *supra*, note 6).

Ajoutons que ce principe de distinction caractère/glyphe est poussé à l'extrême. Ainsi, les ligatures ne sont pas prises en considération par Unicode, pour qui « ct » n'est qu'une variante « glyphique» du couple de caractères « c t ». Pour Unicode, la gestion des ligatures doit alors être confiée à des formats de polices tel OpenType. Ceci est symptomatique : si Unicode conçoit l'édition graphique d'un texte abstrait, il semble ignorer la démarche inverse qui nous préoccupe ici, à savoir l'extraction d'un texte abstrait à partir d'un texte imprimé existant. Cette position est certes défendable, mais on doit toutefois signaler quelques incohérences. En vertu d'un « principe de convertibilité » à l'égard de tous les codages antérieurs à mai 1993, les ligatures ﬀ, ﬁ, ﬂ, ﬃ, ﬄ, « ﬅ » et ﬆ qui figuraient dans l'*Expert Character Set* de Postscript ont été officiellement intégrées à la grille Unicode sous les numéros FB00 à FB06. Selon Haralambous (2004, p. 61), ce principe de convertibilité est celui « qui a causé le plus de tort à Unicode […]. Le fait est que 99,9 % des incohérences d'Unicode sont dues à ce seul principe ». Il serait donc utopique de demander aujourd'hui au consortium l'intégration de ligatures supplémentaires, tels les bigrammes « ct » ou « us » pourtant fréquents dans les fontes anciennes comme celles de Claude Garamont (André, 2011).

La possibilité offerte par Unicode de recourir à des caractères « combinatoires » justifie également le refus d'intégrer un certain nombre de signes composites : ainsi le caractère abréviatif « » (pour *hoc*) est-il considéré par Unicode comme un signe formé des caractères U+0068 (LETTRE MINUSCULE LATINE H) et U+0366 (DIACRITIQUE LETTRE MINUSCULE LATINE O) et non comme une seule et même entité, ce qui rend le travail des codeurs plus difficile.

3.2. MUFI

Conscient des contraintes imposées par ces principes fondamentaux, Unicode offre cependant des plages de codes dans une grille réservée à des usages privés (*Private Use Area*, abrégée en *PUA*). C'est le moyen auquel ont eu recours les universitaires désireux d'utiliser Unicode pour échanger et éditer des textes médiévaux. Des médiévistes ont mis sur pied un projet, MUFI (pour *Medieval Unicode Font Initiative*), destiné à recenser les caractères manquants à Unicode et à proposer leur intégration au consortium[13]. Grâce à la pugnacité du groupe MUFI,

13. Voir MUFI (2001), Haugen (2009) et la contribution de Odd Einar Haugen dans ces actes (Haugen, 2013). Les chercheurs impliqués dans MUFI venaient principalement d'Europe du nord, et peu de Méditerranéens se sont impliqués dans le projet, aussi les ajouts à Unicode concernèrent d'abord les langues nordiques (voir cependant Emiliano et Pedro, 2013). Outre la définition de caractères, MUFI produit également des fontes supportant ces nouveaux caractères (la fonte Cardo, utilisée ici pour les exemples, en fait partie).

Unicode intègre désormais officiellement des caractères abréviatifs tels que « ꝑ » ou « ꝙ ».

Mais certains signes nécessaires aux travaux des médiévistes ne recoupent pas la définition stricte du « caractère » par Unicode et ne peuvent donc être intégrés au codage officiel. Tel est par exemple le cas des ligatures « ﬀ » ou même « ct ». Le consortium MUFI assigne donc à ces caractères des codes spécifiques dans la *Private Use Area*. Sans être reconnus officiellement par Unicode, ces signes peuvent ainsi être représentés et leur codage normalisé. Le consortium MUFI a formulé trois *Character recommendations* successives relatives à l'usage des caractères utilisés dans les textes médiévaux utilisant l'alphabet latin (MUFI, 2009). La ligature « ﬀ » que nous venons de citer est ainsi définie dans une telle zone avec le numéro EED7.

MUFI est un projet vivant ; une liste ouverte accueille les propositions de nouveaux caractères qui sont en attente d'inclusion dans MUFI[14].

3.3. État des lieux

Sous la pression d'utilisateurs, Unicode a donc introduit depuis sa version 5 beaucoup de caractères supplémentaires. De son côté, la version 3.0 de la *Character recommendation* de MUFI complète la grille d'Unicode par environ 1 600 signes identifiés dans les manuscrits médiévaux. Un certain nombre de signes utilisés dans ces écritures manuscrites ont des équivalents typographiques. Par commodité, nous désignons désormais par « MUFI » l'ensemble des caractères Unicode officiels complété du sur-ensemble des caractères médiévaux figurant dans la PUA. Ce codage MUFI offre un jeu de caractères plus important qu'on ne le croit souvent et permet de coder certains des caractères imprimés de la Renaissance. On en trouvera une longue liste dans (André, 2014) ; la figure 3 montre quelques exemples typiques.

Classes de caractères	Unicode	Mufi
Lettres	ɇ ɬ ſ	ꝁ
Ligatures	ﬀ ﬆ	ct ﬃ
Abréviations « latines »	ꝑ ꝓ ꝑ ꝙ ⁹	ꝙ ꝓ ꝕ
Traits d'union, marqueurs,...	⸗ ⸚ ⸛ ⸘ ⁜	
Symboles monétaires et pondéraux	ↄ ₶ ʒ ₥	ȝ ₹ ₸
Signes d'astrologie et du zodiaque	☍ ☾ ♆ ♒	

Figure 3. Exemples de caractères présents dans Unicode ou MUFI

14. Voir le « pipeline » de MUFI : http://www.mufi.info/pipeline/

4. Le typème, « chaînon manquant » entre caractère et glyphe

4.1. Le concept de typème

La définition strictement linguistique du « caractère » telle que formulée par Unicode ne correspond pas à la réalité matérielle des sources anciennes. Il nous paraît donc nécessaire, pour éviter toute ambiguïté, d'introduire un nouveau concept, celui de *typème*, sur lequel nous devons nous appuyer pour garantir la cohérence paléotypographique des transcriptions de textes imprimés anciens, qu'elles soient obtenues manuellement ou automatiquement (par OCR).

Qu'est-ce au juste qu'un *typème* ? Le mot employé est un néologisme[15], car nous ne connaissons aucun terme français correspondant précisément à sa définition. Le terme « caractère » est à écarter car Unicode lui donne un sens trop linguistique. Le lexique spécialisé des typographes (avec les termes « type », « sorte » et « œil ») ne convient pas non plus car ces mots ne permettent pas de regrouper les allographes ; ils ont d'ailleurs une polysémie très forte. Le mot « graphème », à connotation linguistique, ne peut pas non plus être utilisé car les spécialistes ne s'accordent pas sur sa définition : même des linguistes pourtant ouverts à l'imprimé comme Jacques Anis (1988) et Nina Catach (1988) en donnent des définitions différentes. Nous proposons donc d'utiliser *typème*, mot valise formé sur *type* et sur *graphème*, que nous définissons comme « l'unité minimale effective entrant dans la composition d'un texte affiché ou imprimé »[16]. L'adjectif « effective » revêt ici toute son importance : conçu comme une zone intermédiaire entre la pure représentation graphique et l'entité abstraite linguistique, ce concept de typème se fonde sur la dimension technique de la typographie.

4.2. Typographie et typèmes

Un caractère typographique est obtenu par un processus en plusieurs étapes (André et Laucou, 2013, p. 278). D'abord la taille d'un poinçon en acier portant en relief et à l'envers l'image de la lettre. Ce poinçon est employé à frapper une matrice de cuivre dans laquelle il laisse son empreinte (à l'endroit et en creux, figure 6). Cette matrice est placée dans un moule au sein duquel on verse un mélange de plomb, d'antimoine et d'étain, afin d'obtenir un caractère portant la lettre en relief (et à l'envers). Encré, ce caractère pressera une feuille de papier sur laquelle il laissera son image (à l'endroit). Il faut donc bien distinguer la trace imprimée sur le papier (qu'on appelle « œil ») d'une part, et le caractère en plomb (qu'on appelle « type ») d'autre part.

15. En fait ce mot a déjà été utilisé il y a une cinquantaine d'années, à peu près avec le même sens, par Göram Hammarström (1964). Signalons que, de son côté, Janusz Bień (2009) utilise le mot « *textel* ».
16. Définition introduite pour la première fois par André (2011).

Objets physiques (parallélépipédiques), sortis d'un même moule, les caractères d'une fonte possèdent le même corps et peuvent donc être composés les uns à la suite des autres pour former une ligne de texte ; ils ne peuvent, en revanche, en aucun cas se chevaucher[17]. On ne peut donc pas procéder à la combinaison de plusieurs types pour obtenir, par exemple, un *e* surmonté d'un accent aigu. Les caractères accentués (ou soulignés, ou chevauchés) doivent être fondus *tels quels*. Si l'on souhaite assembler plusieurs signes en un même type, c'est *antérieurement* à la fonte qu'il convient de le faire : soit au niveau de la gravure du poinçon, soit au niveau de la frappe de la matrice. Beaucoup de caractères accentués étaient ainsi obtenus en frappant sur une même matrice de cuivre le poinçon d'une lettre et celui de l'accent (voir *infra*, paragraphe 4.3). De même, à moins d'avoir recours à des techniques délicates (et de ce fait peu utilisées) comme le parangonnage, on ne peut pas composer sur une même ligne des types de corps différents. On ne peut donc pas utiliser une fonte de corps inférieur pour composer des petites lettres en exposant : celles-ci doivent être gravées et frappées spécifiquement, avant d'être fondues sur le même corps de caractère que l'alphabet principal.

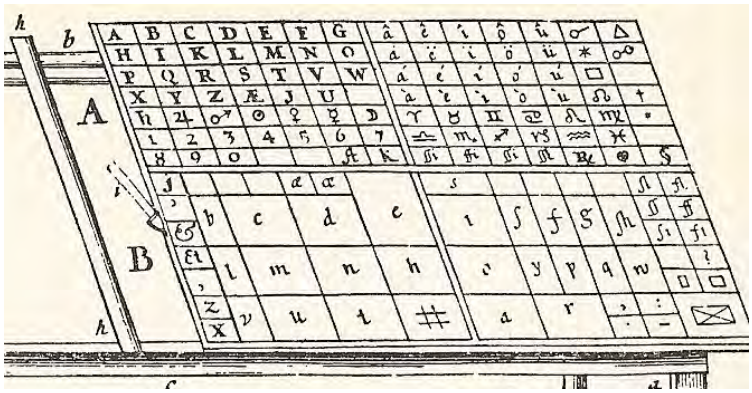


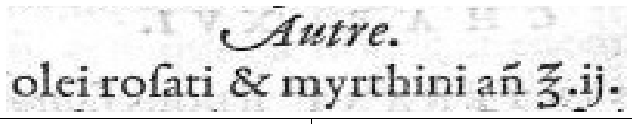
Figure 4. Plan de casse de Moxon (1683). On y voit notamment des lettres (capitales et bas-de-casse), des caractères accentués, des signes de ponctuation, des ligatures (dont un « ſh » spécifique à la typographie anglaise), des signes des planètes et du zodiaque, ou encore les cassetins dédiés aux espaces (représentées par le signe ⌗) et aux cadrats (⌧)

Une fois fondus, les caractères sont rangés dans une « casse », tiroir plat compartimenté en une séries de petites cases, dites cassetins, chacun de ces cassetins

17. Exception faite des techniques basées sur la découpe des types (associée à du parangonnage) ou leur limage (crénage), sur lesquelles nous ne pouvons pas nous attarder ici.

recevant une seule « sorte » de types[18]. On trouve dans la casse typographique toutes les lettres (majuscules, minuscules et petites-capitales), les chiffres, les signes de ponctuation, les ligatures courantes, divers filets et espaces, etc. On retrouve aussi parfois (par exemple dans la casse de Moxon, figure 4) des signes du zodiaque, des abréviations latines, etc.[19] Dès le XVI^{e} siècle, ces plans de casse sont pratiquement invariants (André et Laucou, 2013, p. 81), la plupart des imprimeurs utilisant dès l'origine des modes de classement assez proches, pour tous les types de fontes latines utilisées. Au sein d'une même imprimerie, les fontes romaines, italiques, gothiques étaient classées selon le même plan, ce qui permettait aux ouvriers de composer facilement des textes, quelle que soit la nature de la police utilisée.

Cette normalisation du plan des casses montre que les imprimeurs eux-mêmes reconnaissaient comme une même entité un signe, quel que soit le dessin ou le corps de la police dans laquelle il pouvait être fondu. Voilà qui justifie, selon nous, l'existence du concept de *typème*. Il ne s'agit pas d'une notion purement théorique ; il s'agit pour nous de nommer un concept auquel les anciens typographes avaient inconsciemment recours.



A majuscule	1	a minuscule	2
u minuscule	1	esperluette	1
t minuscule	3	m minuscule	1
r minuscule	3	y minuscule	1
e minuscule	2	h minuscule	1
point	3	n minuscule	1
o minuscule	2	n minuscule avec titulus	1
l minuscule	1	signe once	1
i minuscule	5	j minuscule	1
s long	1	espace	5

Figure 5. Inventaire des typèmes d'un texte donné (extrait de Ambroise Paré, Cinq livres de chirurgie, *Paris, Wechel, 1572)*

18. On appelle *sorte* l'ensemble des types d'un même caractère appartenant à une même fonte (par exemple, la fonte de « garamonde romaine » acquise par Christophe Plantin et présentée figure 18 compte 5 000 a. Ces 5 000 types forment une seule et même *sorte*).
19. Les casses n'ont toutefois pas une taille suffisante pour contenir toutes les sortes de lettres ou caractères. On range alors les caractères supplémentaires dans des casses complémentaires, dites casseaux, mais sans plan prédéfini.

En clair, tout signe formant l'œil d'un type en plomb constitue un typème. Le style d'une police (son allure) et les caractéristiques métriques (taille, position, etc.) n'entrent pas en jeu dans la définition du typème. En revanche, le regroupement de plusieurs signes sur un même caractère de plomb (comme la ligature « ﬅ » ou même la voyelle accentuée « é ») doit être considéré comme un seul typème. La figure 5 illustre ce concept. On y remarque que le « t » minuscule italique de « *Autre* » et ceux (romains) de « roſati » et « myrthini » correspondent à un même typème, que l'abréviation « *ñ* » (n minuscule avec titulus) n'est pas traitée comme un caractère composite mais comme un seul et même typème, que les trois occurrences du point « . » sont identifiées comme un seul et même typème (même si les deux derniers signalent la présence d'un chiffre qu'ils encadrent) et que les chiffres romains minuscules « i » et « j » sont traités comme des lettres minuscules. On ne tient pas non plus compte ici de la forme spéciale du « A » italique (spécifique au style de la fonte, et donc sans connotation).

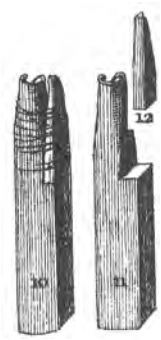
4.3. Caractères composites, caractères simulés

Le concept de typème nous amène à rejeter la notion de « caractères composites » telle qu'Unicode la définit. Certes, nombre de matrices étaient effectivement frappées par une combinaison de deux poinçons : les lettres accentuées pouvaient ainsi être réalisées en frappant, sur un même bloc de cuivre, le poinçon de la lettre de base et celui d'un accent (figure 6). Mais, une fois la matrice frappée, les types qui y sont moulés constituent des entités qu'il n'est plus possible de décomposer[20]. Nous considérerons donc chaque typème comme une seule entité, à laquelle le codage numérique doit attribuer un numéro unique.

Nous proposons ainsi l'entrée dans MUFI d'une LETTRE MINUSCULE LATINE *E* À CROCHET. Ce caractère, apparu chez Conrad Néobar en 1540, a été gravé par Garamont[21]. L'analyse archéologique de la matrice originale conservée au musée Plantin-Moretus montre qu'elle a été frappée à l'aide de deux poinçons différents : celui d'un *e*, puis celui d'un esprit doux grec (figure 7). Il serait pourtant incohérent de composer ce caractère dans une transcription numérique à l'aide des deux codes U+0065 (lettre minuscule latine e) et U+1FBF (esprit doux, ou *psili*, grec), car il s'agit bien d'un seul et même typème.

20. Considérer les typèmes comme des caractères composites n'irait d'ailleurs pas sans poser problème : le caractère LETTRE MINUSCULE LATINE E BARRÉ (U+0247) a généralement pour glyphe un « e barré » : « ɇ ». Mais dans les imprimés du XVI^e^ siècle, le caractère désigné « e barré » par Nina Catach (1968) est en fait un « c » barré (les matrices montrent bien qu'elles ont été obtenues par la frappe d'un c et d'une barre oblique). Il est bien évident qu'on ne va pas coder cet *e* sourd comme la composition d'un c et d'une barre mais bien comme le typème « e barré » !

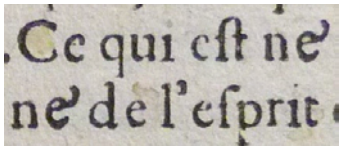
21. Voir Catach (1968, p. 82) et André (2011).



a)

b)

Figure 6. a) utilisation de poinçons d'accents mobiles. Ici le poinçon d'un accent aigu est associé à celui d'un e *pour frapper une matrice (Fournier, 1764, pl. IV). b) poinçons de caractères de civilité (XIX^e^ s.) présentant une encoche destinée à recevoir un accent mobile (Lyon, Musée de l'imprimerie)*



a)

b)

Figure 7. a) deux occurrences du e à crochet dans La Vie de Nostre Seigneur Jesus Christ *(Paris, Néobar, 1540) ; b) la matrice originale du musée Plantin-Moretus. D'après André (2011)*

Envisageons à présent le cas inverse : celui des *caractères simulés* par l'association de plusieurs typèmes. Lorsqu'il leur manquait un caractère spécifique, les imprimeurs pouvaient en simuler l'aspect en associant deux types différents. Tel est bien souvent le cas, au XVI^e^ siècle, des capitales accentuées, tel le *É*, reproduit par la combinaison d'un simple *E* et d'une apostrophe (figure 8). De même, les fontes du début de la Renaissance possèdent souvent la capitale K (utilisée pour le latin *Kalenda*), mais sont généralement dépourvues de son équivalent en bas-de-casse. Par convention, les imprimeurs utilisent alors l'association d'un *l* et d'un *z* pour simuler l'aspect visuel d'un *k* (figure 9). Au niveau typémique, la dimension « matérielle » prévalant sur la valeur sémantique, nous considérons bien ces caractères comme deux entités distinctes, en leur attribuant deux codes différents.

Figure 8. Caractère simulé : É *restitué à l'aide d'un* E *et d'une apostrophe, dans Jean de Serres,* De l'immortalité de l'âme, *Lyon, frères Gabiano, 1596 (Blois, BM)*

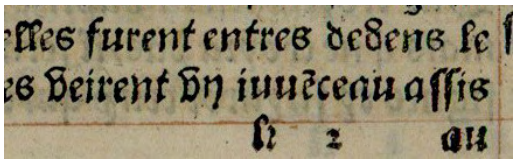


Figure 9. Caractère simulé : signature d'un feuillet « k 2 » *dans laquelle le* k est *restitué par la combinaison d'un* l *et d'un* z *dans* Le Nouveau Testament*, Genève, 1538 (Châteauroux, médiathèque Equinoxe)*

4.4. Variantes glyphiques ou typémiques ?

La distinction qu'instaure Unicode entre glyphe et caractère est opérante en ce qu'elle permet de différencier l'entité linguistique de ses simples manifestations graphiques (variante de style, de taille, graisse, etc.). Il convient cependant de rappeler que, dans les textes imprimés anciens, certaines « variantes graphiques » n'ont pas une simple valeur esthétique. Les lettres à paraphe (ou traits de plume) peuvent souvent avoir un usage spécifique. Le petit-parangon romain gravé par Claude Garamont en 1557 (Vervliet 2010, n° 128) possède ainsi des variantes avec traits de plume pour les lettres a, e, m, n, r, t, et z. On peut être tenté de voir dans ces types de simples « variantes glyphiques » destinées à être utilisées en fin de mot. Mais le *Spécimen* de Le Bé-Moretus (c. 1599) montre clairement (figure 10) que leur utilisation n'est pas généralisée. Dans la perspective paléotypographique qui est la nôtre, partant du postulat méthodologique que tout peut être signifiant, il convient de conserver cette information (qui peut correspondre, par exemple, à une indication phonétique, ou simplement venir seconder la ponctuation pour souligner la syntaxe de la phrase). Nous proposons donc de faire entrer, en plus de la classique LETTRE MINUSCULE LATINE E (U+0065), un second numéro pour LETTRE MINUSCULE LATINE E AVEC PARAPHE, et de même pour les autres lettres[22].

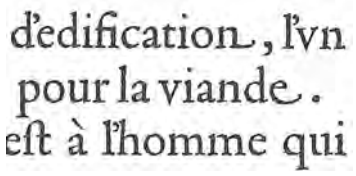


Figure 10. Lettres à paraphe du petit-parangon de Garamont (1557) tel qu'il est présenté dans le Spécimen *de Le Bé-Moretus (ca. 1599). Extrait de (Vervliet, 2010)*

22. Il conviendrait même de distinguer les lettres à paraphe initiales (dont le trait de plume part vers la gauche) des lettres à paraphe finales (qui tirent vers la droite).

Certaines capitales calligraphiques ont été utilisées dès le début du XVI^e siècle comme symboles spéciaux (par exemple les symboles monétaires ou pondéraux) et il est d'usage de leur attribuer un numéro de code spécifique (voir figure 3). De même, les mathématiciens ont dès cette époque eu besoin de symboles facilement repérables. Ainsi Stifel utilise-t-il dans son *Arithmetica integra* de 1 544 des *frakturs* dans un texte romain. De même trouve-t-on dans la traduction française de *L'Arithmétique de Nicolas Tartaglia* (1613) des lettres de civilité utilisées comme symboles mathématiques. Il serait bon de ne pas perdre ces informations sur de tels choix d'auteur ou d'éditeur-imprimeur. La solution semble devoir être celle utilisée aujourd'hui par les mathématiciens, qui ont réussi à faire incorporer dans le bloc U+1D400 d'Unicode tout une série de *Symboles mathématiques alphanumériques*. Unicode précise même que ces caractères ne sont à utiliser « que dans le cas où le sens des variables mathématiques dépendrait de leur œil ». On y trouve par exemple des MINUSCULE MATHÉMATIQUE DE RONDE et MAJUSCULE MATHÉMATIQUE GOTHIQUE, utilisables dès à présent et qu'il faudra compléter par des typèmes de la Renaissance.

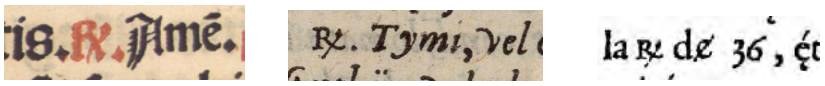
5. Une typologie des niveaux de transcription adaptée à l'imprimé

5.1. Un problème : les glyphes polysémiques

Étroitement attaché à la dimension matérielle de la typographie, le typème est antérieur à toute interprétation sémantique. Or un même typème peut posséder plusieurs significations et correspondre ainsi à plusieurs « caractères » au sens d'Unicode. C'est la raison pour laquelle la définition purement « linguistique » du caractère par Unicode ne recouvre pas celle, plus graphique, du typème. Remarquons cependant que le nommage des caractères par Unicode entretient, de ce point de vue, une certaine ambiguïté. De nombreux caractères Unicode sont, en fait, dotés d'une définition graphique : typiquement le caractère « ę » (utilisé dans les imprimés de la Renaissance pour « æ ») est dénommé « LETTRE MINUSCULE LATINE E OGONEK » ; ou le *titulus* qui se confond avec le tilde pour Unicode. Unicode présente des homographes, c'est-à-dire des « caractères » ayant le même glyphe. Mais comme ils ont des noms différents, ce ne sont pas les mêmes caractères. C'est par exemple le cas de LETTRE MAJUSCULE LATINE A et de LETTRE MAJUSCULE GRECQUE ALPHA qui ont tous deux pour glyphe « A ». Mais, les typographes le savent bien, certains types, gravés dans un but précis, peuvent être réutilisés, surtout s'il n'y a pas ambiguïté, dans d'autres documents avec un sens complètement différent.

Le typème ℟, utilisé dans les ouvrages liturgiques pour indiquer les répons, se rencontre également dans des recettes de préparations médicales, pour désigner l'impératif latin *recipe*, « prenez ». Il s'agit bien d'un seul et même typème polysémique, mais auquel correspondent deux numéros Unicode différents : U+211F (RÉPONS) et U+211E (ORDONNANCES). Et pour bien marquer cette distinction, Unicode invente deux glyphes dissemblables : ℞ (pour ORDONNANCES) et ℟ (pour RÉPONS) alors que, dans les imprimés, on utilise le même typème

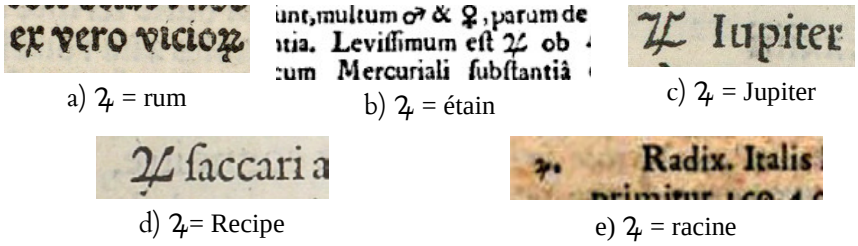
(figure 11). MUFI s'en inquiète d'ailleurs: « *It is questionable whether 211E PRESCRIPTION TAKE and 211F RESPONSE should be recognised as different characters* » (MUFI 2009). D'autres significations viendront s'ajouter à ce typème, vers 1550-1600, lorsque des mathématiciens recycleront ces caractères pour indiquer les racines cossiques chez des algébristes comme Peletier du Mans (Loget, 2012) ou pour signifier *resta* (comme chez Tartaglia), usages qui ne correspondent à aucun des deux caractères Unicode mentionnés (figure 11).



a) ℞ = Répons b) ℞ = Recipe c) ℞ = Racine

Figure 11. Trois significations différentes pour le même typème ℞ :
a) Missale mixtum, *Tolède, 1500 (Blois, BM)*
b) Pierre de Gorris, Formulae remediorum, *Paris, 1569 (Bourges, BM)*
c) Jacques Peletier du Mans, L'Algèbre, *Lyon, 1554 (Lyon, BM)*

Le même phénomène affecte le typème ♃ : utilisé le plus souvent comme caractère abréviatif latin *rum*, il sert également de symbole pour désigner la planète Jupiter ; à partir de 1550 et jusqu'au XIX[e] siècle, on l'emploie également dans des recettes de préparations médicales pour désigner l'impératif latin r*ecipe* « prenez » (fig. 12) ; on l'utilise dans la littérature alchimique pour symboliser l'étain ; enfin vers 1600 certains algébristes l'emploient pour « racine ». Unicode prévoit trois numéros différents correspondant aux trois premiers usages de ce type : U+A75C (LETTRE MAJUSCULE LATINE RUM ROTUNDA), U+2643 (JUPITER) et 1F729 (SYMBOLE ALCHIMIQUE POUR ÉTAIN). Pour le codage des traités de pharmacopée, seul le code U+211E (ORDONNANCES) semble adapté, mais on a vu que l'exemple de glyphe donné par Unicode tendait à rapprocher ce caractère du ℞.



a) ♃ = rum b) ♃ = étain c) ♃ = Jupiter

d) ♃= Recipe e) ♃ = racine

Figure 12. Cinq significations différentes pour le même typème ♃ :
a) Richard de Saint-Victor, Benjamin minor, *Paris, 1489 (Bourges, BM)*
b) Michael Ettmuller, Opera medica theorico practica, *Genève, 1736 (Gand, Univ.)*
c) J. Firmicus Maternus, Astronomicon, *Bâle, 1533 (Tours, CESR)*
d) Ambroise Paré, Oeuvres, *Paris, 1633 (Tours, CESR)*
e) Christophorus Clavius, Algebra, *Genève, 1609 (Lyon, BM)*

5.2. Un échelon supplémentaire : la transcription typémique

Le codage rationnel de la transcription d'un document imprimé ancien ne pose donc pas seulement la question des caractères *manquants* (qu'il convient d'intégrer à la grille MUFI), mais aussi celle des caractères *surnuméraires.* Pour rester fidèle à la matérialité du texte effectivement imprimé, il est indispensable de coder avec un même numéro toutes les occurrences d'un même typème. Ces considérations nous amènent à adapter la typologie des « niveaux de transcription » définis par Robinson et Solopova (cf. *supra*, § 2.2) en y ajoutant un échelon spécifique à la transcription du texte imprimé des XV^{e}-XVIII^{e} siècles. La chaîne de transcription prend ainsi une nouvelle forme (figure 11).

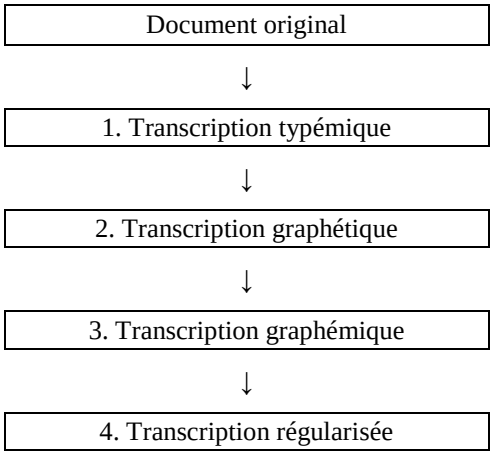


Figure 13. Chaîne de régularisation de la transcription pour l'imprimé ancien

Nous proposons donc d'asseoir l'intégralité de la chaîne de traitement sur un niveau élémentaire, celui de la transcription « typémique ». Celle-ci se distinguerait de la transcription dite « graphétique » telle que la définissent Robinson et Solopova, par l'utilisation d'un codage *univoque*, réduisant chaque typème à un seul et même numéro Unicode+MUFI. Le principe est simple : à une trace imprimée, un numéro ! Nous ne présupposons aucune balise du type <oe> ni aucune « entité » comme « œ ». Une transcription typémique pourra être lue, imprimée, affichée sur écran, à condition de disposer d'une fonte adaptée. Certes, à partir du moment où la transcription « graphétique » est obtenue par l'utilisation d'un codage intégrant un grand nombre d'abréviations, ligatures, etc., il n'y a qu'une différence minime entre les niveaux de transcription typémique et graphétique. Mais, seule l'existence de ce niveau de transcription « typémique » permet de coller au plus près de la réalité concrète du texte imprimé et de mettre en œuvre des recherches paléotypographiques susceptibles d'intéresser linguistes et historiens. L'existence de ce premier niveau de transcription permettra de résoudre de nombreux problèmes

rencontrés par les transcripteurs et fera gagner un temps considérable : ne nécessitant plus aucune part d'interprétation, la transcription pourra être confiée à une machine (logiciel de reconnaissance de caractères) en limitant les pertes d'informations.

Pour assurer la cohérence de ce codage univoque et pour permettre la mise en œuvre, au niveau graphétique, d'une distinction sémantique entre les caractères, nous proposons la mise en place d'une table d'équivalence associant à chaque typème un seul et unique numéro, tout en listant ses éventuelles variantes « sémantiques ». À ce niveau, nous proposons, par convention, d'attribuer le même code U+A75C à tous les typèmes « ꝓ », et le même code U+211F à tous les « ℟ ». La table d'équivalence proposera une définition élémentaire du typème, ainsi que l'inventaire des codes « sémantiques » correspondant, par exemple :

211F = R BARRÉ	**A75C = LETTRE MAJUSCULE LATINE RUM ROTUNDA**
~ U+211F RÉPONS	~ U+2643 JUPITER
~ U+211E ORDONNANCES	~ U+211E ORDONNANCES
~ U+221A RACINE CARRÉE	~ U+221A RACINE CARRÉE
	~ U+1F729 SYMBOLE ALCHIMIQUE ÉTAIN

5.3. En pratique : la gestion des accidents typographiques

La mise en œuvre d'une transcription typémique permet également de traiter des cas spécifiques au monde de l'imprimé et de rendre compte d'un certain nombre d'accidents typographiques qui, s'ils ne sont pas signalés à ce niveau, ne pourront plus l'être par la suite.

C*aractères illisibles ou non-reconnus.* Un caractère peut ne pas être reconnaissable car il est abîmé ou mal encré (figure 14). Il est toutefois possible de coder ce caractère par U+FFFD CARACTÈRE DE REMPLACEMENT, caractère doté d'un glyphe visible spécifique :  (c'est le caractère que devrait utiliser un logiciel d'OCR qui ne reconnaîtrait pas un type avec certitude). Nous proposons de rendre compte de cet accident au niveau de la transcription typémique en utilisant ce caractère , mais de restituer le bon caractère lors du passage à une transcription graphétique.

Caractères pied-en-haut. Il s'agit de types (en plomb) insérés dans la composition avec l'œil en bas (ou le pied en haut !). Il peut s'agir d'un simple accident de composition, mais ce procédé peut également être utilisé volontairement (et provisoirement) par le compositeur, au moment du tirage des épreuves, pour signaler la place d'un caractère manquant dans sa casse (Jimenes, 2012). Ce n'est alors plus l'œil du caractère qui est imprimé, mais la face inférieure du type, qui revêt le plus souvent l'aspect d'un rectangle noir, coupé en son milieu par une épaisse bande blanche (figure 15). On définira donc un numéro de code spécifique CARACTÈRE PIED-EN-HAUT que l'on utilisera au niveau 1 mais auquel on pourra substituer le caractère *ad hoc* au niveau 2.

Caractères renversés. On rencontre parfois des caractères qui n'ont pas été composés dans le bon sens ; il y a en effet quatre possibilités de mettre un caractère dans le composteur, par exemple F⊓dE (les deux formes couchées posant toutefois un problème de parangonnage). Le procédé peut-être accidentel ou volontaire (figure 16). Il conviendra donc de prévoir un code spécifique pour un caractère sans glyphe signifiant « ERREUR DE POSITION » (comme ceux de la série U+202A). Là encore, le caractère sera traité au niveau 2, selon les cas appropriés.

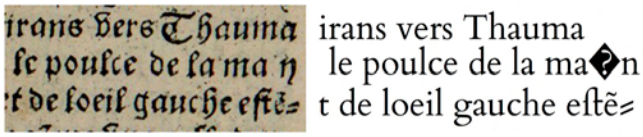


Figure 14. Codage typémique à l'aide du caractère U+FFFD, d'un caractère illisible (le i de main). Fr. Rabelais, Pantagruel, *Lyon, François Juste, 1542, f. 82 (Châteauroux, médiathèque Equinoxe)*

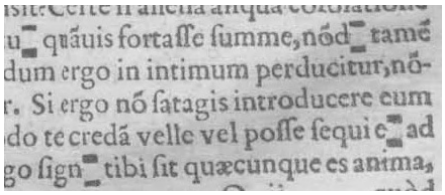
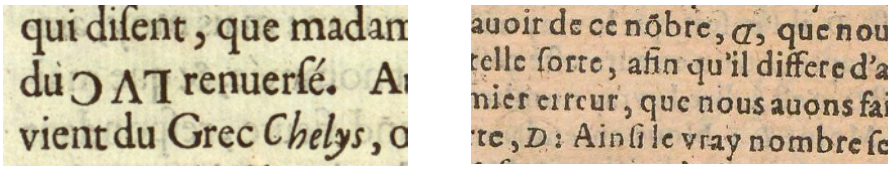


Figure 15. Types retournés pied-en-haut indiquant un caractère manquant (ici le caractère « ũ ») dans les épreuves de S. Bernard, Opera, *Paris, Ch. Guillard, 1551 (Lyon, BM). Extrait de Jimenes, 2012*



a) b)

Figure 16. Lettres volontairement renversées. a) Discours non plus mélancoliques que divers, *Poitiers, Marnef, 1557, p. 97 (Poitiers, BM) ; b) Tartaglia,* L'arithmétique, *Paris, A. Perier, 1613, p. 114 (Zürich, ZB)*

6. Conclusion : un projet d'inventaire des caractères anciens (PICA)

On voit l'importance de disposer d'un codage aussi complet que possible pour la production de transcriptions aux niveaux typémique et graphétique. Le codage

MUFI intègre quelques-uns des typèmes manquant à Unicode. Des essais de transcription de textes imprimés français de la Renaissance montrent toutefois qu'en son état actuel, le codage MUFI n'est pas encore suffisant pour couvrir l'intégralité de nos besoins.

L'étude systématique des caractères gravés par Garamont (André 2011) indique par exemple l'absence dans MUFI de certains caractères composés (voyelles accentuées, lettres avec apostrophes suscrites, abréviations latines), de certaines ligatures, de signes divers comme l'hyphen, les crochets bouclés (comme ceux de Dolet) ou l'obèle originale en forme de broche (figure 17). De même, l'examen d'ouvrages de Peletier du Mans (tels que *L'Algèbre*, Lyon, 1554) ou de Baïf (*Étrénes de poézie fransoeze*, Paris, 1574), a déjà permis d'identifier des caractères qui pourront être introduits dans MUFI[23] (figures 18 et 19). Il en est de même pour certains ouvrages mathématiques en italien (figure 20).

Lettres composées	æ̂ ĉ e̊ ę m̃ q̃ r̃ t̃
Ligatures	ꝭ Qu qv ſb
Abréviation latine	p̃ q́ q̃ q̃ j̃ j̃ q̃ q̃ s̃ x̃
Apostrophes ligaturées	à è m̀ ǹ ò œ̀ s̀ ù v̀ ỳ d̀ h̀ ſ̀ f̀ t̀
Parenthèses, tirets, etc.	(cc ⸗ ⸗ ⸗

Figure 17. Quelques caractères gravés par Garamont et manquants dans MUFI

L'ABÇ, DU LANGAJE FRANSOÉS.

Aa. Bb. Çç. Dd. Ee. Éé. Ęę. Ff. Gg. Jj. Hh. Ii. Kĸ.
Ll. Ļļ. Mm. Nn. Ņņ. Oo. Ww. Pp. Ꝺꝺ. Rr. Ss.
Tt. Uu. Vv. Zz. Ꝛꝛ.

S'anſuivet les noms e valers deſ leſtres nuveles.

Figure 18. Caractères français du XVI^e^ siècle, absents de MUFI : les lettres phonétiques de Baïf (1574)

Çanſæ:ꝯ, Cubæ:ꝯꝯ,

Figure 19. Caractères français du XVI^e^ siècle absents de MUFI : les caractères « cossiques » de Peletier du Mans (1554)

23. Certains de ces caractères ont déjà été signalés par M. Michaud et sont en attente dans le pipeline de MUFI.

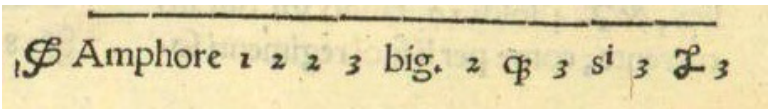


Figure 20. Cet extrait du General trattato di numeri et misure *de Tartaglia (Curtio Troiano, 1560) montre à gauche une abréviation pour* summa *(total) et à droite un L cursif pour* lire. *Aucun de ces caractères n'est dans MUFI*

Mais il convient de systématiser ces relevés. Nous proposons donc la mise en place d'un *Projet d'inventaire des caractères anciens* (PICA), associé au programme de numérisation des Bibliothèques Virtuelles Humanistes, conduit à Tours sous la direction de Marie-Luce Demonet[24]. Les structures institutionnelles et financières d'un tel projet sont en cours de définition. Le cadre scientifique de travail peut en revanche déjà être décrit.

1. *Recensement*. Il s'agit dans un premier temps d'identifier les besoins, en répertoriant les typèmes manquants. On dépouillera pour ce faire quelques-uns des ouvrages de référence fournissant des inventaires de caractères typographiques, tels ceux d'Anatole Claudin (1900) ou d'Hendrik D. L. Vervliet (2008, 2010). On pourra également dépouiller quelques précieux documents d'archives, comme les inventaires de police de l'imprimerie de Christophe Plantin conservés au Musée Plantin-Moretus d'Anvers, qui fournissent l'intégralité des types constituant une même police (figure 21). Les matériels originaux conservés (poinçons, matrices) devront également être listés. Ces premiers dépouillements pourront être complétés par l'examen minutieux des ouvrages originaux. L'association du projet PICA à un programme de numérisation comme celui des *Bibliothèques Virtuelles Humanistes* constituera à cet égard un précieux atout. On se concentrera dans un premier temps sur des ouvrages susceptibles de présenter quelques typèmes rares : traités d'algèbre ou de pharmacopée, ouvrages scientifiques ou alchimiques, almanachs, etc. Cet inventaire des caractères anciens ne devra jamais être clos, chaque transcription étant susceptible de porter à notre connaissance un nouveau typème. Il convient d'associer à chaque typème un nom explicite, fondé non sur une définition sémantique, mais sur une définition graphique (R BARRÉ, LIGATURE FF, etc.), et de renvoyer à ses différents équivalents sémantiques répertoriés par Unicode et MUFI.

24. http://www.bvh.univ-tours.fr/ - Le Projet PICA s'inscrit dans le droit fil des recherches menées sur les matériels typographiques par les BVH depuis plusieurs années. Il complétera les travaux d'identification des matériels typographiques développés dans le cadre de la Base de Typographie de la Renaissance (Jimenes, 2013) et bénéficiera des outils informatiques, notamment les logiciels Agora (analyse de mise en page) et Rétro (clustering et transcription de caractères) développés par le Laboratoire d'informatique de l'Université de Tours (équipe RFAI), sous la direction de Jean-Yves Ramel (PaRADIIT, 2013).

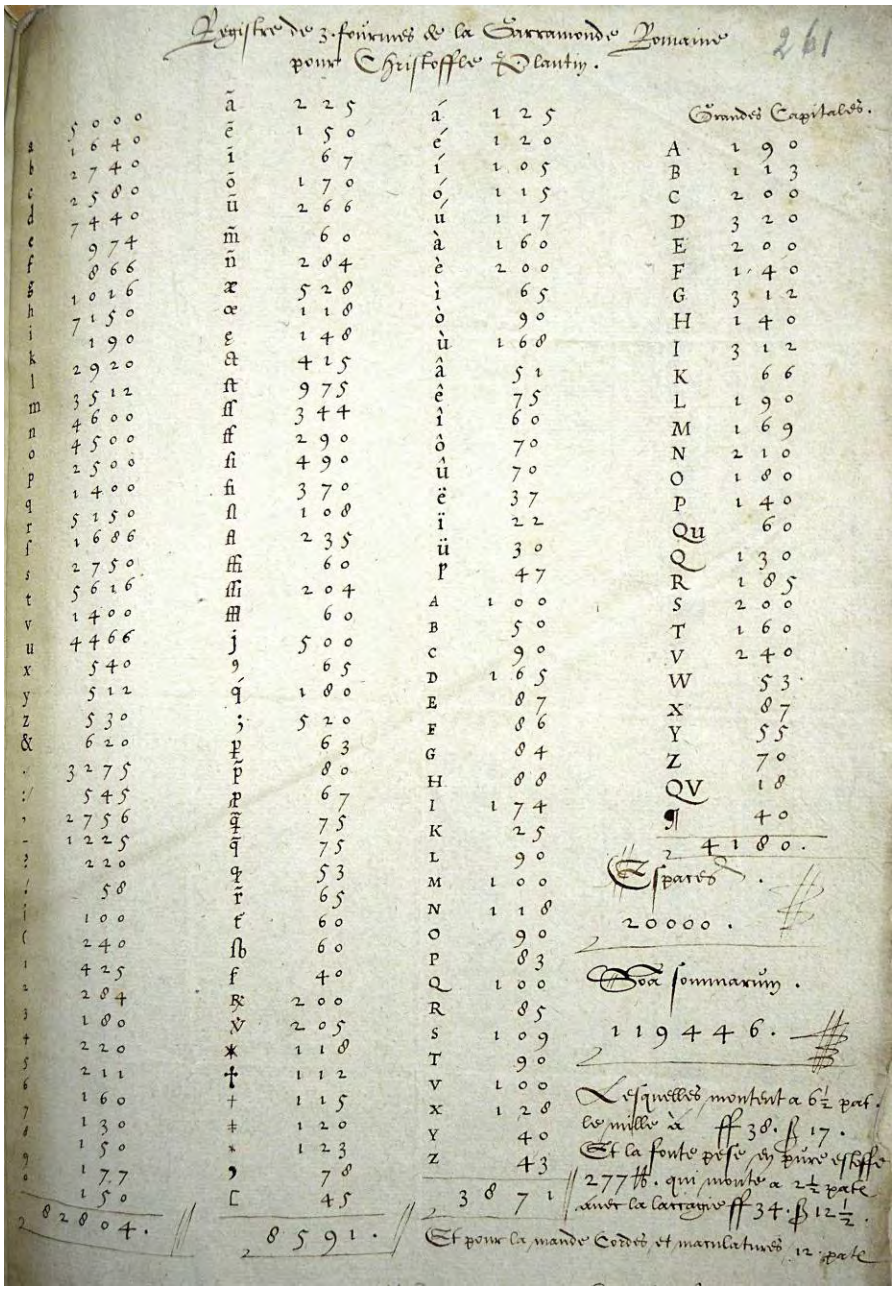


Figure 21. Registre de 3 fourmes de la Garramonde Romaine pour Christoffle Plantin, *inventoriant le nombre de types constituant une fonte. Chaque typème (ligatures, lettres accentuées, abréviations, etc.) y est reproduit à la main. (Anvers, musée Plantin-Moretus, Arch. 36)*

2. *Codage*. Il convient de s'assurer que chaque typème répertorié dispose d'un numéro de code normalisé. On cherchera donc à intégrer le pipeline de MUFI pour proposer au consortium Unicode la création officielle de nouveaux caractères, ou pour fixer, à défaut, un code normalisé au sein de la *Private Use Area*. On veillera enfin à produire une grille spécifique de codage typémique, préconisant pour chaque typème un point de code unique (tout en renvoyant à ses différents équivalents sémantiques possibles en Unicode pour permettre le passage d'un codage typémique à un codage sémantique).

3. Fontes et outils de saisie. Afin de permettre à la communauté scientifique la production de transcriptions typémiques, on s'attachera enfin à proposer aux chercheurs des polices de caractères libres de droit comprenant la totalité des typèmes répertoriés (il faudra d'ailleurs qu'une telle fonte soit disponible dès la constitution de cet inventaire). On s'attachera également à fournir aux utilisateurs différents outils destinés à faciliter la saisie : pilotes de clavier, tables de caractères, etc.

Enfin, cet inventaire ne sera adopté par la communauté des chercheurs que s'il est connu et reconnu (d'où l'importance de le concevoir comme un projet collaboratif).

Remerciements

Nous remercions, pour leurs relectures et leurs conseils, nos collègues Christine Bénévent, Lauranne Bertrand, Jorge Fins, Jean-Yves Ramel et Toshinori Uetani.

Sauf mention contraire, les figures proviennent des fac-similés numérisés dans le cadre du programme Bibliothèques Virtuelles Humanistes (http://www.bvh.univ-tours.fr). Nous remercions M. Pierre Meulepas, qui a bien voulu nous accorder, au nom du musée Plantin-Moretus, l'autorisation de reproduire les figures 7 et 21.

Bibliographie

André J. (2014). *Les caractères latins et français présents dans Unicode et MUFI*. Projet PICA/CESR, Tours. http://www.bvh.univ-tours.fr/batyr/liste_unicode_mufi.pdf

André J. (2011, sous presse). Les typèmes de Garamont. À propos d'un projet de codage des caractères anciens, *Gens du livre & gens de lettres à la Renaissance. Actes du LIV^e^ colloque international d'études humanistes, Tours, 27 juin-1^er^ juillet 2011* (éd. Ch. Bénévent, I. Diu et Ch. Lastraioli), Brepols, Turnhout, p. 369-389 (à paraître).

André J. (2003a). Numérisation et codage des caractères de livres anciens. *Document numérique*, vol. 7, n° 3-4, p. 127-142.

André J. (2003b). The Cassetin Project – Towards an Inventory of Ancient Types and the Relate Standardised Encoding. *TUGboat* vol. 24, n° 3, p. 314-318. http://www.tug.org/TUGboat/tb24-3/andre.pdf

André J. et Hudrisier H., éds. (2002). *Unicode, écriture du monde ?*, numéro spécial de *Document numérique*, vol. 6, n° 3-4.

André J. et Laucou C. (2013). *Histoire de l'écriture tyographique – le XIX^e siècle*. Atelier Perrouseaux, Gap.

Andries P. (2013). *Unicode et ISO 10646 en français*, http://hapax.qc.ca/ListeDesNoms-7.0a.txt

Andries P. (2008). *Unicode en pratique*. Dunod, Paris.

Anis J. (1988). *Théories et descriptions*, avec la coll. de J.-L. Chiss & C. Puech. De Boeck, Bruxelles.

Bettens O. (2006-2008). *Jean-Antoine de Baïf (1532-1589), Œuvres en vers mesurés*, http://virga.org/baif/index.php?item=1

Bień J. (2009). Facilitating access to digitalized dictionaries in DjVu format. *Studia Kognitywne - Études Cognitives*, 9, p 161-170.

Bowers F. (1949, rééd. 1994). *Principles of Bibliographical Description*. Princeton University Press, Princeton.

Bradshaw H. (1889). *Collected Papers*. Cambridge University Press, Cambridge.

Bradshaw H. (1870). *A Classified Index of the Fifteenth Century Books in the collection of M. J. De Meyer*. MacMillan, Londres (repris dans Bradshaw, 1889).

Catach N. (1997). Orthographe de la Renaissance : perspectives d'ensemble. *L'Information grammaticale*, n° 74, 1997, p. 32-38.

Catach N. (1988). *Pour une théorie de la langue écrite*, éditions du CNRS, Paris.

Catach N. (1983). La Graphie en tant qu'indice de bibliographie matérielle. *La Bibliographie matérielle* (éd. R. Laufer), éditions du CNRS, Paris, p. 122-123.

Catach N. (1968). *L'orthographe française à l'époque de la Renaissance.* Droz, Genève.

Claudin A. (1900). *Histoire de l'imprimerie en France au XV^e et au XVI^e siècle*. Imprimerie nationale, Paris, 4 vol.

Emiliano A. et Pedro S. (2013). *The Portuguese Medieval Font Project and the Medieval Unicode Font Initiative*. Centro de Linguística da Universidade Nova de Lisboa, http://www.hit.uib.no/mufi/portuguese/TM_Unicode.pdf.

Fournier P.-S. (1764). *Manuel typographique utile aux gens de lettres*. Barbou, Paris, 1764-1766.

Hammarström G. (1964). Type et typème, graphe et graphème. *Studia neophilologica*, vol. XXXVI, n° 2, p. 332-340.

Haralambous Y. (2004). *Fontes et codages*. O'Reilly, Paris (plusieurs rééditions).

Haugen O.E. (2013). Transcribing and Editing Medieval Nordic Sources. *Revue Document numérique, Gestion informatisée des écritures anciennes*, vol 16 n° 3/2013, p. 97-111

Haugen O.E. (2011). Do we need all these characters? On the transcription and Encoding of Medieval Primary Sources. *Linguistica e filologia digitale: aspetti e progetti* (éd. Paola Cotticelli Kurras), Edizioni dell'Orso, Allessandria p. 101-120.

Jimenes R. (2013). La Base de Typographie de la Renaissance (BaTyR). Un outil pour l'histoire du livre. *Bulletin des Bibliothèques de France*, t. 58, n° 5, p. 18-22. En ligne : http://bbf.enssib.fr/consulter/bbf-2013-05-0018-004

Jimenes R. (2012). Pratiques d'atelier et corrections typographiques à Paris au XVI^e siècle : les œuvres de saint Bernard imprimées par Charlotte Guillard (1551). *Passeurs de texte, imprimeurs et libraires à l'âge de l'humanisme* (Ch. Bénévent, A. Charon, I. Diu et M. Vène éds.), École nationale des chartes, Paris, p. 205-228.

Legros A. (2010). *Montaigne manuscrit*. Garnier, Paris.

Loget F. (2012). Printers and Algebraists in mid-16th Century France. *Philosophica* 87, p. 85-116. http://logica.ugent.be/philosophica/fulltexts/87-3.pdf.

Moxon J. (1683). *Exercises on the whole art of printing*, reedited by Herbert Davis and Harry Carter, Dover, 1962.

MUFI (2009). *MUFI character recommendation. Characters in the official Unicode Standard and in the Private Use Area*, version 3.0. http://www.MUFI.info/specs/MUFI-CodeChart-3-0.pdf

MUFI (2001). *Medieval Unicode Font Initiative*. www.MUFI.info. Dernière mise à jour : juin 2013.

PaRADIIT (2013). *Pattern Redundancy Analysis for Document Image Indexation & Transcription*, https://sites.google.com/site/paradiitproject/

Pedraza-Gracia M., Sánchez Ibáñez J.A. et Larraz L.J. (1998). *El diseño de un tipo para la descripcion bibliographica*, Lemir 2, http://parnaseo.uv.es/Lemir/Revista/Revista2/DICE-HTM/DICE.HTM.

Robinson P. et Solopova E. (1993). Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. *Tne Canterbury Tales Project Occasional Papers*, vol. 5, Oxford, p. 19-52. http://canterburytalesproject.org/pubs/transguide-MI.pdf

Smith M. (2010). *Dictionnaire des abréviations françaises, XII-XVIII^e siècles*, en ligne : http://theleme.enc.sorbonne.fr/dico.php

Stutzmann D. (2013). Ontologie des formes et encodage des textes manuscrits médiévaux. *Revue Document numérique, Gestion informatisée des écritures anciennes*, vol 16 n° 3/2013, p. 81-95.

Stutzmann D. (2010). Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? *Codicology and Palaeography in the Digital Age 2*, Norderstedt, p. 247-277 (et p. 249-251 pour la typologie des transcriptions).

Unicode (2013). *The Unicode Standard : a Technical Introduction*. http://www.unicode.org/standard/principles.html.

Unicode (2013a). Unicode, section française, Index des noms de caractères Unicode. http://www.unicode.org/fr/charts/charindex.html

Unicode (1991). *Unicode 6.2.0*, http://www.unicode.org/versions/Unicode6.2.0/. Dernière mise à jour : juin 2013.

Vervliet H.D.L. (2010). *French Renaissance Printing Types – A Conspectus*. The Bibliographical Society and Oak Knoll Press, Londres and New Heaven.

Vervliet H.D.L. (2008). *The Palaeotypography of the French Renaissance : Selected Papers on Sixteenth-Century Typefaces*, Brill, Leiden, 2 vol.