Le texte dans tous ses états
Philosophie d'encodage du projet Khartês

Nicolas Mazziotta

Université de Liège – 4000 – Belgique
nicolas.mazziotta@ulg.ac.be

Abstract

When editing Medieval sources, one always encounters the same dilemma : how to make the text readable nowadays without spoiling its testimony ? « Cleaning » the text from all that makes it difficult to read for someone erases marks that would look valuable from another point of view. Hence, there is an unavoidable gap between the source and the edition. The *Khartês* project (which intends to edit and study original Old French charters written in Liège before 1300) has encountered this problem. Traditional « solutions » were inadequate and computers opened an interresting path. However, *Khartês* had to think about creating an encoding scheme that would suit its aim and remain easy to use. This is the way KML (the *Khartês Markup Language*, an XML-conformant language) was born, together with a small set of handy encoding rules. A wide range of applications are made possible. For instance, thanks to the way the *corpus* has been created, it has been possible to search a particular document for *marked letters* (which look like our uppercase letters) and to study them.

Résumé

Éditer un document médiéval pour que celui-ci soit appréhendable à l'époque actuelle n'est pas une mince affaire. Constamment tiraillé entre le respect du document et la volonté de le rendre accessible, l'éditeur en est réduit à contenter un type de public en particulier au détriment d'un autre. Son travail se solde toujours par une importante distance entre le support original et l'édition moderne ; c'est inévitable. Ce problème, *Khartês* (projet liégeois d'édition et d'étude linguistique de chartes originales françaises antérieures à 1300) l'a rencontré. Les « solutions » qu'offraient la perspective traditionnelle étaient insuffisantes, et l'ordinateur ouvrait des voies intéressantes. Néanmoins, il a fallu réfléchir à une manière de faire qui convienne tout à fait à notre propos tout en suivant la philosophie d'économie que les contingences matérielles nous obligeaient à suivre. Ce qui ressort de cette expérience, c'est un système de balisage relativement simple baptisé *Khartês Markup Language* (KML, qui est conforme à XML) et une petite série de règles d'encodage simples et légères. Les applications sont multiples. Par exemple, grâce à cette formalisation textuelle, nous avons pu mettre à l'épreuve le corpus et l'interroger au sujet de la répartition des *lettres marquées* (qui ressemblent à nos capitales actuelles) dans un document particulier.

Mots-clés : *représentations textuelles, chartes médiévales, XML, fatware*.

1. Introduction

Signalons pour commencer que notre contribution fait partie des recherches menées dans le cadre du projet *Khartês* — projet d'édition et d'étude linguistique des chartes originales rédigées en français en Wallonie avant 1300. Nous avons ici pour but d'exposer le principe fondamental qui guide l'encodage que nous avons employé pour la constitution du corpus du projet et son utilité dans le cadre d'études spécifiques. Après avoir brièvement présenté ce dernier (2) ainsi que ses objectifs (3) et situé notre travail personnel à l'intérieur de cette entreprise, nous exposerons les difficultés qu'a posées le passage du texte la formalisation des documents (4) et les solutions que nous avons retenues (5). Nous terminerons par un exemple d'application (6) avant de conclure (7).

2. Projet *Khartês* : présentation

Khartês est un projet d'édition et d'étude des chartes originales rédigées en langue française en Wallonie avant 1300. Il a débuté en 1998 à l'initiative de Marie-Guy Boutier (professeur à l'Université de Liège)[1] et trouve sa place dans la vaste entreprise d'édition que constitue la collection des *Documents linguistiques de la France*.

La collection des *Documents linguistiques de la France* est inaugurée en 1974, avec l'édition des documents originaux conservés dans le département de Haute-Marne, préparée par Jean-Gabriel Gigot sous la direction de J. Monfrin (Gigot, 1974). Il s'agit de la première pièce d'un vaste projet destiné à réunir les textes non littéraires rédigés en langue vulgaire dans le domaine français. La collection devait initialement comprendre trois séries couvrant les domaines d'oïl (série française) et d'oc, mais aussi le domaine du franco-provençal. Il y était notamment prévu d'éditer des chartes originales conservées dans les différents fonds d'archives de la France (la série franco-provençale est un recueil d'éditions de textes dialectaux non littéraires).

En 1984, parait le premier volume des *Documents linguistiques de la Belgique romane*, série parallèle à la série française des *Documents linguistiques de la France* et qui comprendra finalement trois volumes distincts : un volume rassemblant les chartes du Hainaut (Ruelle, 1984), un autre pour les chartes flamandes (Mantou, 1987) et un troisième, encore manquant, dédié au chartes de Wallonie. Répondant à l'invitation que lui fit personnellement Jacques Monfrin en 1994, M.-G. Boutier entreprend le projet *Khartês*, destiné dans un premier temps à fournir les matériaux nécessaires à l'élaboration du tome trois des *Documents linguistiques de la Belgique romane*, et donc à couvrir la Wallonie (provinces de Liège, de Namur et de Luxembourg). Le corpus étudié dans le cadre du projet *Khartês* sera constitué pour commencer des chartes originales conservées dans les provinces de Liège et de Namur. Le travail a débuté par l'édition de celles qui sont conservées aux Archives de l'État à Liège. Toutes les chartes originales antérieures à ca 1291 seront étudiées.

3. Objectifs

Pour commencer, voyons concrètement à quoi ressemble une charte (v. fig. 1, p. suivante). *Khartês* se pose d'emblée comme objectif d'éditer ces documents pour qu'ils servent de matériaux pour des recherches linguistiques ; le projet vise donc à « donner un texte à lire »[2] à des linguistes.

Or il se fait que les documents tels que celui que nous venons de montrer ont longtemps occupé une communauté de chercheurs qui n'étaient pas linguistes : les diplomatistes. En matière de typologie textuelle, les chartes appartiennent en effet toutes à une même grande famille[3], celle des textes diplomatiques. Indépendamment du public ciblé, les textes que *Khartês* édite doivent recevoir l'habillage traditionnel qui convient à ce type de texte ; la lecture de divers manuels, des plus vénérables (Giry, 1894) aux plus récents (Guyotjeannin, 1993 ; Vieillard, 2001a ; Vieillard, 2001b), nous a permis de prendre conscience de ces normes déjà bien établies.

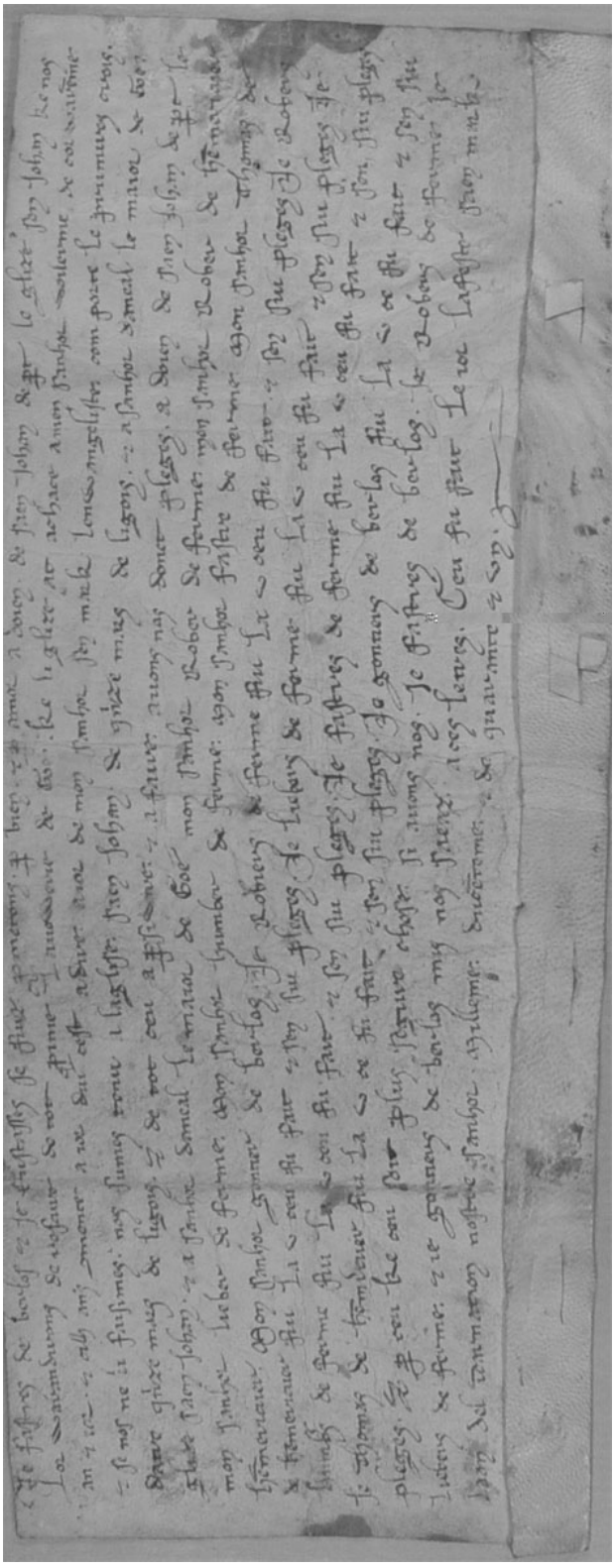
Comme nous sommes linguistes (du moins en formation), nous avons directement entrevu la possiblité de travaux linguistiques : les chartes devaient servir de matériaux pour l'étude d'unités qu'on peut définir intuitivement comme étant des *textes* (études énonciatives), des *phrases* (études syntaxiques), des *mots* (lexicologie, onomastique), des *lettres* (graphétique et graphé-

[1] Nous en profitons pour remercier Marie-Guy Boutier, qui s'est donné la peine de relire attentivement le texte de notre communication.

[2] Selon les mots de Marie-Guy Boutier.

[3] Cela n'exclut pas que l'on puisse raffiner la typologie.

Figure 1. Archives de l'État à Liège, collégiale Saint Jean l'Évangéliste, 1241



mique) et des *signes de ponctuation* (études spécifiques).

Très vite, il nous est apparu que les manuels de diplomatique, élaborés par des historiens, décrivaient des techniques insuffisantes à nos yeux de linguistes, non qu'elles soient mauvaises en soi, mais parce qu'elles réduisent de manière trop importante la richesse linguistique des documents. Nous ne trouvions guère de quoi satisfaire notre appétit dans les manuels de philologie (Foulet, 1979 ; Lepage, 2001), le plus souvent basés sur la tradition. Les conventions de ces guides d'édition permettent de donner des textes à lire à des historiens ou à des littéraires, pas à des linguistes.

Dans cette optique, un questionnement préalable sur les études linguistiques possibles sur ces textes est inévitable. De manière générale, des études de syntaxe, de lexicologie ou de dialectologie historique sont envisageables, mais aussi d'autres, considérant des aspects négligés jusqu'à présent, comme celui de la graphétique ou de la ponctuation.

C'est précisément ce dernier domaine que nous avons choisi d'étudier. Pour que l'édition puisse nous servir de base de travail, il nous fallait *préserver* ces signes jusqu'ici délaissés.

4. Le texte dans tous ses états

4.1. *Perspective traditionnelle*

Au départ, nous pensions naïvement que la question « Comment arriver à nos fins ? » recevait comme réponse « Il suffit de transcrire tout ce dont nous avons besoin ». Malheureusement, l'accumulation de tout ce que nous trouvions « intéressant » nous a vite déçus : d'une part, les documents que nous produisions étaient difficilement lisibles pour les personnes extérieures au projet, et même pour nous, lorqu'il était question de nous relire ou de nous corriger. Nous pensions de manière traditionnelle et n'entrevoyions que deux solutions. La première (a) consistait à imprimer deux versions différentes du texte, l'une en-dessous de l'autre. La seconde (b) consistait à imprimer un seul texte en s'ingéniant à le réduire le moins possible tout en y introduisant des marques d'élaboration par le truchement de marques typographiques. Par exemple, Jacques Monfrin, même s'il ne le fait pas, préconise, dès le début de la série des Documents linguistiques de la France, de conserver la ponctuation originale devant les noms propres[4]. De même, M.-D. Gleßgen marque les petites capitales originales par des grasses (idée que nous avons reprise ci-après).

Revenons au texte présenté ci-dessus (p. 795). Si nous le transcrivons en gardant tous les signes qui nous intéressent dans le cadre d'une étude sur la ponctuation, nous obtenons quelque chose comme (le signe ‹ | › indique une petite espace) :

> [1] **ȷ**e **f**aſtres de **b**erloſ *et* ȷe **e**uſtaſſes ſe fiuz · p*ro*metons · p*or* bien *et* p*or* amor a doien · de ſaen **ȷ**ohan de|p*or*t le glize ſen ȷohan ke nos [2] **l**or warandirons de|reſcure de|tot p*ro*me **l**auowerie de **g**oeꝛ ke li glize at achate a|mon ſanhor wilerme de cor|ware*mm*e · [3] an *et* ior · *et* cil ans *com*mence a ior dui ceſt a|dire · a|ior de mon ſanhor ſen mark lenwangeliſte · com|porte le prumires crois · [4] *et* ſe noſ ne li faiſimesꝛ nos ſumes tenu a lagliſe · ſaen ȷohan · de q*u*inze mars de ligois · *et* a|ſanhor daneal le maíor de **g**oe · [5] datre q*u*inze mars de ligois

Puisqu'il nous serait utile d'avoir accès à la ponctuation moderne également, nous aurions besoin d'une édition qui ressemblerait à :

> [1] Je, Fastrés de Berlos, *et* je, Eustasses, se fiuz, p*r*ometons, p*or* bien *et* p*or* amor a doien de Saen

[4] « Il n'y avait qu'un bien mince intérêt à côté de beaucoup d'inconvénients à tenir compte de[s] signes [de ponctuation originaux] dans la publication. En revanche, je regrette d'avoir supprimé les points (. ou ..) qui précèdent les noms propres dans quelques actes [...] ; je les conserverai à l'avenir. » (Monfrin, 1974 : lxiv)

> Johan de|*por*t le glize Sen Johan, ke nos [2] lor warandirons de rescure de tot p*ro*ime l'avowerie de Goé, ke li glize at achate a|mon sanhor Wilerme de Cor|ware*mm*e, [3] an *et* jor ; *et* cil ans *com*mence a jor d'ui, c'est a|dire a|jor de mon sanhor sen Mark l'Enwangeliste c'om|porte le Prumires Crois. [4] *Et* se nos ne li faisimes, nos sumes tenu a la_glise Saen Johan de q[u]inze mars de ligois, *et* a|sanhor Daneal, le maior de Goé, [5] d'atre q*u*inze mars de ligois.

Cette démarche (a) complique le travail de l'éditeur, du correcteur, et de celui qui veut simplement comparer les deux états du texte, le poussant à effectuer sans-cesse des allers-retours entre eux (ce qui l'amène inévitablement à commettre des erreurs), c'est ce qu'on peut nommer *problème de correspondance*.

D'où l'utilité, si nous désirons comparer la ponctuation originale à la ponctuation moderne, d'une édition qui ressemblerait à :

> [1] **J**e, Faſtrés de Berloſ, *et* ȷe, **E**uſtaſſes, ſe fiuz, · p*ro*metons, · p*or* bien *et* p*or* amor a doien · de Saen **J**ohan dep*or*t le glize Sen Johan, ke nos [2] **l**or warandirons de reſcure detot p*ro*ime **l**'avowerie de **G**oé,ˀ ke li glize at achate a|mon ſanhor Wilerme de Cor|ware*mm*e, · [3] an *et* ȷor ; · *et* cil ans *com*mence a ȷor d'uí, c'eſt a dire · a|ȷor de mon ſanhor ſen Mark l'Enwangeliſte · c'om porte le Prumires Crois. · [4] *Et* ſe noſ ne li faiſimes,ˀ nos ſumes tenu a la_gliſe · Saen Johan · de q[u]inze mars de ligois, · *et* a|ſanhor Daneal, le maíor de **G**oé, · [5] d'atre q*u*inze mars de ligois.

Malheureusement, cette méthode rend fastidieuse la lecture du document pour celui qui n'est pas intéressé par la ponctuation, l'encombrant de bruits, ce qui constitue un *problème de compréhension*.

Bien sûr, nous pouvions raisonner comme le fait Arrigo Castellani (1985 : 246) et privilégier la seconde solution, considérant qu'elle ne fait obstacle qu'au lecteur peu expérimenté[5]. Mais ce n'est pas exclusivement l'ignorance qui rend un texte surchargé difficile d'accès : la surcharge impose une opération de discernement supplémentaire de la part du lecteur ; et les marques de ponctuation moderne peuvent constituer une surcharge pour celui qui ne s'y intéresse absolument pas. À quoi bon fournir une édition qui ne serait pas plus facile à lire que le support original ?

En somme, la solution au problème de correspondance est le document unique et celle qui résout le problème de compréhension tout en préservant les données est le document double. D'un point de vue traditionnel, les deux sont incompatibles.

Donc, nous voici amené à faire trois éditions distinctes d'un même texte en fonction de l'usage que nous voulons en faire. On comprendra aisément que l'édition d'un simple document d'une dizaine de lignes peut rapidement mener à un casse-tête chinois lorsqu'il est question de corriger quelques erreurs de lecture ou d'apporter des précisions par des notes interprétatives. D'autre part, dans la mesure où aucune de ces trois éditions ne conviendrait à un lecteur appartenant au projet, il faudrait en produire une quatrième... La voie traditionnelle mène à l'impasse.

4.2. *Apports des nouvelles technologies*

La perspective traditionnelle ne rend possible aucun lien opératoire : dès lors qu'il en a besoin, l'humain doit toujours construire les relations qui ne figurent pas sur le papier.

[5] « Rimane la scelta tra il testo [di un trattato d'Alessandro Leonardi (1554) come è originalamente uscito] e [un] testo [...] in cui la punteggiatura, la divisione delle parole, le maiuscole e gli accenti sono conformi all'uso moderno. Darei la preferenza al primo, giacché i segni paragrafematici costituiscono un sistema sufficientemente evoluto e coerente : sistema di cui è difficile non tener conto, anche se in parte differisce da quello attuale e se può causare certe difficoltà (o certe ulteriori difficoltà, in più di quelle grafematiche) al lettore sprovveduto. Ma quanti sono i lettori sprovveduti che si dilettano a leggere il trattato di Leonardi ? E per gli altri, non basterà una semplice avvertenza sugli usi grafici e interpuntivi dell'epoca a cui risale la stampa originale ? »

Comme chacune des éditions traditionnelles n'est qu'une façon de regarder un même texte, nous avons pensé à créer un fichier contenant un document unique bardé de notes — solution traditionnelle (b) —, mais dont il serait possible de ne sélectionner qu'une partie en fonction du regard qu'on veut poser sur le document. Nous nommons cette méthode *texte monolythique*, parce qu'elle permet de maintenir l'unité intellectuelle du texte.

Imaginons que la première ligne du document présenté ci-dessus soit sauvegardée suivant un encodage fictif[6] qui marque les résolutions d'abréviation entre parenthèses (pratique par ailleurs très courante) :

```
[1] Je, Fastrés de Berlos, (et) je, Eustasses se fiuz p(ro)metons
p(or) bien (et) p(or) amor a doien de Saen Johan dep(or)t le glize
$en johan ke nos
```

Les parenthèses inélégantes sont utiles pour le sceptique lexicologue, qui veut être certain qu'un mot est bien attesté sous une forme et pas une autre ; elles lui offre un moyen de contrôle supplémentaire. Par contre, elles ne servent à rien pour l'historien. Dans cette optique, un seul fichier contient donc le texte dans tous ses états, entremêlant marques d'élaboration et richesse non réduite sous la forme d'un balisage *ad hoc*[7], mais chacun n'y sélectionne que ce qui l'intéresse. Les deux solutions traditionnelles, *a priori* antagonistes (rendu unique *vs* rendu multiple), sont conciliées.

Le texte monolithique s'impose par sa flexibilité et par la sécurité qu'il offre aux données (en rendant impossible les anomalies de mise à jour). En plus d'offrir tous les avantages combinés des méthodes traditionnelles, la méthode permet une numérotation et un classement aisé des documents.

5. Élaborer le langage

5.1. *Le choix d'XML*

Le choix de la norme XML (*eXtensible Markup Language*, v. W3C, 2000) s'est imposé assez rapidement : lisible, simple, populaire, personnalisable et extensible, mais aussi gratuit, le standard permet de créer des documents qui peuvent en outre être manipulés par une grande quantité de logiciels libres. Enfin, les fichiers associés à des feuilles de style (W3C, 1999) génèrent les formats de sortie désirés[8], ce qui répond à la nécessité de différencier le contenu du formatage.

5.2. *À la recherche de nouvelles solutions*

Du fait de la spécificité du traitement que nous réservons aux données de nos chartes, nous n'avons pas trouvé de solution satisfaisante dans le résultat du travail énorme fourni par la TEI et EAGLES pour définir des règles d'encodage flexibles utilisant XML.

La plupart du temps, les solutions les plus proches de ce que nous cherchions proposaient d'*aligner plusieurs corpus*[9]. La TEI donne des exemples où les variantes de différents manuscrits sont effectivement « alignées » à l'aide le l'apparat critique (TEI Consortium 2001 :

[6] C'est-à-dire un texte tel qu'il pourrait figurer dans le fichier informatique.

[7] Le balisage peut en effet inclure n'importe quel type d'information « Markup can be used to encode many different features within a text. In general, the more markup there is in a text, the more useful that text can be, and the more expensive it is to create. Markup can be used to encode both structural and analytic features. » (Hockey, 1998 : 108)

[8] Les formats ASCII, HTML et RTF ne posent aucun problème ; le format PDF peut être obtenu via un fichier LaTeX — solution que nous avons retenue.

[9] C'est-à-dire de mettre deux textes correspondant suivant un critère choisi en correspondance réciproque ; voir, par exemple, Habert, 1997 : 135, qui présente les corpus alignés dans la perspective de la traduction.

§19) :

> A very simple partial apparatus for the first line of the Wife of Bath's Prologue might take a form something like this :

```
<app>
<rdg wit="El">Experience though noon Auctoritee</rdg>
<rdg wit="La">Experiment thogh noon Auctoritee</rdg>
<rdg wit="Ra2">Eryment though none auctorite</rdg>
</app>
```

La relation de correspondance entre les mots est évidente, mais ce système ne permettrait que difficilement de mettre un caractère en relation avec un autre ou un groupe de caractères. D'autre part, si l'alignement est tout à fait justifié pour le cas de deux manuscrits différents ou d'une traduction, sa validité est on ne peut plus douteuse dans le cas de deux vues possibles d'un texte *unique* ; sans compter qu'il mène à presque tous les désavantages de la première solution traditionnelle (document multiple) : dislocation de l'unité du texte, danger d'anomalies de mise à jour, multiplication des versions en fonction des emplois, etc.

5.3. *Une philosophie : l'économie*

Pour élaborer nos règles d'encodage, nous avons poursuivi deux objectifs : éviter ce qu'on nomme le *fatware* et privilégier la transparence des données.

5.3.1. *Éviter le* fatware

Par *fatware*, on entend « un défaut consistant, pour un logiciel ou un langage, à avoir une taille démesurée par rapport à sa fonctionnalité »[10].

D'après Gary Simons, dans le cadre d'un projet particulier, ciblé et donc nécessitant moins de ressources, l'utilisation de langages très généraux pour encoder des documents peut mener au *fatware* :

> The large SGML DTDs[11] in widespread use (e.g. HTML, DocBook, ISO 12083, CALS, EAD, TEI) offer the advantage of standardization, but for a particular project they often carry the disadvantage of being too large or too general (Simons, 1998 : §1).

Il peut arriver que la DTD TEI ne convienne pas à un projet spécifique :

> For the Sikaiana dictionary project, the TEI DTD proved to be huge in comparison to the subset of elements and attributes that were actually used (Simons, 1998 : §2.3).

D'après lui, la solution consiste à définir soi-même une DTD adéquate, ce qui est rendu aisé par le fait qu'XML est initialement prévu pour ce type de développement rapide d'un balisage *ad hoc* tout en restant conforme à un standard. En adoptant cette position, on simplifie de nombreuses tâches inévitables dans la mise en œuvre d'un projet

> Having a DTD that is limited to just the elements and attributes that are used in a project simplifies many tasks like building project-specific software, specifying stylesheets, shipping the DTD with the data, and documenting markup practice (Simons, 1998 : §2.3).

[10] « Five years ago, a cover story in Byte (Perratore, 1993) decried the problem of « fatware » — software that just keeps getting bigger and bigger with each release without returning commensurate benefit to the user. Niklaus Wirth, in his plea for lean software (Wirth, 1995), sums up the situation thus : « Software's girth has surpassed its functionality. » » (Simons, 1998 : §2.3, nous adaptons les références)

[11] [*Document Type Definition*, grammaire définissant formellement un document SGML.]

Cet écueil à éviter – tout spécialement pour un projet en développement et encodé par des linguistes peu rompus à l'informatique – constitue une des premières raisons qui nous ont menés à « contourner » les recommandations pourtant détaillées et sans cesse améliorées de la TEI (2001) ou d'EAGLES (1999). En fait, les documents que le projet présente sont d'une extrême simplicité par rapport à tout ce qui est prévu par TEI. La conversion du corpus vers ces formats généraux reste possible – notamment grâce aux transformations XSLT[12], qui permettent de convertir facilement un fichier XML en un autre. C'est ce qui nous a amené à définir un langage dédié spécialement à l'édition des chartes : KML (*Khartes Markup Language*)

5.3.2. *Privilégier la transparence*

Dans un projet linguistique à caractère philologique utilisant l'informatique, on peut s'attendre à rencontrer des problèmes récurrents liés principalement à la distance qui existe entre les deux disciplines.

Le projet *Khartês* entend intégrer des étudiants préparant un mémoire de licence (spécialistes en formation) ; et il va de soi que si l'on peut demander au linguiste de devenir philologue, il est beaucoup moins raisonnable de le forcer à être à même de comprendre et manipuler tout ce qui a trait à l'aspect informatique du projet. Néanmoins, on peut lui demander de se servir d'outils existants d'usage accessible pour le non-spécialiste.

Pour rendre l'activité d'encodage facile d'accès, il y a deux solutions : soit fournir une interface efficace à l'utilisateur, soit créer un langage et des modalités de balisage simples.

Des deux solutions, nous avons choisi la seconde, essentiellement parce que le développement d'une interface est un métier en soi et demande du temps (tant pour la conception que pour le test) et des compétences que nous n'avons pas nécessairement le loisir développer.

Nous privilégions ainsi la *transparence de l'encodage*, la capacité qu'a un encodage à signifier de lui-même ce qu'il enregistre, c'est-à-dire à être lisible par un être humain – n'importe quel encodage est lisible par un ordinateur, pourvu qu'on lui fournisse le bon algorithme. Pour un humain, `u0041`, `&#x0041;` et `&#65;` pour signifier *A* sont illisibles, et personne ne reconnaîtra *livre* dans

```
&#x006C;&#x0069;&#x0076;&#x0072;&#x0065;
```

Mais si la solution idéale est de trouver un signe motivé pour tout élément encodé, sa mise en pratique est difficile. Nous devons en effet nous plier à des contraintes purement matérielles (et contingentes), comme le nombre de touches d'un clavier par exemple. Pour la plupart des caractères, cela va sans difficulté, mais lorsqu'il s'agit d'encoder une analyse et que nous avons besoin du balisage, il faut que ce dernier soit bref et explicite ; et ces deux attributs, applicables tant au balisage qu'aux données caractères, sont souvent antagonistes...

L'inconvénient est qu'en évitant les entités Unicode[13], les données ne sont pas encodées à l'aide d'un langage strictement conforme à XML, mais une simple conversion à l'aide d'expressions rationnelles permettra de résoudre le problème lorsque le corpus devra être distibué en dehors des membres du projet.

Il est parfois nécessaire d'encoder des caractères qui n'ont pas d'équivalent Unicode défini. C'est le cas des deux variantes du signe de ponctuation nommé (Parkes, 1992) *punctus elevatus*

[12] C'est-à-dire *eXtensible Stylesheet Language Transformations*, un langage de programmation de très haut niveau destiné à modifier la structure des documents XML ; XSLT est conforme au standard XML (voir W3C, 1999).

[13] Voir http ://www.unicode.org/.

(point renforcé d'un trait ou d'un crochet) : ‹ ⁝ › et ‹ ⁄ ›. Puisque ces signes ressemblent fortement à ‹ ? › et ‹ ! ›, ce sont ces saisies clavier qui servent à les encoder. De la même façon, ‹ * › sert à noter ‹ · › et ‹ / › équivaut à ‹ / ›.

6. Exemple

6.1. *Document*

Nous ne pourrions faire ici une présentation complète de la DTD de KML et des conventions d'encodage pour les caractères spéciaux. On retiendra essentiellement :
– que l'élément c entoure les lettres marquées sur le support,
– que l'élément l marque le début des lignes,
– que les crochets entourent les abréviations résolues,
– que le signe ‹ ˚ › précède les caractères qui doivent être transformés (‹ u › en ‹ v › et vice-versa, ‹ | › en ‹ - ›, etc.) pour qu'un lecteur moderne appréhende plus facilement le texte.

Si nous transcrivons intégralement le document présenté p. 795, nous obtenons la transcription qui suit :

```
<charte fonds="LgSJe" an="1241(29b)"><texte><recto><part /><l /><c>J</c>e, Fa$trés de Berlo$ [et] je, <c>E</c>u$ta$$es, $e fiuz, *
<part /> p[ro]metons, * p[or] bien [et] p[or] amor a doien * de §aen <c>J</c>ohan de˚|p[or]t le glize §en Johan, ke nos <l /><c>l</c>or
warandirons de|re$cure de|tot p[ro]ime <c>l</c>'a˚uowerie de <c>G</c>oé,? ke li glize at achate a|mon $anhor Wilerme de Cor|ware[m]me,
* <l />an [et] ˚ior; * [et] cil ans [com]mence a ˚ior d'uí, c'e$t a|dire * a|˚ior de mon $anhor $en Mark l'Enwangeli$te * c'om|porte
le Prumires Crois. * <l />[Et] $e no$ ne li fai$imes,? nos $umes tenu a la_gli$e * §aen Johan * de q[u]inze mars de ligois, * [et]
a|$anhor Daneal, le maior de <c>G</c>oé, * <l />d'atre q[u]inze mars de ligois. * [Et] de tot ce a p[or]$iwre * [et] a faire * a˚uons
nos donet pleges * a doien de §aen <c>J</c>ohan de-p[or]t le <l />glize §aen Johan * [et] a $anhor Daneal le maior de <c>G</c>oé,
mon $anhor <c>R</c>ober de|Ferme, * mon $anhor <c>R</c>ober de He[m]mericuer, <l />mon $anhor Lieber de Ferme, * <c>m</c>on $anhor
Humber de Ferme, * <c>m</c>on $anghor Fa$tré de Ferme, <c>m</c>on $anhor <c>T</c>homas de <l />He[m]mericuer, * <c>m</c>on $anhor
Gontier de Berlos. * [Et]_<c>je</c><rem>l'abréviation rentre dans le <c>j</c>e</rem>, <c>R</c>obiers de Ferme, fui <c>l</c>a ˚v ceu
fu fait; [et]|$'en $ui pleges. * [Et]_<c>j</c>e, <c>R</c>ober$ <l />de He[m]mericuer, fui <c>l</c>a ˚v ceu fu fait; * [et]|$'en
$ui pleges. * [Et]_<c>je</c>, <c>L</c>iebers de Ferme, fui <c>l</c>a ˚v ceu fu fait; [et] $'en $ui pleges. * [Et]_<c>j</c>e, <l
/>Humb[er]s de Ferme, fui <c>l</c>a ˚v ceu fu fait; * [et] $'en $ui pleges. * [Et]_<c>je</c>, Fa$trés de Ferme, fui <c>l</c>a ˚v ceu
fu fait; [et] $'en $ui pleges. <l /><c>J</c>e, <c>T</c>homas de He[m]m[er]icuer, fui <c>l</c>a ˚v ce fu fait; [et] $'en $ui pleges.
* [Et]_<c>je</c>, Gontiers de Berlos, fuí <c>l</c>a ˚v ce fu fait; [et] $'en $uí <l />pleges. * <part /> [<c>Et</c>] p[or] ceu ke ceu
$oit plus $egure cho$e, $i a˚uons nos, * <c>j</c>e, Fa$trés de Berlos, * <c>j</c>e, <c>R</c>obers de Ferme, * <c>j</c>e, <l />Liebers
de Ferme, * [et] ˚ie, Gontiers de Berlos, mis nos $aeaz * a|ces letres. * <c>C</c>eu fu fait <c>l</c>e ˚ior <c>l</c>a_fe$te $aen Mark,
* <l />l'aan del I[n]carnation No$tre §anhor * <c>m</c>illeme * duce[n]teme * [et] de quarante [et] ˚vn. * <rem>signé à droite</rem>
</recto></texte></charte>
```

6.2. *Exemple d'application*

Prenons un problème concret : nous voulons étudier les lettres marquées[14] apparaissant à l'initiale des noms propres sur le support original. L'encodage d'un texte monolythique permet d'interroger ce dernier de manière à en extraire tout ce qui est formellement défini. Comme on l'a vu ci-dessus, les lettres marquées sont transcrites dans l'élément c et les capitales modernes figurent en majuscules dans le code, ce qui constitue deux marquages compatibles.

En écrivant un simple programme d'extraction, il est aisé de relever tous les mots commençant par une majuscule et de les pointer par un point d'exclamation s'ils comportent une initiale marquée. Nous obtenons la liste suivante :

1. ! Je
2. Fastrés
3. Berlos
4. ! Eustasses
5. Saen
6. ! Johan
7. Sen
8. Johan
9. ! Goé
10. Wilerme
11. Cor|waremme
12. Mark
13. Enwangeliste
14. Prumires
15. Crois
16. Et
17. Saen
18. Johan
19. Daneal
20. ! Goé
21. Et
22. Saen
23. ! Johan
24. Saen
25. Johan
26. Daneal
27. ! Goé
28. ! Rober
29. Ferme
30. ! Rober
31. Hemmericuer
32. Lieber
33. Ferme
34. Humber
35. Ferme
36. Fastré
37. Ferme
38. ! Thomas
39. Hemmericuer
40. Gontier
41. Berlos
42. Et
43. ! Robiers
44. Ferme
45. Et
46. ! Robers
47. Hemmericuer
48. Et
49. ! Liebers
50. Ferme
51. Et
52. Humbers
53. Ferme
54. Et
55. Fastrés
56. Ferme
57. ! Je
58. ! Thomas
59. Hemmericuer
60. Et
61. Gontiers
62. Berlos
63. ! Et
64. Fastrés
65. Berlos
66. ! Robers
67. Ferme
68. Liebers
69. Ferme
70. Gontiers
71. Berlos
72. ! Ceu
73. Mark
74. Incarnation
75. Nostre
76. Sanhor

[14] Lettres se distingant de leurs voisines par leur épaisseur, leur taille ou leur couleur ; elles font intuitivement penser à nos majuscules modernes.

Sur 76 mots comportant une majuscule dans l'édition, 18 mots ont été pointés comme commençant par une lettre marquée dans l'original. Parmis ces 18 mots, trois ne sont pas des noms propres. Donc, la plupart du temps, le scribe ne marque pas les noms de personnes.

Dès lors, l'emploi des lettres marquées est-il lié à la segmentation en phrases ? Un dépouillement similaire montre qu'à part la première phrase du texte, qui commence par *Je*, toutes les phrases commencent par *Et* et une seule par *Ceu*. Les lettres marquées au début de

1. ! Eustasses	4. ! Goé	7. ! Rober	10. ! Robiers	13. ! Je
2. ! Johan	5. ! Johan	8. ! Rober	11. ! Robers	14. ! Thomas
3. ! Goé	6. ! Goé	9. ! Thomas	12. ! Liebers	15. ! Robers

ne correspondent pas à des débuts de phrase. Par contre, ce qui frappe dans cette liste, c'est que les lettres qui commencent ces mots appartiennent à un stock relativement réduit : ‹ e ›, ‹ g ›, ‹ j ›, ‹ l ›, ‹ r ›, et ‹ t › ; un autre dépouillement du même texte montre que ‹ e ›, ‹ r › et ‹ t › n'apparaissent jamais non marquées au début d'un nom propre. Nous décelons donc que la tendance d'un nom propre à être graphié avec une initiale marquée dépend de la nature de cette initiale, fait qui mériterait d'être étudié sur base de textes plus nombreux de la même époque.

7. Conclusion

Avec le développement de la linguistique de corpus et la démocratisation des moyens informatiques, l'édition de texte prend une dimension nouvelle : au lieu de se limiter à « donner un texte à lire » à une catégorie restreinte de chercheurs (voir à un seul lecteur), l'éditeur peut destiner son travail à un plus vaste public, libre de réduire comme il l'entend ce qui ne lui est pas utile. La réduction de la richesse originale du document n'affecte que la *vue* désirée par le lecteur et n'altère en rien le fichier contenant l'édition. Les possibilités de comparaison entre la forme originale et le marquage moderne peuvent être partiellement (sinon totalement) automatisées, ce qui facilite grandement les dépouillements.

Références

Castellani A. (1985). Problema di lingua, di grafia, di interpunzione nell'allestimento dell'edizione critica. In Malato (1985) : 229-54.

EAGLES [= Expert Advisory Group on Language Engineering Standards] (1999). *Corpus Encoding Standard. Version 1.5*. [http ://www.ilc.pi.cnr.it/EAGLES/home.html].

Foulet A. et Speer M.B. (1979). *On editing Old French Texts*. Kansas : The Regents Press of Kansas.

Gigot J. G. (1974). *Chartes en langue française antérieures 1271 conservées dans le département de la Haute-Marne*. Paris : C.N.R.S. (Documents linguistiques de la France, série française 1).

Giry A. (1894). *Manuel de diplomatique*. Paris : Hachette.

Guyotjeannin O., Picke J. et Tock B. M. (1993). *Diplomatique médiévale*. Louvain : Brepols (L'atelier du médiéviste 2).

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin (U Linguistique).

Hockey S. (1998). Textual Databases. *Lawler* : 101-37.

Lawler J. et Aristar Dry H. (dir.) (1998). *Using Computers in Linguistics. A Practical Guide*. London/New York : Routledge.

Lepage Y. G. (2001). *Guide de l'édition de textes en ancien français*. Paris : Champion.

Malato E. (dir.) (1985). *La critica del testo. Problemi di metodo ed esperienze di lavoro. Atti del convegno di Lecce 22–26 ottobre 1984*. Roma : Salerno Editrice.

Mantou R. (1987). *Chartes en langue française antérieures à 1271 conservées en Flandre Orientale et Flandre Occidentale*. Paris : C.N.R.S. (Documents linguistiques de la Belgique romane 2).

Monfrin J. (1974). Introduction. In Gigot (1974) : xi-lxxx.

Parkes M.B. (1992). *Pause and effect. An introduction to the history of punctuation in the West*. Cambridge : Scholar Press.

Perratore E., Thompson T., Udell J. et Malloy E. (1993). Fighting fatware. *Byte* : 98-108.

Ruelle P. (1984). *Chartes en langue française antérieures à 1271 conservées dans la province de Hainaut*. Paris : C.N.R.S. (Documents linguistiques de la Belgique romane 1).

Simons G. F. (1998). Using architectural processing to derive small, problem-specific XML applications from large, widely-used SGML applications. *SIL Electronic Working Papers*. [http ://www.sil.org/silewp/1998/006/silewp1998-006.html].

TEI Consortium. (2001). TEI P4. Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition. [http ://www.tei-c.org].

Vieillard F. et Guyotjeannin O. (dir.) (2001a). *Conseils pour l'édition de textes médiévaux. Fascicule I : Conseils généraux*. Paris : CTHS École des Chartes.

Vieillard F. et Guyotjeannin O. (dir.) (2001b). *Conseils pour l'édition de textes médiévaux. Fascicule II : Actes et documents d'archives*. Paris : CTHS École des Chartes.

W3C [= World Wide Web Consortium]. (2000, 1998). Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommandation 6 October 2000. [http ://www.w3.org/TR/REC-xml].

W3C [= World Wide Web Consortium]. (1999). XSL Transformations (XSLT) 1.0. W3C Recommandation 16 November 1999. [http ://www.w3.org/TR/1999/REC-xslt].

Wirth N. (1995). A plea for lean software. *IEEE Computer* : 64-8.