NLP Tools and the Norman Laws:

equiping a corpus in long diachrony

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 -- The constitution of a European Law: the ConDÉ project

- Research project funded by the Normandy region for a period of three years (December 2018 => September 2021)
- Three objectives:
  - (i) *Patrimonial* dimension: build a database compiling texts representative of the Norman Laws, over a long stretch of time (1250 => end of the 18th century), and establish a digital map of the owners/authors to show the vitality of written texts in Normandy.
  - (ii) *Legal* dimension: facilitate access for historians and law historians to a highly homogeneous legal corpus, characterized by a rich arrestography and a strong tradition of commentary.
  - (iii) *Linguistic* dimension: enriching existing textual corpora with "speciality" speeches.

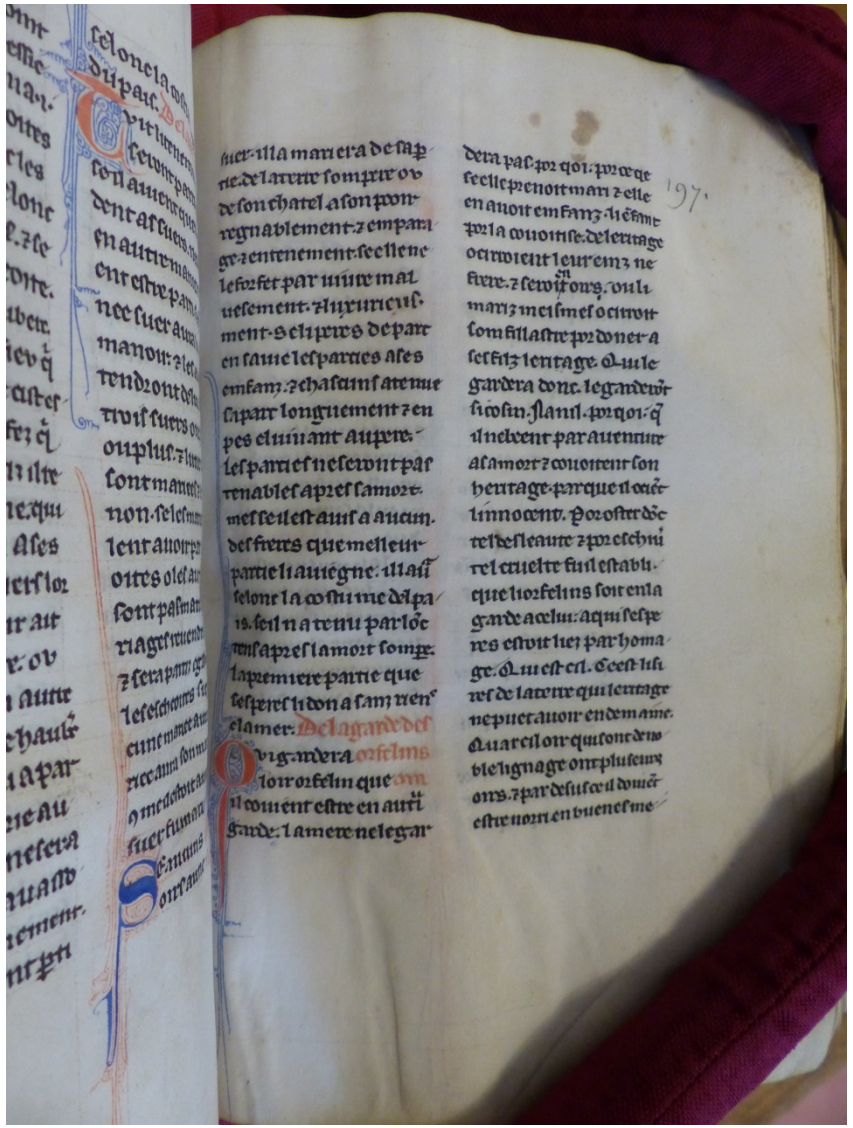
=> We will explore this last point in particular, by taking an interest:

- I. To the texts selected for the corpus, their notable characteristics and our transcription process;
- II. To the XML-TEI structure used for their computer translation;
- III. To the set of POS labels and tokenization rules we followed.

I. Texts selection & transcription methods

- P. Larrivée and G. Cazals identified the representative texts before the project began. They selected about fifteen texts, from the *Très Ancien Coutumier* (mid/late 13th century), the oldest known text of the Custom, to Pesnelle's *Coutume expliquée* (1771), the last major work on Norman Law before the establishment of the Napoleonic Civil Code, which marks the end of French customary law.
- An attempt was made to include one text for every 50 years, knowing that before the sixteenth century, sources were scarce.
- Many texts were already digitized / accessible on large databases, including *Gallica* (<https://gallica.bnf.fr>) and *Google Books*. However, we had to take pictures of some of the manuscripts ourselves, or ask for a new digitization, their initial quality being insufficient for a semi-automated transcription.

- Some noteworthy texts:
  - *Très Ancien Coutumier* (BSG MS1743 [Sainte-Geneviève Library, Paris]).



- Until now, there was no official digital reproduction of the manuscript. There is a transcription by Marnier (1839 : 6-86), very faithful to the original except for a modernized punctuation. Tardif (1881 : XXV) described the manuscript:

« The mss. F.f.2 of the Sainte-Geneviève Library, part of which we print, comes from the Church of Saint-Lo in Rouen [...] ; it is written on vellum, with the greatest care and in-4° format ; enriched with capital letters decorated with lines of different colors. » (personal translation)

- As we can see, there is red ink that is difficult to digitize, but which makes this manuscript remarkable.

- *Rouillé* (1539)

Deiuſticement, Fo.x.

quelles ſuffiſent pour le pourſuir dl a faict le meſfaict/ faiſon ce quil ne lait point faict. Ceſt 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 ꝛ preſumption contre luy pour lors cõme malfaicteur puis quil eſt pourſuy du meſfaict. Et ainſi appert largument ſolut.

¶ Au tiers argument qui argue que meſfaict neſt autre choſe que delict/ſauf la grace de larguant: combien que on le puiſſe bien prendre ainſi eſtroictement/ toutesfois eſt il ſouuẽt prins plus largement comme a ce ppos ou il eſt prins generalement pour toute defaulte de faire droict/ainſi dl peut apparoir en ce chapitre es paraphes enſuyuantz.

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

b ¶ 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 : ꝛ 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. ¶ 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. ¶ 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 ꝛ retardement de leurs rentes auoir: qui ſeroit en leur grand preiudice / ꝛ dont il ſe pourroit enſuyr inconuenient. ¶ 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 ,pmiſes de payer a certain terme qui ne les payeroit/quon en peuſt leuer amẽde/ qui eſt manifeſtement faulx. ¶ 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.

¶ La premiere pour ce que le texte eſt vniuerſel en tout default. ¶ 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 faict/dequoy il ſoit tenu a faire ſatiſfaction.

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

b ¶ Pour terme paſſe 3 eſt hõme iuſticie quãd terme luy eſt aſſis / ꝛ il ne vient pas. Et auſſi quand terme eſt aſſis a aulcũ de payer la rente dl doibt/ꝛ 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 teſpaſſemẽtz de termes ſont appellez defaultes. 4

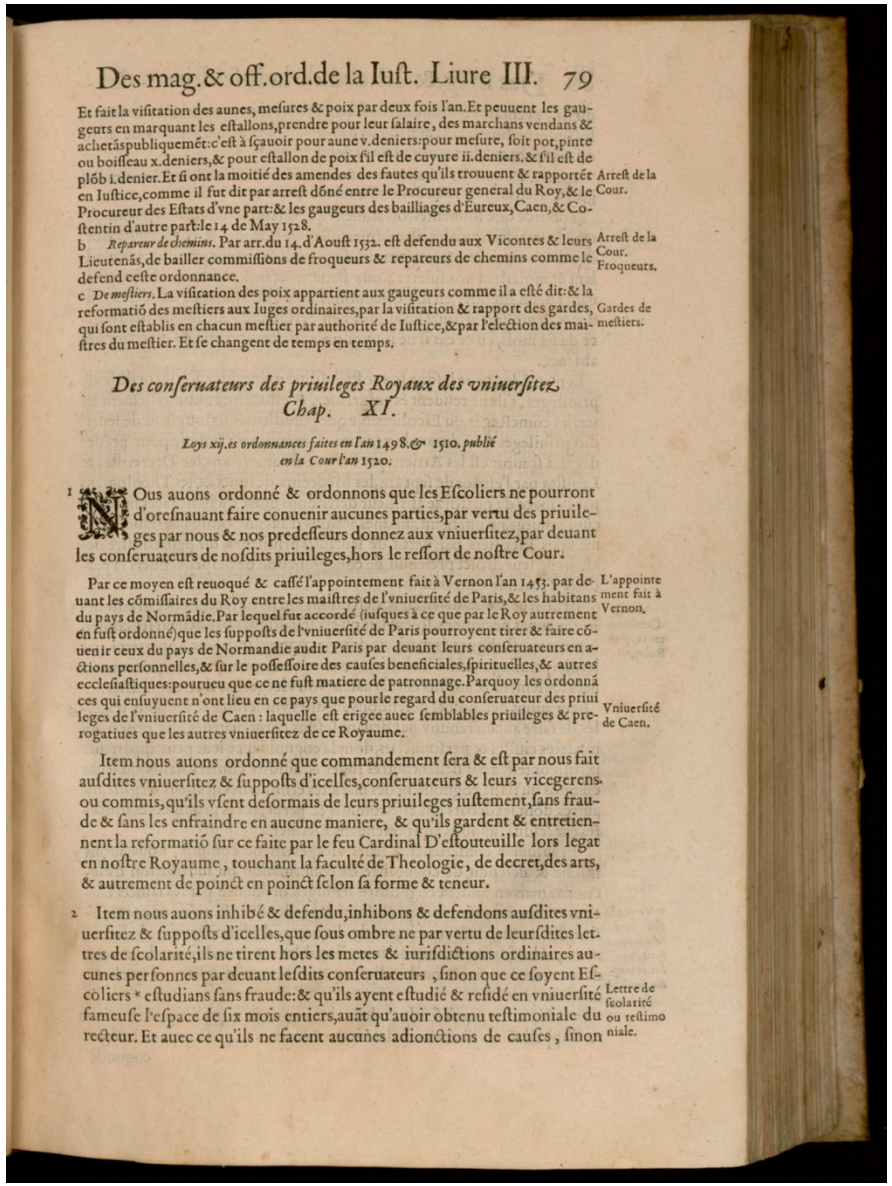
lamẽde de ces hõmes pour nõ eſtre paye de ſes rẽtes au terme/car en ce nauroit point de deſpit de iuſtice/pour ce quil nauroit pl⁹ de iuriſdiction. ꝛ qui alleigueroit q̃ vng hõme q̃ a rẽte ſur vng aultre et iuſticie, cõbie dl ne ait point de iuriſdiction : car luy meſme peut iuſticier ſes hõmes pour ſa rẽte, ꝛ pour ce pourroit auoir amẽdes po⁹ terme paſſe. ¶ Le 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 q̃z 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 peuent 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. ¶ 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 inconuenies 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 ꝛ pl⁹ amplement q̃ les bas iuſticiers comme dict eſt. ¶ A la ſeconde raiſon qui argue contre la queſtion que les bas iuſticiers ne peuent leuer amendes ꝛc. pour ce que ce ſeroit vſure. ¶ Len peut reſouldre que non car vſure ſe faict par cõuenãt accorde de partie/ꝛ 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.

¶ Item ſur ce que deſſus eſt dict des haultz iuſticiers. ¶ 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 ꝛ par luſaige ſur ce notoirement garde/ꝛ ceſta. ¶ Pour la reſponſe du doubte / len peut dire que ladicte piece de terre

b ij

- Available on Gallica, in a very good digitization.
- Despite the blackletter, it’s a print, quite similar to incunabula Bibles, with a framing glose.
- A notable work in the history of Norman law.

- *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 & achetãs 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 *Reparateur 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'electiõn 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.

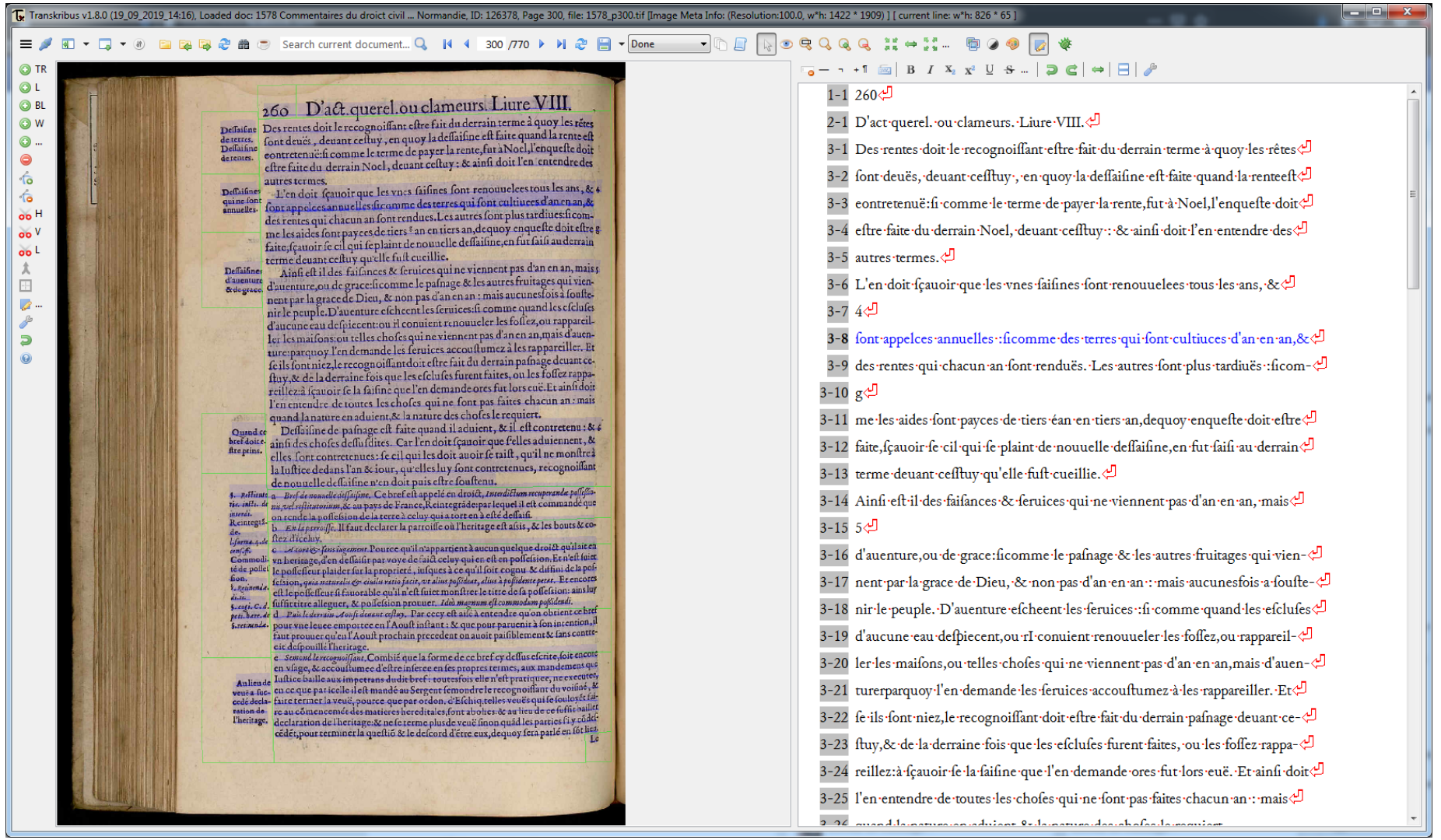
Par ce moyen eſt reuoqué & caſſé l'appointement fait à Vernon l'an 1453. par deuant les cõmiſſaires du Roy entre les maiſtres de l'vniuerſité de Paris, & les habitans du pays de Normãdie. Par lequel fut accordé (iuſques à ce que par le Roy autrement en fuſt ordonné) que les ſuppoſts de l'vniuerſité de 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.

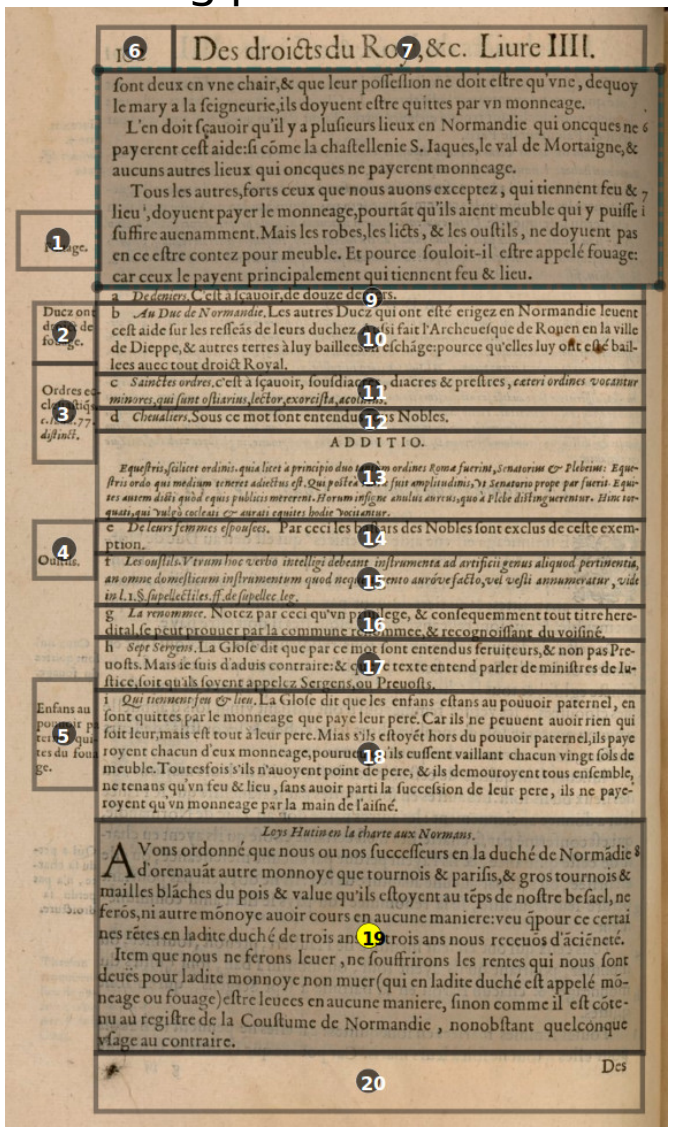
2 Item nous auons inhibé & defendu, inhibons & defendons auſdites vniuerſitez & ſuppoſts d'icelles, que ſous ombre ne par vertu de leurſdites lettres de ſcolarité, ils ne tirent hors les metes & iuriſdictions ordinaires aucunes perſonnes par deuant leſdits conſeruateurs, ſinon que ce ſoyent Eſcõliers * 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.

- Available on Gallica as well.
- It is a printed document, but the complexity of its composition, with networks of notes interwoven into each other and with the main text, has made it a very complicated text to encode (*cf*. Part II.).

- After digitization, we had to transcribe the texts. To do this, we used the OCR/HTR *Transkribus* software <https://transkribus.eu/Transkribus/>, which allows collaborative work and training of handwritten text recognition models.



- We had very good success rates (more than 97/99% character recognition). Even if perfection is unattainable, we managed to transcribe something in the order of 15/20 million characters in less than a year. Transkribus also allows us to assign and locate text zones and produce an XML-TEI output, which was particularly useful during the logical structuring phase.



- Although tedious at the beginning (it was necessary to identify lines, text zones, etc.), this preparation saves us quite a lot of time in the end.
- It should be noted that this attribution of zones is not only a technical choice, but also a scientific analysis, as it corresponds to the internal organization of the text.
- In particular, it should be recalled that in this perspective, any segmentation is significant, and prepares future analysis.

II. XML-TEI Structure

- The general structure will be based on the XML-TEI standard with the TEI root and three types of direct children: a <teiHeader> for metadata, a <text> for the body of the text and, in between, <facsimile> elements. The latter contain information and links to the facsimile of a page and the different text zones it contains.

```
 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>
```

- The structure of the <facsimile> elements is almost identical to this part of Transkribus' XML-TEI export.

- It is divided into « text region » <zone> elements, which are in turn separated into « line » <zone> elements. Attributes indicate the portion of the image they occupy. Each one also has an identifier that allows it to be linked to the text it represents in the text body. Reproducing almost exactly the structure defined in Transkribus will allow us, when needed, to display the text according to the page structure instead of

```
<?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>
```

The internal structure of the <text> element has been modified. It will be based on divisions corresponding to the logical structure of the text: @type="book" containing @type="chapter" , themselves containing @type="section" . Inside the latter are elements <p> (paragraph) and elements <note>.

- This distribution allows us to obtain the same encoding grain on each text to facilitate the navigation. For those who, like the *Très Anciens Coutumier*, do not contain a division into books or chapters, the different sections will be contained in a single chapter, itself in a single book.

```
<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>
```

- The fine segmentation of the pages and the definition of the nature of each text area allowed us to write a first XSL transformation on the XML-TEI output by the software, based on the typing of said areas to them transform into the suitable element.

The auctorial notes will be contained in <note> elements, if possible with an @type attribute, the note call contained in @n attribute and a <label> element containing the reference to the commented main text portion.

- Elements of custom, judgments, ordinances, etc. are the <p> elements and will form the basic unit of the text. They will contain the notes in their logical location.

```
&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_r2l41"/>ſont ſuiets d'aſſembler toutes les rentes ,qui en depédent,&amp;amp; les porter en auât au ſei-
  <lb facs="#facs_127_r2l42"/>gneur feudal. Car ſi aucun tient d'aucun bas Iuſticier vne aiſneſſe:&amp;amp; entre iceluy, &amp;amp;
  <lb facs="#facs_127_r2l43"/>ſes puiſnez ou ſous tenas ſont aucûs debats, le bas Iuſticier en a la cognoiſſance. Mais.
  <lb facs="#facs_127_r2l44"/>ſi vn tiers y mettoit debat,la cauſe reſſortiroit deuant le haut Iuſticier.
</note>
```

Original form of the note thus encoded:

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

- The editorial paratext (headers, signatures, page numbers) will be contained in the facsimile elements. As they are linked to the codex format and the digital format giving new internal reference possibilities, these indications are now superfluous and hinder the logical progression of the text. We have therefore chosen not to include them in the body of the text. However, they are kept to reconstruct the exact transcription of a page if necessary.

We use the <fw> (form work) element of the TEI Consortium with attributes detailing the nature and position of the paratext it contains according to the mandatory values of the TEI guidelines.

```
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>
```

Abbreviations will be resolved using the <choice> element and its children <abbr> (abbreviation) and <am> (abbreviation marker) for the original abbreviation and <expan> for its resolution.
Punctuation will be modernized for the oldest texts, still with the <choice> element but this time with its children <orig> (original form) and <reg> (regularization). Missing characters such as the long S (ſ) will be represented in the same way.

- The inclusion of the original character and its modernization will make it possible to give the choice of a diplomatic, semi-diplomatic or modernized text visualization and, we hope, the juxtaposition of two versions for comparison.

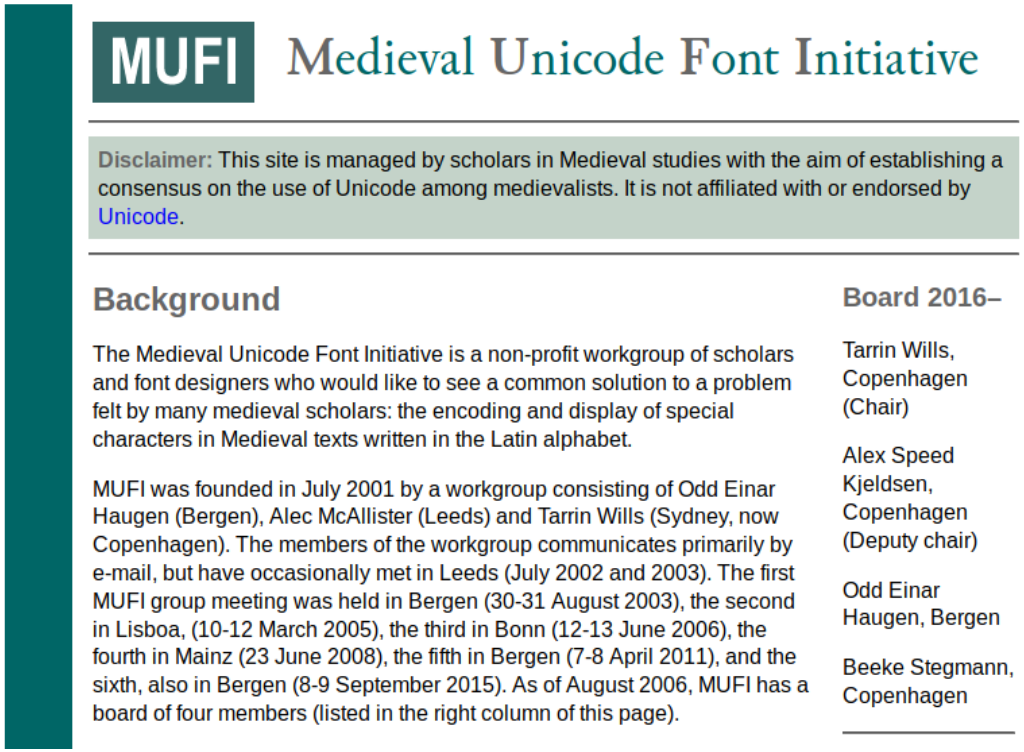
Example :

p̲ſonnes

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

- We use the Junicode font, which supports many antique or medieval characters that are not supported by most fonts, like ꝑ , ⸗ or ℂ. All characters are however declared in the Unicode table and have therefore an objective encoding, whether the font supports them or not.

- As the site is currently under development, the use of Junicode on the site is not yet fully confirmed. On some browsers, the diplomatic display could be "perforated“ as custom font settings could refuse to use Junicode…



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

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

III. PoS labelling, tokenization & lemmatization

- The last step in the encoding is the NPL enrichment, the difficulties of which are known in terms of (i) software training; (ii) the relevance of the label set & the tokenization rules.
- But long-term diachronic labelling also presents its own difficulties:
    - In terms of lemmatization: should a medieval text be lemmatized in the same way as a modern one? What about words that have undergone significant morphological changes?
    - In terms of PoS labelling: does the set of labels have to reflect the evolution of grammatical categories over time?
- There are, for example, famous difficulties in the grammaticalization of some French units. For example, the complex conjunction *parce que*:
    - Nowadays, it’s a single "word". But in old French, wouldn't it be more relevant to segment it as *par ce que*, in three words?
    - Is it relevant to lemmatize the conjunction *pource que* as such, even though it was replaced in Classical French by *parce que*, which occupied the same syntactic roles? More broadly, which lemma must we chose for allomorphs?

- There have been several large-scale diachronic morphosyntactic labelling projects in the past.
- In particular, we hesitated between two models: CATTEX and PRESTO. We finally opted for this second set. CATTEX is certainly very effective for Old French – it was designed specifically for it – but its annotative choices are less relevant for the other language states. PRESTO <presto.ens-lyon.fr>, on the other hand, was designed for this particular perspective.
- We made this choice for two main reasons:

    - Firstly, the target audience. The (future) users of the database will not necessarily be linguists, but also historians and even enthusiasts. It was therefore necessary to use grammatical categories that were strongly in common usage strongly enough to facilitate research, while allowing a grain fine enough to allow advanced research.
    - Secondly, the cost / benefit ratio of the process. Given the size of our texts, we needed a set that limited human intervention and produced an acceptable noise / silence rate, at least at first.

- Among the PRESTO choices to remember:
  - A dedicated label for *être* & *avoir*.
  - A PAG label for “Participe, Adjectif verbal, Gérondif". This label was notably made for the "-ant forms" and past participles. This label prevents the annotator to have to decide between the identity of one specific form.
  - A tokenization / segmentation / concatenation process that takes into account the end point of linguistic evolution. Therefore, *parce que* is tokenized and analyzed as a single word, through every one of its mention.
  - The ability to navigate between a minimum set (with only first level categories, 13 of them) and the complete set (around 50 labels – or example, from the "Verb" label, we add the sub-labels “Infinitive" / "Tensed", etc.)

=> Of course, these choices are questionable from a linguistic point of view. But we have to remember that a set of labels is not meant to solve analytic problems.

=> Moreover, this PoS enrichment is only one more entry among others, especially the research of exact words and/or lemmas. The aim is thus to refine the research process and not to propose an absolute, indelible grammatical categorization of the texts.

=> Finally, one last advantage: the PRESTO set is proposed by *Frantext*, alongside its usual set. This allows corpora interoperability, and the sustainability of our project.

- The annotation was made thanks to:
  - A dictionary using various archaization rules;
  - The collaborative annotation software ANALOG, developed by Marie-Hélène Lay (Université de Poitiers, France)

- The dictionaries are available for download on the PRESTO website <presto.ens-lyon.fr>
- They are presented in a ".dff "format, readable with a txt editor. Each line corresponds to a word. The general format is the following: <Word / PoS 1 / PoS 2 / Lemma 1 /

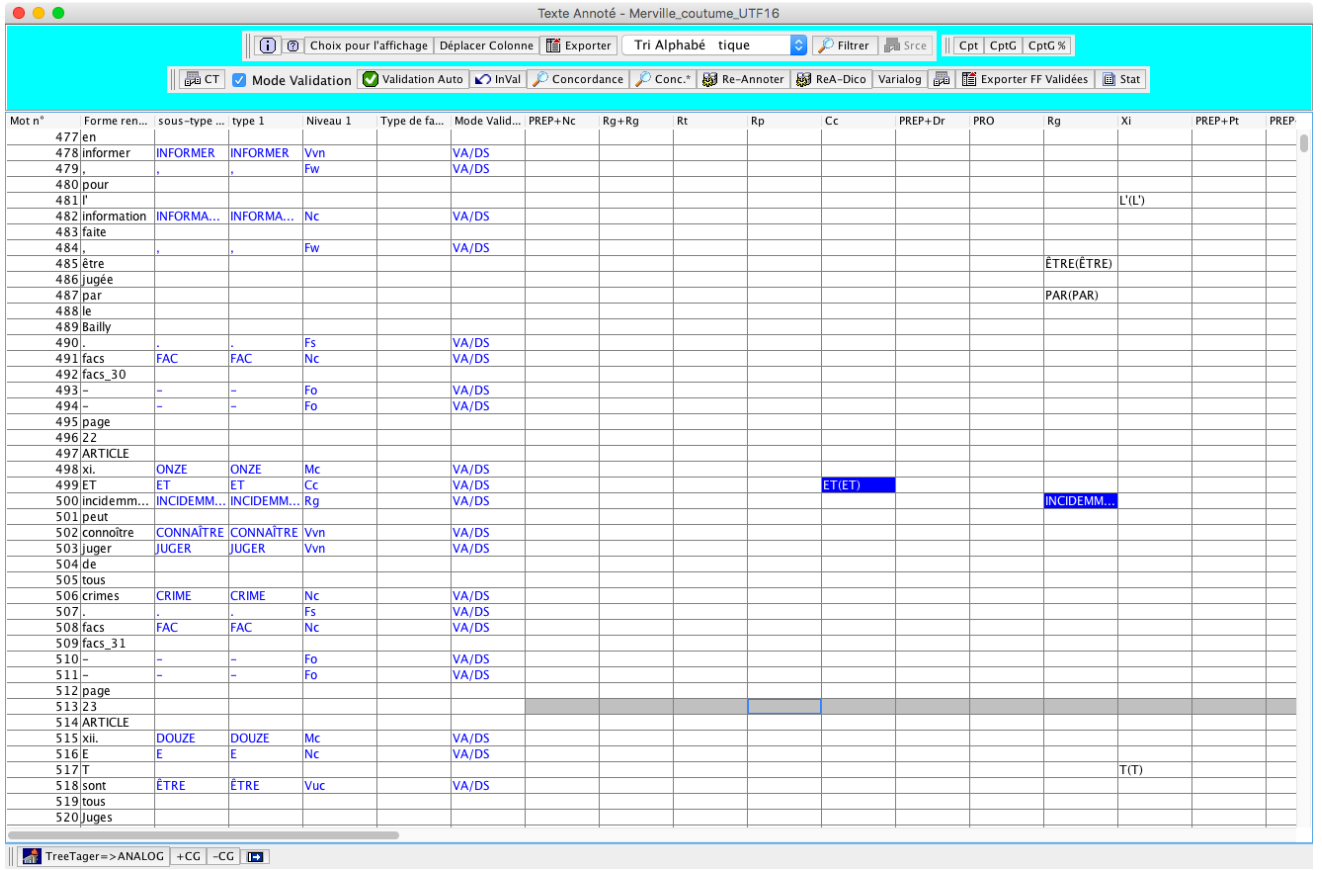
```
1284629 déculpabilisans/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284630 déculpabilisant/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284631 déculpabilisante/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284632 déculpabilisantes/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
1284633 déculpabilisantez/PAG/Ga/DÉCULPABILISER/DÉCULPABILISER/INC
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
```

- The first label refers to the minimal set, the second to the complete one.
- Similarly, lemmatization can be refined, for example by distinguishing derived/compound words. In this example, Lemma 1 can thus be "Déculpabiliser" and Lemma 2 "Coupable", or "Culpabiliser"... depending on the scientific choices made.
- However, we have chosen to remain in a modern lemmatization perspective here, for the same reasons as before

- The dictionary was automatically generated using online dictionaries such as *Morphalou*, the *DMF* and the *TLFI*, and archaicization rules were then applied to consider potential diachronic forms.
- We still need to continue this annotation phase, but the first results on transcribed texts (i.e., from the 16th century onwards) are very promising. However, we still need to write decision rules to clear ambiguities (such as the analysis of *que,* between adverb, pronoun, conjunction…) in order to automatize the verification of forms.

=> Here is one of the main advantages of working on a generically homogeneous corpus: it is likely that words that could be ambiguous must always be analyzed in the same way. For instance, the word *bailly*, which could be a tensed form of the verb *baillir*, will likely be resolved as a substantive, given the nature of our texts.

- The ANALOG software, kindly given to us by Marie-Hélène Lay (FoReLLIS, Université de Poitiers), makes these annotation operations particularly efficient. The software then exports the annotated text in a CSV file, which is then easily transformed into an XML one.



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																

- The color blue indicates automatically validated words. Obviously, the richer the dictionary is, the more potential ambiguities increase.
- It is possible to add words to the dictionary "on the fly", and share the annotation to work collaboratively.

- We structured these informations in the TEI-XML file as such:

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

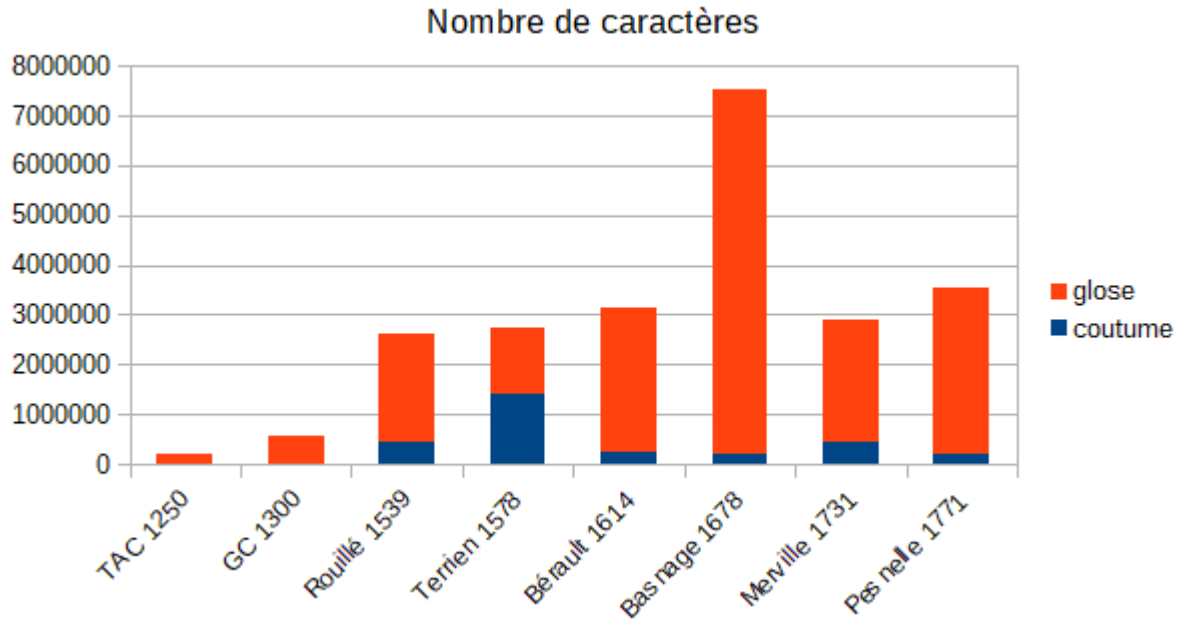
=> Each token is encapsulated in a <w> tag with three attributes:

- *n*, which corresponds to the location of the word in the linearity of the text - at least in the linearity of the file provided to Analog for the lemmatization.
- *lemma*, obviously.
- *pos*, or “Part of Speech”, which corresponds to the grammatical labelling.

Conclusions

- As of today, we finished around 75% of the enrichment of our corpus. We still have to transcribe some manuscripts, and to “clean” the automatic transcription of the others, but we have today the full TXT transcript of the main part of our corpus.
- As for the NPL annotation, we have to apply and check the disambiguation of problematic forms. Once again, the first results are promising.
- Some data about our corpus, and some early results:
  - Firstly, regarding the size of each texts. We distinguish the text from the law (blue, “Coutume”) and the comment (red, “Glose”). As we see, Basnage 1678 is, by far, the biggest of our texts. We have about 23 M characters in total. These numbers will not increase much: the remaining manuscripts are variants of the *Grand Coutumier* (GC 1300), perhaps with some original marginalia, but nothing more. Even the written forms are quite similar.

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



- Secondly, some results regarding the expression of subject pronouns in medieval texts. They confirm our first results included in previous papers, notably (i) the role of the grammatical person, (ii) the parameter of expressivity and (iii) how early these legal texts are compared to the expected change process. These results will be discussed in-depth at the CMLF 2020.

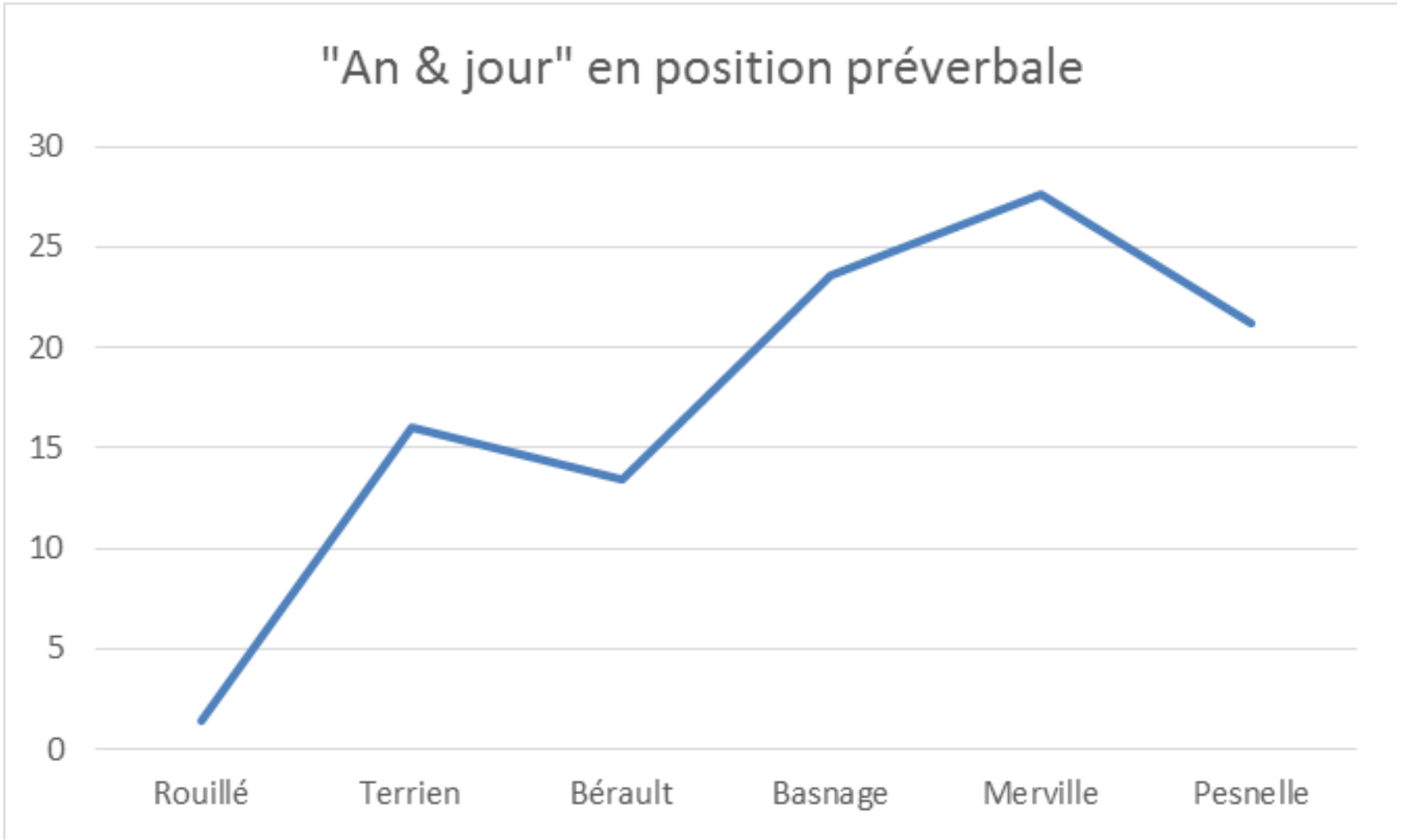
	P1	P2	P3	P4	P5	P6	Total
TAC (410 prop.)	0/0 (0%)	1/0 (100%)	79/6 (93%)	2/0 (100%)	0/0 (0%)	20/1 (95,2%)	102/7 (93,6%)
GC (302 prop.)	4/0 (100%)	0/0 (0%)	45/4 (91,8%)	16/1 (94,1%)	0/0 (0%)	14/1 (93,3%)	79/6 (93%)
Total	4/0 (100 %)	1/0 (100%)	124/10 (92,5%)	18/1 (94,7%)	0/0 (0%)	34/2 (94,4%)	

Tableau 1a – Fréquence (en %) des sujets pronominaux exprimés/non-exprimés dans notre corpus

	P1	P2	P3	P4	P5	P6
TAC	16/0 (100%)	5/0 (100 %)	594/80 (88,1%)	25/6 (80,6%)	7/0 (100%)	19/2 (90,5%)
GC	63/0 (100%)	18/0 (100 %)	889/183 (82,9%)	32/0 (100%)	0/0 (100%)	189/20 (90,4%)

Tableau 1b – Fréquence (en %) de l'antéposition/postposition des sujets pronominaux exprimés

- At last, some results regarding the phrase “An & jour”, as a semantic frame. As time goes by, the phrase ascends to the beginning of the sentence. It seems to be linked to the periodic structure of the discourse, but further analysis must be made. However, first results confirmed some hypothesis on the evolution of infratextual units, and this of reading practices.



- To sum up:
  - An unprecedented corpus;
  - High degree of generic, spatial, linguistic homogeneity;
  - Non-literary texts, which give a closer look of the chronology of change.

=> Delivery date: spring/summer 2020

=> Corpus available on demand: <mathieu.goux@unicaen.fr>