Cahiers de recherches médiévales et humanistes
Journal of medieval and humanistic studies

20 | 2010
Idylle et récits idylliques à la fin du Moyen Âge

Outils informatiques pour l'édition et le traitement des textes, des images, du langage

Mattia Cavagna

Édition électronique
URL : http://journals.openedition.org/crm/12246
DOI : 10.4000/crm.12246
ISSN : 2273-0893

Éditeur
Classiques Garnier

Édition imprimée
Date de publication : 30 décembre 2010
Pagination : 357-390
ISSN : 2115-6360

Référence électronique
Mattia Cavagna, « Outils informatiques pour l'édition et le traitement des textes, des images, du langage », *Cahiers de recherches médiévales et humanistes* [En ligne], 20 | 2010, mis en ligne le 30 décembre 2013, consulté le 13 octobre 2020. URL : http://journals.openedition.org/crm/12246 ; DOI : https://doi.org/10.4000/crm.12246

Outils informatiques pour l'édition et le traitement des textes, des images, du langage

Abstract : This paper is a detailed account of a symposium held at the Université catholique de Louvain on computer science tools used in the field of human sciences, in particular but not exclusively, in the edition of ancient and medieval texts.

Résumé : La présente contribution est le résumé détaillé d'une journée d'étude qui s'est tenue à l'Université catholique de Louvain et qui était consacrée aux outils informatiques utilisés dans la recherche en sciences humaines, tout particulièrement, mais non exclusivement, dans les éditions des textes anciens et médiévaux.

Le 24 avril 2009 à l'Université catholique de Louvain s'est tenue une journée d'étude consacrée à un choix d'outils informatiques actuellement utilisés dans la recherche en sciences humaines[1]. Plusieurs chercheurs de l'UCL et d'autres institutions étrangères ont profité de cette occasion pour confronter leurs méthodes et leurs outils de travail et pour présenter les caractéristiques et les possibilités d'exploitation des logiciels qu'ils utilisent dans le cadre de leurs travaux. Plusieurs chercheurs en différentes disciplines et plusieurs étudiants de deuxième et troisième cycle ont participé à la rencontre ; plusieurs autres chercheurs, étant dans l'impossibilité de joindre Louvain-la-Neuve m'ont sollicité par courrier électronique en manifestant leur intérêt et en me suggérant de laisser une trace écrite de cette rencontre. Vu l'intérêt et la qualité des interventions proposées, j'ai cru opportun d'accueillir cette suggestion et ai décidé de rédiger le présent écrit, qui se veut une sorte de bilan ou de compte rendu de la journée.

Le titre que j'ai choisi, *Outils informatiques pour le traitement des textes, des images, du langage*, suggère avant tout la pluralité des approches et des disciplines concernées. Si l'édition et le traitement des textes ont été au centre de la plupart des interventions, nous n'avons pas négligé le traitement des images, notamment des reproductions numérisées des manuscrits médiévaux, et du langage, notamment des documents sonores, dans la perspective de l'analyse sociolinguistique. Finalement, une intervention qui n'avait pas été prévue à l'origine est venue enrichir le programme de la journée grâce à un logiciel pour le traitement informatique des livrets d'opéra, alliant le traitement du texte à celui de la partition musicale.

Les critères fondamentaux qui ont guidé mon choix des interventions sont la qualité et la fiabilité des outils informatiques proposés. J'ai donné la priorité, voire l'exclusivité, aux langages et aux formats standard (XML, TEI) et aux logiciels qui offrent des garanties sur le plan de la fiabilité, de la portabilité (la compatibilité avec les différents systèmes d'exploitation) et, dans plusieurs cas, de la modularité, à

[1] La journée a bénéficié des financements du Fonds National de la Recherche Scientifique belge, de la Faculté de Lettres et Philosophie et du Département d'Études Romanes de l'Université catholique de Louvain que je tiens à remercier chaleureusement.

savoir la possibilité d'adapter le logiciel aux différentes exigences du chercheur à travers l'ajout de « modules » (cf. les logiciels oXygen, UNITEX et LaTeX).

Ce type de caractéristiques et de garanties me semble important pour plusieurs raisons. Tout d'abord, l'utilisation de formats et d'encodages standard permet le partage des données et ouvre d'importantes possibilités de collaborations entre les chercheurs et les spécialistes des différentes disciplines. D'autre part, le choix d'un outil informatique fiable et approprié permet d'abolir les risques liés à l'utilisation et au développement de logiciels « artisanaux », à savoir que les contraintes et les limites techniques prennent le dessus par rapport aux méthodes envisagées au départ.

Loin d'avoir des prétentions d'exhaustivité et loin de vouloir imposer des choix, cette rencontre avait un but très clair et très modeste : il s'agissait tout simplement de présenter et de confronter un certain nombre d'outils informatiques actuellement en cours d'utilisation, afin que chacun des participants puisse en tirer profit pour ses propres recherches en cours ou à venir.

1. Logiciels pour le traitement des textes et l'édition électronique

La première séance, présidée par Craig Baker, professeur à l'Université Libre de Bruxelles, est consacrée aux langages et aux outils informatiques utilisés principalement pour les éditions électroniques et pour l'encodage des textes en vue de leur traitement informatique.

1.1. James Cummings (U. d'Oxford), Editing TEI XML with oXygen

La séance est ouverte par James Cummings, médiéviste et spécialiste du traitement informatique des éditions critiques. En tant que chercheur qualifié du *Research Technologies Service* de l'Université d'Oxford, il intervient dans le comité directeur de plusieurs projets, comme la TEI (*Text Encoding Initiative*), dont il sera question dans sa présentation, et le *ENRICH project*, qui a pour but de réunir et d'harmoniser le système de catalogage et les données numériques concernant les manuscrits de plusieurs fonds européens[2].

James Cummings propose d'abord une brève introduction au langage XML qui, dans les dernières années, s'est imposé parmi les principaux langages standard dans le traitement de textes et aussi dans la création de documents pour la publication sur le web. Le système d'encodage, ou d'annotation (*markup*) propre à ce langage repose sur un postulat très simple : toutes les annotations doivent être insérées – et visualisées – dans le corps du texte sous forme de balises (*tags*), puisque seules les annotations explicites peuvent être reconnues dans les étapes successives de transformation d'un document. Les balises XML sont des données

[2] Cf. le site http://enrich.manuscriptorium.com. Le présent résumé est basé en grande partie sur la documentation que James Cummings a rendue accessible sur le site suivant : http://tei.oucs.ox.ac.uk/Oxford/2009-04-24-louvain/2009-04-24-louvain.pdf.

qui peuvent être facilement réutilisées à l'intérieur d'autres environnements informatiques[3].

STRUCTURE DU DOCUMENT XML

Le document XML présente une structure hiérarchique explicite déterminée par une série de balises selon le principe de l'arborescence et de l'encapsulation. Les balises fonctionnent par couple, l'une ouvrante, l'autre fermante, et sont définies à l'aide de chevrons :

```
<ref>content</ref>
```

Dans le jargon XML, les blocs délimités par les couples de balises sont nommés « éléments » (*element*). Un élément peut contenir d'autres éléments et / ou du texte. Certaines balises contiennent aussi un attribut (*attribute*) et un élément qui définit la valeur (*value*) de cet attribut. Voici un modèle théorique de document XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
<element attribute="value"> content </element>
<!-- comment -->
</root>
```

Dans la première ligne se trouve une balise qui permet d'indiquer la version du XML utilisée et le jeu de codage (*encoding*) des caractères choisi, dans ce cas il s'agit de UNICODE (UTF-8) ; le point d'interrogation qui suit le crochet indique qu'il ne s'agit pas d'une balise d'encodage, mais d'une sorte de prologue du document. La deuxième ligne contient la base ou racine (*root*) de l'arborescence, l'intégralité du document XML étant comprise entre les balises <root> et </root>. La troisième ligne présente un bloc ou élément – désigné ici par *content* – pourvu d'un attribut et d'une valeur. Au moment de la conversion du fichier, cette valeur permettra facilement d'identifier l'élément et de le traiter selon les paramètres définis par la feuille de style choisie (cf. le paragraphe suivant). À l'avant-dernière ligne se trouve un commentaire qui sera lisible exclusivement dans le document de travail : le point d'exclamation qui suit le crochet distingue, à l'instar du point d'interrogation, cette indication d'une balise d'encodage. Dans un logiciel d'édition XML (cf. ci-dessous, *oXygen*), toutes ces composantes sont affichées en différentes couleurs, afin de mieux les distinguer.

Toutes les composantes qui structurent le document (balises, éléments, attributs, valeurs) doivent être « déclarées » à l'intérieur d'un fichier qui constitue une sorte de grammaire ou de tableau récapitulatif servant à garantir le bon fonctionnement du document XML. James Cummings conseille l'utilisation d'un

[3] Par exemple, la présentation citée ci-dessus a été réalisée par James Cummings en langage XML, puis été convertie en LaTeX afin de générer des documents en format PDF.

schéma RELAX NG, qui utilise le même langage XML[4]. Un tel schéma décrit la structure de l'arborescence, énumère les noms des éléments, des attributs et des valeurs, et gère l'association des attributs et des valeurs.

```
<element name="texte">
    <element name="rubrique">
      <attribute name="glose">
        <text/>
      </attribute>
      <attribute name="source">
        <text/>
      </attribute>
    </element>
</element>
```

Ce schéma sert à valider la balise <rubrique> qui sera nécessairement insérée à l'intérieur de l'élément <texte> </texte>, en précisant que cette balise pourra contenir deux types d'attributs, à savoir "glose" ou "source". À l'intérieur du fichier XML, la balise rubrique pourra donc avoir deux variantes : <rubrique type="glose"> </rubrique> ou <rubrique type="source"> </rubrique>. Au moment de la transformation du fichier, les deux variantes seront traitées différemment, par exemple en leur associant des couleurs ou des polices différentes.

TRANSFORMATION DU DOCUMENT XML

Le langage XML sert uniquement à définir la structure du document, c'est-à-dire à encoder le document d'une façon cohérente et exploitable par toutes sortes d'outils informatiques. Afin d'afficher le document et de le rendre facilement lisible pour l'être humain, il est nécessaire de le convertir sous un autre format (par exemple HTML pour la publication sur le web ou PDF et RTF pour l'impression sur papier). Cette conversion se fait à travers des feuilles de style CSS (*Cascading Style Sheets*) permettant de préciser les propriétés qu'on veut attribuer aux éléments du schéma XML lors de la conversion. Dans la feuille de style CSS, chacun des éléments du schéma XML est accompagné d'un bloc, délimité par les accolades {}, comprenant les propriétés qui lui seront attribués.

```
glose{font-size:50pt;font-family:Helvetica,sans-serif}
```

Les feuilles de style CSS prennent en charge individuellement les éléments du document XML, mais ne permettent pas de réaliser des transformations de structure. C'est pourquoi il faut avoir recours au langage XSL (*eXtensible Stylesheet Language*), qui permet la création d'un document HTML, auquel on associe une ou plusieurs feuilles de styles CSS.

[4] Autrement on peut utiliser le système de la DTD (*Definition Type Document*), mais James Cummings le considère comme fort dépassé. Le système RELAX NG est plus synthétique et flexible et n'impose pas l'apprentissage d'un autre langage.

James Cummings préfère ne pas donner trop de détails à ce propos, afin de ne pas multiplier les exemples. Il se contente de souligner que ce système est basé en somme sur la séparation du contenu et de la mise en forme ; il permet très facilement d'adapter la présentation d'un document sur la base des exigences, soit d'une maison d'édition, soit d'une publication sur le web.

TEXTUAL ENCODING INITIATIVE

Si l'utilisateur du langage XML a la possibilité de créer et de définir ses propres balises, James Cummings insiste sur l'importance d'utiliser un encodage standard qui permet un maximum de compatibilité et garantit le « bon usage » du langage XML. C'est ici qu'entre en jeu la TEI, un organisme international qui développe et supervise la standardisation des balises pour l'encodage des textes en forme numérique. La TEI définit une norme de balisage et fournit un manuel qui précise les méthodes d'encodage dans les sciences humaines, dans les sciences sociales et dans la linguistique. James Cummings met en garde le public contre une mauvaise interprétation du concept de standardisation : il s'agit moins d'imposer des choix que de partager un langage qui facilite l'intercompréhension. Pour reprendre sa formule : *standardization should not mean "Do what I do", but rather "Explain what you do in terms I can understand"*.

Loin d'imposer des choix restrictifs, l'utilisation de l'encodage standard donne également la possibilité de personnaliser (*customise*) le choix de balises à utiliser. La plateforme ROMA (http://tei.oucs.ox.ac.uk/Roma/) permet à l'utilisateur d'établir son propre schéma, tout en opérant ses choix à l'intérieur du standard défini par la TEI.

Le chercheur a la possibilité de choisir les paramètres fondamentaux de son système, en supprimant les composantes qu'il ne souhaite pas utiliser et en ne retenant que les modules qui s'adaptent le mieux aux exigences de son projet. James Cummings propose une liste des apports et des éléments que la TEI met à disposition du chercheur :

- une structure indépendante (*framework*) pour définir les langages d'annotations
- un système très simple, qui a été élaboré de façon pragmatique sur la base d'une large concertation, pour organiser et structurer les ressources textuelles ou autres…
- …qui peut être enrichi et personnalisé d'une façon idiosyncratique ou hautement spécialisée
- un répertoire très riche en composantes spécialisées pour décrire presque tous les types de phénomènes textuels
- une série intégrée de feuilles de style standard pour produire des textes, des schémas et de la documentation encodées selon les principes de la TEI et les transformer en différents formats et langages
- une communauté d'utilisateurs vaste, active, ouverte et transparente

James Cummings ajoute que la plateforme ROMA est également pourvue d'un système de contrôle (*sanity checker*) permettant de vérifier la structure du schéma XML et de repérer les éventuelles erreurs de syntaxe qui empêchent son utilisation.

oXygen

James Cummings présente ensuite un logiciel fournissant un environnement idéal pour l'édition des textes en XML. Son choix porte sur le logiciel oXygen qui présente plusieurs avantages par rapport à d'autres éditeurs XML. Tout d'abord, il intègre les balises de la TEI. Ensuite, il fournit à son tour une série de modèles de documents contenant une structure prédéfinie, qui peuvent être utilisés comme point de départ pour la création de nouveaux documents. L'utilisateur peut également créer ses propres modèles et les partager avec d'autres utilisateurs, par exemple dans le cadre d'un travail d'équipe.

Au moment de la saisie du texte, l'insertion des éléments se fait à l'aide de menus déroulants qui offrent un choix de balises et des attributs en fonction du schéma (RELAX NG ou DTD) associé au fichier. Le logiciel permet de choisir entre plusieurs types d'affichages. Au moment de la saisie du texte, l'utilisateur choisira de visualiser toutes les balises. Les différents éléments, les balises, les attributs et tous les éléments de la syntaxe XML sont pourvus de couleurs différentes afin de mieux distinguer entre l'encodage informatique et le texte proprement dit. Sur la barre verticale à la droite de l'écran s'affiche la numérotation des lignes comprenant une série d'indicateurs permettant de distinguer les différents blocs en repérant facilement les deux indicateurs situés aux extrémités (la balise ouvrante et la balise fermante). Une fois le document complété et correctement encodé – les éventuelles erreurs de syntaxe ayant été éliminées grâce au *sanity checker* – il pourra ensuite être transformé selon les paramètres des feuilles de style choisies qui seront associés au fichier.

James Cummings rappelle enfin que des stages d'initiation à l'utilisation du XML et de la TEI sont organisés tous les étés en juillet au *Research Technologies Service* de l'Université d'Oxford. Le stage de cette année aura lieu du 20 au 24 juillet 2009.

1.2. Cédrick Fairon (UCL), UNITEX – une boîte à outils pour l'analyse de textes

La séance se poursuit avec l'intervention de Cédrick Fairon, professeur de linguistique et directeur du Centre de traitement automatique du langage (CENTAL), à l'Université catholique de Louvain, centre qui est à l'origine de plusieurs applications informatiques largement diffusées aussi en dehors du cadre académique, telles la correction orthographique automatique, la reconnaissance de la parole, la traduction automatique[5]. À la fois linguiste et informaticien, Cédrick Fairon est spécialiste de la description linguistique et du traitement informatique des langues vivantes.

Cédrick Fairon présente un outil informatique appelé UNITEX (développé à l'Université de Marne-la-Vallée), actuellement utilisé dans un certain nombre de projets, en partie liés à l'UCL et au CENTAL, permettant de réaliser des recherches

[5] Pour une description du laboratoire et de ses activités, voir l'adresse suivante : http://cental.fltr.ucl.ac.be.

et des analyses complexes sur des corpus textuels de grande envergure. Il s'agit d'un logiciel libre (*open source*) multi-plateforme et largement modulable et compatible[6].

Plusieurs caractéristiques distinguent UNITEX des concordanciers traditionnels et des systèmes de traitement de corpus. La première caractéristique est la possibilité d'intégrer des ressources linguistiques, sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire.

Un ensemble de ressources linguistiques est déjà intégré au logiciel, mais il est possible d'en ajouter d'autres, par exemple pour les langues anciennes et médiévales et surtout de développer ses propres ressources linguistiques, à partir du travail de *parsing* réalisé par UNITEX[7].

La deuxième caractéristique est la possibilité d'engendrer des moteurs de recherches capables de repérer des expressions complexes en croisant plusieurs critères. Les critères de recherche sont visualisés à l'écran sous forme de graphes, ce qui permet, d'un côté, de mieux définir le parcours accompli par le logiciel et, de l'autre, d'apporter facilement des modifications en intervenant sur l'ordre et sur la hiérarchie des critères proposés.

Cédrick Fairon propose une démonstration pratique à partir d'un texte en français moderne. Au moment où le texte est importé dans UNITEX, il est soumis à un prétraitement : le parseur découpe le texte en phrases en insérant automatiquement la balise {S} pour « phrase » (*sentence*) en se basant sur des indicateurs tels la ponctuation et la présence des majuscules. On passe ensuite au processus de reconnaissance automatique des mots. À travers un système de menus déroulants, l'utilisateur applique les ressources lexicales appropriées (*apply lexical resources*), en l'occurrence le dictionnaire de français moderne. UNITEX classe alphabétiquement les mots du fichier. Il est rapide et il évite les répétitions puisqu'il ne retient qu'une seule entrée par mot.

Les mots sont classés et présentés selon le formalisme DELA (Dictionnaires Electroniques du LADL), permettant de décrire les entrées lexicales simples et composées en leur associant des informations grammaticales, sémantiques et flexionnelles. Voici un exemple tiré d'un relevé lexical automatique présenté selon le formalisme DELA :

```
chapitres,chapitre.N+z1:mp
```

– `chapitres` est la forme fléchie de l'entrée telle qu'elle a été relevée dans le texte.
– `chapitre` est la forme canonique (ou *lemme*) de l'entrée ; elle est séparée de la forme fléchie par une virgule ; pour les noms et les adjectifs, il s'agit de la forme au masculin singulier, pour les verbes la forme de l'infinitif.
– `N+z1` est la séquence d'informations grammaticales et sémantiques. N désigne un nom. Le sigle z1 est un indicateur de fréquence dans la langue et désigne un mot tout à fait courant. Les sigles z2 et z3 permettent d'identifier des mots appartenant

[6] Le logiciel peut être téléchargé à l'adresse suivante :
http://www-igm.univ-mlv.fr/~unitex/download.html
[7] Cf. ci-dessous l'intervention de Bastien Kindt.

au langage spécialisé et très spécialisé. Ces indicateurs sont particulièrement utiles lorsqu'il s'agit de distinguer des homographes.
– :mp est la séquence d'informations flexionnelles décrivant le genre et le nombre pour les substantifs ou les adjectifs (dans ce cas, masculin pluriel) ou, pour les verbes, les temps et modes de conjugaison.

Après avoir effectué l'analyse linguistique – opération extraordinairement rapide –, UNITEX ouvre une fenêtre divisée en trois parties proposant trois listes de mots. La première section, en haut à gauche, énumère les entrées lexicales simples (*simple-word lexical entries*), selon le modèle présenté ci-dessus. La deuxième section, placée juste au-dessous de la première, propose les mots et les entrées lexicales composées (*compound lexical entries*). Voici un exemple d'entrée composée :

```
au commencement,au commencement.ADV+PCDN+z1
```

Le code PCDN identifie une classe d'adverbes qui ont une structure et un comportement syntaxique particulier[8].

La troisième section, qui occupe à elle seule la colonne de droite, présente les mots simples qui ne sont pas reconnus par le dictionnaire (*unknown simple words*). Cette fonction est particulièrement utile dans la mesure où elle permet de repérer les éventuelles erreurs et coquilles et, dans le cas des langues anciennes, de repérer toutes les formes qui ne sont pas encore représentées dans le ou les dictionnaire(s) appliqué(s) au texte.

Une fois que le fichier a été analysé et que tous les mots ont été identifiés et classés, le texte peut être soumis à toutes sortes d'analyses. Il est possible de créer des moteurs de recherche fort complexes, pour identifier des expressions ou des séquences de mots, en appliquant les mêmes codes grammaticaux, sémantiques et flexionnels propres au formalisme DELA, en faisant appel aux informations contenues dans les dictionnaires du texte. L'utilisateur créera des « masques lexicaux » sous cette forme :

– <N> : reconnaît toutes les entrées qui ont le code grammatical N.
– <être> : reconnaît toutes les entrées qui ont être comme forme canonique.

La recherche suivante « <être> en <N> » permettra donc de repérer toutes les expressions comprenant une forme fléchie du verbe être suivie de la préposition en et d'un substantif (*suis en colère, étaient en retard,* etc.).

Il est possible de raffiner la recherche en combinant plusieurs codes grammaticaux ou sémantiques, séparés par les caractères + ou –. Par exemple le masque <N–z1> permet d'exclure de la recherche tous les substantifs les plus communs et de ne retenir que les substantifs appartenant au langage spécialisé (z2) et très spécialisé (z3). Les contraintes flexionnelles sont indiquées après deux points, par exemple le masque <A:mp:f> reconnaît un adjectif qui est soit au masculin pluriel soit au féminin.

[8] Cf. M. Gross, *Grammaire transformationnelle du français*, 3 : *Syntaxe de l'adverbe*, Paris, ASSTRIL, Université Paris 7, 1990.

Il est également possible d'utiliser des « méta–motifs », afin de distinguer, par exemple, entre les minuscules et les majuscules : <MIN>/<MAJ>, ou afin d'identifier les suites contiguës de chiffres : <NB>. De plus, il est possible d'appliquer des « filtres morphologiques », afin de repérer des caractères ou des suites de caractères en début, à la fin ou à l'intérieur des mots. Par exemple le code <<^a>> identifie un mot commençant par la lettre *a* ; <<ent$>> identifie un mot se terminant par la désinence *–ent* ; <<ss>> identifie une suite de deux *s* à l'intérieur d'un mot ; <<a.ss>> identifie un *a* suivi d'un caractère, suivi de *ss* ; <<ée?>>: contient *é* suivi par un *e* facultatif.

Cédrick Fairon propose enfin un exemple de graphe complexe qui prend en compte les différentes façons d'exprimer le concept lié à « raconter des blagues », y compris les expressions les plus haut en couleur :

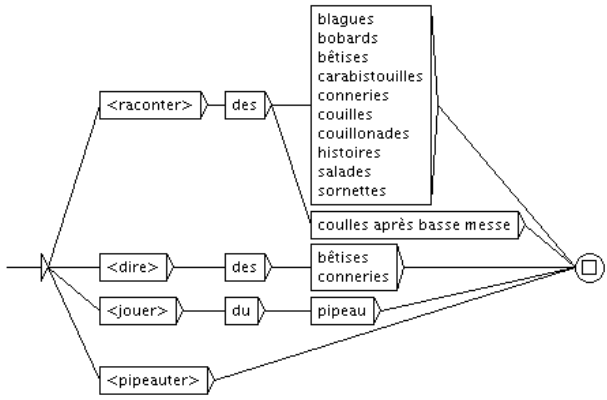


Figure 1 : exemple de graphe complexe UNITEX

Le graphe a l'avantage de présenter les différents types de parcours que le logiciel peut / doit accomplir afin de repérer, voire de recomposer, toutes les expressions acceptables du point de vue syntaxique, en associant notamment les verbes (*raconter, dire jouer, pipeauter*) aux articles partitifs (*des, du*) et aux substantifs (*blagues, bobards,* etc.) et aux expressions de saveur dialectale (*coulles après basse messe*). Une fois encore, le logiciel repèrera toutes les formes fléchies des verbes en question.

Ce système, extrêmement synthétique, fiable et performant a été adopté, entre autres, par le projet de recherche en lexicologie grecque (Institut Orientaliste de l'UCL), qui produit des concordances lemmatisées d'auteurs grecs, principalement des Pères de l'Église et des historiens byzantins, dont il sera question ci-dessous grâce à l'intervention de Bastien Kindt[9].

[9] Cf. aussi le corpus de textes latins traité par le Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège : http://www.cipl.ulg.ac.be/Lasla/

1.3. Bastien Kindt (UCL, Brepols Publishers) : UNITEX et autres outils centaliens pour le traitement et l'exploitation de corpus en grec ancien

Pour des raisons personnelles, Bastien Kindt n'a pas pu être présent à la journée, mais il a eu la bienveillance de me transmettre le présent résumé écrit de sa communication. Je le remercie donc tout particulièrement de sa compétence et de sa générosité.

Bastien Kindt, doctorant à l'Institut orientaliste de l'UCL et collaborateur scientifique de la maison d'édition Brepols Publishers, est attaché au *Projet de recherche en lexicologie grecque* (*PRLG*) dirigé à l'Institut orientaliste par le Professeur Bernard Coulie. Ce projet a pour vocation de produire des concordances d'auteurs grecs, principalement des Pères de l'Église et des historiens byzantins. Ces réalisations, entièrement lemmatisées et désambiguïsées, sont publiées dans le *Thesaurus Patrum Graecorum* (*TPG*), une série du *Corpus Christianorum*[10]. Deux versions différentes d'UNITEX sont utilisées dans le cadre de ces travaux et recherches.

D'abord, une *station de travail* permettant d'annoter lexicalement les *corpus* traités a été conçue à partir de la version 1.1. d'UNITEX. Puisque ce développement original est le fruit de la collaboration avec le CENTAL, il a été baptisé UNITEX_cental. Cette *station de travail* fonctionne avec un dictionnaire de grec ancien propre au projet, le *DAG* mis au format DELAF (cf. paragraphe précédent), et diverses ressources linguistiques qui lui sont associées. Ces dernières assurent le traitement automatique des particularités lexicales du grec ancien, principalement les crases et les formes élidées. Elles permettent également d'éliminer certaines ambiguïtés lexicales facilement résolues grâce à une analyse contextuelle locale des formes concernées. Cet outil, désormais à la base de la production des volumes du *TPG*, a déjà été décrit dans d'autres publications[11]. Quand un *corpus* est parfaitement traité, les textes et les informations de lemmatisation qui leur sont attachés migrent vers des bases de données.

---

[10] Sur le *P.R.L.G.*, voir B. Coulie, « La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs », *Byzantion*, 66, 1996, p. 35-54, ainsi que B. Kindt, « La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. Les principes de formulation des lemmes du Dictionnaire Automatique Grec (D.A.G.) », *Byzantion*, 74, 2004, p. 213-72. Voir aussi le site Internet du projet sous l'adresse http://tpg.fltr.ucl.ac.be (page consultée le 18 juin 2009). Sur le *T.P.G.*, voir B. Coulie, « Corpus Christianorum. Thesaurus Patrum Graecorum », dans *Corpus Christianorum 1953-2003. Xenium Natalicium. Fifty Years of Scholarly Editing*, éd. J. Leemans, Turnhout, Brepols, 2003, p. 169-72.

[11] Voir en particulier S. Deodati, B. Kindt, « La lemmatisation automatisée des sources en grec ancien : présentation de ressources linguistiques et d'outils de traitement », dans *Atti del XII Congresso Internazionale di Lessicografia*. *Proceedings XII EURALEX International Congress* (Torino, 6-9. Sept 2006), éd E. Corino, C. Marello, C. Onesti, vol. II, Alessandria, Edizioni dell'Orso, 2006, p. 1137-43.

Des générateurs de rapports et des logiciels de publication assistée par ordinateurs – autres développements réalisés par le CENTAL – puisent dans ces bases de données les éléments nécessaires à l'édition de concordances, d'index lemmatisés ou de listes spécialisées (index inverses, tables fréquentielles, etc.).

La version 1.2. d'UNITEX est utilisée pour élaborer et évaluer des outils de désambiguïsation lexicale et flexionnelle. Ces outils sont des règles (écrites par des experts humains) opérationnelles sous le module ELAG (Elimination of Lexical Ambiguities by Grammars) implémenté dans cette version d'UNITEX ainsi que dans les versions postérieures. Les lignes qui suivent illustrent brièvement l'utilisation faite de ELAG pour le traitement du grec ancien[12].

Techniquement, les règles utilisées sous ELAG sont des automates. Appliqués sur un *corpus*, elles réalisent une double opération de lecture et de réécriture. Par commodité, ces automates sont appelés des « grammaires », chacune définissant un ensemble concret de contraintes grammaticales qui doivent être respectées pour qu'une analyse lexicale et flexionnelle soit correcte. Comme l'illustre l'exemple de la figure 2, de telles grammaires présentent deux chemins : un *chemin supérieur* balisé de points d'exclamation (<!>) et un *chemin inférieur* ponctué de signes « égal » (<=>). Cette grammaire assume une partie du traitement de l'ambiguïté lexicale de la forme τοῦ. En grec ancien, ce petit mot répond à deux analyses : 1) un déterminant article au génitif masculin ou neutre singulier (<DET:Gms:Gns>) ; 2) un pronom interrogatif au génitif masculin, féminin ou neutre singulier (<PRO+Int:Gs>). La fréquence d'apparition de cette forme τοῦ est très élevée dans les textes. Dans un tel cas, un traitement automatique représente un gain de temps appréciable.

La grammaire doit être lue de la manière suivante : le *chemin supérieur* présente une condition (« SI ») ; le *chemin inférieur* présente une conséquence qui se réalise si les termes de la condition sont rencontrés *in textu* (« ALORS »).

[12] Sur ELAG, voir É. Laporte, A. Monceau, « Elimination of lexical ambiguities by grammars : the Elag system », *Lingvisticae Investigationes*, 22, 1998-99, p. 341-67 ; É. Laporte, « Reduction of lexical ambiguity », *Lingvisticae Investigationes*, 24/1, 2001, p. 67-103 ; A. Dister, *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*, Université catholique de Louvain, Thèse de Doctorat, Louvain-la-Neuve, 2007, p. 394-9. L'implémentation d'ELAG sous UNITEX a été réalisée par Olivier Blanc.

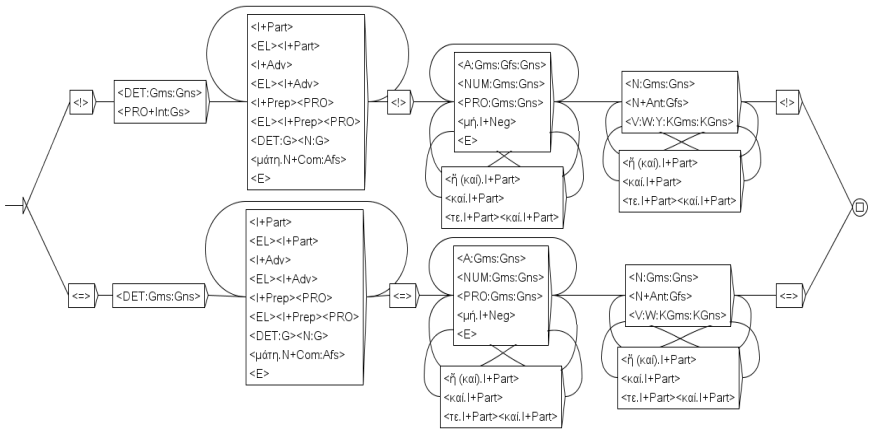


Figure 2 : la grammaire DET_Gmns_01.grf

SI, en parcourant le texte, la grammaire rencontre une forme pouvant être soit un déterminant-article au génitif masculin ou neutre singulier, soit un pronom interrogatif au génitif singulier, suivie d'abord, facultativement, d'une particule, d'un adverbe, d'un couple constitué d'une préposition et d'un pronom (ou les formes élidées de ces mots ou groupes de mots), suivie ensuite, toujours de manière facultative, d'un adjectif, d'un déterminant numérique ou d'un pronom (ces trois mots au génitif singulier) ou de la négation μή, et suivie enfin, de manière obligatoire cette fois, d'un nom au génitif masculin ou neutre singulier, d'un verbe (un infinitif, un impératif ou un participe au génitif singulier), …

ALORS, le mot à la tête du syntagme est un déterminant-article (et non le pronom interrogatif) et le dernier mot de la séquence ne peut être qu'un nom ou qu'un verbe à l'infinitif ou au participe, au génitif singulier, masculin ou neutre (et non à l'impératif).

En d'autres termes, cette grammaire formalise la structure des syntagmes nominaux au génitif masculin ou neutre singulier. Ce syntagme est borné en amont par un article et en aval par un nom ou un verbe au participe ou à l'infinitif (puisqu'en grec ancien l'infinitif connaît, accompagné de l'article, des emplois nominaux). Ainsi, le chemin supérieur de la grammaire reconnaît des séquences telles que τοῦ ανδρός / de l'homme, τοῦ μεγάλου Δίος / du grand Zeus, τοῦ Ααρων / d'Aaron, τοῦ Ελληνίζειν / du fait de parler grec (ou d'être païen), τοῦ σπειράντος / de celui qui sème, τοῦ εν ημιν Πνεύματος / de l'Esprit qui est en nous. Dans tous ces exemples, l'analyse pronominale de la forme initiale τοῦ est à exclure. Dans sa phase de réécriture, la grammaire enlève cette possibilité. Les grammaires ELAG sont mises en œuvre sur des textes prétraités, après l'application des dictionnaires et des grammaires de normalisation (voir ci-dessus, la présentation de Cédrick Fairon). Elles interviennent non pas sur le texte mais sur l'automate du texte. Les données de départ sont donc conservées et demeurent, tout comme les résultats du traitement, visualisables par l'utilisateur.

La grammaire présentée ci-dessus, nommée DET_Gmns_01.grf, est la première d'un jeu de six grammaires portant chacune le même nom, mais indicées de _01 à _06, et utilisées pour la désambiguïsation lexicale et flexionnelle des formes τοῦ rencontrées dans les textes. Ce jeu de grammaire, nommé DET_τοῦ, a été développé et appliqué sur un *corpus d'expérimentation*, un *sous-corpus* des œuvres de Grégoire de Nazianze, Père de l'Église du IV^e^ s. ap. J.-C.[13] Le tableau 1 fournit les résultats de l'application de ce jeu de grammaires sur ce dernier corpus d'expérimentation. Pour les besoins de l'expérimentation, le *corpus* complet a été divisé en deux parties : un *corpus de construction* — exploité exclusivement afin d'écrire et d'expérimenter les grammaires ; un *corpus d'évaluation* — qui a servi à en évaluer l'efficacité.

<DET: Gms:Gns>	**A Occurrences**	**B Occ. traitées**	**C Traitements corrects**	**D Trait. erronés**	**Rappel**	**Précision**
Corpus de construction	189	177	177	0	93,65%	100%
Corpus d'évaluation	198	173	172	1	87,37%	94,42%
***Corpus* complet**	387	350	349	1	90,43%	99,71%

Tableau 1 : évaluation du jeu de grammaires DET_ τοῦ

L'efficacité de tels outils s'évalue en termes de *Rappel* et de *Précision*. Dans le cas illustré ici (la désambiguïsation de la forme τοῦ), le *Rappel* est le pourcentage de formes τοῦ traitées par le jeu de grammaires (B) par rapport au nombre de toutes les formes τοῦ réellement présentes dans le texte (A). La *Précision* est la proportion de formes correctement traitées (C) par rapport aux formes prises en compte par le jeu de grammaires (B). Dans le cadre d'un traitement automatique visant à lemmatiser les textes, le *Rappel* peut ne pas être optimal, un résidu pouvant toujours être traité manuellement. Par contre, tout résultat fautif est prohibé. En d'autres termes, le philologue qui supervise le traitement tolèrera du *silence* (l'absence de décision par le jeu de grammaires), mais pas le *bruit* (la formulation d'une réponse inadéquate). Dans ce contexte, on comprend que le soin apporté à la conception des grammaires revêt une importance particulière. Dans le cas présent, un résultat est considéré comme satisfaisant si l'interprétation pronominale de la forme τοῦ est écartée de

[13] Ce corpus expérimental comprend six textes du Nazianzène et totalise 30129 mots–occurrences regroupant 9396 formes simples différentes. L'analyse lexicale du *corpus* complet des œuvres de Grégoire de Nazianze a inauguré la série du *TPG* ; voir J. Mossay et CETEDOC, *Thesaurus Sancti Gregorii Nazianzeni*, vol. I. *Enumeratio lemmatum, Orationes, Epistulae, Testamentum* (*Corpus Christianorum*. *Thesaurus Patrum Graecorum*), Turnhout, Brepols, 1990 ; J. Mossay, B. Coulie et CETEDOC, *Thesaurus Sancti Gregorii Nazianzeni*, vol. II. *Enumeratio lemmatum, Carmina, Christus Patiens, Vita* (*Corpus Christianorum*. *Thesaurus Patrum Graecorum*), Turnhout, Brepols, 1991.

l'automate du texte. Un traitement partiel, mais éliminant l'étiquette < τοῦ, τίνος:Gms:Gfs:Gns> est donc considéré comme correct.

Le taux de *Rappel* obtenu pour le *corpus de construction* (93,65%) est plus élevé que celui du *corpus d'évaluation* (87,37%). Ce résultat était attendu puisque les grammaires ont toutes été écrites pour résoudre les ambiguïtés lexicales observées dans ce *corpus de construction*. La *Précision* est optimale dans le premier cas, pour les mêmes raisons[14].

Ces outils de désambiguïsation viennent en appui au travail de lemmatisation supervisé par des experts. La priorité est donnée à l'analyse lexicale. Mais les informations flexionnelles ne sont pas négligées pour autant car leur prise en compte dans les traitements permet d'améliorer l'analyse des données purement lexicales[15]. Le tableau 2 indique le nombre de formes τοῦ ayant fait l'objet d'un traitement simplement lexical (élimination, dans l'automate du texte, de l'étiquette pronominale <PRO+Int:GS>) ou d'un traitement tant lexical que flexionnel (rationalisation de toutes les étiquettes flexionnelles du syntagme). Ces chiffres complètent ceux de la colonne D (*Traitements corrects*) du tableau 1.

Traitements corrects (détails)	***Corpus de construction***	***Corpus d'évaluation***	***Corpus* complet**
Lexical	62	60	122
Lexical et flectionnel	115	112	227

Tableau 2 : effectifs des traitements lexicaux et flexionnels
effectués par le jeu de grammaires DET_τοῦ

Bastien Kindt utilise actuellement une série de 98 grammaires. Appliquées sur le *corpus d'expérimentation*, elles lèvent plus de 35% des ambiguïtés du texte. Les interfaces d'UNITEX et de ELAG permettent aux linguistes de « dessiner » des grammaires sous forme de graphes. Il n'est donc pas nécessaire de connaître un langage de programmation particulier. La compilation des grammaires — leur transformation en automates exploitables par l'ordinateur — est assurée par UNITEX. À chaque étape de son travail, l'utilisateur peut visualiser « en clair » les données littéraires et les informations grammaticales qu'il manipule. Les développements futurs devraient donc poursuivre un triple objectif : 1) compléter les jeux de grammaires ; 2) accroître la taille du *corpus d'expérimentation*[16] ; 3) faire évoluer UNITEX_cental vers les versions plus récentes d'UNITEX (1.2., 2.0 et 2.1) afin que la *station de travail* utilisée pour la lemmatisation des textes grecs puisse directement tirer profit des outils d'analyse mis en œuvre sous ELAG.

[14] Il existe bien évidemment des cas, dont il ne sera pas question ici, qui ne sont pas traités ou qui engendrent des erreurs dans le système.

[15] Voir la conclusion d'un article antérieur : L. Kevers, B. Kindt, *Traitement automatisé de l'ambiguïté lexicale en grec ancien. Première approche par application de grammaires locales*, *Lingvisticae Investigationes*, 28, 2 (2005), p. 251.

[16] Dans cette optique, Saulo Delle Donne (Université de Lecce) effectue le même travail que Bastien Kindt sur le *Hiéron* de Xénophon et sur le *Solon* de Plutarque. Enrichi de ces deux textes, le *corpus d'expérimentation* passera de 30129 à 44.581 mots-occurrences.

2. Logiciels pour la mise en page des éditions imprimées

La deuxième séance, présidée par Baudouin Van den Abeele, chercheur qualifié du FNRS et professeur à l'UCL, comprend la présentation de deux logiciels, LaTeX et Classical Text Editor, déjà largement diffusés dans le domaine de l'édition des textes. Les deux outils présentent de nombreux points en commun et offrent des solutions très intéressantes en matière d'ecdotique et en particulier dans la gestion et dans la présentation des apparats critiques à plusieurs étages. Tous les intervenants dans cette séance sont impliqués dans des projets d'éditions portant sur des textes encyclopédiques médiévaux et sont rattachés, à titres différents, aux départements d'études romanes, d'histoire et d'études classiques et orientales de l'UCL.

2.1. Laurent Brun (U. de Stockholm) LaTeX – *L'édition du* Miroir historial *de Jean de Vignay.*

Après une thèse de doctorat consacrée au *Miroir historial* de Jean de Vignay à l'Université de Stockholm, Laurent Brun prend ses fonctions en juillet 2009 à l'Université d'Ottawa. Ses champs de recherche privilégiés sont la philologie et l'édition de textes, la traduction inter– et intralinguale, la littérature médiévale romanesque et encyclopédique. Il est le concepteur et le responsable du site bibliographique ARLIMA (Archives de Littérature du Moyen Âge). Il prépare actuellement la première édition intégrale du *Miroir historial*, traduction française du *Speculum historiale* de Vincent de Beauvais, réalisée par Jean de Vignay autour des années 1320-1330[17]. L'édition est effectuée avec LaTeX, si bien que la présentation offerte par Laurent Brun repose sur une pratique très approfondie du logiciel.

LaTeX est un système de création de documents libre, gratuit et polyvalent qui, à l'instar du langage XML, nécessite l'utilisation d'un éditeur de texte fournissant une interface. Créé en 1984 par Leslie Lamport sur la base de TeX, LaTeX est né du besoin des ingénieurs et mathématiciens de disposer d'un logiciel permettant des mises en page à la fois très complexes et très précises, ce que les logiciels de traitement de texte traditionnels ont généralement beaucoup de difficulté à offrir. Tandis que les possibilités infinies du logiciel sont bien connues dans le monde des sciences dites dures, son existence est à peu près inconnue dans celui des lettres et sciences humaines.

---

[17] Je codirige, avec Laurent Brun, ce projet d'édition qui aboutira tout d'abord à une publication dans la collection de la Société des Anciens Textes Français. Nous bénéficions actuellement de la collaboration de Nathalie Bragantini-Maillard, boursière post-doc à l'UCL, qui travaille à l'édition des quatre premiers livres de l'encyclopédie. Pour les détails philologiques du projet, cf. L. Brun, M. Cavagna, « Pour une édition du *Miroir historial* de Jean de Vignay », *Romania*, 124, 2006, p. 378-428 et « Das *Speculum historiale* und seine französische Übersetzung durch Jean de Vignay », dans *Übertragungen, Formen und Konzepte von Reproduktion im Mittelalter und früher Neuzeit*, Actes du colloque de Göttingen, (juin 2004), éd. B. Bussmann *et alii*, Berlin – New York, De Gruyter, 2005, p. 279-302 [Trends in Medieval Philology, 5].

Les éditeurs de texte conseillés pour l'utilisation de LaTex sont TeXMaker (pour Windows, Mac, Linux), TeXShop (Mac) et Kile (Linux), mais il est tout à fait possible d'utiliser d'autres éditeurs de fichiers texte comme Notepad, BBEdit, vim[18]. L'avantage des éditeurs spécialisés comme TeXMaker, TeXShop et Kile est qu'ils permettent de transformer en un seul clic de la souris le document LaTeX en fichier PDF, tandis que, pour les éditeurs basiques, il faut plutôt passer par une interface en ligne de commande.

Le document LaTeX comporte deux parties: 1. un en-tête définissant la plupart de ses caractéristiques, à savoir le type de document qu'on souhaite réaliser (livre, article, etc.), le format de la page, le caractère et la taille de la police choisie, ainsi que les modules (*packages*) qu'on souhaite utiliser pour des utilisations plus spécifiques. Ces modules constituent en fait des collections de macros créées pour des fonctions précises, par exemple le choix d'un type de police, la création de colonnes et d'index ou encore la mise en page d'un apparat critique. 2. le corps du document, commençant par \begin{document} et se terminant par \end{document}.

Laurent Brun présente d'abord un exemple de structure basique pour la réalisation d'un livre, comportant une organisation hiérarchique en chapitres et en sections :

```
\documentclass[a4paper,12pt]{book}
\usepackage{ledmac}
\begin{document}
\chapter{Titre du chapitre}
texte
\section{Titre de la section}
texte
\subsection{Titre de la sous-section}
texte
\subsubsection{Titre de la sous-sous-section}
texte
\end{document}
```

Il existe plusieurs centaines de commandes standard qui permettent de définir et modifier la structure du document et la mise en forme du texte. Il existe plusieurs manuels d'instructions, également disponibles gratuitement sur plusieurs sites internet qui répertorient et expliquent toutes ces commandes. Voici quelques exemples pour la modification de l'apparence des caractères d'un texte :

`\textbf{texte...}` > **texte en gras**
`\textit{texte...}` > *texte en italique*
`\textsc{texte...}` > TEXTE EN PETITES CAPITALES
`\emph{texte...}` > texte en *italique* ou en romain en fonction de l'environnement

[18] Voici le principal site de référence conseillé par L. Brun : http://www.latex-project.org.

Laurent Brun insiste sur le fait que, même si LaTeX offre une énorme quantité de commandes diverses pour la mise en page des documents, il est aussi possible, voire essentiel, de personnaliser l'utilisation de LaTeX en créant ses propres commandes en fonction des exigences du texte traité et de l'utilisation envisagée. Ainsi, au lieu d'utiliser des commandes qui décrivent le résultat de la mise en forme du document (gras, italique, petites capitales, etc.), il est vivement conseillé de créer des commandes qui décrivent la nature de l'objet sur lequel une mise en forme particulière sera ensuite appliquée.

Ainsi, pour prendre un exemple simple, au lieu d'utiliser la commande \emph{...} pour mettre les titres d'ouvrages cités en italique, il est beaucoup plus judicieux de créer une commande \titre{...} qui effectuera la mise en forme de notre choix. Cela a un double avantage :

1. Au cas où l'on déciderait de changer la mise en forme de l'élément en question (par exemple, souligner les titres plutôt que les mettre en italique), il suffit simplement de modifier la définition de la commande dans le préambule. Comme la commande \emph{...} peut servir à mettre en italique bien d'autres choses que des titres, on s'épargne ainsi la peine de devoir vérifier si chaque commande \emph{...} s'applique à un titre ou non.

2. Si l'on veut transformer le document en un autre format (HTML, XML, etc.), il est alors très facile de le faire, car on aura ainsi clairement indiqué la structure du document et la nature de ses éléments.

Pour prendre un exemple encore plus concret et lié à la philologie, voici un exemple de commande créée pour indiquer le sigle d'un manuscrit :

```
\newcommand*{\ms}[1]{\textbf{#1}}
```

La commande \newcommand{...} permet de créer la balise \ms{...} que le philologue utilisera librement dans la création de son document, notamment dans l'introduction à son édition, mais aussi, éventuellement, dans son apparat critique. La commande \textbf{...} définit ici la présentation du texte compris dans les accolades, qui apparaîtra donc en gras dans le document PDF :

```
Source :      Le manuscrit \ms{Or1} présente des traits
linguistiques…
```

Document PDF : Le manuscrit **Or1** présente des traits linguistiques…

La création de ces commandes intervient principalement sur les paramètres typographiques, mais elle offre également la possibilité d'ajouter ou de supprimer des éléments textuels. Voici, par exemple, une commande indiquant le changement de feuillet dans un manuscrit :

```
\newcommand*{\folio}[1]{\textbf{[f.~#1]}}

Source:       Et il \folio{207va} traversa Asie
```

PDF : Et il **[f. 207va]** traversa Asie

L'utilisation de la commande \folio{...} permet d'insérer automatiquement, dans la version à imprimer, les deux crochets carrés avant et après, l'abréviation « f. » ainsi qu'un espace insécable (indiqué à l'aide du signe « ~ » en LaTeX) en plus de mettre le texte en gras. Encore une fois, il est possible de modifier à tout moment ce genre de paramètres de façon rapide et systématique afin, par exemple, de répondre aux critères de la collection qui accueillera le texte édité (éliminer le gras, noter la mention « fol. », remplacer les crochets par des parenthèses, etc.).

L'inconvénient de ce système est lié au fait qu'il n'existe pas encore d'équivalent de la TEI pour LaTeX, à savoir un organisme qui gèrerait et superviserait la création et la gestion de commandes standardisées, si bien que chaque document LaTeX aura un jeu de balises qui lui sont propres. Toutefois, le système est parfaitement cohérent et toutes les commandes créées *ex novo* par l'utilisateur sont déclarées dans l'en-tête du document, ce qui permet de comprendre rapidement le système employé par le créateur d'un document donné.

Au-delà des commandes de base définissant la structure du document et la mise en forme du texte et celles que l'on peut créer soi-même, il existe des commandes parfois beaucoup plus complexes, qui sont contenues à l'intérieur de modules (*packages*) et qui offrent justement des palettes de commandes répondant à des besoins variés. Parmi ces modules, Laurent Brun présente Ledmac, conçu pour la mise en page d'éditions de textes et notamment pour la création et la gestion des apparats critiques sur plusieurs niveaux et renvoyant automatiquement aux numéros de ligne, que le texte soit en vers ou en prose[19].

Pour la création de l'apparat critique, tous les paramètres doivent être définis dans l'en-tête. L'utilisateur déclare notamment les modules et les commandes qu'il souhaite utiliser. Le module Ledmac met à disposition la commande \edtext{...}{...}, qui définit le lieu variant (ici appelé « lemme »), lequel sera repris dans le ou les apparats en bas de page et qui, à l'aide des commandes \Afootnote, \Bfootnote, \Cfootnote, etc., permet de spécifier à quel étage pour définir autant d'étages que l'on veut à l'intérieur de l'apparat présenté en bas de page. Laurent Brun insiste sur la possibilité de personnaliser ce système en remplaçant, tout simplement, la balise commande par une commande qui décrit la nature des notes présentées sur un étage donné. Voici, par exemple, le système adopté pour l'édition du *Miroir historial* :

```
Préambule :

\documentclass[a4paper,12pt]{book}
```

[19] Il existe d'autres modules (par exemple, poemscol et ednotes) qui peuvent répondre aux besoins des philologues, mais Laurent Brun a choisi ici de présenter celui qui, selon lui, est le plus achevé et répond le mieux à presque tous les besoins en matière de mise en page d'un apparat critique. Il signale également que Ledmac est accompagné de deux autres modules optionnels : Ledpar, qui sert à mettre en page des éditions parallèles (par exemple, texte original avec traduction en regard), et Ledarab, qui offre la possibilité de mettre en page des éditions de textes adaptées aux exigences typographiques de la langue arabe.

```
\usepackage{ledmac}
\let\rj=\Afootnote
\let\var=\Bfootnote

Source : Et prist \edtext{de son suegre}{\rj{du pere sa
femme (J)}} la \edtext{cure}{\var{\emph{om.} A}} de
nourrir ses bestes
```

PDF : texte : 10 et prist de son suegre la cure de nourrir ses bestes.

Apparat A : 10 de son suegre] *ms* : du pere de sa femme (J)

Apparat B : 10 cure] *om*. A

Dans l'en-tête du document, les commandes par défaut \Afootnote et \Bfootnote ont été remplacées par \rj et \var, deux commandes plus synthétiques indiquant respectivement la leçon rejetée du manuscrit de base – qui est présentée au premier étage de l'apparat – et la variante d'un autre témoin, qui est insérée à l'étage inférieur. Les deux commandes sont insérées dans le corps du texte, à la suite de la commande \edtext{...}{...}, qui définit le lemme repris tel quel en bas de page. Dans cet exemple, l'éditeur a choisi de rejeter la leçon *du pere de sa femme*, offerte par son manuscrit de base, et de la remplacer par *de son suegre*, conservée dans le manuscrit J (signalé entre parenthèses), qui est visiblement une *lectio difficilior*[20]. Cette correction est signalée à l'étage supérieur de l'apparat. Le manuscrit A omet le terme *cure*, si bien que la variante (indiquée à l'aide de la commande \var) est présentée à l'étage inférieur de l'apparat (sans parenthèse).

Au niveau ecdotique, le résultat est certainement tout aussi satisfaisant que celui qu'on peut obtenir avec Classical Text Editor (dont il sera question ci-dessous), puisque la numérotation des lignes et l'alignement du texte et de l'apparat sont gérés de façon automatique. Il faut quand même noter que, comme on le voit ci-dessus, le texte dans le document de travail (source) est moins lisible, même si, dans tous les éditeurs de LaTeX, toutes les commandes sont colorées et ressortent assez clairement par rapport au texte. En revanche, la distinction entre le texte et l'apparat critique est beaucoup moins claire et nette, contrairement à CTE, par exemple, qui présente l'un et l'autre dans sa fenêtre propre.

Par rapport à CTE, LaTeX se distingue en outre par sa flexibilité puisqu'il offre à l'utilisateur la possibilité de définir, de modifier et de configurer tous les paramètres concernant à la fois la source (création, substitution et gestion des commandes) et le document de sortie (le plus souvent un fichier PDF). LaTeX permet de composer des documents à la mise en page très complexe et, surtout, ces documents peuvent être de diverses natures, allant de l'article bref à une édition critique en plusieurs volumes en passant par la thèse de doctorat ou le manuel technique. Un autre avantage non négligeable de LaTeX est la possibilité est son

[20] Le terme *suegre*, traduit littéralement le latin *socero* (à l'ablatif) « beau-père », qui est justement son étymon [FEW XII, 15b *socer*].

interopérabilité relativement aisée avec d'autres langages et logiciels courants, par exemple:

- convertir en XML ou en tout autre langage balisé (et vice-versa) ;
- mettre en forme en LaTeX des données extraites d'une base de données et converties à l'aide de simples scripts composés en Perl ou PHP ;
- modifier un document LaTeX à l'aide d'« expressions rationnelles », qui offrent des possibilités extrêmement avancées de recherche et de remplacement.

2.2. Iolanda Ventura et Sébastien Moureau (UCL), Classical Text Editor – expériences d'édition de textes arabes et latins

Iolanda Ventura bénéficie actuellement d'une bourse de recherche post-doctorale à l'UCL et étudie la transmission du savoir médical entre les encyclopédies et les traités scientifiques. Spécialiste de philologie médiolatine, en particulier de la production littéraire scientifique, elle collabore à plusieurs projets internationaux, en lien avec l'Atelier Vincent de Beauvais de Nancy, avec les Universités de Münster et de Salerne. Elle est co-responsable, entre autres, de l'édition critique du *De proprietatibus rerum* de Barthélémy l'Anglais, dont le premier volume, réalisé avec Classical Text Editor, est sorti en 2007[21].

Sébastien Moureau fait une thèse de doctorat à l'UCL portant sur l'édition, la traduction et le commentaire du *De anima in arte alchemiae*, texte faussement attribué à Avicenne. Ayant une connaissance approfondie à la fois du moyen arabe et du latin, Sébastien Moureau s'intéresse aux encyclopédies médiévales et à la transmission du savoir arabe dans l'Occident médiéval.

À l'instar de LaTeX, Classical Text Editor (dorénavant CTE) est un outil très performant, surtout au niveau ecdotique, et permet de gérer automatiquement des apparats critiques fort complexes et stratifiés. Il permet en outre de gérer simultanément plusieurs colonnes du texte – par exemple dans l'affichage d'un texte et de sa traduction, ou de plusieurs variantes du même texte – avec en plus des apparats de gloses marginales. À l'instar de LaTeX, il est conçu pour générer des documents en format PDF prêts pour être donnés aux imprimeurs (impression en *camera-ready*), mais depuis quelques années il permet aussi de créer des documents en format HTML pour la publication sur le web ou en format XML pour des traitements électroniques plus élaborés. La transformation en XML se fait à travers l'insertion des balises TEI.

Le logiciel est payant, mais une version d'essai est téléchargeable gratuitement et est exploitable sans limite de temps à tous les niveaux, sauf pour la publication des documents[22].

La démonstration pratique offerte par Iolanda Ventura et Sébastien Moureau est très éclairante et s'ouvre sur la définition des feuilles de style (appelées ici

---

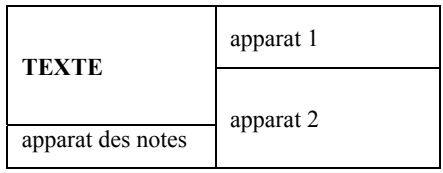
[21] Bartholomaeus Anglicus, *De proprietatibus rerum*. Vol. VI : Liber XVII, éd. I. Ventura, Turnhout, Brepols, 2007 [De diversis artibus 79, n.s. 42] ; cf. aussi le volume précédent : Bartholomaeus Anglicus, *De proprietatibus rerum*. Vol. I : Libri I-IV, éd. B. Van den Abeele *et alii*., Turnhout, Brepols, 2007 [De diversis artibus 78, n.s. 41].

[22] Une version de démonstration du logiciel est disponible sur le site suivant : http://www.oeaw.ac.at/kvk/cte/. On y trouve d'excellentes instructions pour l'utilisation du programme, ainsi qu'une liste, régulièrement mise à jour, des éditions réalisées avec CTE.

templates) qui permettent de définir tous les paramètres du document, tels que la taille du texte, la police, les caractères et le sens de l'écriture, le miroir de la page (les marges sont calculées par rapport au format de papier choisi), le nombre d'apparats, etc.

La première différence qui saute aux yeux, par rapport à LaTeX, est le caractère « rassurant » de l'interface qui permet de gérer ces paramètres à travers un système d'onglets, de cases à cocher et de menus déroulants, système qui ressemble beaucoup à celui du traitement de texte Word. L'autre différence est l'utilisation du multi-fenêtrage qui présente plusieurs avantages, notamment dans la gestion simultanée des différentes composantes de l'apparat critique et des notes.

L'exemple de document proposé par Iolanda Ventura et Sébastien Moureau comporte un double apparat critique plus un apparat de notes. À l'écran sont affichées quatre fenêtres : le texte critique occupe les deux tiers en largeur et les trois quarts en hauteur, sur la droite, le troisième tiers de l'écran est occupé par les deux fenêtres accueillant les deux étages de l'apparat critique ; en bas, une petite fenêtre accueille les notes critiques[23] :



Une fois ces paramètres définis, on procède à l'importation d'un extrait de texte, à partir d'un autre environnement. Cette opération peut être effectuée soit d'une façon automatique – la touche *import* permet d'effectuer une recherche dans les répertoires de l'ordinateur – soit à travers un simple copier-coller. Le logiciel est compatible avec les traitements de texte les plus répandus, notamment avec Word, et supporte le format UTF-8.

Le texte importé doit être structuré avec l'insertion d'une série d'identifiants (de chapitres, de paragraphes, de sous-sections). Ceux-ci n'apparaîtront pas dans la version de sortie, mais sont bien visibles – ils sont marqués en jaune – dans la version de travail.

La création de l'apparat critique passe à travers l'encodage des sigles des manuscrits qui permet ensuite l'automatisation de différentes fonctions dans le traitement des variantes. L'insertion d'une variante dans l'apparat est particulièrement simple et efficace : il suffit de sélectionner dans le texte le lieu variant et dans une fenêtre latérale, l'entrée d'apparat (lemme) est générée de façon automatique. Pour intervenir sur le lemme, il faut passer par des commandes qui gèrent sa présentation, par exemple, lorsqu'il comprend plusieurs mots, le logiciel en supprime automatiquement quelques-uns en les remplaçant par des petits points. L'utilisateur peut intervenir sur cette réduction automatique du lemme en opérant d'autres choix.

Iolanda Ventura attire l'attention sur une fonction particulièrement utile, permettant de repérer les éventuelles ruptures de lien entre le texte et l'apparat

[23] L'organisation de ces fenêtres est très facile à configurer.

critiques, qui peuvent être provoquées par des interventions successives sur le texte. Une double croix apparaît immédiatement afin de signaler cette rupture. La rupture peut être réparée dans l'immédiat, mais il est également possible d'intervenir par la suite, avec un moteur de recherche permettant de repérer toutes les erreurs éventuelles.

CTE permet de gérer la synchronisation automatique entre deux fichiers de texte, par exemple pour présenter le texte et sa traduction en langue moderne. Il est possible de choisir entre plusieurs options de mise en page : synchronisation verticale ou horizontale, sur plusieurs colonnes dans une même page ou sur deux pages qui se font face. De même, il est possible de choisir entre une synchronisation automatique, basée sur la division en chapitres ou en paragraphes ou l'insertion d'indicateurs manuels, dans le corps du texte. Cette fonction est également utile pour présenter les apparats de gloses ou des notes marginales.

Le programme propose également la création d'index sous forme de documents texte (éditable avec n'importe quel éditeur de texte). Trois méthodes fondamentales sont proposées, qui peuvent également être combinées l'une avec l'autre. La première consiste en l'indexation de tous les mots du texte, sans discernement. La seconde, plus intéressante, fonctionne sur un système de lemmatisation et de sélection des termes à indexer au moyen d'un fichier texte dans lequel l'utilisateur insère les mots qu'il veut voir apparaître dans l'index (avec possibilité de troncature). Ainsi, une ligne « indur* indurare » permettra au programme de reprendre tous les termes commençant par *indur–* sous l'entrée *indurare*. La troisième technique est l'encodage manuel d'entrées d'indexation, semblable aux systèmes utilisés dans les éditeurs de texte les plus connus (comme Word). La création de l'index peut porter sur les différentes parties du texte (corps de texte, apparat, notes, etc.), avec possibilité de les combiner.

Sébastien Moureau souligne la possibilité d'utiliser différents jeux de caractères et en particulier différents types d'écritures. La barre des langues doit être configurée à la fois dans le système d'exploitation et dans le programme. Une fonction spécifique permet de régler l'orientation du texte et aussi de la modifier à l'intérieur du document, par exemple pour insérer des citations arabes dans un texte en langue romane ou vice-versa. Il est également possible d'utiliser le logiciel Multikey, qui est parfaitement compatible. La coupure des mots peut être gérée (manuellement ou automatiquement) à travers la commande *hyphenation*, qui permet de définir préalablement la coupure en syllabes selon la langue du texte (toutes les langues ne sont pas encore prises en charge).

Avec un point de vue très pragmatique, parfaitement en ligne avec l'esprit de la journée, Iolanda Ventura et Sébastien Moureau proposent enfin une liste d'avantages et d'inconvénients basée exclusivement – ils tiennent à le préciser – sur leur propre expérience et sur leur utilisation personnelle du logiciel.

Avantages :

- Le logiciel est constamment amélioré et de nombreuses mises à jour sont proposées gratuitement à tous les utilisateurs.
- Le concepteur et responsable du logiciel, Stefan Hagen, philologue classique de formation, est très disponible à l'égard des utilisateurs et fournit un support technique de grande valeur.

- Système de sauvegarde automatique. Conseil : sauvegarder toutes les étapes de l'édition.
- L'interface utilisateur est plutôt agréable et, en raison du fait qu'elle se rapproche de Word, assez rassurante. Le logiciel est relativement simple à utiliser et opère des copies de sauvegarde automatique.
- Le système du fenêtrage multiple facilite la visualisation de l'apparat critique en tant qu'unité compacte, mais en revanche, il comporte des risques de rupture de liens (risques pourtant limités par le système de vérification automatique).

Limites :

- Risques de plantages, sauvegardes assez lentes, ce qui peut provoquer des états d'angoisse (conseil : laisser le temps qu'il faut, sans bloquer).
- Flexibilité limitée. Même si les possibilités de mise en page sont très nombreuses et variées, l'utilisateur est lié au formatage du logiciel et aux paramètres préétablis. En outre, les possibilités sont parfois limitées au niveau du traitement du texte et de la compatibilité avec les autres logiciels. Cependant, la récente fonction d'exportation en format XML/TEI pallie cette limitation et permet l'utilisation d'outils informatiques complexes.
- Le logiciel est disponible seulement en anglais et ne permet pas de personnaliser la barre d'outils. Les utilisateurs de MacOs et de Linux doivent utiliser un émulateur Windows.

3. Logiciels pour le traitement des images et de la parole

Avec la troisième séance, présidée par l'organisateur de la journée, le champ de l'enquête est élargi et partiellement réorienté. D'un côté, l'attention est concentrée sur le support matériel du texte ancien et médiéval, à savoir le manuscrit, qui est considéré tout particulièrement dans ses composantes codicologiques et para–textuelles ; de l'autre, il est question d'aller au-delà du support et même du texte, pour concentrer l'analyse sur la parole et le traitement des données orales. La présentation qui clôt la journée porte sur un logiciel qui permet de traiter simultanément la notation textuelle et la notation musicale notamment pour les livrets et les partitions des opéras.

Le présent document montre ici toutes ses limites puisque au moment où les données textuelles cèdent la place, ou plutôt se voient enrichies, par les données visuelles, sonores et musicales, le support papier s'avère fort inadéquat. Le caractère plus synthétique des résumés qui suivent tient exclusivement à cette circonstance. J'essaierai tout de même de rendre compte, dans la mesure du possible, de la richesse des outils informatiques proposés, tout en renvoyant, comme je l'ai fait pour les précédents, aux sites Internet de référence.

3.1. Peter Ainsworth (U. de Sheffield), Virtual Vellum – Le traitement des images dans quelques manuscrits contenant les Chroniques de Jean Froissart

Peter Ainsworth est directeur du projet « Froissart en ligne » au Département de français de l'Université de Sheffield. Spécialiste de littérature historiographique du Moyen Âge, il est l'éditeur des *Chroniques* de Jean Froissart, responsable d'une

nouvelle édition du troisième Livre de celles-ci dont le premier tome est paru chez Droz en 2007 (collection des TLF) et d'une édition bilingue réalisée en collaboration avec George Diller et Alberto Varvaro, parue en deux volumes dans la collection Lettres Gothiques (2001 et 2004). Il collabore aussi à un projet reliant le *Dictionnaire du Moyen Français* (laboratoire ATILF, Université de Nancy 2), le *Froissart en-ligne* et le *Queen's Manuscript* (manuscrit des œuvres de Christine de Pizan à la British Library, Université d'Edimbourg) et subventionné par la British Academy et le CNRS.

Depuis plusieurs années, Peter Ainsworth dirige une campagne de numérisation des manuscrits enluminés contenant les *Chroniques* de Jean Froissart. Grâce à des conventions établies avec plusieurs bibliothèques d'Europe (BNF, KBR, Bibliothèques municipales de Toulouse et Besançon, Bibliothèque de Stonyhurst Collège, Lancashire) et grâce à la collaboration d'un photographe professionnel, il a rassemblé une banque d'images de très grande qualité, conservée au format TIFF (Tagged Image File Format, fichiers de 150MO en moyenne) et convertie par la suite au format JPEG2 (fichiers de 10MO environ) pour les exploitations scientifiques évoquées ci-dessous.

Outil libre et gratuit, *Virtual Vellum* est un logiciel permettant de visualiser et d'exploiter au maximum ces images[24]. Au contraire des logiciels génériques et de grande diffusion (genre *PowerPoint* de Microsoft), *Virtual Vellum* – conçu par des médiévistes, pour les médiévistes – permet, à travers un système de fenêtrage multiple, l'affichage synoptique de plusieurs documents numérisés aujourd'hui conservés dans des bibliothèques fort éloignées les unes des autres, mais relevant d'une source originale commune (ateliers de copistes et d'artistes à Paris). L'orientation des fenêtres peut être verticale – orientation idéale pour la comparaison des miniatures – ou horizontale, pour faciliter une lecture comparée des textes ou un travail de collationnement. Un manuel est à la disposition de l'utilisateur qui peut télécharger le logiciel sans frais et sans adjonction d'autres 'plug-in', exception faite toutefois pour *Java Runtime*.

L'un des points forts de ce logiciel est la souplesse de son utilisation et le fait qu'il permet de gérer facilement et rapidement des images d'un très grand format qui seraient très difficiles à manier dans un autre type d'environnement. Il permet d'agrandir les images à un tel degré de précision qu'il permet de saisir des détails difficilement perceptibles à l'œil (grattages, corrections ou indications pour le rubriqueur ou pour l'enlumineur qui n'ont pas été complètement effacées). Loin de constituer un simple outil de visualisation, *Virtual Vellum* présente donc des avantages concrets pour la recherche, tant au niveau de l'analyse en détail des images qu'au niveau de l'affichage de la page manuscrite, ce qui facilite le travail d'analyse, de comparaison, de collationnement enfin. L'étape la plus récente du développement du logiciel comporte la prise en charge simultanée de l'alignement du texte édité avec l'image du facsimilé numérique et avec une traduction en anglais moderne.

[24] Il est téléchargeable au site suivant :
http://www.shef.ac.uk/hri/projects/projectpages/vv/downloads.html.

Relayé par 'Storage Resource Broker', *Virtual Vellum* fait partie intégrante d'un environnement de recherche 'en temps réel' (sur Access Grid) permettant à des chercheurs dans différents pays du monde de participer à des ateliers de collaboration scientifique consacrés à l'analyse de documents manuscrits numérisés. *Virtual Vellum* fait partie d'une initiative patronnée par la National Science Foundation et le Engineering and Physical Sciences Research Council du Royaume Uni réunissant les Universités de Sheffield et d'Illinois à Urbana-Champaign autour de la notion de 'art connoisseurship' (identification scientifiquement probante des artistes responsables des miniatures de tel ou tel manuscrit enluminé).

Virtual Vellum a contribué aussi au développement d'un logiciel appelé *Kiosque* qui partage certaines de ses fonctions mais en vue, cette fois, d'expositions publiques d'objets matériels, y compris les livres manuscrits (expositions de Leeds, Royal Armouries, décembre 2007-avril 2008, et de Paris, Musée de l'Armée, mars-juillet 2010). *Virtual Vellum* fut choisi en septembre 2009 par le directeur du Engineering and Physical Sciences Research Council du Royaume Uni pour illustrer les capacités de la *e-Science* en Arts et Lettres[25]; le logiciel a représenté ces disciplines lors d'un symposium international tenu à Oxford en décembre 2009.

3.2. Florence Clavaud (École nationale des chartes), Le projet THELEME et l'édition électronique des documents médiévaux

Florence Clavaud est directrice des nouvelles technologies et de l'informatique à l'École nationale des chartes (ENC). Médiéviste de formation, archiviste-paléographe, elle s'est spécialisée en informatique documentaire et applications XML pour les sources primaires.. Son équipe s'occupe des projets informatiques de l'Ecole des chartes ; elle a aussi la responsabilité technique, avec le service éditorial et des publications électroniques de l'IRHT (Institut de Recherche et d'Histoire des Textes) du centre de ressources numériques TELMA (Traitement Electronique des Manuscrits et des Archives),.

Florence Clavaud propose tout d'abord une présentation de l'ENC, prestigieuse institution fondée au XIX^{e} siècle, en insistant sur sa double vocation d'enseignement et de recherche[26]. L'enseignement concerne la formation des chercheurs en histoire et les conservateurs des archives, des bibliothèques, des monuments historiques et des musées ; la recherche comprend les activités d'une équipe de recherche pluridisciplinaire, spécialisée dans les sciences historiques et philologiques, notamment la paléographie, la diplomatique, les langues anciennes, la philologie, l'archivistique, l'histoire du livre et des media, l'histoire du droit, l'histoire de l'art.

En dépit de la primauté accordée à des disciplines orientées vers le passé, cette institution est tout à fait à l'avant-garde dans l'utilisation des outils informatiques. En effet, depuis plusieurs années, l'ENC a choisi de considérer le

---

[25] Voir à ce propos P. Ainsworth, M. Meredith : « e-Science for medievalists : options, challenges, solutions and opportunities », *Digital Humanities Quarterly* (sous presse).

[26] Le présent compte rendu se fonde en large partie sur le support écrit de la conférence de Florence Clavaud, qu'elle a eu la bienveillance de me transmettre au lendemain de notre rencontre. Je la remercie chaleureusement pour son amicale disponibilité.

développement Web comme une activité stratégique pour diffuser des ressources et les outils de référence, pour publier les travaux réalisés par ses chercheurs et des chercheurs associés à partir de documents primaires, pour renouveler les moyens et les méthodes de la recherche en histoire.

Florence Clavaud présente le site Web de l'ENC, en attirant l'attention sur les nombreuses ressources à vocation pédagogique qui ont comme objectif d'aider à aborder correctement les sciences auxiliaires de l'histoire, en particulier, la diplomatique, l'histoire du livre, la paléographie[27].

Parmi ces ressources, elle choisit de présenter l'application THELEME (Techniques pour l'Historien en Ligne : Études, Manuels, Exercices), dont la création remonte à l'année académique 2005-2006[28]. Cette application, destinée principalement aux étudiants du premier cycle et à leurs enseignants, est constituée de trois parties comprenant un volet « cours », un volet « bibliographies » et un volet « dossiers documentaires » qui présente actuellement nonante dossiers constitués autour d'autant de documents médiévaux (quinze extraits de manuscrits littéraires, septante-cinq actes médiévaux), dont la plupart est basées sur la collection de fac-similés de l'ENC .

Tous les dossiers sont répertoriés et la liste est accessible par un sommaire, mais il est également possible d'effectuer des recherches en combinant plusieurs critères. À l'intérieur des dossiers, chacun des fac-similés, est accompagné d'une notice descriptive, d'une édition du texte normalisé, avec un apparat de notes historiques et, éventuellement d'une traduction en français, d'un commentaire paléographique, diplomatique et, éventuellement, linguistique. Mais l'un des atouts principaux de ce site concerne la présentation même du document qui est affiché sur écran en forme de fac-similé numérique interactif. Le document affiché réagit au passage du curseur : il suffit de déplacer celui-ci sur une zone de l'image pour faire apparaître la transcription du texte reproduit dans cette zone. Ce système interactif permet en somme de résoudre, d'une façon à la fois très précise et dynamique, le problème de l'alignement texte – image.

Florence Clavaud révèle la technologie qui se situe à la base de cet outil très performant : il s'agit d'un système hybride qui combine plusieurs composantes, mais qui est fondé essentiellement sur l'encodage en XML, conformément à la TEI P5, du dossier documentaire et comporte notamment les étapes décrites ci-dessous.

Tout d'abord, la déclaration de l'image numérique du document édité et de zones dans cette image :

```
<facsimile>
<surface>
   <graphic url="fax.jpg" width="685px" height="1000px" xml:id="fax"/>
   <zone  xml:id="zone-1" ulx="279" uly="228" lrx="555" lry="305"/>
   <!-– etc. pour les autres zones -->
     </surface>
</facsimile>
```

[27] Le site est accessible à l'adresse suivante : http://www.enc.sorbonne.fr.
[28] http://theleme.enc.sorbonne.fr/.

La balise <graphic> permet de déclarer l'image et comprend, en tant qu'attributs, le nom et l'identifiant du fichier graphique contenant le fac-simile (ici fax.jpg), ainsi que ses dimensions (largeur et hauteur). La balise <zone> porte un identifiant (`xml:id="zone-1"`) et d'autres attributs permettant de délimiter la zone précise de l'image qui sera associée à une portion de texte dans la transcription.

L'étape suivante comprend la déclaration du lien entre la zone de l'image choisie (ici, `"zone-1"`) et la portion de texte dans la transcription :

```
<div type="facsimile" facs="#fax">
  <p>
    <seg facs="#zone-1">Colosenses et hi sicut Laodicenses sunt Asia-</seg>
<!— etc. -->
</p>
</div>
```

Finalement, le document XML sera transformé, à l'aide d'un programme XSLT (cf. ci-dessus, l'intervention de James Cummings), afin de générer dynamiquement la page Web en langage HTML, pour disposer d'une image réactive, et d'une carte des coordonnées des zones sensibles dans cette image[29] :

```
<img usemap="#map" src="99/fax.jpg" border="0"/>
<map name="map">
    <area shape="rect" coords="279,228,555,305" href="#"

OnMouseOver="AffBulle(' Colosenses et hi sicut Laodicenses sunt Asia-',279)"
OnMouseOut="HideBulle()" OnClick="return false"/>
</map>
```

Les fonctions Javascript AffBulle() et HideBulle() permettent d'afficher et de faire disparaître la fenêtre (ou bulle) au passage de la souris sur la zone définie ci-dessus.

L'encodage XML/TEI du dossier documentaire est actuellement réalisé à l'aide du logiciel oXygen présenté par J. Cummings ; la définition des zones des images numériques et le relevé en XML de leurs coordonnées se font grâce au logiciel libre Image Markup Tool développé par le *Humanities Computing and Media Centre*, *University of Victoria* (Canada)[30].

Cette application continue d'évoluer (contenus en accroissement, améliorations techniques, mise en place de solutions plus simples de production des dossiers documentaires). Un projet d'Album de Diplomatique Européenne En Ligne a par ailleurs été lancé par l'ENC, pour constituer sur les mêmes bases (objectifs pédagogiques, dossiers présentant édition critique et images de documents originaux), à l'échelle européenne et avec une interface multilingue, un florilège de

[29] Le dossier documentaire dont on a tiré le segment XML ci-après est consultable en HTML à l'adresse suivante : http://theleme.enc.sorbonne.fr/dossiers/notice99.php.
[30] http://www.tapor.uvic.ca/~mholmes/image_markup/.

documents de l'histoire européenne, reproduits, transcrits, traduits et commentés par des historiens, diplomatistes et archivistes.

Cette explication éclairante nous permet de découvrir la technologie qui se trouve au fondement du système des fac-similés interactifs et nous permet de mieux apprécier la diversité et l'efficacité des applications du langage XML.

Florence Clavaud présente enfin le Master nommé « nouvelles technologies et histoire » qui offre à de jeunes chercheurs en histoire une double formation aux sciences auxiliaires de l'histoire et aux nouvelles technologies appliquées aux sources primaires[31].

3.3 Anne Catherine Simon (UCL), MOCA – Multimedia Oral Corpora Administration : un système de gestion et d'annotation de données orales

Professeur de linguistique à l'UCL, Anne Catherine Simon est spécialiste de sociolinguistique, d'analyse du discours et de linguistique de corpus. Elle est membre du Centre de recherche VALIBEL (Variétés linguistiques du français en Belgique) de l'UCL, qui se distingue pour l'utilisation des moyens informatiques et technologiques au service de la recherche sur les données orales. Anne Catherine Simon a collaboré, entre autres, à la création d'une plate-forme e-Learning pour un master international et a développé un CD-ROM pour l'apprentissage de l'analyse linguistique du français parlé.

Anne Catherine Simon présente d'abord la banque de données textuelles orales qui a été constituée par le Centre de recherche VALIBEL au fil de 20 années de recherches. Cette base comprend actuellement une quarantaine de corpus différents pour un total de 550 heures d'enregistrement auprès de 700 informateurs originaires de Bruxelles et de la Wallonie, pour un total d'environ cinq million de mots, transcrits et encodés sur support informatique[32]. La base VALIBEL est actuellement la plus importante ressource de données orales pour la langue française.

L'exploitation de ces données orales impose avant tout un travail de transcription, dans la mesure où la parole ne peut devenir objet d'étude et d'analyse qu'à travers un processus de codification écrite. Anne Catherine Simon insiste sur l'importance cruciale de ce processus de codification. Loin d'être une activité neutre et purement mécanique, la transcription des données orales comporte un premier degré d'interprétation, voire même d'analyse. D'où l'importance de définir des critères de transcription très précis, afin d'harmoniser les pratiques des différents chercheurs travaillant sur le même corpus, et de rendre ces critères explicites, afin de bien distinguer entre les différents niveaux de l'analyse[33].

---

[31] http://www.enc.sorbonne.fr/parcours-master.html.

[32] Les principaux renseignements concernant le Centre VALIBEL sont disponibles à cette page : http://valibel.fltr.ucl.ac.be/. Cf. aussi l'article récent : A. Dister, M. Francard, Ph. Hambye, A.C. Simon, « Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009) », *Cahiers de Linguistique* 33/2, 2009, p. 113-29.

[33] Il n'existe pas un modèle standard de transcription, même s'il existe plusieurs initiatives en ce sens. Cf. par exemple NERC (Network of European Reference Corpora) ; EAGLES

Anne Catherine Simon présente brièvement les différentes tendances de la linguistique de corpus, en précisant que les critères utilisés pour la transcription varient considérablement en fonction des objectifs de la recherche. Elle distingue notamment entre les approches quantitatives et les approches qualitatives. Les premières approches recherchent avant tout des corpus de taille importante (plusieurs millions de mots), sous la forme de texte brut, non annoté, afin de faire ressortir des phénomènes de fréquence ou de collocations. L'unité d'analyse est souvent le « mot ». Ce type d'analyse, parce qu'il est automatisé, requiert des transcriptions en orthographe standard, qui permettront d'identifier les unités lexicales de manière non ambiguë pour les compter[34]. Les approches plus qualitatives considèrent la transcription comme contenant déjà une part d'analyse, que ça soit par la notation de phénomènes liés au médium oral (prononciations non standard, chevauchements de parole entre locuteurs, pauses et silences ou phénomènes paraverbaux, etc.) ou par l'ajout d'annotations à la transcription (annotation morpho-syntaxique, sémantique, pragmatiques, etc.). Il est en tout cas acquis, aujourd'hui, qu'une base de données textuelles orales doit répondre aux besoins suivants : transcriptions alignées sur le son / image ; possibilité d'ajouter des annotations (manuelles ou semi-automatiques) ; atteindre des corpus d'une taille importante (qui se mesure en millions de mots) ; posséder les métadonnées et les rendre accessibles ; rendre les données « partageables » entre équipes et entre systèmes informatiques.

Cherchant à rendre opérationnelles le plus d'exploitations possibles de ses corpus, le Centre VALIBEL a mis au point une série de conventions de transcription qui répondent aux contraintes suivantes :

- Utilisation de l'orthographe standard pour rendre possible l'analyse automatique des données textuelles ; les phénomènes de prononciation (liaison, élisions, accent social ou régional du locuteur, allongements de syllabes pour cause d'hésitation) ne sont pas notés dans la transcription orthographique de base mais peuvent faire l'objet d'un codage secondaire ;
- Notation de phénomènes liés à l'interaction orale (chevauchements de parole, pauses et silences, interruptions et reprises)
- Anonymisation des données.

Les conventions de transcription définies à l'origine ont été récemment modifiées et mises à jour en fonction des nouvelles possibilités offertes par les outils informatiques et notamment par la création du logiciel MOCA (Multimedia Oral Corpora Administration). MOCA est un logiciel d'administration de corpus et une interface de consultation, d'interrogation et d'archivage de données orales, sous leur forme sonore et transcrite. L'interface de consultation permet d'aligner la transcription et l'enregistrement et de rendre les données sonores immédiatement accessibles à partir de la transcription.

(Export Advisori Group on Language Engineering Standards) et aussi la TEI (cf. ci-dessus) qui a consacré un chapitre au problème de *transcription of spoken text*.

[34] Dans ce type de recherche, l'existence de plusieurs orthographes générerait des erreurs (par ex. *il y a* vs *y a* ; *tu as* vs *t'as* ; *petit* vs *p'tit*).

Cette facilité d'accès des données sonores permet de proposer une transcription qui va assez loin dans le processus de standardisation. Les nouvelles normes permettent d'alléger le travail des transcripteurs au profit d'une interaction plus importante entre l'écrit et l'oral. VALIBEL a donc confirmé l'option d'une transcription orthographique conventionnelle. Ces conventions de transcription sont illustrées dans l'extrait suivant, issu d'une conversation informelle entre deux jeunes locuteurs :

```
01 blaJV1: oui c'est ça ça m'a fait du bien quoi ça m'a
fait du bien
02         (silence)
03          quand même mis un quart d'heure pour monter
chez moi /
04          tellement j'étais énervé ça m'a bien défoulé
/
05 blaND0: (rire) c'est |- plutôt speed ça
06 blaJV1: {bacord au/} / record -| battu quoi /
07         un quart d'heure |- quoi
08 blaND0: attends -| tu as couru quoi
09 blaJV1: non non je marchais
10         ah non non j'ai pas couru /
11            et j'ai fait encore un détour pour aller
trouver une clope /
12         chez au truc à pitas quoi /
13         c'était fermé donc j'ai fait même un détour /
14         jusqu'à l'Espérance
15         puis je suis revenu
16 blaND0: {c'est} incroyable quoi
17 blaJV1: un quart d'heure
[Valibel; 2004; conversation; blaJV1l]
```

Chaque enregistrement sonore est accompagné d'un apparat de métadonnées ou fiches d'identification précisant le profil sociologique de chaque locuteur (âge, degré de scolarité, profession, etc.) et le contexte de l'enregistrement (durée, nombre de locuteurs, degré de formalité, etc.) et les objectifs de recherche pour lesquels le corpus a été constitué. Ces métadonnées sont interrogeables via l'interface MOCA. L'utilisateur peut donc ainsi se constituer un corpus « sur mesure », en fonction du genre de texte qu'il recherche, du profil des locuteurs qu'il souhaite analyser, de la date des enregistrements, etc. Dans son ensemble, la banque de données VALIBEL ne constitue donc pas un corpus *équilibré*[35] du français parlé en Belgique, mais chaque chercheur peut se constituer un corpus représentatif et équilibré pour son

[35] On définit un corpus comme équilibré lorsque les différents types de textes et les profils de locuteurs représentés sont dans un rapport quantitativement équilibré, et que la taille moyenne des échantillons de parole est identique (en durée ou en nombre de mots).

objet d'investigation, à partir des informations linguistiques et métalinguistiques sur les données.

L'intérêt d'un tel logiciel pour le domaine de la sociolinguistique et de la linguistique de corpus consiste principalement en la mise en parallèle de l'analyse linguistique fondée sur la transcription – analyse syntaxique, lexicale, stylistique, etc. – avec les métadonnées sociales et situationnelles. D'autre part, la présentation d'Anne Catherine Simon engendre une réflexion méthodologique profonde, concernant tout chercheur qui se penche sur l'analyse et l'édition du texte ancien. Elle permet de souligner, tout d'abord, que le processus d'encodage informatique se situe au bout d'une chaîne dont la première étape est justement la codification écrite de la donnée orale. Les supports écrits, manuscrits ou imprimés, que nous avons inévitablement tendance à confondre avec *le* texte, ne sont en réalité que les tentatives – plus ou moins abouties – de codifier quelque chose qui a existé, qui a circulé, qui a été transmis et qui s'est développé en dehors de l'écrit, et surtout en dehors – en amont – de la norme orthographique.

3.4. Johan Wijnants (UCL), Libropera – style sheets et scripting au service de l'accessibilité des "textes" de musique vocale

Johan Wijnants est actuellement doctorant auprès du Centre d'Études Italiennes de l'UCL où il prépare une thèse sur la relation drame – poésie – musique dans les premiers opéras (1600-1637) sous la direction de Costantino Maeder. Il a accompli une formation en langues et littératures romanes à l'Université de Leuven en présentant un mémoire sur les Orphées de Poliziano et Striggio – Monteverdi. Il a ensuite approfondi ses recherches dans ce domaine durant des séjours d'étude à Florence, Bologne et Rome. À côté de la formation académique, Johan Wijnants a des compétences musicales théoriques et appliquées, outre une familiarité précoce avec les outils informatiques, ce qui lui permet d'étudier l'opéra, et la musique vocale en général, de manière effectivement interdisciplinaire.

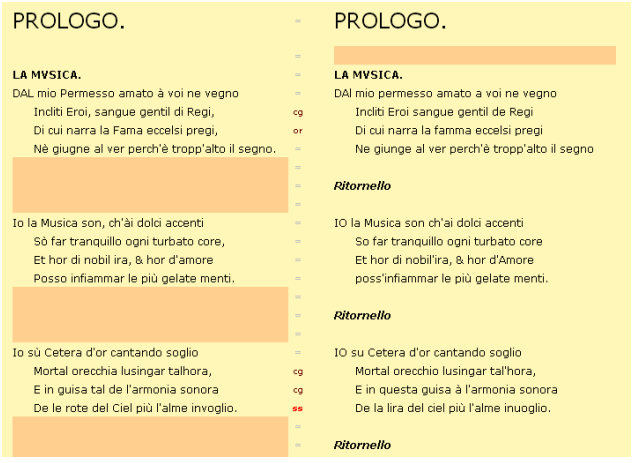
C'est essentiellement pour approfondir ses recherches doctorales que Johan Wijnants a créé le logiciel LibrOpera. Ce logiciel, basé sur la technologie PHP en combinaison avec une base de données MySQL, est pour l'instant d'usage privé, mais l'environnement de travail a été choisi dans l'optique de le rendre accessible, dans un futur très proche, à la communauté scientifique[36]. Pour la musique vocale, qui se transmet à travers des supports partiels (livret) ou non équilibrés (partition), il reste donc une forte barrière au niveau de l'accessibilité des données pertinentes pour l'analyse de l'ensemble de l'oeuvre. Le logiciel permet d'encoder et de traiter les textes et les notations musicales en appliquant à cet encodage les principales méthodes d'analyse de la philologie et de la musicologie.

Johan Wijnants propose une démonstration pratique très éclairante en partant du texte de l'*Orfeo* de Claudio Monteverdi (1607). Il importe dans son logiciel le texte des différentes versions du livret, y compris les versions présentes dans les partitions musicales. Les caractéristiques typographiques des textes sources sont

[36] Quelques visualisations statiques du logiciel sont disponibles à l'adresse suivante : http://perso.uclouvain.be/johan.wijnants/libropera.

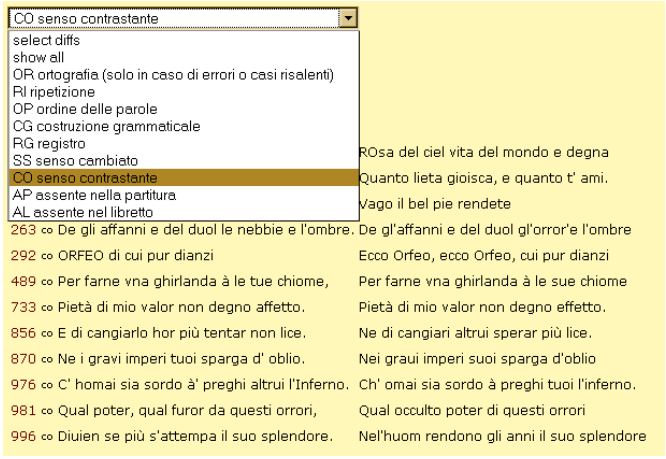
respectées soigneusement (alternance entre les graphies v et u / i et j, ponctuation ancienne etc.), si bien que l'encodage se fait sous forme d'édition diplomatique. L'insertion d'une série d'étiquettes (*tags*) permet d'isoler les éléments du paratexte, tels les titres, les indications scéniques, les didascalies musicales, etc. en les distinguant très nettement par rapport au texte. Les différentes versions sont alignées d'une façon automatique, de sorte que le même étiquetage vaut pour toutes les versions emportées les unes après les autres.

Grâce à une visualisation synoptique et à l'application de différentes CSS (Cascading Style Sheets) le logiciel met en évidence tous les écarts et les variantes entre les différentes versions, ce qui facilite le travail de collation et d'analyse critique des variantes.



L'analyse critique est ensuite opérée par l'utilisateur qui institue un système de hiérarchisation afin de pondérer les lieux variants en fonction de leur importance. Les différences moins significatives, par exemple les écarts purement graphiques, sont ainsi distinguées des variantes plus importantes qui relèvent des différentes sensibilités des auteurs des livrets et des partitions. Une fois instituée cette hiérarchisation, une sélection critique des variantes peut être facilement visualisée grâce à la base de données relationnelle.

Au moyen d'un menu déroulant, l'utilisateur choisit le type de variantes qu'il souhaite visualiser. Dans l'exemple ci-dessous, le sigle CO (*senso contrastante*) permet d'afficher la liste de tous les lieux variants impliquant un décalage de sens :



Cette méthode permet en somme de considérer le livret et la partition musicale comme les différents témoins d'un même texte et de les analyser selon les méthodes de la philologie traditionnelle. Johan Wijnants souligne le fait que l'insertion du texte du libretto à l'intérieur de la partition musicale implique souvent un certain nombre d'adaptations, voire de mutilations qui le défigurent. Il avance donc l'idée de reconstituer une sorte de texte idéal – dans une perspective visiblement lachmanienne – qui permettrait de concilier la structure poétique fortement élaborée du libretto avec la pensée définitive du compositeur et ses éventuels ajouts textuels en musique.

À côté de ces démarches philologiques, le logiciel permet d'effectuer bien d'autres opérations automatisées, comme le calcul du nombre de vers et de syllabes, l'analyse de la structure rimique et de l'accentuation et la mise en évidence de structures sémantiques basées sur des éléments récurrents. L'analyse de l'accentuation permet de faire le lien avec la notation musicale, car il permet de vérifier jusqu'à quel point le compositeur a respecté la déclamation des vers.

Pour sa thèse de doctorat, Johan Wijnants se contentera de prendre en charge la composante textuelle, mais il a déjà élaboré une première version des modules qui permettent d'aborder l'analyse musicale et notamment le calcul des intervalles de la mélodie, la visualisation du rythme, celle de l'harmonie et finalement du rythme harmonique.

Ces opérations faciliteront le travail, souvent artisanal, du musicologue et lui permettront de se consacrer davantage à interpréter les textes qu'à les déchiffrer.

Conclusion – Perspectives de recherche

Si le support papier constitue indéniablement une limite pour la présentation des logiciels et des outils informatique, il m'a paru tout de même utile de laisser une trace écrite de cette initiative. Les nombreux renvois aux sites Internet de référence permettront facilement à nos lecteurs d'avoir accès à la fois aux logiciels

téléchargeables et aux documents ou aux manuels qui constitueront les supports techniques de leur apprentissage.

Le choix d'adresser cet écrit principalement aux spécialistes de la période médiévale et humaniste n'est pas anodin, et se justifie par le fait que le corpus des textes français des XIV^e^ et du XV^e^ siècle fait l'objet, depuis quelques années, d'éditions et d'études de plus en plus orientés vers le traitement informatique. À côté de l'édition du *Miroir historial* de Jean de Vignay, dont il a été question ici grâce à la présentation de Laurent Brun, je peux citer, par exemple, le projet d'édition du manuscrit Harley 4431, contenant les œuvres de Christine de Pizan, dirigé par James Laidlaw (Université d'Edimbourg), basée sur le langage XML, et surtout l'édition de la *Cité de Dieu* de saint Augustin dans la traduction de Raoul de Presles, dirigée par Olivier Bertrand (Université de Savoie, Chambéry – CNRS, Nancy). Ce dernier a pris le parti de faire transcrire le manuscrit de base par ses collaborateurs directement en XML, à l'usage du logiciel oXygen, si bien que la transcription et l'encodage – réalisé selon le standard TEI – se font d'une manière simultanée. Une fois terminée la transcription, pour la relecture et l'édition imprimée, les fichiers XML seront directement convertis et mis en page à l'aide du logiciel ouvert et gratuit Open Office. Je profite de cette occasion pour saluer ce projet comme un modèle à suivre par les nouvelles générations d'éditeurs et notamment pour les textes de grande envergure[37].

Pour ce qui est des perspectives ouvertes par la rencontre du 24 avril, je ne peux citer, à l'heure actuelle, que le cas de l'édition du *Miroir historial*, dans lequel je suis directement impliqué. Si d'un côté Laurent Brun et moi-même sommes entièrement satisfaits du système LaTeX et de ses performances au niveau de la mise en page de l'édition, nous avons pris conscience, de l'autre, de l'importance d'encoder le texte en XML, afin d'ouvrir de nouvelles perspectives dans l'analyse et dans le traitement de cette importante base textuelle.

Mattia Cavagna
Université catholique de Louvain

[37] Je remercie Olivier Bertrand de m'avoir transmis la documentation concernant son projet.