Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage

Nicolas Mazziotta

Édition électronique

URL : <http://journals.openedition.org/corpus/1517>

DOI : 10.4000/corpus.1517

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 10 novembre 2008

ISSN : 1638-9808

Référence électronique

Nicolas Mazziotta, « Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage », *Corpus* [En ligne], 7 | 2008, mis en ligne le 13 novembre 2009, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/1517> ; DOI : <https://doi.org/10.4000/corpus.1517>

Traiter les abréviations du français médiéval Théorie de l'écriture et pratiques d'encodage

Nicolas MAZZIOTTA
Université de Liège

Introduction

La constitution de nouveaux corpus électroniques de textes français médiévaux¹ conduit parfois les éditeurs qui s'intéressent aux phénomènes graphiques² à « donner à lire » le texte au travers d'une édition imitative du manuscrit. Dans cet article, nous concentrons nos efforts sur les questions soulevées par le traitement informatique des abréviations dans ce type de corpus.

Pour le linguiste qui se sert d'une édition de texte ancien, la relation entre forme abrégée et forme non abrégée est fondamentale, notamment dans le cadre d'études lexicographiques ou scriptologiques ; il est alors essentiel que le lecteur mû par de telles questions puisse faire le départ entre les « données » du texte et les « lettres » suppléées par l'éditeur³. Les recommandations classiques actuelles rencontrent ces préoccupations. Il est ainsi préconisé de marquer les « lettres restituées par de l'italique » (voir par exemple Viellard &

1 Voir Kunstmann, 2000, pour le premier exposé critique sur une partie des ressources disponibles – toujours utile malgré son âge, en particulier d'un point de vue méthodologique. Voir en outre le site du *Consortium pour les Corpus de Français Médiéval* (CCFM, <http://ccfm.ens-lsh.fr/>), qui rassemble les entreprises d'édition ayant choisi de mutualiser leurs ressources.

2 Segmentation des mots, usage des majuscules, ponctuation, etc. Pour ne citer que les travaux les plus récents, voir notamment l'article de Suzan Baddeley et Liselotte Biedermann-Pasques, 2004 (segmentation), celui de Guillaume François, 2006 (majuscules), ainsi que les thèses sur la ponctuation d'Alexei Lavrentev (en préparation) et de Nicolas Mazziotta (2007).

3 Peu d'éditions problématisent la résolution des abréviations ; pour un exemple, voir Monfrin, 1974 : lxxviii-lxix.

Corpus n°7 « Constitution et exploitation des corpus d'ancien et de moyen français » (2008), 191-218

Guyotjeannin, 2001 : 31). D'un point de vue philologique, c'est donc d'abord la manière dont l'éditeur interprète les procédés abrégatifs qui est décrite. Cependant, cette approche fait peu de cas des procédés et de leur fonctionnement, que nous traiterons dans le détail.

Notre exposé sera divisé en trois parties. Dans un premier temps, nous présenterons la manière dont les abréviations médiévales sont traditionnellement décrites (1). Nous tâcherons ensuite, en partant d'abréviations attestées dans le corpus de chartes que nous éditons, de construire un modèle de description original (2). Nous mettrons ensuite à profit ce modèle pour évaluer les techniques d'encodage électronique proposées par la *Text Encoding Initiative* (3), qui font office de standard à l'heure actuelle.

1. Approches traditionnelles

Dans cette première section, nous commençons par résumer le classement que les historiens ont fait des différents types d'abréviations médiévales (1.1) ; nous abordons ensuite brièvement la description grammaticale des abréviations modernes (1.2).

1.1. Approche paléographique

Les abréviations présentes dans les manuscrits médiévaux ont longtemps intéressé les paléographes, qui en ont décrit le fonctionnement de manière empirique. Nous ne traiterons pas ici l'histoire du classement des abréviations dans les traités, qu'ils soient anciens ou modernes, mais nous nous centrerons plutôt sur la vue traditionnelle actuelle. Selon cette dernière, les abréviations employées dans les manuscrits écrits en latin et dans les écrits romans sont classées en cinq catégories (voir Stiennon, 1991 : 145sv.⁴). La première de ces catégories comprend les abréviations dites *par contraction*, lesquelles « consiste[nt] à prélever au sein d'un mot une ou plusieurs lettres, [et] s'accompagne[nt] d'habitude d'un signe d'abréviation placé au-dessus du mot ». Deuxièmement, il existe des

4 Voir par ailleurs Prou 1924 : 109-165, Bischoff 1985 : 169-186 et le dictionnaire d'Adriano Cappelli (1929).

abréviations dites *par suspension*, qui « affecte[nt] la fin du mot, privé ainsi d'une ou plusieurs lettres et muni d'un point, point-virgule ou d'un autre signe qui marque l'arrêt ». Les abréviations *par lettres suscrites*, ce qui « consiste à placer une lettre de petit module comme exposant ou coiffé à une ou plusieurs lettres principales », forment une catégorie à part. Les abréviations *par signes spéciaux* sont des « signes d'abréviation de valeur précise ou déterminée ». Enfin, les abréviations *par graphismes* sont des « symboles représentant un mot ».

Adriano Cappelli, dans l'introduction de son dictionnaire des abréviations latines et italiennes (1929), propose par ailleurs de distinguer, parmi les signes spéciaux, ceux dont la résolution est contextuelle (comme c'est le cas du tilde dans les manuscrits français, qui peut être résolu en <n>⁵, <ue>, <st>, etc. suivant le contexte) et ceux dont la résolution est indépendante du contexte (comme <9>, qui est toujours résolu par <com>⁶).

Ce classement paléographique, très concret, permet déjà de distinguer efficacement différents types de procédés, mais ne replace pas ces abréviations dans le cadre de la linguistique de l'écrit. Pour qu'il puisse être opérationnalisé, le classement devrait se baser sur une série de critères descriptifs réduits et formalisés.

1.2. Approche grammaticale

Sauf erreur de notre part, il n'y a pas eu jusqu'à présent de description linguistique du système abréviatif médiéval. D'autre part, le champ des abréviations du français moderne a également été assez peu exploré dans le cadre du système graphique. On en trouve néanmoins une description dans les différentes éditions du *Bon usage* (voir, pour la dernière édition, Grevisse & Goosse, 2008 : §§ 111-113). L'approche y est comparable à celle des paléographes, si ce n'est que la notion de norme y est introduite.

⁵ Nous transcrivons les unités discrètes du système graphique – que nous nommons *grammèmes* ci-dessous (voir 2.1 c) – entre simples chevrons.

⁶ Voir cependant 2.3 ci-dessous.

Ainsi, André Goosse, après avoir distingué les abréviations dans les notes « qu'on prend pour soi » (non décrites), caractérisées par la liberté individuelle, et celles « qui n'apparaissent pas dans de véritables phrases », énumère trois procédés d'abréviation : 1/ le mot est réduit à son début, suivi d'un point (si la forme abrégée n'est pas réduite à une lettre, elle doit se terminer par une consonne précédant une voyelle dans la forme non abrégée du mot) ; 2/ le mot est réduit à son début et à sa fin (généralement placée en exposant dans un caractère plus petit) ; 3/ le mot est réduit à la lettre initiale à laquelle est adjointe une lettre intérieure.

Le classement est donc plus *grammatical* que *linguistique*. En conséquence, il convient peu à notre objet, principalement constitué d'unités à décrire.

2. Description systématique des abréviations

Dans cette section, nous reprenons progressivement la description des unités graphiques que la tradition a rangées sous l'appellation commune d'*abréviation*. Nous partons de données concrètes (reproduction de manuscrits) pour analyser ces unités de manière systématique et interne à l'espace graphique, sans référence à la manière dont le texte est oralisé. Cette position correspond à une vision *autonomiste* de la langue écrite, selon laquelle il est possible de décrire le fonctionnement du système graphique indépendamment de la langue orale⁷.

Nous partirons toujours de faits attestés dans le corpus que nous étudions et préparons dans le cadre de *Khartès*⁸, projet d'édition et d'étude linguistique de chartes originales rédigées en français en Wallonie avant 1300 (dir. Marie-Guy Boutier, Université de Liège), pour en abstraire une modélisation plus générale. La méthode de classement, inductive et empirique, conduira à mobiliser des concepts fondamentaux de description des unités de la langue écrite. Il nous paraît essentiel que cet effort de conceptualisation soit soutenu par une terminologie homogène, c'est pourquoi nous ferons usage de celle que nous

⁷ Voir Anis *et al.*, 1988 : 77, qui situe cette position par rapport à d'autres.

⁸ Voir Mazziotta, 2004 pour une brève présentation.

avons développée dans notre thèse (Mazziotta, 2007 : ch. 2), à laquelle nous renvoyons le lecteur intéressé par une approche plus complète des unités de la langue écrite. Cette terminologie n'étant pas courante, nous prendrons la peine de la présenter progressivement.

La présente section est subdivisée en deux parties. Dans la première, nous décrivons les procédés abrégatifs. Dans la deuxième, nous traitons de la relation entre ces procédés et ce que la tradition désigne généralement sous le terme de *résolution* des abréviations.

2.1. Procédés abrégatifs

Les abréviations sont des unités particulières à l'intérieur du système graphique, qu'elles contribuent à construire. Elles fonctionnent avec les autres unités dans une relation paradigmatique spécifique (2.1.1). Cette relation est parfois complexe et implique, au plan syntagmatique, des interférences entre unités de fonctionnement différent (2.1.2). Par ailleurs, au delà de la nature des unités discrètes, la manière dont ces dernières sont positionnées contribue également au fonctionnement des abréviations (2.1.3).

2.1.1. L'abréviation dans le plurisystème graphique

Partons d'une approche intuitive (a), qui nous permettra de mettre en évidence la relation de synonymie qui lie les formes abrégées à celles qui ne le sont pas (b), avant de systématiser nos découvertes (c).

a) Approche intuitive. Amorçons notre modélisation à partir d'un cas concret, fréquent et relativement simple. L'illustration ci-dessous comporte un signe que les lecteurs habitués à la fréquentation des manuscrits reconnaîtront aisément comme « une abréviation médiévale mise pour *et* »⁹ :

⁹ Document, 1247-06 : 3. Il faut lire le texte représenté comme suit : *les cens* [≈ « revenu fonciers »] *et lez chapons* [= « coq castré »] ; nous négligeons ici toute description graphique pour éviter de compliquer inutilement l'exposé.

Il s'agit du signe qui ressemble à un <z> barré¹⁰, que nous transcrivons <ꝛ>. En reprenant la terminologie présentée par Jacques Stiennon (voir 1.1 ci-dessus), nous qualifierions donc ce signe d'*abréviation par graphisme*. En effet, il s'agit bel et bien d'un « symbole représentant un mot ». Vu de manière traditionnelle, il y a une relation de représentation qui unit le graphisme présent à un mot absent. Si l'on considère que le signe se définit par l'association d'une forme d'expression à une forme de contenu, cette dernière est un autre signe.

b) *Synonymie et plurisystème*. Si nous adoptions cette vision des choses, nous pourrions dire que l'abréviation est une sorte de *métasigne* au sein du système de la langue écrite : une abréviation est *mise pour* d'autres signes qu'elle remplace et que l'éditeur restitue. Cependant, cette conception est issue de l'approche philologique ; dans la visée descriptive qui est la nôtre, il ne s'agit pas d'appréhender les opérations qui nous permettent de comprendre le texte (en l'occurrence, le remplacement d'unités par d'autres), mais de décrire le fonctionnement d'une partie de l'espace graphique. De ce point de vue, tout ce que nous pouvons dire est que chaque abréviation intègre un paradigme et peut commuter sans altération sémantique (mais peut-être bien avec une altération sur le plan énonciatif ou pragmatique) avec des groupes d'unités qui ne fonctionnent pas de la même manière (les « lettres »). Il existe une sorte de choix entre deux formes d'expression de contenu similaire, qui peut être lié à des contraintes extralinguistiques. Ainsi, certains scribes abrègent systématiquement les mots de la liturgie par respect ou crainte (Stiennon, 1991 : 145-146). La différence est similaire à celle qui existe entre les deux formes de la première personne de l'indicatif présent de *pouvoir* (*peux* et *puis*), qui sont employées

10 Plus exactement, ce signe ressemble à un <z> moderne barré : la forme du <z> dans les mss du corpus est généralement différente.

dans des situations différentes, la seconde relevant d'une langue plus affectée. Si l'on reconnaît un signe distinct pour chacune de ces deux formes, il nous semble que les formes «et» et «ɛ» doivent être considérées comme deux signes indépendants, liés par une relation de synonymie.

D'un point de vue spatial, les unités qui composent «et» et celles qui composent «ɛ» sont très différentes. Les premières, «e» et «t», ont deux caractéristiques : 1/ elles ne sont pas intrinsèquement significatives ; 2/ elles sont concaténées suivant l'axe sur lequel se déploient et s'enchaînent toutes les unités graphiques significatives. Ainsi, «et» se décompose en «e» et «t», qui se combinent sur le même axe que celui où «et» peut lui-même s'articuler avec d'autres unités. Par contre, si tant est qu'on veuille décomposer le signe «ɛ», on sera forcé de constater que les unités qui le composent ne s'enchaînent pas sur ce même axe. Ces signes, composés d'unités qui s'enchaînent sur un axe différent de celui où se combinent les unités significatives, et entretenant une relation de synonymie avec d'autres unités composées suivant ce même axe, correspondent à une partie de ce que nous nommons *abréviations*¹¹. Il y a donc deux types d'unités distincts, qui fonctionnent ensemble dans un système au fonctionnement mixte qu'on peut nommer *plurisystème graphique*¹².

c) *Synthèse*. Synthétisons cette première approche. Nous donnons aux unités discrètes de la langue écrite le nom de *grammème*^{13,14}, classe que nous pouvons subdiviser en deux

11 Cette relation de synonymie est importante pour distinguer les abréviations des symboles intraduisibles en langue écrite (par exemple, les manicules) et pour dissocier le problème des abréviations de celui de la description de la ponctuation (voir les références citées dans la note) et celui des écritures dites *idéographiques*.

12 Nous reprenons cette notion à Nina Catach (notamment 1988b : 248 sv.), sans toutefois la suivre dans son modèle d'analyse de l'écriture, trop fortement lié à l'oralité pour convenir à notre objet.

13 Nous discutons ce terme plus précisément dans notre thèse de doctorat, où nous définissons les *grammèmes* de manière négative par rapport à l'organisation spatiale, que nous nommons *topème* (voir Mazziotta, 2007 : ch 2, en particulier 39-41). Il nous a paru judicieux d'éviter le terme *graphème*, dont la polysémie est trop importante (voir à ce sujet

sous-classes : 1/ celle des *caténogrammes*, comprenant les grammèmes non signifiants organisés suivant le même axe que les unités signifiantes, en *chaîne* ; 2/ celle des *logogrammes*, grammèmes signifiants dont les parties constitutives ne sont pas forcément organisées suivant le même axe que celui où s'enchaînent les unités signifiantes. Schématiquement¹⁵ :

La relation d'*abréviation* lie un caténogramme, significatif, à un logogramme synonyme. Nous pouvons qualifier le rapport paradigmatique par le verbe *porter*. Nous dirons que la *portée* des logogrammes impliqués dans un mécanisme d'abréviation consiste en un caténogramme. Schématiquement,

Il est important de noter que la notion d'*abréviation* devient dès lors relationnelle : une abréviation s'envisage toujours *par rapport* à une autre unité qui n'est pas abrégée. D'autre part, la relation de *portée* est potentielle et non obligatoire¹⁶.

Pellat 1988) et mène parfois à la confusion concernant l'approche choisie (voir la référence citée dans la note 6, p. 194).

14 Nous forgeons également l'adjectif *grammémique* et procédons similairement pour les termes qui suivent dans notre exposé.

15 Ce diagramme de classes simple est dessiné en suivant la norme UML (voir Hamilton & Miles, 2006 pour une introduction à la version 2.0 du standard). Les flèches dont la pointe est blanche expriment une relation hiérarchique qui se lit « est un » (*op. cit.* : § 5.1.2).

16 La flèche, indiquant une relation orientée dans le diagramme ci-dessus, ne comporte pas d'information concernant la cardinalité de la relation (Hamilton & Miles, 2006 : § 5.1.5).

Pour que cette première synthèse soit tout à fait complète, nous devrions donner un nom aux parties constitutives des caténogrammes et des logogrammes. Nous aborderons le sujet dans la section ci-dessous.

2.1.2. Interférences entre les unités du plurisystème

a) Approche intuitive. Les « lettres » peuvent se mêler à des formants qui, sans être signifiants, ne s'organisent pas non plus nécessairement sur le même axe. Prenons à présent le cas suivant¹⁷ :

Dans cette reproduction, un tilde surplombe directement un <e>. En d'autres termes, il est commun que les éléments du plurisystème se mélangent. Nous pouvons raisonnablement penser que le <e> et le tilde forment ici un signe à part entière, lui-même en relation de synonymie avec <est>. Selon la définition proposée ci-dessus, ce signe est un logogramme. Sur le plan de l'expression, on retrouve un élément <e> commun aux deux signes et un segment <~> qui commute avec <st> sans modification de la forme de contenu. De ces deux dernières formes d'expression, seule la seconde n'emploie que des « lettres » en présence. Nous pouvons observer un cas similaire dans la reproduction ci-dessous¹⁸ :

le prénom d'un des protagonistes de la charte citée est représenté par une « lettre » seule entourée de points, que le lecteur rapprochera, au vu du reste du document, du prénom

17 Document, 1285-10-24 : 10. Il faut lire : *ki desor* [= « ci-dessus »] *est nomee*.

18 Document, 1244-01-19 : 13. Lire : *mon saignor Oston*.

Oston. Au plan du contenu, la relation qui unit ce « O » à la suite de caractères «Oston» est également d'ordre synonymique.

Il existe une différence importante entre la relation de «et» à son abréviation et celle d'«Oston» ou «est» à la leur : dans ces deux derniers cas, une partie de l'abréviation se retrouve dans la forme qui n'est pas abrégée : une « lettre », respectivement «O» et «e». Ainsi, ces deux exemples montrent que des unités de nature différente peuvent se combiner à un niveau non signifiant pour construire ensemble des unités qui font sens. Nous verrons plus loin (2.2.1) que cette différence peut être décrite en termes de *portée*.

b) Superpositions. La combinaison d'unités de nature différente peut être compliquée par le fait que ces unités peuvent, accidentellement¹⁹ ou non, se superposer et se recouvrir les unes les autres. Ainsi, dans la reproduction²⁰,

le troisième mot commence par un «p» dont la hampe est barrée. Cette organisation de traits, qui commute en l'occurrence avec «par», ne peut être décrite comme nous venons de le faire que parce que, d'une part, nous avons conscience de la *dynamique* impliquée dans le tracé et que, d'autre part, nous *reconnaissons* «p». Ces connaissances qui permettent l'analyse nous paraissent suffisantes pour accepter qu'il ne faut pas considérer ce «p» barré comme une unité indécomposable – comme le sont le signe «ɾ» ou le tilde. Il y a donc une commutation possible entre la barre et la chaîne de linéogrammes «ar».

c) Synthèse. Les points *a* et *b* ci-dessus mettent en évidence les lacunes du schéma de la figure 2, page 8 : pour synthétiser les procédés abrégatifs, nous avons besoin de pouvoir décrire les unités graphiques non signifiantes qui

¹⁹ En règle générale, nous ne traiterons pas des « accidents » de l'écriture manuscrite. Voir cependant 2.1.3.

²⁰ Document, 1247-06 : 9. Lire : *ki sunt pardevant dites*.

forment les grammèmes. Il est évident que la séparation que nous avons posée dans le schéma, qui permet de discerner les *logogrammes* des *caténogrammes*, trouve un écho au niveau purement distinctif. Nous proposons de poser tout d'abord une distinction entre les grammèmes significatifs, que nous nommerons *plérégrammes*²¹ (cette classe comprenant les logogrammes et les caténogrammes), et les grammèmes uniquement *distinctifs*, que nous nommerons *cénégrammes*. Soit, schématiquement :

Cette nouvelle séparation mène à une autre : les cénégrammes qui se combinent sur le même axe que celui où s'enchaînent les caténogrammes (comme «a», «e», etc.) sont distincts de ceux qui obéissent à des règles d'organisation spatiale qui ne peuvent être réduites à une concaténation sur un axe unique (comme le tilde). Nous nommerons les premiers *linéogrammes* et les seconds *périgrammes*. Nous pouvons donc revoir la hiérarchie des classes comme suit :

21 Cette terminologie est issue de Hjelmslev, 1954 : 48-49. Voir Mazziotta, 2007 : 41, note 60 pour une discussion.

La conceptualisation des mécanismes d'abréviation s'en trouve également légèrement modifiée. Nous pouvons regrouper les périgrammes et les logogrammes dans une même classe comprenant les procédés abrégatifs :

2.1.3. Altérations positionnelles

Les paragraphes qui précèdent recouvrent également les cas traditionnellement rassemblés dans la classe des abréviations par « lettres suscrites » (a), mais ces dernières posent un problème de modélisation (b).

a) Approche intuitive. Partons à nouveau de la réalité concrète : dans la reproduction ci-dessous²²,

la forme d'expression « conqⁱs » est en relation de synonymie avec « conquis », mais comporte moins de linéogrammes. À ce stade, nous pouvons dire que la première abrège la seconde. Ces

22 Document, 1265-04-15 : 11. Lire : *ki avoit conquis*.

modifications de taille et de position des linéogrammes renforcent la prégnance du procédé d'expression. Cependant, on ne peut pas dire qu'elles constituent automatiquement des marques d'abréviation. Ainsi, dans²³ :

la position du <r> n'est pas un procédé d'abréviation, puisque <propre>, synonyme de <p^ropre>, ne comporte pas plus de linéogrammes que ce dernier, qui n'est donc pas une abréviation.

b) *Problème de modélisation.* Ces deux exemples montrent que l'on ne peut pas véritablement distinguer les linéogrammes des périgrammes à l'aide d'une opposition intrinsèque. Les deux classes de cénégrammes sont proches. À strictement parler, on devrait plutôt distinguer un processus de combinaison périgrammique d'un processus linéogrammique plutôt que deux types de grammèmes différents.

Néanmoins, si nous comparons le <i> de <conqⁱs> à celui de <ki>, nous pouvons déceler un surcroît de valeur dans le premier cas, qui n'apparaît pas dans le second : précisément le phénomène d'abréviation. D'autre part, il nous semble raisonnable d'admettre que le premier <r> de <p^ropre> est *accidentellement* placé au-dessus des autres linéogrammes, parce qu'il a été omis par inadvertance. Le surcroît de valeur qu'il pourrait comporter est extralinguistique : il renseigne sur la dynamique de l'acte d'écriture. Il faut donc admettre deux notions : 1/ celle de perméabilité des classes (dans des conditions strictement déterminées par une valeur spécifique accordée au changement de classe) ; 2/ celle d'accident (difficilement formalisable, mais inévitable dans le cadre défini par l'écriture manuscrite).

23 Document, 1280-05-05 : 15. Lire : *propre*.

2.2. Portées

La description des procédés abrégatifs ne suffit pas à épuiser l'analyse. Il faut également expliquer comment la relation de synonymie s'établit entre les procédés et les linéogrammes avec lesquels ils commutent. Cette étude de la *portée* est ici menée en deux temps. Dans un premier temps, nous opposons les formes signifiantes entièrement abrégées et celles qui ne le sont que partiellement (2.2.1). Ensuite, nous distinguons les portées en fonction de la répartition, contiguë ou non, des linéogrammes commutant avec les procédés (2.2.2).

2.2.1. Portée globale vs portée partielle

Reprenons l'illustration 1, page 6. Nous avons vu qu'une combinaison de périgrammes était utilisée pour tracer un signe <□>, synonyme du caténogramme <et>. En observant la relation entre le plan du contenu et celui de l'expression, nous constatons que ces deux signes commutent et que la commutation n'est possible que si l'on considère dans les deux cas l'ensemble des deux mots. Les analyses respectives des formes d'expression <et> et <□> ne correspondent pas ; nous dirons que la portée du procédé abrégatif mobilisé est *globale*. Il n'est pas possible de faire commuter une partie du logogramme avec un des linéogrammes de la forme qu'il abrège. En pratique, une portée ne pourra donc être globale que si l'abréviation est un logogramme.

À l'inverse, si nous prenons la forme <guïers> dans la reproduction suivante²⁴,

<□> n'est pas un plérégramme, mais un cénégramme (il n'est pas intrinsèquement signifiant) ; il commute avec la chaîne <con> et s'intègre dans le reste du mot graphique. Dans ce cas, nous dirons que la portée est *partielle* : les procédés mobilisés dans la forme abrégée correspondent à une partie des linéogrammes de

24 Document, 1265-04-15 : 3. Lire : *Warniers, conviers* [= « frère convers »] et *maïestres*.

la forme non abrégée. Il en est de même pour « O » et « è » ci-dessus.

2.2.2. Portée continue vs portée discontinue

Lorsqu'elle est partielle, la portée d'un procédé n'est pas toujours une simple suite de linéogrammes contigus, même si cela est généralement le cas en ancien français. Par exemple, dans « uiers » que nous venons d'examiner, les linéogrammes remplacés par le tilde sont la chaîne « con », c'est-à-dire trois linéogrammes qui se suivent. Nous dirons dans ce cas que la portée du procédé abrégatif est *continue*.

Il existe néanmoins des cas où le procédé commute avec un ensemble de linéogrammes qui ne se suivent pas forcément. Par exemple, dans²⁵ :

le tilde commute avec le groupe formé par « eua » et « ie », la forme non abrégée étant « cheualier ». Ce cas s'oppose au précédent et nous parlerons de portée *discontinue*.

On pourrait objecter que le logogramme constitué à partir des linéogrammes et du tilde est une seule unité indécomposable, que les lecteurs sont capables d'interpréter comme un synonyme de « chevalier ». Ce logogramme, ainsi considéré comme un agrégat de procédés abrégatifs, aurait donc, à l'instar du « æ » étudié ci-dessus, une portée globale. Néanmoins, nous ne pouvons souscrire à une telle description, qui n'est pas compatible avec le modèle que nous proposons. Nous avons en effet reconnu l'existence d'unités purement distinctives (les linéogrammes) que nous retrouvons concaténées sous le tilde de la forme abrégée – « c », « h », etc. sont attestés dans d'autres plérégrammes. Postuler que les linéogrammes en question ne sont pas isolables quand ils sont mobilisés de la sorte est artificiel et sort du système que nous avons exposé. La question de la perception globale des mots, qui relève de la psychologie de la lecture, nous paraît peu

25 Document, 1260-02-29 : 3. Lire : *chevaliers*.

pertinente dans le cadre d'une description structurelle de la langue écrite.

2.3. *Synthèse*

Les phénomènes abrégatifs peuvent être représentés par le diagramme :

En accord avec ce schéma, la description d'une abréviation, pour être complète, doit :

1/ identifier les procédés logogrammiques, périgrammiques et positionnels – sans nécessairement préciser le niveau d'articulation (distinctif ou significatif) ;

2/ identifier la portée des procédés, c'est-à-dire la chaîne de linéogrammes avec laquelle ils commutent ;

3/ mettre en relation les procédés et les linéogrammes de manière précise, de façon à ce que chacun des termes de la relation présentée dans la figure 14 soit clairement identifiable.

D'autre part, nous avons vu ci-dessus qu'Adriano Cappelli distinguait les abréviations en fonction de la prédétermination ou non de leur portée²⁶. En pratique, la situation n'est jamais aussi simple : «g» peut très bien être résolu par «cum», «con», «cun», etc., en fonction des autres formes attestées dans le manuscrit. Même une abréviation « simple » comme «r» peut se révéler problématique (voir Monfrin, 1974 : lxix). Nous ne nous étendons pas sur le sujet, mais cela nous semble suffisant pour souligner que chaque portée d'abréviation doit être traitée de manière contextuelle (en discours) et non *a priori* (en langue).

3. Encoder les abréviations médiévales

Dans la mesure où l'éditeur juge nécessaire de traiter les abréviations médiévales, la manière dont il les encode permet

26 Voir 1.1 ci-dessus ; nous transposons sa description dans les termes que nous proposons.

plus ou moins bien de reconstituer l'ensemble des termes de la relation représentée dans la figure 14. Il est évident que les objectifs de l'édition (le public auquel elle est destinée) déterminent le degré de précision de l'encodage et la nécessité d'employer l'une ou l'autre des techniques que nous allons présenter. Dans cette section, nous adopterons le point de vue selon lequel nous recherchons à représenter le mieux possible chacun des termes de la figure 14. Les possibilités d'encodage « traditionnel » seront exposées en premier (3.1) ; ensuite, nous montrerons comment les nouvelles techniques d'encodage permettent de traiter le problème (3.2).

3.1. Traitement traditionnel courant

La philologie traditionnelle a longtemps été amenée à négliger ces phénomènes à cause des restrictions imposées par le *medium* (le papier), se limitant la plupart du temps à marquer les linéogrammes correspondant à la portée des abréviations par la typographie, voire à ne pas les marquer du tout – malgré les recommandations pratiques désormais en vigueur (Vielliard & Guyotjeannin, 2007 : 31).

Il en est de même pour un certain nombre de corpus informatisés d'ancien français – même ceux qui essayent de fournir une transcription diplomatique, comme Holtus *et al.*, 2003 ou les éditions en cours du projet *Khartès*, dont sont tirés nos exemples. Ainsi, nous transcrivons <conuïers> le plérégramme qui nous intéresse dans la figure 12.

La *Text Encoding Initiative* propose plusieurs alternatives d'encodage, dont une est proche de celle que nous venons d'observer (Burnard & Bauman, 2007 : 343) :

<expan><ex>con</ex>uïers</expan>

L'élément *expan*, constitué d'une balise ouvrante <expan>, d'une balise fermante </expan> et de l'ensemble des caractères entre ces deux balises, délimite un mot contenant des linéogrammes correspondant à une portée d'abréviation. Ces linéogrammes sont identifiés et délimités par l'élément *ex*.

Cette solution a l'avantage d'expliciter la distinction entre les données et l'annotation²⁷. Elle décrit l'ensemble comme un caténogramme commutant avec une forme abrégée et délimite les linéogrammes correspondant à la portée par un élément spécifique.

Dans tous ces cas²⁸, que le format soit traditionnel ou électronique, standard ou non, l'expression de la relation de la figure 14 est mise à mal. Ces éditions permettent en effet au lecteur de repérer les linéogrammes liés à une portée de procédés abrégatifs, mais en aucun cas elles ne permettent de retourner aux procédés mobilisés, et *a fortiori* de leur lier la portée en question. Nous représentons ces manques par la couleur grisée sur le schéma suivant :

Cette pratique ne permet pas non plus de comprendre, dans le cas où plusieurs portées relevant de groupes de procédés différentes se suivraient, quels procédés se combineraient pour une seule portée.

La tradition des éditions « diplomatiques » mimétiques de ce qui est observable dans le parchemin²⁹ – reproduisant les procédés par une typographie riche, comme nous l'avons fait plusieurs fois ci-dessus –, a exactement l'inconvénient inverse, limitant l'expression de la relation à celle des procédés :

27 Cette distinction, qui touche le problème de la relation entre les objets perçus et la manière dont ils sont conçus n'est pas aussi nette que l'on pourrait le penser et mériterait une étude détaillée qui sortirait du cadre de cet article.

28 Nous verrons ci-dessous que les recommandations de la TEI ne se limitent pas à cette possibilité.

29 Voir l'exposé de synthèse qu'Alexei Lavrentev en fera dans sa thèse en préparation.

Aucune des deux situations n'est satisfaisante dans l'optique d'une représentation complète.

3.2. Encodage synoptique

Traditionnellement, l'édition peut, plus rarement, être constituée d'une double transcription synoptique, alliant la reproduction « fidèle » du manuscrit à une édition critique. La première reproduit les abréviations, la seconde les résout en explicitant leur portée – voir par exemple l'édition des Serments de Strasbourg d'Albert Henry (1957 : 1-2) ou les pages éditées dans Careri *et al.*, 2001. Les procédés et les linéogrammes remplacés sont ainsi exprimés, mais la relation qui les unit doit encore être faite par l'humain. Voyons comment le *medium* électronique permet de traiter ce problème.

Les nouvelles techniques d'encodage, fondées sur XML, rapidement devenu un standard *de facto*, permettent, par ajout d'annotation, de faire apparaître simultanément des informations de nature différente au niveau de l'encodage des données. Cette expression peut se manifester soit en intégrant les différentes couches d'annotation directement dans les données, soit, par le truchement de pointeurs, dans un autre document qui se réfère à celui qu'il décrit. Pour désigner ces deux pratiques d'annotation, Sylvain Loiseau emploie respectivement les épithètes *embarquée* et *débarquée*. Il parle ainsi, par exemple, d'*annotation embarquée* (voir 2007 : 167-173 pour un exposé très clair). Nous faisons nôtre cette terminologie et abordons successivement ces deux démarches.

3.2.1. Formats embarqués

Les annotations XML embarquées, directement intégrées à l'édition électronique dans sa linéarité, permettent un enrichissement évident, mais sont intimement liées au format et à ses limitations. La situation varie en fonction du niveau de granularité du codage de la relation qui décrit le phénomène

d'abréviation : ce codage peut se faire au niveau plérégrammique ou au niveau cénégrammique.

a) *Niveau plérégrammique*. À l'heure actuelle, il est possible d'encoder conjointement procédés abrégatifs et portée de ces procédés. La dernière version des recommandations de la *Text Encoding Initiative* propose d'autres alternatives d'encodage des abréviations que celle exposée ci-dessus. Ainsi, l'élément exprimant une alternative, *choice*, est défini comme suit :

groups a number of alternative encodings for the
same point in a text (Burnard & Bauman 2007 :
800)

On encode alors conjointement une forme abrégée (élément *abbr*), c'est-à-dire l'ensemble constitué des procédés et des linéogrammes qui forment un logogramme, et le caténogramme qui y correspond (élément *expan*), soit (cf. Burnard & Bauman, 2007 : 343) :

```
<choice>  
  <abbr>ml&#x303;t</abbr>  
  <expan>moult</expan>  
</choice>
```

où le code `̃` est le code unicode correspondant à un tilde se combinant avec le linéogramme précédent³⁰. D'un point de vue plus général, cette technique se rapproche des techniques d'alignement qui mettent en correspondance différents niveaux d'élaboration éditoriale du manuscrit, de la plus imitative à la plus critique (Haugen, 2004 : § 3).

Cette pratique permet de traiter précisément les procédés et les linéogrammes commutant. Toutefois, elle n'objective pas clairement la relation entre les deux. Certes, les deux formes synonymes sont encodées conjointement et un humain est capable de mettre en relation les périgrammes de l'abréviation avec les linéogrammes correspondant dans la

30 Voir les tables fournies par le consortium, sur le site <http://www.unicode.org>. La question de la combinaison d'un « diacritique » à plusieurs linéogrammes n'est pas résolue.

forme non abrégée, mais le traitement logiciel de ce type d'encodage devient rapidement complexe, en particulier si plusieurs procédés correspondant à des portées distinctes sont présents dans une même forme. Schématiquement, l'encodage conjoint au niveau plérégrammique ne permet pas d'exprimer la portée :

Il est néanmoins vrai que les cas ambigus sont rares dans notre corpus ; les abréviations qu'on rencontre dans les manuscrits français sont généralement d'un niveau de complexité inférieur à celui de celles qu'on rencontre dans les manuscrits latins, où l'interprétation exacte de la portée de chaque procédé peut se révéler plus difficile, même pour un humain entraîné³¹. Par ailleurs, dans le cas d'une édition en cours, la répétition dans l'encodage des linéogrammes <m>, <l> et <t> peut mener à des erreurs. En effet, chaque correction de lecture de ces unités devant être reportée à deux endroits différents, la probabilité de faire une erreur dans l'une ou l'autre des transcriptions est forcément supérieure³².

Qu'elles permettent ou non d'identifier la portée, ces solutions d'encodage conjoint³³ ne fournissent pas de possibilité pour encoder les portées discontinues sans rompre la relation entre les procédés abrégatifs et leur portée.

b) Niveau cénégrammique. Une ébauche de solution a été proposée récemment³⁴ (cf. Burnard & Bauman, 2007 : 344). Elle consiste à « descendre » l'annotation à un niveau inférieur à celui des unités signifiantes et fait entrer en jeu deux éléments, *am* et *ex*, qui balisent respectivement les procédés abrégatifs et les linéogrammes correspondants :

31 Voir de nombreux exemples de cas complexes dans Cappelli 1967.

32 Voir à ce propos Mazziotta, 2004 : 797-798.

33 Similaire à celles prônées dans Haugen, 2004 : § 3.

34 Contrairement aux encodages présentés ci-dessus, la pratique qui suit n'apparaît pas dans les éditions antérieures des recommandations de la TEI.

```
m<choice>
  <am>&#x303;</am>
  <ex>ou</ex>
</choice>lt
```

À première vue, tous les éléments impliqués dans la relation qui nous intéresse sont exprimés et l'encodage répond ainsi à nos exigences. Les éléments `am` et `ex` ont le même parent, dont la sémantique est précisément assimilable à une possible substitution. Malheureusement, sa structure arborescente empêche de traiter efficacement les portées discontinues. Ainsi, face à la forme abrégée de «chevaliers» (examinée sous 2.2.2), on constate que cette méthode rend impossible l'expression d'un seul `am` correspondant à plusieurs `ex`, puisque le lien entre les deux éléments est exprimé par leur cointégration au sein de l'élément `choice`. Par ailleurs, les conventions définies par le consortium unicode obligent les « diacritiques » comme le tilde à se trouver directement à la suite du linéogramme sur lequel ils portent ; il est ici impossible de respecter cette convention.

Dans la continuité de ce qui précède, nous ne pouvons souscrire à la description qui fait de «p» un grammème indécomposable sur le plan technique et nous suivons les propositions de MUFI, qui le définit de manière compositionnelle:

LATIN SMALL LETTER P WITH STROKE
THROUGH DESCENDER (2006 : 71).

3.2.2. *Formats débarqués*

Au contraire des annotations embarquées, les annotations débarquées ne sont pas directement mêlées aux données qu'elles décrivent. Grâce à un mécanisme de *pointeur* (les éléments XML étant identifiés par leur *Uniform Resource Identifier*, ou URI, qui est une adresse unique qui permet de retrouver cet élément en question), l'annotation est reportée dans un document indépendant des données, auxquelles elle fait référence.

a) *Expression exacte de la portée et des linéogrammes.*

L'annotation débarquée permet d'atteindre le niveau d'expressivité nécessaire à la description précise de la portée. Notre proposition s'appuie sur le mécanisme de *pointeur* adapté

à la philosophie de la TEI. Une possibilité pour encoder les procédés à portée multiple serait d'expliciter la relation entre ces procédés et les différentes chaînes de linéogrammes qui leur sont liées. Les recommandations de la TEI permettent de noter cette information à l'aide de l'attribut `@corresp`, dont la valeur consiste en un lien vers un autre élément identifié. Ainsi, une manière d'encoder serait :

```
<choice>
  <abbr>
    chl<am xml:id='a1'>&#x303;</am>s
  </abbr>
  <expan>
    ch<ex corresp='#a1'>eua</ex>l
    <ex corresp='#a1'>ie</ex>r
  </expan>
</choice>
```

où nous voyons que l'identifiant du procédé (`@xml:id`) est mis en correspondance avec les linéogrammes qui commutent avec lui.

En utilisant les éléments de la TEI, le regroupement des procédés et de leur portée doit nécessairement se faire par le biais de l'élément exprimant une alternative. Plus contraignant encore, il est nécessaire de remonter l'encodage au niveau plérégrammique, ce qui mène au même type de problème que ceux que nous rencontrons ci-dessus (3.2.1 b).

En outre, si la sémantique de l'élément `choice` est relativement claire, celle de l'attribut `@corresp` est on ne peut plus floue :

`@corresp` (corresponds) points to elements that
correspond to the current element *in some way*.
(Burnard & Bauman 2007 : 543, nous
soulignons)

On peut se poser la question de la transparence de l'encodage. Comme la TEI ne propose pas une ontologie à strictement parler, la permissivité et la flexibilité extrême de certaines recommandations sont susceptibles de poser de grandes difficultés d'échange de données. Comment savoir si l'attribut `@corresp` a la même valeur d'un document à l'autre, voire d'une occurrence à l'autre ? Un humain aura tôt fait de

comprendre la relation, mais que dire d'un automate ? Il est de plus imaginable qu'une unité soit liée à plusieurs autres de différentes manières...

Il y a donc deux objectifs à poursuivre pour résoudre ces problèmes : 1/ redescendre l'encodage au niveau cénégrammique ; 2/ rendre univoque le lien entre les éléments. Nous proposons donc : 1/ de nous débarrasser simplement des éléments *choice*, *expan* et *abbr*, qui deviennent redondants du fait de la mise en relation explicite des éléments *am* et *ex* ; 2/ de créer un attribut spécialisé dans la mise en correspondance de ces deux éléments ; il apparaîtrait par exemple dans l'élément *ex*, porterait le nom *@am* et contiendrait une URI. En l'occurrence, les données seraient encodées comme suit :

```
ch
<ex am='#a1'>eua</ex>l<ex am='#a1'>ie</ex>
r<am xml:id='a1'>&#x303;</am>
s
```

b) Expression exacte des procédés. La description exacte des procédés peut, dans le cas de périgrammes ou de logogrammes spécifiques, être réduite à la simple utilisation d'entités caractères comportant un code unicode. Il existe en outre un mécanisme débarqué plus complet pour la description de nouveaux « caractères » (Burnard & Bauman, 2007 : ch. 5), qui consiste à définir le grammème quelque part – nous passons le détail de cette description – et à y faire référence par son URI, à l'intérieur d'un élément *g*. Par exemple, en imaginant que le tilde est défini ailleurs et non comme un caractère unicode :

```
ch
<ex am='#a1'>eua</ex>l<ex am='#a1'>ie</ex>
r<am xml:id='a1'><g ref='#titulus'></g></am>s
```

l'élément *g* (mis pour *gaiji*, mot japonais signifiant « caractère non standardisé », cf. Burnard et Bauman, 2007 : 910) contient une référence à une unité distincte (*#titulus*).

Pour que la description tienne compte de nos objectifs, il est absolument nécessaire de distinguer les grammèmes de leur position ; cela permettra non seulement de gérer adéquatement le cas des linéogrammes placés au-dessus des autres, mais aussi de décrire les combinaisons « diacritiques ».

Ainsi, la forme abrégée reproduite dans la figure 14, p. 15 comporte un linéogramme dont la position a été altérée, cette

position constituant le procédé d'abréviation. À notre connaissance, il n'y a à ce jour aucun moyen de décrire une position comme une unité distincte et nous voudrions souligner ce manque. Une possibilité pour contourner le problème serait de faire usage de l'élément `rendition`, défini comme suit par la TEI :

supplies information about the rendition or
appearance of one or more elements in the source
text (Burnard & Bauman, 2007 : 1081)

Cet élément, qui contiendrait la description du changement de position, serait ensuite intégré à l'élément `am` et ferait référence au linéogramme qu'il modifie par un mécanisme de pointeur (l'élément `c` entoure le linéogramme auquel il est fait référence) :

```
cong
<ex am="#a1">u</ex>
<c xml:id="#c1">i</c>
<am xml:id="a1">
  <rendition ref="c1" scheme="free">
    superscript
  </rendition>
</am>
```

De manière plus générale, il nous semble que le format XML atteint ici ses limites. L'expression des relations dans un modèle de données précisément conçu pour gérer ce type d'information, comme le *Resource Description Framework* (RDF, voir Becket, 2004 pour sa syntaxe XML), permettrait sans doute de s'affranchir de ces difficultés de manière plus efficace.

4. Conclusions

Nous avons suivi une progression relativement simple, qui nous a permis de voyager des pratiques traditionnelles aux pratiques modernes sans négliger la description théorique de notre objet. Au delà des propositions conceptuelles et pratiques que nous avons présentées dans cet article, nous espérons que notre exposé a mis en évidence l'importance de l'élaboration d'un modèle précis préalablement à la recherche d'un format d'encodage. Dans le champ par trop négligé du système

graphique des manuscrits anciens, il nous semble absolument nécessaire que de telles réflexions soient menées à tous les niveaux : ponctuation, forme des linéogrammes, mise en page, etc. Il y a là un travail important à fournir.

D'autre part, nous espérons avoir réussi à démontrer au lecteur que la modélisation ne peut se faire efficacement que sur une base foncièrement *inductive*, partant de données attestées dans un système particulier. Les principes dégagés pourront être appliqués de manière *déductive*, en tentant de les étendre à d'autres systèmes, dont nous savons par ailleurs qu'ils sont relativement proches.

Lorsqu'ils furent amenés à noter les textes vernaculaires, les copistes adoptèrent naturellement le système graphique du latin auquel ils étaient habitués, c'est-à-dire simultanément un alphabet et un code d'abréviations bien rodé et efficace, dont les principes remontaient [xxxi] à l'Antiquité et les outils avaient été perfectionnés par les scriptoria monastiques, avant que la pratique n'en soit bouleversée, au XIII^e siècle, par les méthodes de la scolastique. (Careri *et al.*, 2001 : xxx-xxxi)

Ainsi, dans le cas des abréviations, il est probable que notre approche servira également à traiter au moins en partie le latin et, *a fortiori*, d'autres langues romanes que l'ancien français.

Références bibliographiques

Documents

Tous les documents cités sont conservés aux Archives de l'État à Liège : Document (1244-01-19), 19 janvier, Cathédrale Saint-Lambert. – Document (1247-06), juin, Couvent de Robermont. – Document (1260-02-28), 28 février, Collégiale Saint-Jean-en-Isle. – Document (1265-04-15), 15 avril, Abbaye du Val-Benoît. – Document (1280-05-05), 5 mai, Abbaye du Val-Benoît. – Document (1285-10-24), 24 octobre, Abbaye du Val-Saint-Lambert.

Travaux

- Anis J., Chiss J.-L. & Puech Chr. (1988). *L'écriture. Théories et descriptions*. Bruxelles : De Boeck.
- Baddeley B. & Biedermann-Pasques L. (2004). « Histoire des systèmes graphiques du français à travers des manuscrits et des incunables (IX^e-XV^e siècle) : segmentation graphique et faits de langue », *Revue de linguistique romane* 68 :181-201.
- Becket D. (éd.) (2004). *RDF/XML Syntax Specification (Revised)*. W3C Recommendation 10 February 2004. [en ligne : <http://www.w3.org/TR/rdf-syntax-grammar>]
- Bischoff B. (1993). *Paléographie de l'Antiquité romane et du Moyen Âge occidental*, Paris : Picard.
- Burnard L. & Bauman S. (éds) (2007). *Guidelines for Electronic Text Encoding and Interchange*. [en ligne : <http://www.tei-c.org/release/doc/tei-p5-doc/html/index.html>].
- Cappelli A. (⁶1929, 1899). *Lexicon abbreviaturatum. Dizionario di abbreviature latine ed italiane [...]* Milano : Ulrichs Hoepli. [réimpression anastatique]
- Careri M., Fery-Hue F., Gasparri F., Hasenor G., Labory G., Lefèvre S., Leurquin A.-F. & Ruby C. (2001). *Album de manuscrits français du XIII^e siècle. Mise en page et mise en texte*. Roma : Viella.
- Catach N. (éd.) (1988a). *Pour une théorie de la langue écrite. Actes de la Table Ronde internationale C.N.R.S.-H.E.S.O. Paris, 23-24 octobre 1986*. Paris : C.N.R.S.
- Catach N. (1988b). « L'écriture en tant que plurisystème, ou théorie de L prime », in Catach, 1988a : 243-256.
- François G. (2006). « L'émergence de la majuscule dans *La chanson de Roland* », *Revue de linguistique romane* 70 : 41-52.
- Grevisse M. & Goosse A. (¹⁴2008). *Le bon usage : grammaire française*. Bruxelles : De Boeck & Larcier.
- Hamilton K. & Miles R. (2006). *Learning UML 2.0*. O'Reilly Media [édition électronique].

- Haugen O. E. (éd.) (2004). *The Menota handbook : Guidelines for the electronic encoding of Medieval Nordic primary sources. Version 1.1* [en ligne : <http://www.hit.uib.no/menota/guidelines>].
- Hjellslev L. (1954). « La stratification du langage ». In Hjellslev, 1971 : 45-77.
- Hjellslev, L. (1971). *Essais Linguistiques*. Paris : Minuit.
- Holtus G., Overbeck A. et Völker H. (2003). *Luxemburgische Skriptastudien. Edition und Untersuchung der altfranzösischen Urkunden Gräfin Ermesindes (1226-1247) und Graf Heinrichs V. (1247-1281) von Luxemburg*. Tübingen : Niemeyer.
- Kunstmann P. (2000). « Ancien et moyen français sur le Web : textes et bases de données », *Revue de linguistique romane* 64 : 17-42.
- Lavrentev A. (en préparation). *Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIII^e au XV^e siècle* [titre susceptible de modifications mineures]. Lyon : thèse de doctorat en cours de rédaction (dir. Christiane Marchello-Nizia).
- Loiseau S. (2007). « CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus* 6 : 153-186.
- Mazziotta N. (2004). « Le texte dans tous ses états. Philosophie d'encodage du projet *Khartès* », in G. Purnelle, C. Fairon & A. Dister (éds) *Le poids des mots*. Louvain : Presses universitaires de Louvain, 793-803.
- Mazziotta N. (2007). *Ponctuation et syntaxe dans la langue française médiévale. Étude d'un corpus de chartes originales écrites à Liège entre 1236 et 1291*. Liège : thèse de doctorat inédite (dir. Marie-Guy Boutier) [table des matières et résumé : <http://bictel.ulg.ac.be/ETD-db/collection/available/ULgetd-12032007-103637/>]
- Monfrin J. (1974). « Introduction », in J.-G. Gigot *Chartes en langue française antérieures 1271 conservées dans le département de la Haute-Marne*. Paris : C.N.R.S, xi-lxxx.

Traiter les abréviations du français médiéval

- MUFI = Medieval Unicode Font Initiative (2006). *Character recommendation Characters in the official Unicode Standard and in the Private Use Area for Medieval texts written in the Latin alphabet. Part 1 : Alphabetical order. Version 2.0 (22 December 2006)* [en ligne : <http://www.mufi.info/specs/MUFI-Alphabetic-2-0.pdf>]
- Pellat J.-C. (1988). « Indépendance ou interaction de l'écrit par rapport à l'oral ? Recensement critique des définitions du graphème », in Catach, 1988a : 133-146.
- Prou M. (⁴1924). *Manuel de paléographie latine et française*. Paris : Auguste Picard.
- Stiennon J. (²1991). *Paléographie du moyen âge*. Paris : Armand Colin.
- Vielliard F. & Guyotjeannin O. (2001). *Conseils pour l'édition des textes médiévaux. Fascicule I. Conseils généraux*. Paris : École des Chartes.