Le projet CONDÉ : présentation

Les défis d'un corpus de textes en diachronie longue

Mathieu Goux <mathieu.goux@unicaen.fr>

Morgane Pica <morgane.pica@unicaen.fr>

https://conde.hypotheses.org

RÉGION NORMANDIE

UNIVERSITÉ CAEN NORMANDIE

Centre de Recherches Inter-langues sur la Signification en COntexte

E.A. 4255

Introduction - La CONstitution d'un Droit europÉen : le projet CONDÉ

- Projet RIN financé par la région Normandie pour trois années (décembre 2018 => septembre 2021)
- Objectifs triples :
  - (i) dimension *patrimoniale* : construire une base de données compilant sur temps long (1250 => fin du 18e siècle) des témoins représentatifs de la Coutume de Normandie, et établissement d'une carte numérique des possesseurs/auteurs de la coutume montrant la vivacité de l'écrit en Normandie.
  - (ii) dimension *juridique* : faciliter l'accès, pour les historiens et les historiens du droit, à un corpus juridique d'une grande homogénéité et qui se caractérise par une riche arrestographie et une forte tradition du commentaire.
  - (iii) dimension *linguistique* : enrichir les corpus textuels existant de discours « de spécialité », transcrits et annotés morphosyntaxiquement.

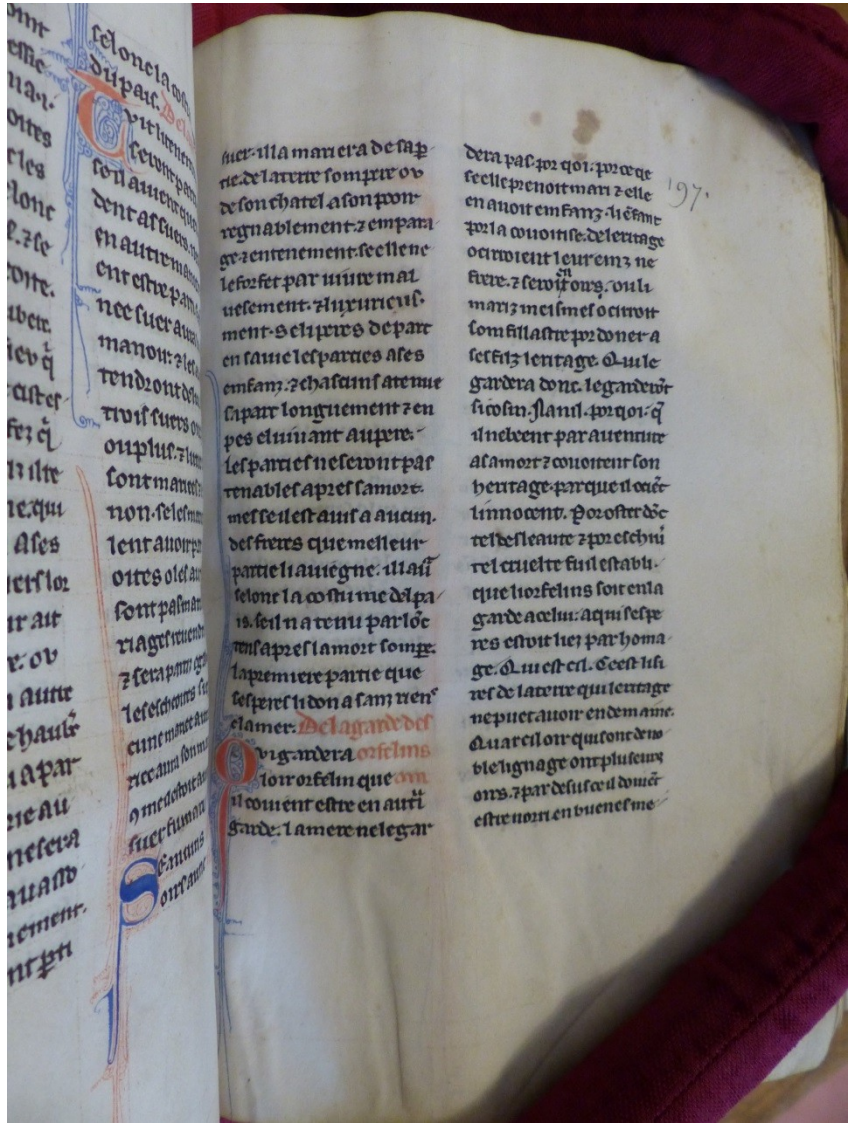
- => Nous explorerons notamment ce dernier point, en nous intéressant :

  - I. Aux témoins retenus pour le corpus, leur caractéristiques textuelles notables et notre dispositif de transcription ;
  - II. À la structure XML-TEI retenue pour leur traduction informatique ;
  - III. Au jeu d'étiquettes POS et aux règles de tokenisation que nous avons suivis.

I. Témoins CONDÉ et méthodes de transcription

- Le repérage des coutumiers représentatifs a été opéré par P. Larrivée et G. Cazals, en amont du projet. Il a été sélectionné une quinzaine de témoins, du *Très Ancien Coutumier* (mi/fin 13e siècle), plus vieux texte connu de la Coutume, à la *Coutume* de Pesnelle (1771), le dernier ouvrage d'importance sur la Coutume normande avant l'établissement du code civil napoléonien, qui marque la fin du droit coutumier français.
- L'on a tâché de prendre un témoin par tranche de 50 ans, sachant qu'avant le seizième siècle, les sources sont rares.
- Beaucoup de textes numérisés / accessibles sur les grandes bases de données, dont *Gallica* et les sites des bibliothèques. Nous avons cependant dû faire des prises de vues de certains manuscrits nous-mêmes, ou redemander une numérisation de certains textes, car leur qualité initiale était insuffisante pour une transcription semi-automatisée.

- Quelques témoins notables de notre corpus :
  - Le *TAC* (BSG MS1743).



- Il n'y avait jusqu'à présent aucune reproduction numérique officielle du manuscrit. On en trouve une transcription chez Marnier (1839 : 6-86), très fidèle nonobstant une ponctuation modernisée. Le manuscrit a été fidèlement décrit par Tardif (1881 : XXV) :

« Le mss. F.f.2 de la Bibliothèque Sainte-Geneviève, dont nous imprimons une partie, provient de l'Église de Saint-Lo à Rouen […] ; il est écrit sur vélin, avec le plus grand soin et d'un format in-4° ; enrichi de lettres majuscules ornées de traits de différentes couleurs. »

- Comme on le voit, il est de l'encre rouge qui sort difficilement à la numérisation, mais qui rend ce témoin remarquable dans sa fabrique.

- Le *Rouillé* (1539)

Deiuſticement, Fo.x.

quelles ſuffiſent pour le pourſuir q̃l a faict le meſfaict/ iaſoit ce quil ne lait point faict. C̃eſt a entendre qui ſuffit pour lappꝛehender et tenir priſonnier iuſques a ce quil ſoit trouue innocent : ou quil baille pleige de ſoy purger du meſfaict : Et ainſi il neſt point mis en priſon ſans meſfaict, Car il y a apparence ⁊ preſumption contre luy pour lors cõme malfaicteur puis quil eſt pourſuy du meſfaict. Et aſſi appert largument ſolut. C̃ Au tiers argument qui argue que meſfaut neſt autre choſe que delict/ ſauf ia grace de larguant: combien que on le puiſſe bien prendre ainſi eſtroictemẽt/ toutesfois eſt il ſouuẽt prins plus largement comme a ce ꝓpos ou il eſt prins generalement pour toute defaulte de faire droict/ ainſi q̃l peut apparoir en ce chapitre es paraphes enſuyuantz.

faict/ dequoy il ſoit tenu a faire ſatiſſaction. Trois choſes ſõt par quoy homme doibt eſtre iuſticie : quand il paſſe terme a faire ce q̃l doibt/ quãd il deſpite iuſtice / Et quand il faict tort a aultruy.

b C̃ Pour terme paſſe eſt hõme iuſticie quãd terme luy eſt aſſis / ⁊ il ne vient pas. Et auſſi quand terme eſt aſſis a aulcũ de payer la rente q̃l doibt/ ⁊ il ne la paye au terme et ne loffre: il doibt eſtre iuſticie tant quil ait faict gre auenaument / ou quil ait donne pleges deſter a droict : et telz deſpaſſementz de termes ſont appellez defaultes.

C̃ Item le texte met eu tiers paraphe de ce chapitre.

b C̃ Que pour terme paſſe doibt homme eſtre iuſticie ꝛc. Par ce paraphe peut apparoir quil y a deux manieres de defaulte. Lune eſt/ quãt terme eſt aſſis a aulcun de venir/ et il ne vient au terme : ⁊ lautre quand terme eſt aſſis a aulcũ de payer/ et il ne paye point. Sur quoy len peut faire vne telle queſtion. Scauoir ſe les bas iuſticiers peuuent leuer amende de leurs hommes ſilz ne leur payent au terme leurs rentes. C̃ Len peut arguer que non/ pour deux cauſes. La premiere pour ce que les haultz iuſticiers nẽ prẽnẽt point/ qui ont greigneur pouoir que les bas. La ſeconde pour ce que ce ſeroit prendre argent pour allongement de terme / qui ſeroit vſure. C̃ Len peut reſpondre a ceſte queſtion que les bas iuſticiers peuuent leuer amẽde de leurs hommes ſilz ne payent au terme leurs rentes / car aultrement il ſen pourroit enſuyr ⁊ retardement de leurs rentes auoir: qui ſeroit en leur grand ꝓtudice / ⁊ dont il ſe pourroit enſuyr inconuenient. C̃ Item par ce texte appert meſme que default de payement au terme eſt appelle defaulte/ et par la couſtume eſcripte. Tout default doibt eſtre amende pour deſpit de iuſtice. Et ſe aulcun vouloit dire que le texte de couſtume qui met que tout default doibt eſtre amende pour deſpit de iuſtice / ne ſentent fors des defaultz de nõ venir a court: car il ſenſuyuroit que de toutes debtes ꝓmiſes de payer a certain terme qui ne les payeroit/ quon en peuſt leuer amẽde/ qui eſt manifeſtement faulx. C̃ Len peut a ce reſpõdre que iacoit ce que le texte ſoit plus proprement declaire au regard des default de non venir a court: toutesfoys ſentent il que le default que len faict de non payer la rẽte aux bas iuſticers au terme/ doibt eſtre amende: et ne ſentent pas ſeulement en lautre cas/ pour deux cauſes.

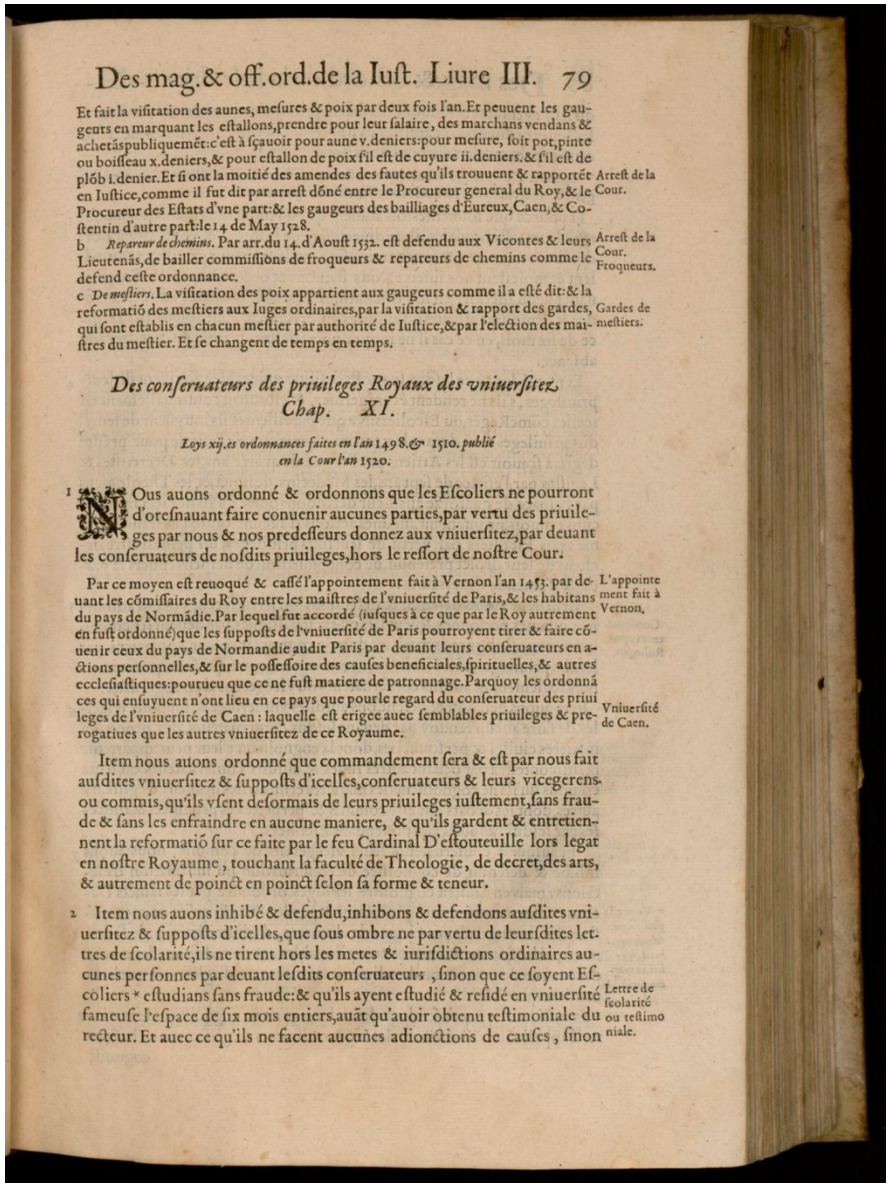
C̃ La premiere pour ce que le texte eſt vniuerſel en tout default. C̃ La ſeconde pour ce que ceſt deſpit de iuſtice : car le bas iuſticier repꝛeſente iuſtice comme il appert eu chapitre de iuriſdiction cy deuant: et ne ſentẽt pas par le texte allegue de default de nõ auoir paye aulcune choſe promiſe payer a certain terme: car ce neſt pas deſpit de iuſtice: pour ce que telles choſes ne ſont pas deues par raiſon de ſeigneurie iuſticiaire. Et auſſi iacoit ce q̃ aulcun deuſt au bas iuſticier argent pour preſt ou pour aultre telle cauſe / quil luy euſt promiſe payer a certain terme/ ſil ne le payoit il ny auroit point damende: car ce ne luy eſt point deu par raiſon de ſa ſeigneurie iuſticiaire: mais en ſon nõ ſeulement. Et ſe vng bas iuſticier auoit perdu ſa iuriſdiction il nauroit plus lamẽde de ces hõmes pour nõ eſtre paye de ſes rẽtes au terme/ car en ce nauroit point de deſpit de iuſtice/ pour ce quil nauroit pl⁹ de iuriſdiction. ⁊ qui alleſgueroit q̃ vng hõme q̃ a rẽte ſur vng aultre et iuſticie, cõbien q̃l ne ait point de iuriſdiction : car luy meſme peut iuſticier ſes hõmes pour ſa rẽte, ⁊ pour ce pourroit auoir amẽdes po⁹ terme paſſe. C̃ Lé po⁹ roit reſpondre q̃ le texte allegue q̃ met. Que tout default doibt eſtre amende pour deſpit de iuſtice/ ne prend pas iuſtice en celle maniere : mais la prend pour iuriſdictiõnaire. Et quãt aux raiſons qui arguent contre la reſpõſe de la queſtion len peut ainſi reſpõdre. A la premiere q̃ argue des haultz iuſticiers / il eſt vray q̃ilz ne prẽnẽt point damẽdes pour rẽtes nõ payees au termes: mais ceſt pour ce q̃ les haultz iuſticiers peuent iuſticier leurs hommes pour leurs rentes par tout et plus amplement que les bas iuſticiers : car ilz peueut pour la rente dune piece de terre que leur doibt vng de leurs hommes iuſticier ſur toutes les aultres pieces de leur fiefz: dont iceluy homme tiẽt : iacoit ce quilz ne ſoient pas ſubiectes a la rẽte deſſuſdicte mais les bas iuſticiers non. C̃ Et ſe larguant repliquoit que neantmoins ceſte ſolution : il ſenſuyroit deſpit de iuſtice qui eſt la cauſe pourquoy amende doibt eſtre leuee. En tel cas len peut reſpondre q̃ lamende neſt pas ſeulement pour ceſte cauſe/ mais pour eſcheuer pluſieurs inconuenſens qui ſen pourroient enſuyr au regart des bas iuſticiers et non pas au regart des haultz iuſticiers/ pour ce quilz ont pouoir de iuſticier par tout ⁊ pl⁹ amplement q̃ les bas iuſticiers comme dict eſt. C̃ A la ſeconde raiſon qui argue contre la queſtion que les bas iuſticiers ne peuent leuer amendes ꝛc. pour ce que ce ſeroit vſure. C̃ Len peut reſoulure que non car vſure ſe faict par cõuenãt accorde de partie/ ⁊ eſt daultre eſſence / car ceſte maniere de prendre amendes neſt pas prinſe pour allongement de terme/ mais eſt vne contrainte et punition iuſticiare pour punir le default.

C̃ Item ſur ce que deſſus eſt dict des haultz iuſticiers. C̃ Len pourroit faire vne tel doubte / ſcauoir ſe vng hault iuſticier a poſſeſſion quarante ou cinquante ans ſur ſon homme et ſur vne piece de terre vaulcune rente en laquelle rente la dicte piece de terre ſur quoy il a eu poſſeſſion neſt pas ſubiecte / mais eſt deue ſur vne aultre piece de terre que tient ſont dict homme. Se la dicte piece de terre ſur quoy ledict ſeigneur a eu poſſeſſion demourra touſiours ſubiecte a ladicte rente. Len peult arguer que ouy / car poſſeſſion de quarante ans ſuffit/ et vault pour tout tiltre et acquerir droicture en poſſeſſion et en propꝛiete affin de heritage / comme peut apparoir par la chartre aux normandz ⁊ par luſaige ſur ce notoirement garde ⁊ ceta. C̃ Pour la reſponſe ou doubte / len peut dire que ladicte piece de terre

b ij

- Disponible sur *Gallica*, dans une numérisation de très bonne qualité.
- Malgré l'écriture gothique, un imprimé. Ressemble aux bibles incunables, avec un principe de glose encadrante.
- Ouvrage notable dans l'histoire de la coutume normande.

- Le *Terrien* (1578)



Des mag. & off. ord. de la Iuſt. Liure III. 79

Et fait la viſitation des aunes, meſures & poix par deux fois l'an. Et peuuent les gaugeurs en marquant les eſtallons, prendre pour leur ſalaire, des marchans vendans & achetás publiquemẽt: c'eſt à ſçauoir pour aune v. deniers: pour meſure, ſoit pot, pinte ou boiſſeau x. deniers, & pour eſtallon de poix ſ'il eſt de cuyure ii. deniers. & ſ'il eſt de plõb i. denier. Et ſi ont la moitié des amendes des fautes qu'ils trouuent & rapportẽt en Iuſtice, comme il fut dit par arreſt dõné entre le Procureur general du Roy, & le Procureur des Eſtats d'vne part: & les gaugeurs des bailliages d'Eureux, Caen, & Coſtentin d'autre part: le 14 de May 1528. Arreſt de la Cour.

b *Reparer de chemins.* Par arr. du 14. d'Aouſt 1532. eſt defendu aux Vicontes & leurs Lieutenãs, de bailler commiſſiõs de froqueurs & repareurs de chemins comme le defend ceſte ordonnance. Arreſt de la Cour. Froqueurs.

c *De meſtiers.* La viſitation des poix appartient aux gaugeurs comme il a eſté dit: & la reformatiõ des meſtiers aux Iuges ordinaires, par la viſitation & rapport des gardes, qui ſont eſtablis en chacun meſtier par authorité de Iuſtice, & par l'election des maiſtres du meſtier. Et ſe changent de temps en temps. Gardes de meſtiers.

Des conſeruateurs des priuileges Royaux des vniuerſitez. Chap. XI.

Loys xij. es ordonnances faites en l'an 1498. *&* 1510. *publié en la Cour l'an* 1520.

1 NOus auons ordonné & ordonnons que les Eſcoliers ne pourront d'oreſnauant faire conuenir aucunes parties, par vertu des priuileges par nous & nos predeſſeurs donnez aux vniuerſitez, par deuant les conſeruateurs de noſdits priuileges, hors le reſſort de noſtre Cour.

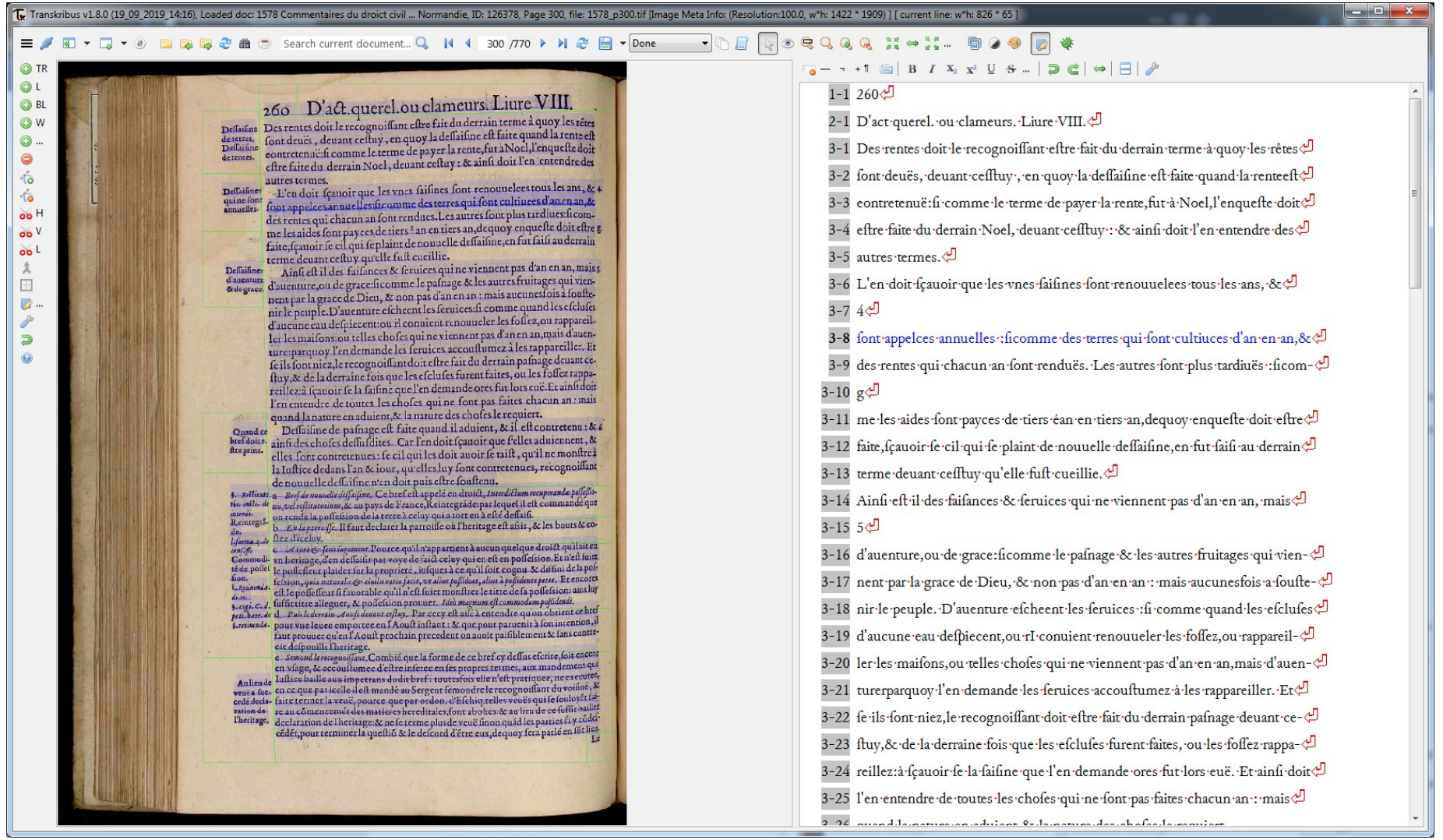
Par ce moyen eſt reuoqué & caſſé l'appointement fait à Vernon l'an 1453. par deuant les cõmiſſaires du Roy entre les maiſtres de l'vniuerſité de Paris, & les habitans du pays de Normãdie. Par lequel fut accordé (iuſques à ce que par le Roy autrement en fuſt ordonné) que les ſuppoſts de l'vniuerſité de Paris pourroyent tirer & faire cõuenir ceux du pays de Normandie audit Paris par deuant leurs conſeruateurs en actions perſonnelles, & ſur le poſſeſſoire des cauſes beneficiales, ſpirituelles, & autres eccleſiaſtiques: pourueu que ce ne fuſt matiere de patronnage. Parquoy les ordonnãces qui enſuyuent n'ont lieu en ce pays que pour le regard du conſeruateur des priuileges de l'vniuerſité de Caen: laquelle eſt erigee auec ſemblables priuileges & prerogatiues que les autres vniuerſitez de ce Royaume. L'appointement fait à Vernon. Vniuerſité de Caen.

Item nous auons ordonné que commandement ſera & eſt par nous fait auſdites vniuerſitez & ſuppoſts d'icelles, conſeruateurs & leurs vicegerens ou commis, qu'ils vſent deſormais de leurs priuileges iuſtement, ſans fraude & ſans les enfraindre en aucune maniere, & qu'ils gardent & entretiennent la reformatiõ ſur ce faite par le feu Cardinal D'eſtouteuille lors legat en noſtre Royaume, touchant la faculté de Theologie, de decret, des arts, & autrement de poinct en poinct ſelon ſa forme & teneur.

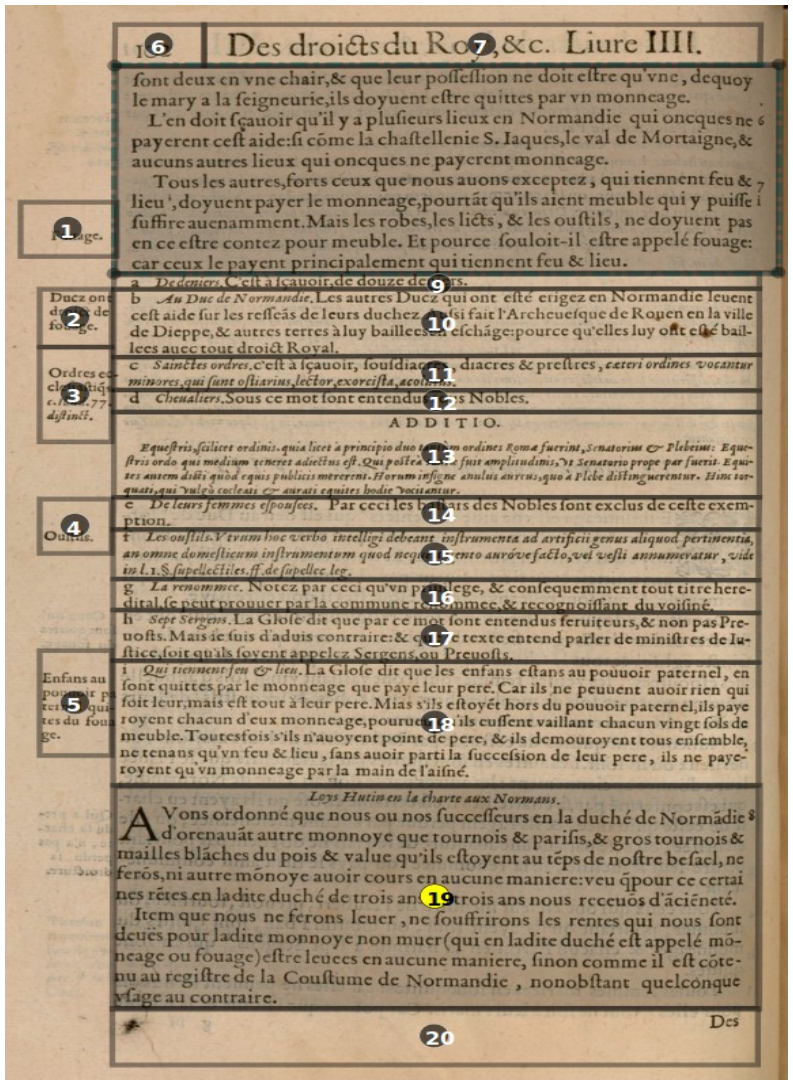
2 Item nous auons inhibé & defendu, inhibons & defendons auſdites vniuerſitez & ſuppoſts d'icelles, que ſous ombre ne par vertu de leurſdites lettres de ſcolarité, ils ne tirent hors les metes & iuriſdictions ordinaires aucunes perſonnes par deuant leſdits conſeruateurs, ſinon que ce ſoyent Eſcõliers * eſtudians ſans fraude: & qu'ils ayent eſtudié & reſidé en vniuerſité fameuſe l'eſpace de ſix mois entiers, auãt qu'auoir obtenu teſtimoniale du recteur. Et auec ce qu'ils ne facent aucunes adionctions de cauſes, ſinon Lettre de ſcolarité ou teſtimoniale.

- Disponible sur *Gallica* de même.
- Il s'agit d'un imprimé, mais la complexité de sa composition, avec des réseaux de notes imbriquées les unes dans les autres, en a fait un texte très compliqué à encoder (cf. Partie II.).

- Après numérisation, il nous a fallu transcrire les témoins. Nous nous sommes servis pour ce faire du logiciel d'OCR/HTR *Transkribus*, qui permet le travail collaboratif et l'entraînement de modèles de reconnaissance d'écriture.



- Nous avons eu de très bons taux de réussite (plus de 97/99% de reconnaissance de caractères). Quand bien même la perfection serait inatteignable, nous avons réussi à transcrire quelque chose de l'ordre 15/20 millions de caractères en moins d'une année. *Transkribus* permet également d'attribuer et de repérer des zones de texte et de produire une sortie XML-TEI, ce qui nous a été particulièrement utile lors de la phase de structuration logique.



- Bien que fastidieuse au commencement (car il a fallu repérer lignes, zones de textes, etc.), cette préparation nous fait gagner énormément de temps *in fine*.
- On notera que cette attribution n'est pas seulement un choix technique, *mais* également déjà une analyse scientifique dans la mesure où elle correspond à une hiérarchisation du propos et de la glose coutumière.
- On rappellera notamment que dans cette perspective de médialité, toute découpe d'un texte est signifiante ; et cette découpe prépare, avec le changement de support, l'analyse future.

II. Structure XML-TEI

- La structure générale reprendra le standard XML-TEI avec la racine TEI et trois types d'enfants directs : un <teiHeader> pour les métadonnées, un <text> pour le corps du texte et, entre les deux, des éléments <facsimile>. Ces derniers contiennent chacun les informations et liens vers le facsimile d'une page et les différentes zones de texte qu'elle comporte.

```
26  <TEI xmlns="http://tei-c.org/ns/1.0">
27    <teiHeader xml:lang="fr"> [92 lines]
120   <facsimile xml:id="facs_1"> [36 lines]
157   <facsimile xml:id="facs_2"> [52 lines]
210   <facsimile xml:id="facs_3"> [53 lines]
264   <facsimile xml:id="facs_4"> [58 lines]
323   <facsimile xml:id="facs_5"> [42 lines]
      [...]
107567  <facsimile xml:id="facs_767"> [144 lines]
107712  <facsimile xml:id="facs_768"> [102 lines]
107815  <facsimile xml:id="facs_769"> [125 lines]
107941  <facsimile xml:id="facs_770"> [79 lines]
108021  <text> [49987 lines]
158009  </TEI>
```

- La structure des éléments de <facsimile> reprend presque telle-quelle cette partie de l'export XML-TEI de Transkribus.

- Elle comporte une division en éléments <zone> de texte, elles-mêmes divisées en <zone> de ligne. Des attributs renseignent la portion d'image qu'elles occupent. Chacune possède également un identifiant qui permet de la relier au texte qu'elle représente dans le corps du texte. Reproduire presque à l'identique la structure produite par Transkribus nous permettra de reprendre la structure de la page source lorsque l'affichage le demandera.

```
<?xml version='1.0' encoding='UTF-8'?>
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
    <teiHeader/>
    <facsimile xml:id='facs_209'>
        <surface ulx='0' uly='0' lrx='1422' lry='1886'>
            <graphic url='1578_p209.tif' width='1422px' height='1886px'/>
            <zone ulx='60' uly='53' lrx='838' lry='137' subtype='header'
                xml:id='facs_209_TextRegion_1'>
                <zone ulx='129' uly='87' lrx='838' lry='152' xml:id='facs_209_line_1'/>
            </zone>
        </surface>
    </facsimile>
    <text>
        <body>
            <pb facs='#facs_209' n='209'/>
            <p facs='#facs_209_TextRegion_19'>
                <lg>
                    <l facs='#facs_209_line_1'>De la difference des biens, &amp;amp;c. Liure V.</l>
                </lg>
            </p>
        </body>
    </text>
</TEI>
```

- La structure interne de l'élement <text> a, quant à elle, été modifiée. Elle sera basée sur des divisions correspondant à la structure logique du texte : des de @type="livre" contenant des de @type="chapitre", contenant elles-mêmes des de @type="section". À l'intérieur de ces dernières se trouvent des éléments <p> (paragraphe) et des éléments <note>.
- Cette distribution nous permet d'obtenir le même grain d'encodage sur chaque témoin pour faciliter le fonctionnement de l'interface. Pour ceux qui, comme le *Très Ancien Coutumier*, ne contiennent pas de division en livres ou chapitres, les différentes sections seront contenues dans un seul et unique chapitre, lui-même dans un unique livre.

```
<text>
  <front/>
  <body>
    <div type="livre" n="2" xml:id="terrien-2">
      <div type="chapitre" n="1" xml:id="terrien-2-1">
        <div type="section" n="4" subtype="coutume" xml:id="terrien-2-1-4">
          <head facs="#facs_56_r2153">Au chapitre de fuite de femmes.</head>
          <p>
            <lb facs="#facs_56_line_1558439494998_11771"/>FEmmes ne doyuent pas eſtre receuës à fuyr cauſes criminelles,ne à les
            <lb facs="#facs_56_line_1558439494998_11770"/>defendre. Mais les hommes peuuent fuyr des meffaits qui ont eſté faits
            <lb facs="#facs_56_r2157"/>à leurs femmes,&amp;amp; les defendre ſ'elles en ſont appelees.
          </p>
        </div>
      </div>
    </div>
  </body>
</text>
```

- La segmentation fine des pages et le typage des différentes zones de texte nous ont permis d'écrire une première transformation XSL sur la sortie XML-TEI du logiciel, nous fondant sur le typage desdites zones pour transformer les différentes zones de texte en un élément adapté.
- Les notes auctoriales seront contenues dans des éléments <note> si possible typés, l'appel de note contenu en attribut @n et un élément <label> contenant la référence à la portion de texte principal commentée.
- Les éléments de coutume, arrêts, ordonnances, etc. sont les éléments <p> et formeront l'unité de base du texte. Ils contiendront les notes à leur emplacement logique.

```
&amp;amp; de larcin. Et nul qui tient ſon fief par vil ſeruice,
<note type="in-text" place="in-text" n="e">
  <label facs="#facs_127_line_1560418925482_550">Par vilſeruice.</label>
  <lb facs="#facs_127_line_1560418925477_549"/>Comme vauaſſouries roturieres, &amp;amp; aiſneſſes qu'aucuns tiennent, &amp;amp;
  <lb facs="#facs_127_r2141"/>ſont ſuiets d'aſſembler toutes les rentes ,qui en depédent,&amp;amp; les porter en auât au ſei-
  <lb facs="#facs_127_r2142"/>gneur feudal. Car ſi aucun tient d'aucun bas Iuſticier vne aiſneſſe:&amp;amp; entre iceluy, &amp;amp;
  <lb facs="#facs_127_r2143"/>ſes puiſnez ou ſous tenas ſont aucûs debats, le bas Iuſticier en a la cognoiſſance. Mais.
  <lb facs="#facs_127_r2144"/>ſi vn tiers y mettoit debat,la cauſe reſſortiroit deuant le haut Iuſticier.
</note>
```

Forme originelle de la note ainsi encodée :

e *Par vil ſervice.* Comme vauaſſories roturieres, & aiſneſſes qu'aucuns tiennent, & ſont ſuiets...

- Le paratexte éditorial (entêtes, signatures, numéros de page) sera contenu dans les éléments de facsimile. Étant liées au format livre et le format numérique donnant de nouvelles possibilités de références internes, ces indications nous sont superflues et gênent la progression logique du texte. Nous avons donc choisi de ne pas les inclure dans le corps du texte. Elles sont cependant gardées pour avoir si besoin la transcription exacte d'une page.
- Nous utilisons pour cela l'élément <fw> (form work) de la TEI avec les attributs détaillant la nature et la position du paratexte qu'il contient selon les valeurs obligatoires du schéma TEI.

```
4333 <facsimile xml:id="facs_61">
4334   <surface ulx="0" uly="0" lrx="1422" lry="2030">
4335     <graphic url="1578_p061.tif" width="1422px" height="2030px"/>
4336     <zone ulx="977" uly="101" lrx="1061" lry="157" rendition="TextRegion" type="pageNum" xml:id="facs_61_pageNum">
4337       <fw place="top" type="pageNum">21</fw>
4338     </zone>
4339     <zone ulx="126" uly="101" lrx="977" lry="157" rendition="TextRegion" type="header" xml:id="facs_61_header">
4340       <fw place="top" type="header">Du droict &amp;amp; eſtat des perſ. Liure II.</fw>
4341     </zone>
4342     <zone ulx="1060" uly="1613" lrx="1154" lry="1717" rendition="TextRegion" type="marginalia" xml:id="facs_61_marginalia_n_1"> [13 lines]
4356     <zone ulx="126" uly="157" lrx="1060" lry="1477" rendition="TextRegion" subtype="coutume" xml:id="facs_61_TextRegion_1558441170372_13324"> [39 lines]
4396     <zone ulx="97" uly="1473" lrx="1060" lry="1519" rendition="TextRegion" subtype="titre" xml:id="facs_61_TextRegion_1559833854152_2914"> [2 lines]
4399     <zone ulx="97" uly="1519" lrx="1060" lry="1725" rendition="TextRegion" subtype="coutume" xml:id="facs_61_TextRegion_1559833854152_2913"> [10 lines]
4410     <zone ulx="126" uly="1725" lrx="1060" lry="1795" rendition="TextRegion" subtype="note_interne" xml:id="facs_61_TextRegion_1558441164599_13314"> [5 lines]
4416   </surface>
4417 </facsimile>
```

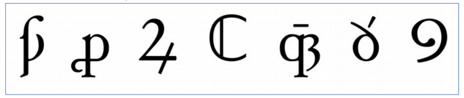
- Les abréviations seront résolues en utilisant l'élément <choice> et ses enfants <abbr> (abbreviation) et <am> (abbreviation marker) pour l'abréviation originelle et <expan> pour sa résolution.
- La ponctuation sera modernisée pour les témoins les plus anciens, toujours avec l'élément <choice> mais cette fois avec ses enfants <orig> (original form) et <reg> (regularization). Les caractères disparus comme le S long seront traités de la même manière.
- L'inclusion de l'original et de sa modernisation permettra de proposer un affichage, au choix, diplomatique, semi-diplomatique ou modernisé et, nous l'espérons, la juxtaposition de deux versions pour comparaison.

Exemple : pſonnes

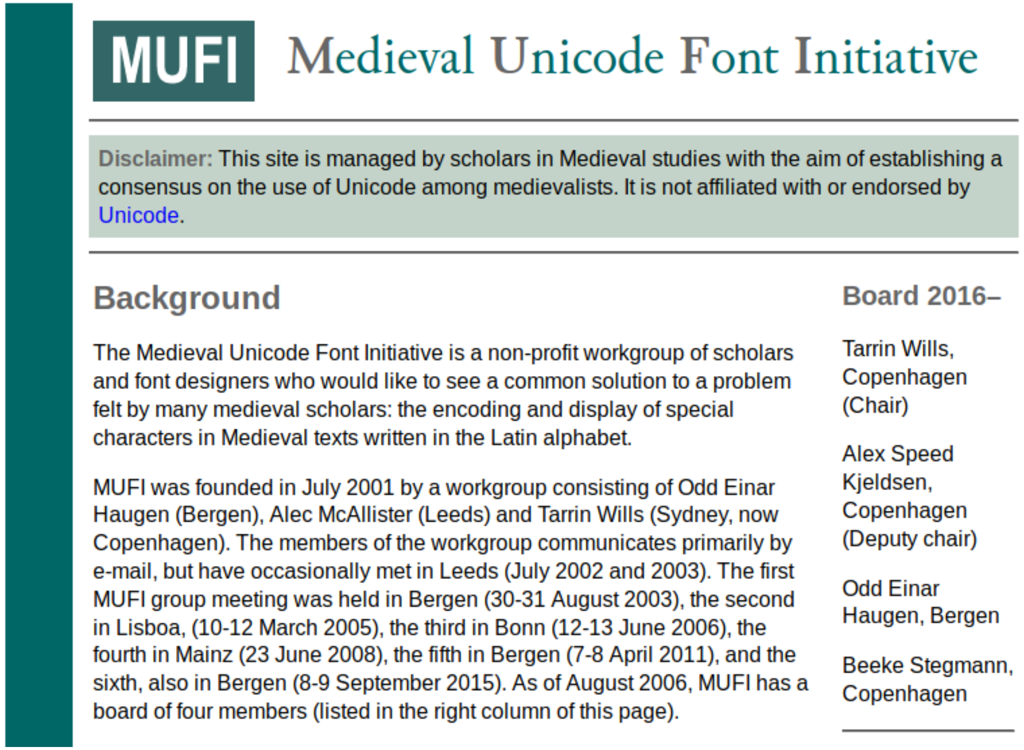
```
<w>
    <choice>
        <am>p</am><expan>per</expan>
    </choice>
    <choice>
        <orig>ſ</orig><reg>s</reg>
    </choice>
    onnes
</w>
```

RÉGION NORMANDIE

- Nous utilisons la police Junicode, qui prend en charge de nombreux caractères anciens ou médiévaux non pris en compte par la plupart des polices.

- Quelques caractères présents en Junicode utiles pour nos témoins plus anciens : 

- Le site étant en cours de conception, l'utilisation qui y sera faite de Junicode n'est pas encore tout-à-fait au point. Certains navigateurs, certains paramètres personnalisés peuvent imposer une autre police, et l'affichage diplomatique sera alors « troué ».



MUFI Medieval Unicode Font Initiative

Disclaimer: This site is managed by scholars in Medieval studies with the aim of establishing a consensus on the use of Unicode among medievalists. It is not affiliated with or endorsed by Unicode.

Background

The Medieval Unicode Font Initiative is a non-profit workgroup of scholars and font designers who would like to see a common solution to a problem felt by many medieval scholars: the encoding and display of special characters in Medieval texts written in the Latin alphabet.

MUFI was founded in July 2001 by a workgroup consisting of Odd Einar Haugen (Bergen), Alec McAllister (Leeds) and Tarrin Wills (Sydney, now Copenhagen). The members of the workgroup communicates primarily by e-mail, but have occasionally met in Leeds (July 2002 and 2003). The first MUFI group meeting was held in Bergen (30-31 August 2003), the second in Lisboa, (10-12 March 2005), the third in Bonn (12-13 June 2006), the fourth in Mainz (23 June 2008), the fifth in Bergen (7-8 April 2011), and the sixth, also in Bergen (8-9 September 2015). As of August 2006, MUFI has a board of four members (listed in the right column of this page).

Board 2016–

Tarrin Wills, Copenhagen (Chair)

Alex Speed Kjeldsen, Copenhagen (Deputy chair)

Odd Einar Haugen, Bergen

Beeke Stegmann, Copenhagen

Pour Junicode, cf. <https://folk.uib.no/hnooh/mufi/> & <https://sourceforge.net/p/junicode/>

III. Étiquetage PoS, tokenisation & lemmatisation

- Dernière étape de l'encodage du corpus, l'enrichissiment TAL dont on connaît les difficultés en termes (i) d'entraînement des logiciels ; (ii) de pertinence du jeu d'étiquettes / des règles de tokenisation.
- L'étiquetage en diachronie longue présente également des difficultés propres :

    - En termes de *lemmatisation* : doit-on lemmatiser un texte médiéval de la même façon qu'un texte de français moderne ? Que fait-on des mots qui ont subi des modifications morphologiques notables ?
    - En termes de *tokenisation* : comment gérer les unités qui ont fait l'objet d'une grammaticalisation au fur à mesurer du temps ?
    - En termes d'étiquetage *PoS* : est-ce que le jeu d'étiquettes doit rendre compte de l'évolution des catégories à travers le temps ?

- Il y a, ainsi des unités fameuses par leur complexité de traitement en français. Par exemple, la locution conjonction *parce que* :
    - En FM, elle est grammaticalisée en un seul « mot ». Mais en AF/MF, ne serait-il pas plus pertinent de la segmenter en *par ce que* ?
    - Est-il pertinent de lemmatiser la variante *pource que*, qui a été remplacée en FC par *parce que* et qui occupait les mêmes rôles syntaxiques ? Plus largement, que fait-on des allomorphes ?

- Il y a eu par le passé plusieurs projets d'ampleur d'étiquetage morphosyntaxiques en diachronie.
- Nous avons notamment hésité entre deux modèles : CATTEX et PRESTO. Nous avons finalement opté pour ce second jeu : CATTEX est certes très efficace pour l'AF – il a été pensé pour --, mais ses choix annotatifs sont moins pertinents, il nous a semblé, pour les autres états de langue. PRESTO, en revanche, a été pensé pour cette perspective.
- Deux paramètres notamment ont orienté notre sélection :

    - D'une part, le public visé. Les (futur.e.s) utilisateurices de la base de données ne seront pas nécessairement des linguistes, mais aussi des historien.ne.s voire des amateurices. Il fallait donc en appeler à des catégories grammaticales assez fortement ancrées dans l'usage pour faciliter les recherches, tout en autorisant un grain suffisamment fin pour permettre des recherches avancées.
    - D'autre part, le rapport coût/bénéfice de l'opération. Compte tenu de la taille de nos textes, il nous fallait un jeu limitant les vérifications humaines et produisant un taux de bruit/de silence acceptable, du moins dans une première campagne d'annotation.

- Parmi les choix de PRESTO qui sont à retenir :
  - Une étiquette consacrée pour les verbes *être* et *avoir*.
  - Une catégorie « PAG » pour « Participe, Adjectif Verbal, Gérondif ». Elle recoupe les « formes en –ant » et les participes passés, dont l'histoire linguistique est particulière en français. Cette étiquette empêche ainsi de trancher entre l'identité de la forme analysée, verbe, adjectif ou autre.
  - Un processus de tokenisation / segmentation / concaténation qui prend en compte le point d'arrivée de l'évolution linguistique : *parce que* sera ainsi toujours tokenisé et analysé comme un seul mot, y compris en AF.
  - La possibilité de naviguer entre un jeu minimum (avec que des catégories de premier niveau) et le jeu complet, bien plus fin (par exemple, à partir de la catégorie « Verbe », ajouter les catégories « Verbes à l'infinitif » / « Verbes tensés », etc.)

=> Ces choix sont, bien entendu, discutables du point de vue linguistique. Mais dans la mesure où, de toutes façons, les grammairiens ne sont pas d'accord entre eux « dans l'absolu », ce n'est pas le rôle d'un jeu d'étiquettes de résoudre ces problématiques diverses.

=> De plus, cet enrichissiment PoS n'est jamais *qu'une entrée de plus* parmi les autres accès au texte, par mot-forme ou par lemme, sans même parler des expressions régulières. Il s'agit ainsi d'affiner la recherche et non pas de proposer une catégorisation absolue, indéboulonnable, du corpus.

=> Enfin, dernier avantage : le jeu PRESTO est un jeu proposé par *Frantext*, au côté de son jeu habituel. Cela permet l'interopérabilité entre les corpus, ce qui est d'autant plus intéressant que *Frantext* a moins de 10% de textes non littéraires.

- L'annotation a été faite grâce :
  - À un dictionnaire mettant à profit des règles d'archaïsation diverses ;
  - Au logiciel d'annotation collaborative Analog.

- Les dictionnaires sont disponibles au téléchargement sur le site de l'ANR PRESTO <presto.ens-lyon.fr>
- Ils se présentent sous un format .dff. Chaque ligne correspond à un mot-forme. Le format général est de type : <MOT_FORME / Etiquette1 / Etiquette2 / Lemme1 / Lemme2 / INC>.

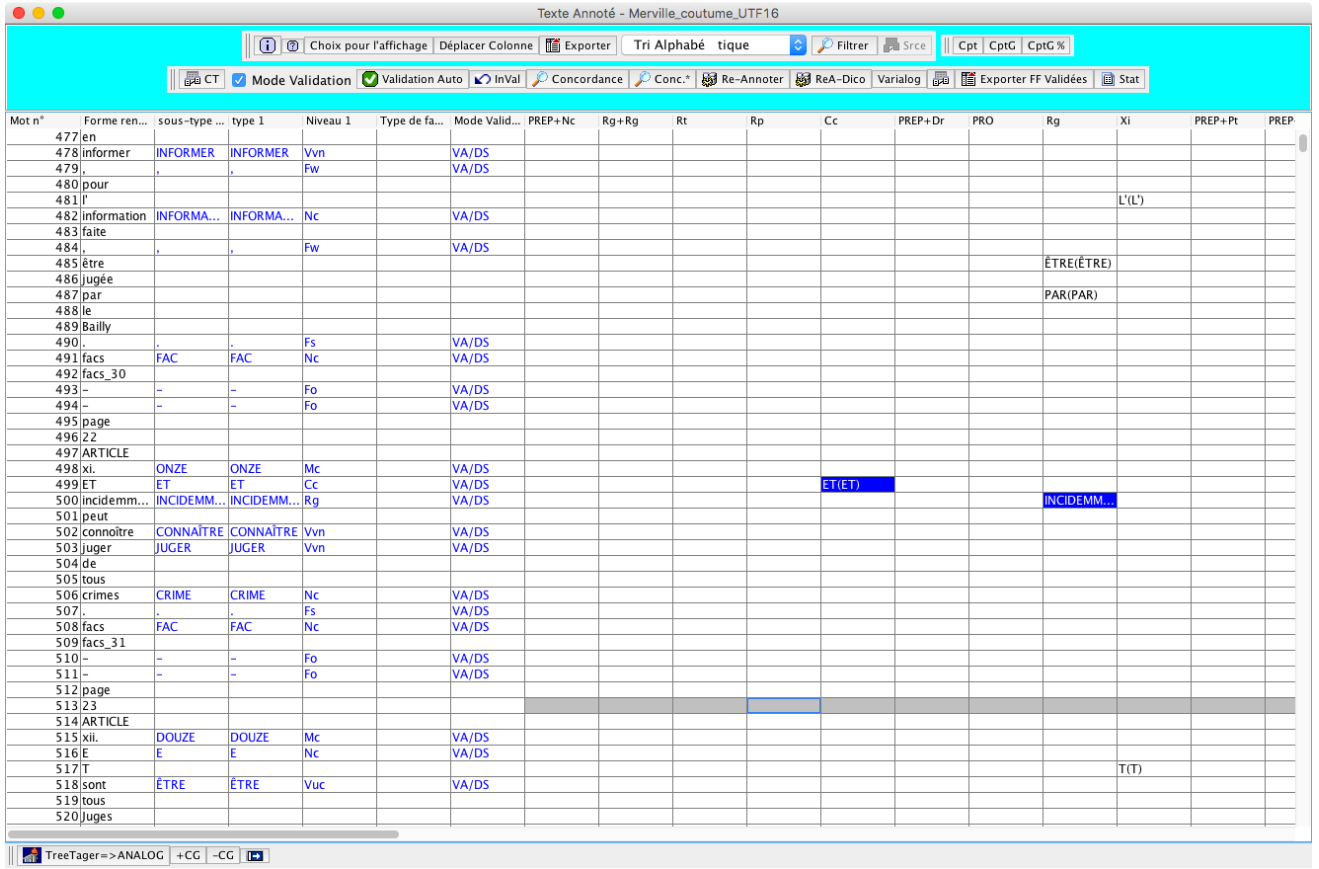
```
1284629 déculpabilisans/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284630 déculpabilisant/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284631 déculpabilisante/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284632 déculpabilisantes/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284633 déculpabilisantez/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284634 déculpabilisants/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284635 déculpabilisantz/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284636 déculpabilisanz/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284637 déculpabilisarent/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284638 déculpabilisas/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284639 déculpabilisasmes/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284640 déculpabilisasmez/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284641 déculpabilisasmés/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284642 déculpabilisasse/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284643 déculpabilisassent/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284644 déculpabilisasses/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
1284645 déculpabilisassez/VER/Vvc/DÉCULPABILISER/DÉCULPABILISER/INC
```

- La première étiquette renvoie au jeu simple, la seconde au jeu complet.
- De même, l'on peut affiner la lemmatisation, par exemple en distinguant les mots dérivés / composés. Dans cet exemple, le Lemme 1 peut ainsi être « Déculpabiliser » et le Lemme 2 « Coupable », ou « Culpabiliser »… selon les choix faits.
- Nous n'avons cependant choisi de rester dans une perspective de lemmatisation moderne ici.

- Le dictionnaire a été automatiquement généré grâce aux dictionnaires en ligne tels *Morphalou*, le *DMF* et le *TLFI*, et des règles d'archaïsation ont ensuite été appliquées pour envisager des formes potentielles en diachronie.
- Il nous faut encore continuer cette phase d'annotation, mais les premiers résultats sur les textes transcrits (soit, à partir du 16^{e} siècle) sont très prometteurs. Il nous faut cependant encore écrire des règles de décision de levée des équivoques, pour automatiser la vérification des formes.

  ⇒ C'est là aussi l'avantage de travailler sur un corpus génériquement homogène : il est vraisemblable que des termes qui pourraient être ambigû doivent s'analyser constamment de la même façon. Par exemple, le mot *bailly*, qui peut être une forme conjuguée du verbe *baillir*, aura tout intérêt à être résolu en substantif compte tenu de la nature des textes. En second temps seulement, des vérifications manuelles permettront de repérer les occurrences atypiques.

- Le logiciel ANALOG, que nous a aimablement confié Marie-Hélène Lay (FoReLLIS, Université de Poitiers), rend ces opérations d'annotation particulièrement efficace. Il exporte ensuite le texte annoté dans un format .csv, qu'il est ensuite aisé de transformer en un .xml que l'on peut, ensuite, intégrer au texte lui-même.



Mot n°	Forme ren...	sous-type ...	type 1	Niveau 1	Type de fa...	Mode Valid...	PREP+Nc	Rg+Rg	Rt	Rp	Cc	PREP+Dr	PRO	Rg	Xi	PREP+Pt	PREP
477	en																
478	informer	INFORMER	INFORMER	Vvn		VA/DS											
479	,	,	,	Fw		VA/DS											
480	pour																
481	l'														L'(L')		
482	information	INFORMA...	INFORMA...	Nc		VA/DS											
483	faite																
484	,	,	,	Fw		VA/DS											
485	être													ÊTRE(ÊTRE)			
486	jugée																
487	par													PAR(PAR)			
488	le																
489	Bailly																
490	.	.	.	Fs		VA/DS											
491	facs	FAC	FAC	Nc		VA/DS											
492	facs_30																
493	–	–	–	Fo		VA/DS											
494	–	–	–	Fo		VA/DS											
495	page																
496	22																
497	ARTICLE																
498	xi.	ONZE	ONZE	Mc		VA/DS											
499	ET	ET	ET	Cc		VA/DS					ET(ET)						
500	incidemm...	INCIDEMM...	INCIDEMM...	Rg		VA/DS								INCIDEMM...			
501	peut																
502	connoître	CONNAÎTRE	CONNAÎTRE	Vvn		VA/DS											
503	juger	JUGER	JUGER	Vvn		VA/DS											
504	de																
505	tous																
506	crimes	CRIME	CRIME	Nc		VA/DS											
507	.	.	.	Fs		VA/DS											
508	facs	FAC	FAC	Nc		VA/DS											
509	facs_31																
510	–	–	–	Fo		VA/DS											
511	–	–	–	Fo		VA/DS											
512	page																
513	23																
514	ARTICLE																
515	xii.	DOUZE	DOUZE	Mc		VA/DS											
516	E	E	E	Nc		VA/DS											
517	T														T(T)		
518	sont	ÊTRE	ÊTRE	Vuc		VA/DS											
519	tous																
520	Juges																

- Le bleu indique les mots validés automatiquement. Évidemment, plus le dictionnaire est riche, plus les ambiguïtés potentielles augmentent.
- Il est possible d'ajouter des mots au dictionnaire « à la volée », et de partager l'annotation pour travailler collaborativement.

- Le TEI est organisé quant à ces informations :

<w n="18087" lemma="LIEU" pos="NOM">lieu</w>

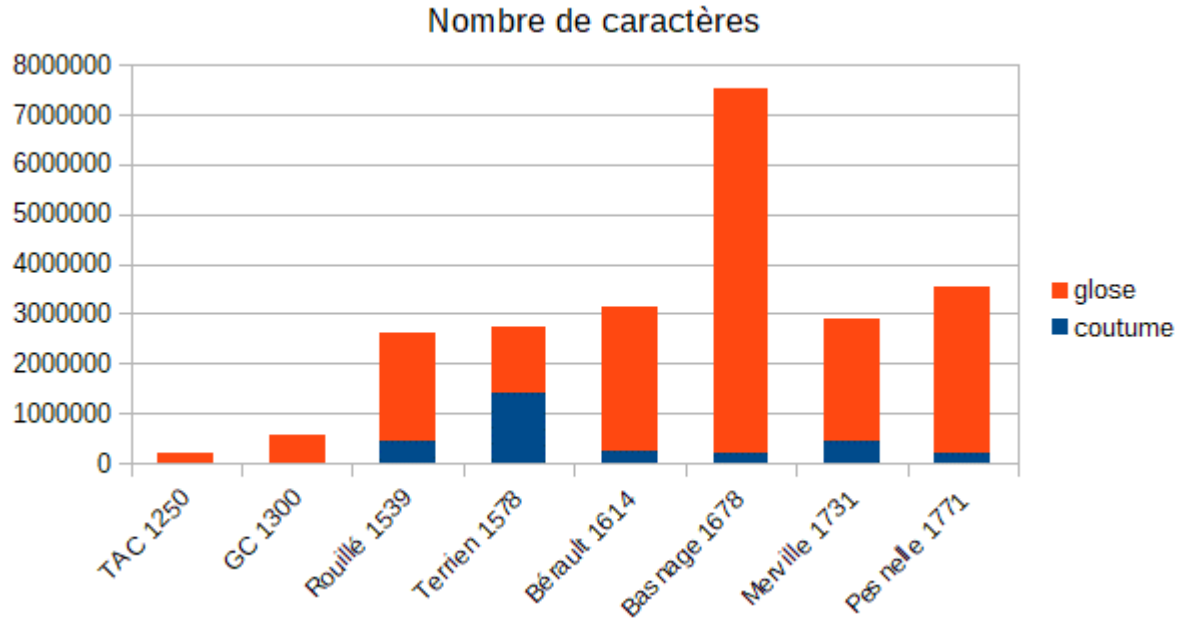
=> Chaque token est encapsulé dans une balise <w> qui possède trois attributs :

- *n*, qui correspond à l'emplacement du mot dans la linéarité du texte – du moins, dans la linéarité du fichier fourni à Analog pour la lemmatisation. C'est l'attribut qui permet ensuite la synchronisation entre les fichiers.
- *lemma*, le lemma.
- *pos*, qui correspond à l'étiquetage grammatical.

Conclusions

- Le projet ConDÉ est particulier par la dimension de son corpus : à l'heure actuelle, nous sommes à 23 millions de caractères. Il nous manque encore des variétés du *Grand Coutumier* et des notes marginales, mais cela rajoutera finalement peu de masse à ce que nous avons déjà.

Texte	Coutume	Glose	Total
TAC 1250		216092	216092
GC 1300		559434	559434
Rouillé 1539	435946	2187708	2623654
Terrien 1578	1401194	1321287	2722481
Bérault 1614	227297	2923600	3150897
Basnage 1678	200706	7337214	7537920
Merville 1731	429019	2454870	2883889
Pesnelle 1771	181903	3356470	3538373
			23232740



- L'expérience du projet PRESTO notamment, nous permet d'accélérer la phase de lemmatisation et d'annotation. En revanche, les structures des plus imbriquées de ces ouvrages de spécialité au regard des textes littéraires nous a obligé à développer des stratégies inédites d'encodage.
- Tous ces choix ne sont, évidemment, pas de simples choix techniques mais des choix scientifiques, qui déterminent l'accès au corpus. Ce ne sera, alors, qu'une fois la base de données accessibles que l'on verra si ces choix sont cohérents avec nos visiteurs et les recherches que nous voudront produire, et les modifications que nous ferons ultérieurement.