DOI : https://doi.org/10.25364/19.2022.8.7 ISSN : 2663-9815

Harmoniser le corpus *ConDÉ*
De l'image à la ressource linguistique

Harmonizing the *ConDÉ* corpus
From the image to the linguistic resource

Morgane L. Pica
École normale supérieure de Lyon (Lyon, France)
morgane.pica@ens-lyon.fr

Reçu le 14/10/2020, accepté le 28/2/2021, publié le 7/10/2022 selon les termes de la licence *Creative Commons Attribution 4.0 International* (CC BY 4.0)

Pour citer cet article
Pica, Morgane L. 2022. Harmoniser le corpus *ConDÉ*. De l'image à la ressource linguistique. *Studia linguistica romanica* 2022.8, 131-154. https://doi.org/10.25364/19.2022.8.7.

Résumé
Le corpus du projet RIN *ConDÉ* comporte douze sources de référence du droit coutumier normand, du 13e au 19e siècle. Bien qu'homogène dans son sujet, ce corpus présente une grande hétérogénéité dans le format et donc la structure des données textuelles. Le traitement des données, fondé majoritairement sur un HTR par *Transkribus*, des transformations informatiques en langages Python et XSLT, une lemmatisation par *AnaLog* et le modèle d'encodage TEI, a donc dû comporter une phase de réflexion sur la structure permettant de restituer les structures et systèmes de référence des différents témoins, ainsi que concilier six cents ans d'évolution linguistique sous un seul jeu d'étiquettes de lemmes et catégories morpho-syntaxiques. Le choix a été d'élaborer une structure sur trois niveaux (partie > chapitre > section), et a nécessité quelques prises de parti ponctuelles afin de rendre les témoins véritablement comparables.

Abstract
The corpus compiled for the RIN *ConDÉ* project consists of twelve reference sources on Norman customary law, from the 13th to the 19th century. Despite dealing with the same subject, the texts in this corpus are very heterogeneous in terms of format and structure. The texts were processed with the HTR tool *Transkribus*; Python and XSLT languages were employed for automated transformations; lemmatization was performed by *AnaLog* and the data was encoded using the TEI encoding model. Processing the data required a stage of reflection to identify the best means of restoring the structures and reference systems and to devise a set of lemma and part-of-speech tags that would work for texts covering six centuries of linguistic evolution. To make the texts maximally comparable, it was eventually decided to create a three-level structure (part > chapter > section).

Sommaire

1 Introduction

[1] Le projet *Constitution d'un droit européen : six siècles de coutumiers normands* (*ConDÉ*) soutenu par la région Normandie, compte parmi ses objectifs l'établissement d'une base de données textuelle constituée d'éditions numériques de coutumiers normands marquants pour l'histoire du droit. Ces éditions numériques ne doivent en aucun cas être comprises comme une version numérique d'un travail réalisable à l'identique sans l'informatique. Au contraire, nous avons cherché à mettre à profit les ressources particulières que nous offre la machine afin de mettre sur pied des restitutions numériques des textes originaux de manière à en rendre explicites la structure logique[1] et/ou la nature des éléments de texte par le biais d'un enrichissement, ici en langage TEI. Cet enrichissement n'est pas destiné à être lu tel-quel par l'œil humain, mais plutôt à rendre compte de la nature concrète de la source, à servir de base à un affichage intelligent, ainsi qu'à une interrogation fine et précise du texte.

[2] La sélection de deux textes par siècle autant que possible permet à chacun d'agir à la fois comme source juridique et comme échantillon linguistique de sa période. Les témoins principaux s'étalent ainsi du 13e au 19e siècle, un espace de temps très long qui a vu les standards de formatage du codex évoluer, peu à peu mais radicalement, du manuscrit au livre moderne.

[3] Dans ces conditions, le premier défi d'une telle base de données est de proposer un formatage homogène de textes hétérogènes, à la fois sur les plans linguistique, codicologique et structurel. Nous proposons donc ici de donner à voir la réflexion qui a mené cette réalisation technique, et nous montrerons, à travers la considération des différents témoins, comment une base de données textuelle doit être le fruit d'un aller-retour constant entre codicologie et analyse linguistique d'un côté, et de l'autre, entre source matérielle et objectif numérique.

[4] Nous reviendrons pour cela tout d'abord sur les choix généraux du projet *ConDÉ* en termes de corpus et de méthodologie, puis nous aborderons la manière dont les structures hétérogènes des témoins ont été harmonisées pour pouvoir être interrogées de concert, pour enfin terminer par les modalités et choix d'encodage pour la microstructure : la résolution des graphies anciennes et le processus d'étiquetage des lemmes et catégories grammaticales.

2 Corpus et méthodologie

2.1 Le corpus *ConDÉ*

[5] De nombreuses bases de données textuelles lemmatisées conséquentes sont déjà au point. Parmi elles, citons notamment la grande base diachronique

1 La structure logique est la structure inhérente à l'œuvre telle que montrée par le document, indépendante, elle, des contraintes matérielles, c'est-à-dire la structure en parties, sous-parties, chapitres, paragraphes, phrases, etc., le grain de précision dépendant des besoins de l'encodeur. Elle s'oppose à la structure dépendant du support, généralement basée sur les folios pour les manuscrits et sur la pagination pour les imprimés.

Frantext, ainsi que le *Nouveau Corpus d'Amsterdam* (*NCA*) ou le corpus *Modéliser le changement : les voies du français* (*MCVF*), développé à l'Université d'Ottawa, qui contient 2,5 millions de mots sur des ressources du 11e au 16e siècle pour le sous-corpus de textes anciens, et jusqu'au 19e siècle pour les textes spécifiquement canadiens. Certains corpus se spécialisent dans un état de langue particulier comme le *Syntactic reference corpus of medieval French* (*SRCMF*) focalisé sur l'ancien français, qui rassemble des parties de la *BFM* et du *NCA*, de 842 à la fin du 13e siècle, pour un total de plus de 251000 mots annotés manuellement, ou la *Base de français médiéval* (*BFM*) (Guillot-Barbance, Heiden & Lavrentiev 2017), un ensemble de textes écrits entre le 9e et la fin du 15e siècle. La plupart des bases de données textuelles contenant une grande majorité de textes littéraires et épistolaires, aucune ne permettait de réunir un corpus suffisant pour étudier le discours juridique en diachronie longue. Un tel corpus offrirait une excellente ressource pour étudier l'émergence de normes terminologiques dans les textes coutumiers, d'autant plus que la performativité caractérise la définition des pouvoirs municipaux comme des pouvoirs seigneuriaux (Cazals, à paraître).

[6] De décembre 2018 à novembre 2021, le projet *ConDÉ* s'est donc donné pour but de produire une base de données textuelle en diachronie longue procurant aux historiens du droit des textes anciens de référence sur la coutume de Normandie, et aux linguistes diachroniciens une base de données de langue spécialisée inédite. Cette base de données contient, tant que cela est réalisable, un témoin par tranche de cinquante ans, entre le *Très Ancien Coutumier* (*TAC*), daté de la moitié du 13e siècle, et les *Ruines de la coutume de Normandie* (Pannier 1856).

[7] *ConDÉ* met à disposition du public un site internet basé sur l'outil *MaX*, permettant la consultation et l'interrogation simple ou fine[2] de la base de données, la production de tableaux de concordances, ainsi qu'une large bibliographie sur la coutume de Normandie. Le corpus est également disponible au téléchargement sous plusieurs formes via un dépôt GitHub.

[8] Notre corpus noyau est ainsi constitué des principaux textes de référence en histoire du droit pour la coutume de Normandie, c'est-à-dire quatre manuscrits et sept imprimés, pour un total de quatre millions et demi de mots. Le tableau 1 illustre la distribution des témoins dans le temps et montre l'inégalité frappante dans la quantité de matériau textuel selon les témoins. Le tout premier, qui est effectivement le plus ancien texte conservé de la coutume de Normandie, est fragmentaire, d'où le petit nombre de mots qu'il comporte. À l'inverse, l'œuvre de Basnage (1678) comporte à elle seule plus d'un tiers du corpus en l'état. Notre but étant cependant de produire un témoin par période et non pas l'équilibrage du corpus, cette disparité était inévitable.

2 Recherche simple : en texte brut, par mot-forme ; recherche fine : par entrée lexicale et/ou catégorie grammaticale (cf. PDN).

Date	Titre	Auteur	Source conservée par	Nombre de mots (env.)
env. 1250	*Très Ancien Coutumier de Normandie*	anonyme	Bib. Sainte-Geneviève, Paris	10000
Fin 13e s.	*Grand Coutumier de Normandie*	anonyme	Harvard Law School Library, Harvard MA	63000
1386-1390	*Instrucions et ensaignemens*	anonyme	Bib. nationale de France, Paris	10000
15e s.	*Notes sur le Grand Coutumier*	Charles Morisse	Bib. municipale de Rouen	5000
1539	*Le grand coustumier du pays et duché de Normendie*	Guillaume Le Rouillé	Bib. nationale de France, Paris	331000
1578	*Commentaires du droict civil*	Guillaume Terrien	Médiathèque Jacques Chirac, Troyes	540000
1614	*La coustume reformée du pays et duché de Normandie*	Josias Bérault	Bib. nationale de France, Paris	637000
1678	*La coutume reformée du païs et duché de Normandie*	Henry Basnage	Bib. municipale de Lyon	1525000
1731	*Décisions sur chaque article de la coutume de Normandie*	Pierre Biarnoy de Merville	Centre de documentation en Droit, Caen	566500
1771	*Coutume de Normandie*	Pesnelle	Bib. nationale de France, Paris	714000
1856	*Les ruines de la coutume de Normandie*	Victor Pannier	Bib. nationale de France, Paris	23500

Tableau 1 : Les témoins principaux du corpus *ConDÉ*

[9] Le corpus se situe donc dans une démarche de production de données pour la recherche et son identité est principalement marquée par la thématique inédite des données produites. Notre démarche ne se résume cependant pas à l'aspect linguistique des textes rassemblés car nous avons choisi de conserver des données intermédiaires potentiellement utiles à d'autres champs disciplinaires. Ainsi les informations liées exclusivement au format 'livre' qui nous servent à accéder à la structure d'une œuvre.

2.2 Quelles informations encoder

[10] L'habitude du format 'livre' a formé nos attentes envers le texte autour de l'objet, ce qui rend omniprésent le système de référence lié au format codex, ne serait-ce que dans l'utilisation de numéros de page. Le format, en effet, dépend en grande partie du support. Des manuscrits médiévaux dépendent ainsi de codes visuels engendrés entre autres par des contraintes matérielles, que l'arrivée du pa-

pier, d'abord, et de l'imprimerie, ensuite, ont peu à peu compensées. La page a pu être repensée pour être rentabilisée autrement et son organisation interne, en conséquence, a radicalement changé, ce qui, nous le verrons, complique la définition d'une structure homogène pour la base de données.

[11] Les formats utilisés par les auteurs et compositeurs de nos différents témoins doivent donc être compris comme chacun dépendant des normes culturelles particulières de son époque. Bien qu'ils nous aident en exposant chacun une organisation interne, et donc un système de références, qui lui sont propres, il est nécessaire de les dépasser. Si nous utilisons le format originel comme accès à la nature du texte, un nouveau format demande un nouveau système de référence conçu pour lui. En effet, ainsi que l'écrit Benoist (1995 : 30) :

> Reste que cet objet physique qu'est le livre, au moins sous sa forme classique, est marqué par son idéalité, c'est-à-dire aussi bien par ce qui le constitue de l'intérieur comme absence d'objet. Le livre comme tel – ce à quoi on accède par l'exemplaire – n'est aucun de ces exemplaires ; il se tient au-delà d'eux, dans le dépassement possible de l'exemplaire par son propriétaire. Ce dépassement précisément lui ouvre ce qu'est le livre comme tel, dans la confrontation à ce qui en fait un livre : le « sens », la « pensée » de l'auteur, consciente ou non, subjectivement voulue ou objectivement déposée dans le livre.

Sans trahir les sources, car nous basant sur leur format, nous pouvons donc « dépasser » (Benoist 1995 : 30) la forme pour créer une nouvelle structure propre à transmettre le sens de chaque témoin.

[12] Les numéros de page, entêtes, réclames, etc, bien qu'importants pour la codicologie, ne nous renseignent pas sur le sens du texte. Même si nous les avons enregistrés dans nos fichiers, ils n'ont aucune influence sur leur structure. En revanche, la division d'un texte en livres, parties, chapitres, etc. nous a permis d'accéder aux structures respectives des témoins.

[13] La numérisation d'un corpus pose, avant les questions techniques, la question de ce que l'on souhaite transposer exactement sous forme numérique. La résolution en est bien sûr différente pour chaque champ disciplinaire. Cependant, la question de la réutilisabilité des données de la recherche ne doit pas être négligée. Nous citerons, sur ce point, Galleron & Idmhand (2020 : § 9, 19) qui déplorent le fait que « les grilles d'analyse existantes […] considèrent le texte numérique comme point d'aboutissement plutôt qu'en tant que possible point de départ », et proposent de « penser la réutilisabilité en tant que principe de conception et d'analyse des éditions numériques et plus largement des ressources créées par les sciences du texte ». Bien que cette démarche fût présente dès le début de notre travail, cet article particulier nous a paru la formaliser exactement et nous procurer des clés supplémentaires pour la faire aboutir.

[14] Bien que non pertinentes à l'analyse linguistique, ces informations graphiques et de formatage peuvent être utiles à nombre de projets, fournissant notamment des modèles d'entraînement pour des HTR permettant à des projets ulté-

rieurs de gagner du temps sur leurs propres transcriptions, comme le prévoit le projet *HTR-United* sur GitHub, une initiative d'Alix Chagué et Thibault Clérice pour centraliser les informations sur les datasets et modèles d'HTR déjà prêts et d'utilisation libre.

[15] Notre utilisation du standard TEI découle de ces principes, ainsi que la conservation d'une version complète dite *base* de la transcription, contenant notamment les informations graphiques sur les facsimiles produites lors de l'établissement du texte, les entêtes et numéros de pages, ou les tables des matières, hors sujet dans le nouveau format, convocables automatiquement et sans intérêt du côté linguistique car répétant les titres déjà pris en compte. Dans le même but de réutilisabilité, notre dépôt GitHub dispose de plusieurs versions : une version dite *lighter* sans les graphies anciennes ni informations graphiques, une version lemmatisée compatible avec le logiciel *TXM* et une version PDF. Chaque version garde la structure de base et comporte des identifiants pour chaque partie et sous-partie, ainsi qu'une mention des différentes versions mises en ligne pour permettre à qui souhaiterait réutiliser un fichier de pouvoir à l'avenir le situer parmi les mises à jour du projet. Cette démarche définie, nous évoquerons ensuite les outils utilisés pour la mener à bien.

2.3 La chaîne de numérisation

[16] On appelle chaîne de numérisation (fig. 1) l'ensemble des étapes de travail permettant d'élaborer un fichier numérique exploitable représentant un objet matériel. La première étape consiste généralement en la production d'autant d'images numériques de l'original qu'il est nécessaire pour en permettre une visualisation virtuelle. Pour des textes, ces images sont en deux dimensions et, lorsqu'on numérise un ouvrage entier, on produit une image par page ou par double-page, en fonction de l'état de la source.

[17] Pour la majorité de nos témoins, nous avons pu bénéficier de numérisations existantes et généralement d'excellente qualité, les imprimés (16e-19e siècle) étant déjà disponibles sur les sites Gallica, Numelyo ou de la Bibliothèque David Houärd. Certains manuscrits ont nécessité une campagne de photographies personnelles ou menées par les institutions. L'avantage de posséder des facsimile numériques était majoritairement la possibilité d'entraîner un modèle d'HTR afin d'accélérer la récupération du texte sous format numérique.

[18] Les étapes suivantes consistent en effet en l'acquisition du texte sous forme de chaînes de caractères numériques, ce pourquoi nous avons choisi d'utiliser *Transkribus*, un outil de transcription automatique des documents avec interface complète développé à l'origine à l'Université d'Innsbruck. C'est le seul outil que nous avons choisi d'utiliser, afin de conserver la plus grande liberté possible dans le traitement des données. Les fonctionnalités de découpage et typage[3] des

3 Cette fonctionnalité de typage, à présent automatisable par entraînement, était alors entièrement manuelle.

zones nous permettaient de gagner beaucoup de temps sur l'encodage : en séparant déjà les textes de coutume et de commentaire du paratexte lié uniquement au format (entêtes, etc.) et en donnant à chaque zone un type correspondant à sa nature, nous nous donnions les moyens d'automatiser la recherche de ces différents types de texte et leur traitement en vue du format final.

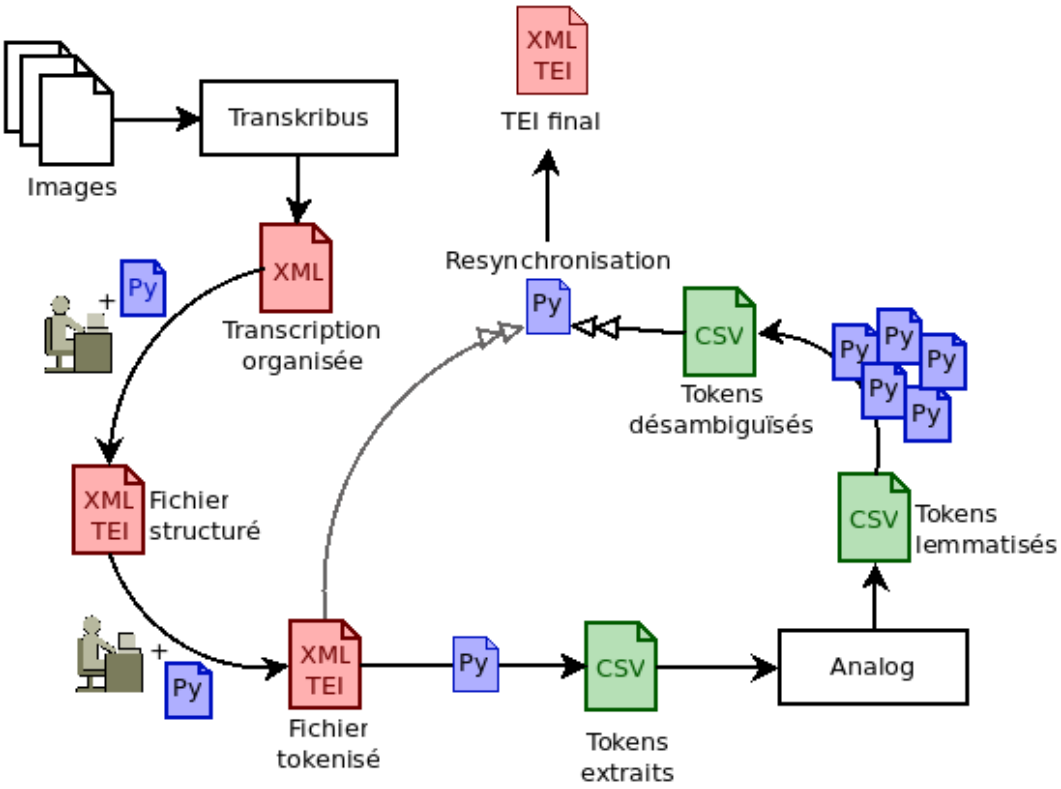

Figure 1 : La chaîne de numérisation des témoins, de l'image au TEI final

C'est cette fonctionnalité qui, en 2019, nous a fait préférer *Transkribus* à notamment *Kraken*, le premier étant encore gratuit et l'interface HTML du second n'étant alors pas encore suffisante pour gérer des mises en page complexes comme celle de Le Rouillé (1539) avec glose encadrante.

[19] Bien qu'il soit possible de faire la majorité de son travail de mise en forme directement sur *Transkribus*, qui gère notamment l'italique et les corrections éditoriales, nous avons choisi de séparer totalement la définition de la structure logique de la mise en forme et n'avons donc exploité que l'HTR et le découpage des zones de texte du logiciel.

[20] La création d'une base de données enrichie d'informations linguistiques rendait la production de fichiers en texte brut contre-productive. L'utilisation de la syntaxe XML s'est imposée naturellement, ainsi que le modèle de la Text encoding initiative (TEI). Le modèle TEI contient tout le nécessaire pour à la fois

structurer un document, renseigner ses métadonnées et détailler la nature lexicale et morpho-sémantique de chaque mot dans le texte. La recherche par lemme et la production de tableaux de concordances montrant les formes y étaient donc parfaitement compatibles et largement rentabilisés par la communauté scientifique.

[21] *Transkribus* utilisant un format XML propre pour stocker les données des transcriptions et permettant un export au format TEI portant toutes les indications de type et mise en forme données par le transcripteur, le travail de transformation vers notre propre schéma était minimal. Nous avons pour ce faire choisi d'utiliser une combinaison de scripts Python et XSLT, bien plus rapide que la transformation manuelle via une interface.

[22] Le choix de Python pour la transformation principale s'est fait majoritairement pour la rapidité de l'exécution du code et sa simplicité d'écriture : sur les deux tomes du Basnage (1678) dont la version finale de *base* pèse 96 Mo, la transformation prenait quelques secondes, pour deux cents lignes de code. D'autres transformations ponctuelles et plus exigeantes en XPath ont ensuite été gérés avec XSLT ou des expressions régulières. Concevoir ces transformations nécessitait bien entendu de savoir quelle forme allait prendre la base de données. Cette forme se devait d'être homogène, malgré des structures très différentes entre les témoins du corpus.

3 Structures logiques à homogénéiser

3.1 L'hétérogénéité des structures

[23] Nos manuscrits, du 13e au 15e siècle, ont ainsi une structure généralement plus simple, basée sur la rubrication[4] et les pieds-de-mouche[5]. Le *TAC* étant fragmentaire et notre seul exemplaire de cet état particulier de la coutume, il est difficile d'être définitif sur sa structure dans son ensemble. Cependant le fragment dont nous disposons ne semble pas établir de macrostructure ou de hiérarchie autre que la rubrique, contrairement au *Grand Coutumier de Normandie* (*GC*), dont nous possédons la liste des titres, que l'auteur rassemble sous des *Distinctions*.

[24] Étonnamment, les notes marginales sont rares dans ces manuscrits, généralement écrites dans des mains plus récentes, probablement par les praticiens dont ils étaient la référence[6], telles les notes de Charles Morisse. De même, texte de coutume, illustrations par l'exemple ou commentaires sont rédigées à la suite, dans la continuité textuelle recherchée dans les manuscrits. Cette structure se complexifie considérablement dans nos témoins imprimés. Les deux premiers, nommément Le Rouillé (1539) et Terrien (1578), n'ont graphiquement en commun que d'être les premiers à visuellement séparer coutume et commentaire.

4 Une rubrique, c'est-à-dire un intitulé tracé à l'encre rouge, marque le début d'une nouvelle division de texte.

5 Pied-de-mouche (*capitulum*) : signe ressemblant à un *c* barré d'un trait vertical, souvent traduit, dans nos éditions actuelles, par un changement de paragraphe.

6 Pour une synthèse sur l'histoire du livre juridique, voir en particulier Mellot (2005).

De deliurance de namps. Fo. xvi.

a noter que se A. dict q̃ B. a prins ses nãps en iusticiãt et B. dict que non. le sergent doibt prendre plege de B. datendre lenqueste. ⁊ se par lenqueste il est cõuaincu et attaint dauoir prins les namps / il doibt estre en grãt amende pour ce que cest aulcune espece de larcin.

¶ Et se lenqueste dict quil ne print point les nãps / cil qui laccusoit a tort le doibt griefmẽt amẽder. Cestassauoir par amende pecuniaire / et non pas par prisson / ⁊ se lẽqueste le met a nõ scauoir / si demourra il en amende pour ce quil a failly a prouuer sa plaite: mais il poursuyura ses nãps cõme chose adiree / sil ne les treuue: cestassauoir sil ne scait ou ses nãps sont. ¶ Item en suyt apres en texte.

q ¶ Aulcũs tiennent de leur seigneur nu a nu, ⁊ aulcũs ont moyẽ ⁊c. par ce texte appert q̃ le chief seigñr peut faire sa iustice sur tous les fiefz qui sont tenuz de luy soit nuemẽt ou p moyen pour la droicture qui luy est deue du fief. et peut bien tenir les ples eu fief qui est tenu de luy soit nuement ou par moyen iacoit ce quil soit noblement tenu / car pour aultre cause ne les y pourroit pas tenir.

¶ Aps ẽsuyt en texte

r ¶ Mul ne peut tenir les namps a ceulx q̃ ne tiennent de son fief ⁊c. par ce texte est a noter que faulcũ est trouue en dommage dault truy en pẽsent meffaict cil a qui le dõmage est soit noble tenãt ou aultre le peut iusticier par ses nãps : ou larrester sil nauoit nãps par la detention de son corps / iusques a ce q̃l eust baille nãps ou plege: ⁊ se on ne trouuoit le malfaicteur en prefẽt meffaict / sil le pourroit on suyuir / ⁊ le iusticier par ses nãps / ou arrester cõme dict est / supposé q̃l fust hors du fief ou il auroit faict le dõmage pourueu que ce fust de prõpte poursuyte. ¶ Apres ensuyt eu texte.

s ¶ Se le querelle apres la deliurance de ses namps se default au iour assigne ⁊c. Par ce texte est a noter q̃ se aulcun iusticie pour sa rente / et lautre faict deliurance de ces nãps et il a iour assigne se celuy q̃ fist deliurance se default tant quil soit mis en amende : lautre aura attaint resaisine de ses namps pour exploiter pour ses arrerages: et nen pourra lautre auoir deliurance : supposé que se soit auant veue faicte / et se cil qui fist la iustice se default tant quil fust mis en amende: lautre auroit attaint a soy en aller sans iour en la saisine de ces nãps: ⁊ ne pourroit depuis cil qui faict la iustice refaire vne aultre iustice pour les arrerages de ladicte rẽte. Mais se iustice estoit faicte pour dommages / ⁊ cil qui auroit faict deliurance se defailloit tant quil fust mis en amende: lautre auroit attaint a prouuer le dõmage par iugemẽt / et aussi se cil q̃ auroit iusticie se defailloit tant quil fust mis en amende : lautre nauroit attaint fors a sẽaller sãs iour en la saisine ỏ ses nãps et pourroit lẽbiẽ vne autre fois faire approcher p action pour lesdictz dommages.

¶ Item par la fin du texte qui met.

t ¶ Et si pourra pour chasser ptre luy les dõmages ⁊c. ¶ Lẽ peut noter que se les nãps estoient empirez pour cause de la prinse comme p la raison du lieu ou pour la maniere de les tenir: ou par telles manieres. ¶ Il a q̃ les namps seroient pourroit bien pourchasser son dommage vers cil qui print les namps.

¶ Sur ce chapitre len peut mouuoir plusieurs doubtes. Le premier est si tous gaige pleges sõt entẽdus sur ce chapitre : et q̃ doibt auoir la saisine de ce q̃ pend eu gage plege le proces pendant. ¶ A ce doubte len peut respondre q̃ tous gages pleges soiẽt pour descort de iustice iurisdiction de edifice ou aultres choses sõt cõpris et entenduz sur ce presẽt chapitre. ¶ Et oultre que cil des parties qui auroit este possesseur par an et iour de la chose descordable demourra sayst le proces pendant. ¶ Et sil estoit descord de cil q̃ auroit este possesseur de la chose litigieuse: elle seroit sequestree en main de iustice: le proces pẽdant iusques a ce q̃ le proces sur la possession soit discute ou que prouision y soit donnee par iustice. ¶ Et le proces fini la saisine soit rendue a cil q̃ auroit este trouue possesseur : ⁊ puis plaideroit on sur la proprietẽ qui vouldroit. Et y a ordonnance deschiquier en ces termes ⁊ en telle substance. Le secõd doubte est se aulcun met gage plege a lẽcontre dun noble tenant: disant quil na iustice iurisdiction ne pouoir aulcun de iusticier sur le lieu iusticie. Et il en echiet se il forfaict le fons. ¶ A ce doubte len peut respõdre et doibt len dire q̃ ouy: puis quil y auroit desauoue de seigneurie. Cestassauoir de court et de vsage. Jacoit ce que le

Figure 2 : Exemple de folio de Le Rouillé (1539) imprimé avec glose encadrante

[25] L'ouvrage de Le Rouillé (1539), imprimé en caractères gothiques avec une glose encadrante sur deux colonnes, ressemble fort à une Bible de Gutenberg (fig. 2), glosé par des notes françaises et des *additio* latines signées de l'auteur. Les caractères étant réguliers, l'HTR n'a eu aucun mal à transcrire le texte français. Le texte latin, en revanche, contenait un grand éventail d'abréviations latines. Faute d'assez de modèles pour chaque abréviation dans ses pages de référence, le pourcentage d'erreurs de l'HTR est assez élevé dans les *additio* latines. En conséquence, nous faisons le choix, dans le temps qui nous est imparti, de ne pas lemmatiser les *additio* de Le Rouillé (1539).

[26] Quarante ans plus tard, Terrien (1578) utilise les caractères romains et a abandonné glose encadrante et colonnes. Il possède cependant le système de références le plus ardu de ce corpus. Suivant la tendance générale de multiplication du matériel paratextuel, il comporte six types de notes et/ou commentaire, parfois uniquement distinctes du matériau législatif par la taille légèrement inférieure de la fonte, sans saut de ligne permettant de facilement repérer le passage d'un type d'information à l'autre. Sur certaines pages, un travail rapide est même impossible pour les séparer efficacement. La composition elle-même fait parfois défaut, tant il y a d'appels à enregistrer sur la page, oubliant ou confondant certains numéros de note.

[27] Ainsi, malgré une macrostructure claire, ces deux témoins ont nécessité plus de travail manuel que certains manuscrits. Une typologie précise des notes de Terrien (1578) s'avère trop ambitieuse. Il nous a donc fallu ajuster notre chaîne de traitement et notre modèle d'encodage en fonction de la précision des informations disponibles.

[28] La structure des témoins modernes, c'est-à-dire des 17e et 18e siècles, est déjà bien plus proche de nos habitudes de lecture. Fidèles à la tradition de la glose médiévale, les auteurs accordent une grande importance à rendre explicite l'intertextualité dans leur œuvre, mentionnant, citant ou critiquant leurs prédécesseurs, les références juridiques antiques ou le matériau coutumier. Cela se matérialise notamment par une séparation visuelle plus évidente bien que toujours basée sur une différence de taille des fontes, entre citation de la coutume réformée ou d'ordonnances, en général en tête de section, et le commentaire de l'auteur, souvent assorti de jugements de l'échiquier de Normandie.

[29] La page devient une véritable unité et les éléments superflus pour notre édition se font nombreux : entêtes, numérotation des pages, réclames[7] doivent être triés pendant le formatage des données. Les manchettes jouent ici souvent le rôle d'une rubrique permettant de rapidement repérer le sujet de la portion de texte locale, cependant pour notre usage elle répète l'information déjà donnée par les titres de livre et de chapitre, voire de sous-chapitre.

7 On appelle *réclame* le fait d'imprimer, à la fin d'un cahier, le·s premier·s mot·s du suivant, afin de ne pas se tromper dans l'ordre des cahiers.

3.2 Structurer la base de données

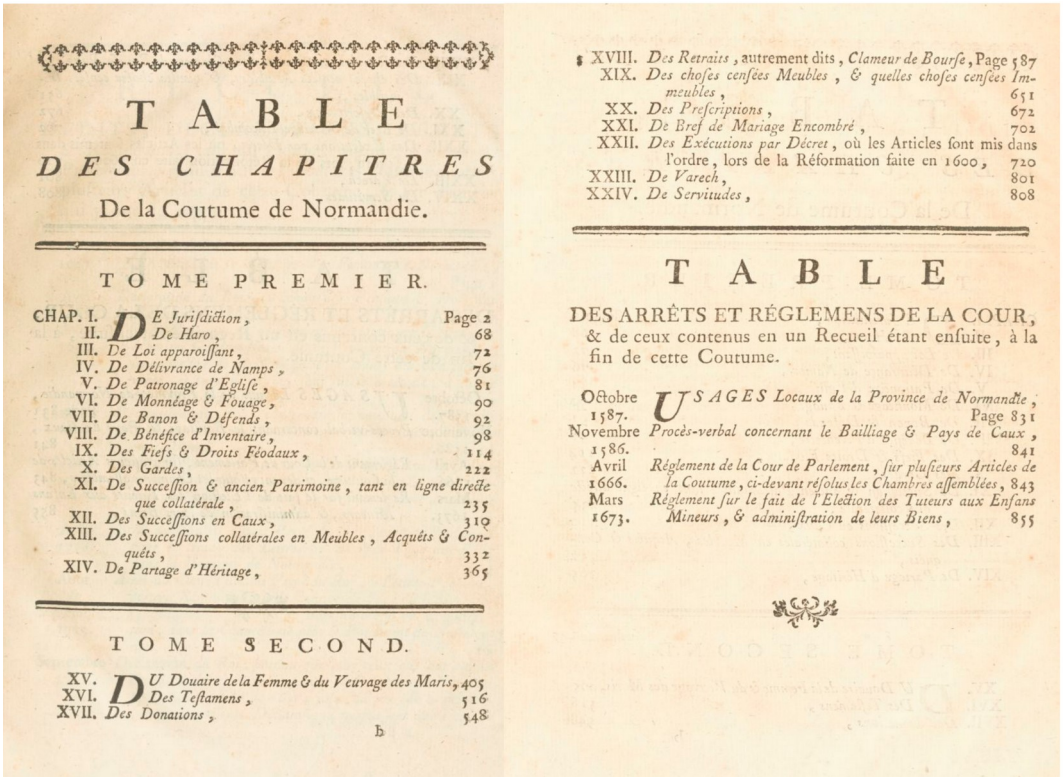

TABLE
DES CHAPITRES
De la Coutume de Normandie.

TOME PREMIER.

TOME SECOND.

b

TABLE
DES ARRÊTS ET RÉGLEMENS DE LA COUR, & de ceux contenus en un Recueil étant enſuite, à la fin de cette Coutume.

Figure 3 : Première double-page de la table des chapitres de Pesnelle (1771)

[30] Malgré cette grande hétérogénéité, une structure homogène était nécessaire pour pouvoir interroger la base de données sur un système d'arborescence fixe. Il s'agissait donc de définir le nombre de divisions internes de nos fichiers et leur typologie. Nous ne comptons pas ici en tant que niveau le tome car les deux auteurs ayant écrit plusieurs tomes, Basnage (1678) et Pesnelle (1771), ont conçu les deux tomes comme une seule unité dont la taille seule nécessitait la séparation en deux volumes, ainsi que l'on peut le constater dans leurs tables des matières. Dans celle de Pesnelle (1771), la division thématique prime en effet graphiquement sur la division matérielle. La coutume vient en premier, remplit d'abord le premier tome et continue au début du second, puis viennent les autres matières abordées (fig. 3). La décision d'organiser les fichiers sur trois niveaux est venue du constat que c'était le nombre maximal de niveaux que l'on pouvait trouver dans nos différents témoins.

[31] Le tableau 2 montre les niveaux détectés dans les principaux témoins, alignés sur le niveau le plus fin (niveau 3) :

Témoin	Niveau 1	Niveau 2	Niveau 3
TAC	/	/	[Rubriques]
GC	/	Distinctions	[Titres]
Morisse *Notes*	[Sauts de page]	Distinctions	[Titres]
Le Rouillé (1539)	Livres	Distinctions	Chapitres
Terrien (1578)	Livres	Chapitres	[Numérotation]
Bérault (1614)	/	[Titres]	[Numérotation]
Basnage (1678)	/	[Titres]	[Numérotation]
Merville (1731)	/	[Titres], Titres	[Numérotation]
Pesnelle (1771)	[Trois tables des matières distinctes]	Chapitres	[Numérotation]

Tableau 2 : Comparaison des niveaux de structure

La seule exception est une section de niveau 3 particulière contenant plusieurs titres, utilisée par plusieurs des imprimés. Cependant, comme ces titres ont été placés sous le même numéro et le cas ne se présentant que trois fois dans le corpus et systématiquement sur le même élément de coutume, nous avons compris que pour les auteurs des sources, ils étaient une unique exception à une structure autrement stable. Nous avons donc choisi de traiter ce cas à part et garder la structure à trois niveaux désignés, dans la structure TEI, par un même type d'élément : (*division*). Les peuvent être imbriquées : une division peut contenir d'autres divisions. On les distingue donc par le nombre d'autres parentes, mais aussi par la valeur de leur attribut @type.

[32] C'est ici qu'intervient la seconde opération d'homogénéisation. Le tableau 2 montre effectivement à quel point la typologie est particulière à chaque témoin : bien que certains termes reviennent, ils ne sont pas toujours appliqués aux mêmes niveaux de structure. Pour définir notre standard, nous avons d'abord considéré Le Rouillé (1539) et Terrien (1578), qui sont les seuls à explicitement nommer leur niveau supérieur (fig. 4). Terrien (1578) classe en effet ses différents sujets généraux en autant de livres, puis opère une classification plus fine dans ses chapitres et, finalement, numérote ses références commentées à l'intérieur de ses chapitres. C'est ce principe que nous avons choisi de garder pour chaque témoin.


DV LIVRE I.
Qui eſt.
De la Iuſtice, & du Droict des Normans.
Du Droict & de Iuſtice. chap. j. pag. 1.
Des parties dõt noſtre droict eſt compoſé. chap. ij. 9
De couſtume, & des loix, vſage & ſtyle. chap. iij. 10
De l'obſeruance des ordonnances. chap. iiij. 12

Figure 4 : Extrait de la table des matières de Terrien (1578)

[33] La typologie des divisions se fera en anglais, pour faciliter le partage des fichiers, sur des termes les moins ambigus possibles. Ainsi le premier niveau portera-t-il *part* comme valeur d'@type, le deuxième niveau *chapter*, suivant la dénomination de Terrien (1578) et Pesnelle (1771), et le troisième niveau *section*, car il s'agit d'une découpe plus fine du texte, sans changement thématique. Cette organisation permet de traduire toutes les structures présentes dans le corpus, y compris lorsque moins de divisions sont utilisées dans la source. Le *TAC*, par exemple, n'ayant qu'un niveau de découpe du texte, le fichier final comporte une seule div[@type="part"], qui comprend elle-même une seule div[@type="chapter"], qui contient l'intégralité du texte transcrit. L'unité de base reste la même : la section.

3.3 Découpage des sections

[34] Chaque témoin présentant différemment ses informations, les div[@type="section"] sont donc hétéroclites selon les témoins. Nous avons cependant pu isoler plusieurs types d'informations généralement utilisés : parfois des titres, du matériau législatif – généralement coutume ou ordonnance –, un commentaire général et des notes ponctuelles. Les titres de section ont sans hésitation été conservés dans des éléments *head*. S'appuyant sur les règles TEI, une structure en paragraphes assortis de notes nous est apparue comme logique. En encodant les notes à leur place au sein du texte, nous nous donnions la possibilité de les placer, lors de l'affichage, comme il nous semblerait alors bon. Considérant que le commentaire concernait le matériau législatif, nous avions d'abord envisagé de faire du texte coutumier ou des ordonnances royales le texte principal et du commentaire.

[35] Toutefois, nous avons remarqué que dans certains témoins, la référence elle-même n'était utilisée que comme base de réflexion et non comme le centre d'une section. Afin de refléter ce fait, tout autant que l'exposition de la coutume comme élément principal de la section, nous avons choisi de marquer ces citations comme telles, dans des éléments *quote*. Le commentaire général sur une citation serait donc représenté par des paragraphes <p> et les notes par des éléments <note>. Nous avons choisi de conserver les numéros des appels de note dans les attributs @n de chaque note, ce qui nous permet de les restituer ou de les ignorer à l'affichage selon les besoins.

3.4 Adaptations et partis pris

[36] La structuration du corpus nous a également permis de repérer des similitudes dans l'organisation des imprimés modernes, bien qu'elles ne soient pas toujours explicites. Le Rouillé (1539) sépare le *liure coutumier de Normendie* et les autres références législatives (Charte aux Normands, serments des avocats normands, ordonnances de l'Échiquier…) (fig. 5). Bien qu'il les expose sans distinction, Bérault (1614) (fig. 6), quant à lui, commence par les différents chapitres de coutume, puis passe en revue les usages locaux et enfin donne la *Charte aux Normands* et autres édits et procès verbaux. Cette organisation est assez constante par la suite. Basnage (1678), dans sa table des matières, donne les différents titres abordés : *A quoy sont ajoutez les Vsages Locaux, la Charte aux Normands, le Procez Verbal, & les Reglemens de la Cour du Parlement de Normandie*. Merville (1731) donne explicitement les différents *Tit[res]* de coutume, puis écrit : *Les Usages Locaux*, *Articles placitez sur la Coutume, de 1666*, etc., sans la mention de *Titre*. Pesnelle (1771 : viii-xv), enfin, divise son sommaire en trois : une *Table des chapitres de la Coutume de Normandie*, s'étendant sur les deux tomes, puis une *Table des arrêts et réglemens de la Cour*, et enfin *Recueil d'édits, déclarations, arrêts et Réglemens*.

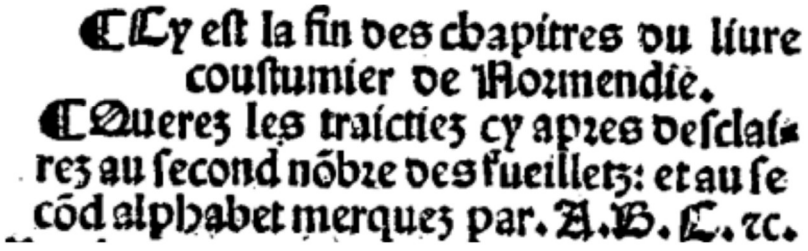
CY eſt la fin des chapitres du liure
couſtumier de Normendie.
CQuerez les traictiez cy apres deſclai-
rez au ſecond nõbre des fueilletz: et au ſe
cõd alphabet merquez par. A.B. C. ꝛc.

Figure 5 : Extrait de la table des matières de Le Rouillé (1539)[8]

[37] Pour Pesnelle (1771), chaque div[@type="part"] correspond à l'une de ces trois tables. Afin de permettre une consultation plus facile sur la base de par-

8 « Cy est la fin des chapitres du livre coustumier de Normendie. / Querez les traictiez cy apres desclairez au second no[m]bre des fueilletz : et au seco[n]d alphabet merquez par. A.B.C. etc. »

ties de texte et considérant que ces distinctions étaient déjà faites dans les regroupements thématiques au sein du témoin, tout comme les termes utilisés dans les tables des matières, nous avons choisi de matérialiser ces différences dans les autres imprimés. On compte dans nos témoins ces différentes parties implicites : exposition et/ou discussion sur la coutume, usages locaux, arrêts, puis chartes anciennes et extraits des recueils du Parlement de Normandie. Nous avons donc opéré quelques divisions nouvelles dans les imprimés, tout en les signalant comme ajouts éditoriaux grâce à un attribut @resp. Bérault (1614) contient donc à présent trois div[@type="part" and @resp], la première correspondant aux chapitres de coutume, la deuxième aux différents usages locaux, la dernière aux éditions de documents anciens suivant les usages locaux, tandis que chaque titre de la table des matières est représenté par une div[@type="chapter"], à l'intérieur de l'une des trois parties.

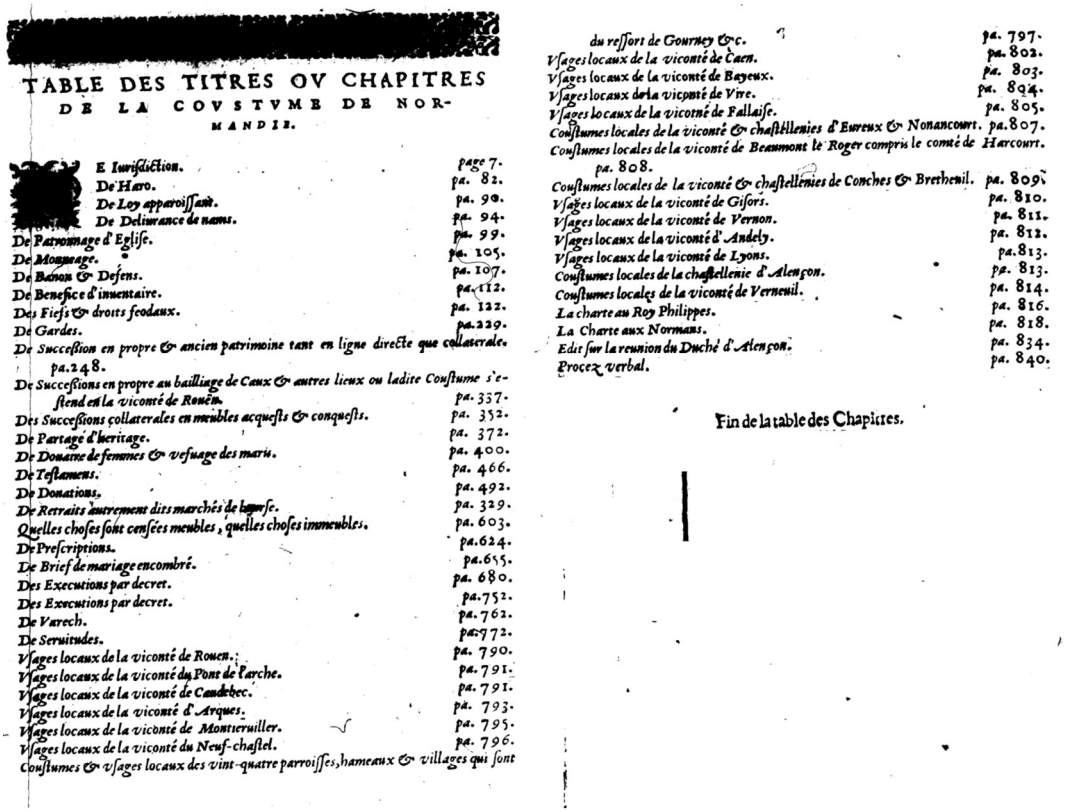

TABLE DES TITRES OV CHAPITRES DE LA COVSTVME DE NORMANDIE.

Fin de la table des Chapitres.

Figure 6 : Table des matières de Bérault (1614)

[38] Nous avons pris ce parti afin de faciliter la navigation dans les chapitres sur l'interface graphique. Dans le même but, afin de pallier l'absence de titres, lorsqu'une partie est faite sur un type commun au corpus (usages locaux, chartes

royales...), nous utilisons l'attribut @subtype pour renseigner son thème avec des valeurs unifiées. Cela nous permet, lors de la génération automatique des sommaires sur *MaX*, d'afficher un titre de substitution signalé par des crochets carrés.

[39] La structure a donc pu être homogénéisée sans que les choix éditoriaux puissent être confondus avec la transcription de la source. Cette adaptation était nécessaire, initialement parce que le corpus devait être utilisable selon les mêmes termes, mais fondamentalement à cause du changement de format. Ces mêmes considérations ont également guidé notre définition de la microstructure : l'encodage au mot, voire au caractère, du contenu textuel.

4 Graphie, modernisation et lemmatisation

4.1 Graphies anciennes, modernisation et XML

[40] L'encodage des caractères, comme leur transcription, demande de prendre en compte la graphie, les caractères anciens, mais également parfois des abréviations. Notre orthographe étant majoritairement plus récente qu'aucun de nos témoins, un logiciel d'OCR basé sur un dictionnaire récent peut considérer comme fautif un caractère juste pour les normes graphiques de son temps. La distinction entre les caractères *u* et *v*, par exemple, dans Le Rouillé (1539), est une question de position dans le mot et non de prononciation : si la lettre est l'initiale du mot, elle sera écrite *v*, sinon ce sera un *u*. Ainsi lit-on *liure*, ce qui peut consister, selon le lecteur, une difficulté d'accès au texte, et doit pourtant être conservé comme représentation exacte de l'état de la source, selon les principes de réutilisabilité des données énoncés plus haut. Cette difficulté est plus grande encore lorsqu'il s'agit d'abréviations, à l'origine utilisées pour gagner de la place sur la ligne et que l'on transcrit traditionnellement entre crochets.

[41] Le processus d'HTRisation se base sur le caractère. La machine devant considérer un caractère comme une seule unité, qu'il s'agisse ou non d'une abréviation, cela signifie de pouvoir donner au modèle un caractère unique pour représenter cette unité. Toutes les abréviations présentes dans Le Rouillé (1539) sont déclarées comme caractères Unicode, bien que la plupart des polices d'écriture ne les connaisse pas et ne puisse donc pas les afficher. Ce n'est pas le cas de *Junicode*, une police d'écriture développée par Peter S. Baker de l'Université de Virginie, pour les transcripteurs et éditeurs de textes anciens et médiévaux. L'utilisation de cette police nous donne la possibilité de représenter une abréviation dans une version typographiée, bien qu'elle soit majoritairement échapée en code HTML dans les fichiers XML de base.

[42] Nous avons donc fait le choix de garder ces caractères dans le texte, mais d'en proposer également la résolution. Le modèle TEI nous permet en effet d'utiliser l'élément <choice> à l'emplacement d'un caractère. Ce dernier permet de stocker deux alternatives pour un caractère. À l'intérieur du <choice>, pour un caractère ancien, un élément <orig> peut contenir la transcription exacte du caractère, tandis qu'un élément <reg> contient une résolution du caractère ancien :

```
con<choice><orig>ſ</orig><reg>s</reg></choice>titué
```

Extrait de code simplifié : *constitué* avec *s* long et *s* rond.

Nous avons maintenu l'utilisation de l'élément <choice> pour un caractère tel que le *s* long, facilement résolvable automatiquement, par cohérence avec le système utilisé pour nombre d'autres cas que nous souhaitions signaler tels que l'alternance entre le *d* rond et le *d* droit jusqu'à Le Rouillé (1539), le *i* pointé ou non dans les manuscrits, ainsi que certaines lettres comme les *é* majuscules réalisés comme un *e* suivi d'une apostrophe. Bien que le temps nous ait manqué, nous souhaitions également utiliser cette méthode d'encodage pour moderniser l'utilisation du *i/j* et du *u/v*.

[43] L'élément <choice> peut également accueillir d'autres couples d'éléments comme <am> et <expan>, permettant d'encoder ensemble une abréviation et sa résolution :

```
autrem<choice><am>ē</am><expan>en</expan></choice>t
```

Extrait de code simplifié : *autrement* avec *en* abrégé et résolu.

[44] Ce système d'encodage nous donne ainsi la possibilité de garder l'état originel de la source, tout en en proposant une version standardisée qui peut donc être utilisée, notamment, pour unifier la graphie des résultats de recherche sur l'interface, ou, dans l'idéal, donner à l'utilisateur le choix de la version qu'il souhaite lire, en distinguant modernisations et résolutions tout en conservant l'élément <choice> comme cadre commun.

[45] Nous utilisons les <choice> au caractère car un même mot peut ainsi donner à la fois archaïsmes et abréviations, comme :

```
empe
<choice><orig>ſ</orig><reg>s</reg></choice>
ch
<choice><am>ā</am><expan>an</expan></choice>
t
```

Extrait de code simplifié : *empeschant* avec un *s* long et *an* abrégé et résolu

Lorsque plusieurs du même type se suivent, nous avons cependant rejoint les deux afin d'alléger les fichiers déjà très lourds (la version *lighter* de Basnage (1614), ne contenant ni informations graphiques ni élément <choice> car entièrement modernisée, pèse à elle seule 68 Mo !) :

<choice><orig>ſegno2</orig><reg>segnor</reg></choice>

Extrait de code simplifié : *segnor* issu du *TAC*

4.2 Tokenisation et lemmatisation

[46] En plus de cet encodage des caractères particuliers, le projet *ConDÉ* a souhaité mettre en place un encodage morpho-syntaxique et la lemmatisation des différents tokens. Cet encodage rend ainsi possible l'interrogation du corpus par lemmes ou catégories grammaticales plutôt que par mot-forme. Pour ce faire, nous avons tout d'abord tokenisé le texte dans les fichiers TEI déjà structurés.

[47] Le terme *token* désigne dans ce cas l'unité lexicale de base : la tokenisation d'un fichier textuel signifie la définition des unités selon lesquelles il sera analysé. Ces unités sont en général des mots, mais peuvent inclure également les signes de ponctuation ou des chiffres. En TEI, la tokenisation signifie isoler chaque unité dans son propre élément <w> (*word*)[9] numéroté.

[48] La tokenisation s'est faite sur des bases graphiques, majoritairement pour des raisons pratiques, c'est-à-dire à la fois pour pouvoir automatiser le processus avec des expressions régulières, et pour assurer la compatibilité des données avec le logiciel choisi pour la lemmatisation : *AnaLog*, avec le jeu d'étiquettes *PRESTO*. Lors d'une rencontre avec un caractère non-alphanumérique, *AnaLog* change en effet de token avant et après lui, ce qui décalait toute notre numérotation de tokens et rendait les informations incompatibles : un token *vingt-cinq*, initialement numéroté 51, suivi de *livres*, numéroté 52, se retrouvait par exemple divisé en tokens no. 51 (*vingt*), no. 52 (-) et no. 53 (*cinq*). Les informations produites lors de la lemmatisation, pour le nouveau no. 52 (-) auraient été associées au token no. 52 (*livres*) originel, décalant les informations de tous les tokens suivants. Le seul moyen de pouvoir réaligner les informations après lemmatisation était donc de faire cette segmentation nous-mêmes auparavant, afin d'avoir une numérotation fixe du début à la fin.

[49] Dans la mesure où le post-doctorant du projet avait une expérience avec le logiciel d'annotations *AnaLog*, qu'il avait utilisé lors de sa participation au projet *PRESTO*, et qu'il l'avait utilisé encore récemment lors du projet *EPELE* basé à Unicaen, nous avons repris ces outils pour l'annotation du corpus. Le jeu d'étiquettes *PRESTO*, pensé pour la diachronie longue du français, est du reste compatible avec la base de données *Frantext*, ce qui assure l'interopérabilité du corpus et son futur partage avec ce dernier. Nous avons également préféré utiliser ces outils qui nous permettaient un meilleur contrôle des étapes d'annotation, dans la mesure où nous ne savions pas si le genre juridique des textes d'une part, le jargon de la coutume d'autre part et la langue normande enfin pour les textes les plus anciens,

9 L'élément <w> « représente un mot grammatical (pas nécessairement orthographique) » (TEI 2022, annexe C).

n'allaient pas perturber les logiciels, entraînés à travailler sur des textes littéraires et sur des langues standardisées.

[50] Quelques scripts en langage Python nous ont permis d'automatiser toutes les étapes intermédiaires : la numérotation des tokens, l'extraction desdits tokens dans un fichier au format CSV lisible par *AnaLog* et, après la lemmatisation elle-même, l'intégration des informations produites par *AnaLog* aux tokens originels en TEI. Ces deux dernières opérations s'appuient sur la numérotation préalable des tokens, garantissant que le script puisse comparer les numéros des tokens TEI à ceux du fichier CSV produit par *AnaLog*, qui gardait scrupuleusement l'ordre des tokens.

[51] Le résulat final est, pour chaque titre ou chapitre, la succession d'éléments <w>, chacun assorti d'attributs @n, @lemma et @pos. Le premier contient le numéro attribué au token matérialisé par l'élément <w>, le deuxième contient le lemme qui lui correspond, et le troisième sa catégorie grammaticale (POS). Ainsi cette suite de quatre tokens issue du premier tome de Basnage (1678) :

```
<w lemma="À+LE" n="844" pos="S+Da">aux</w>
<w lemma="MEUBLE" n="845" pos="Nc">meubles</w>
<w lemma="ET" n="846" pos="Cc">&amp;</w>
<w lemma="ACQUÊT" n="847" pos="Nc">
    acque
    <choice><orig>ſ</orig><reg>s</reg></choice>
    ts
</w>
```

Figure 7 : Exemple de tokens lemmatisés encodés en TEI : *aux meubles & acquests*

4.3 Désambiguïsation semi-automatisée

[52] Il est cependant à signaler qu'un même mot-forme pouvant parfois relever de plusieurs lemmes et/ou catégories morpho-syntaxiques, *AnaLog* nous en donnait toutes les possibilités repérées dans son dictionnaire de base, sans faire lui-même de choix. Cette ambiguïté formelle peut parfois être facilement résolue grâce à de simples règles binaires. Afin d'affiner la lemmatisation, nous avons donc traduit en langage Python les règles les plus fiables et qui ne demandaient aucune supervision humaine.

[53] Ce travail demande de synthétiser de manière binaire des règles grammaticales, en se basant sur une suite de tests que la machine doit effectuer. Ces règles doivent être agencées par ordre de précision et d'utilité. Par exemple, trier les déterminants devait être fait avant le travail sur les pronoms car il permettait de résoudre d'emblée un certain nombre d'ambiguïtés. À l'intérieur du script sur les déterminants, on teste d'abord les formes dont on sait qu'elles doivent systéma-

tiquement être interprétées d'une manière précise, puis on traite les cas les plus fréquents, et ainsi de suite.

[54] Pour certains termes spécifiques au vocabulaire juridique et donc potentiellement inconnus d'*AnaLog*, il suffisait d'en dresser des listes. Le script peut ainsi s'appuyer sur elle pour déterminer la nature exacte de certains termes, s'il les trouve dans une des listes. Il en est de même pour des termes ambigus mais relevant du vocabulaire juridique qui, dans un corpus de cette spécialité, devaient être compris comme tels. Par exemple, le terme *bailly* peut, en règle générale, être analysé comme appartenant aux verbes *bailler* ou *bâiller*. Dans notre corpus, cependant, il s'agit systématiquement de l'officier responsable d'un baillage et la forme doit donc être analysée comme nom commun.

[55] Dans l'idéal, la majorité des règles de désambiguïsation seraient traduisibles en code Python pour la machine. Cependant, pour des raisons d'efficacité nous nous sommes tenus à cinq scripts, à exécuter dans cet ordre précis : désambiguïsation 'simple', c'est-à-dire comparaison du mot-forme avec les différentes listes associées à des catégories grammaticales certaines ; définition des tokens étrangers, majoritairement latins, lorsqu'ils étaient encadrés par deux mots déjà définis comme étrangers ; examen des candidats-déterminants ; examen des candidats-pronoms ; enfin, vérification sur les adverbes pouvant introduire des conjonctions de subordination. Nous synthétisons dans le tableau suivant les différentes vérifications automatisées :

Ordre des vérifications	Base
1	Comparaison de chaque mot aux listes de formes certaines et réécriture du lemme et du POS si nécessaire.
2	Si le mot est encadré par deux mots catégorisés comme de langue étrangère (généralement du latin), on le définit comme tel.
3	Vérification des déterminants. Si un mot peut être un déterminant mais sans certitude et que le mot suivant est un nom commun, on vérifie le reste du contexte. Si les vérifications échouent, on laisse les valeurs de POS telles quelles.
4	Vérification des pronoms. Même principe que la précédente.
5	Vérification des conjonctions. Même principe que les précédentes.

Tableau 3 : Résumé des opérations de désambiguïsation en Python

Avant désambiguïsation automatique, 29,5 % des tokens avaient été validés automatiquement par *AnaLog*, le reste étant donc soit ambigu, soit inconnu du dictionnaire. La désambiguïsation a permis d'en traiter environ 40 % supplémentaires selon les témoins. Après quelques vérifications manuelles subséquentes sur les tokens encore ambigus ou inconnus, le taux de lemmes certains s'élève à plus de 90 % pour les imprimés, jusqu'à 96 % pour le Terrien, bien que 2-5 % d'erreurs puisse subsister à cause d'erreurs d'HTR rendant l'analyse difficile.

5 Conclusion

[56] Le projet *ConDÉ* a produit une base de données textuelle pour les recherches d'historiens, historiens du droit et linguistes sur les textes de la coutume normande, du 13e au 19e siècle. Sa démarche était, ce faisant, de conserver les données intermédiaires pour favoriser leur réutilisabilité. Afin de mettre sur pied les éditions numériques de ces différents témoins sur une structure commune de corpus, nous avons choisi l'outil *Transkribus*, permettant d'entraîner un HTR et de diviser les images en zones de texte typées, le standard TEI pour l'encodage final des données textuelles et le langage Python pour effectuer les diverses transformations, d'un état de fichier au suivant.

[57] La base de données a été structurée de manière homogène pour des raisons pratiques liées à son exploitation. L'organisation interne des différents témoins du projet *ConDÉ*, basée sur des normes culturelles, étant donc particulière à chacun : les structures contiennent un à trois niveaux de division interne que nous avons alignés sur le niveau le plus fin, ainsi qu'entre deux et huit types de contenu (titre, citation, commentaire, notes). En XML, cela s'est traduit par trois niveaux de imbriquées avec @type *part* > *chapter* > *section*. Afin d'homogénéiser au mieux les structures, des divisions du niveau supérieur (*part*) ont été ajoutées dans certains témoins modernes qui n'en possédaient pas explicitement, mais groupaient les chapitres selon des thèmes récurrents de témoin en témoin.

[58] La structure interne aux divisions inférieures (*section*) a, quant à elle, été encodée de manière standard, avec le paragraphe comme unité principale. Les citations de la coutume ou d'ordonnances ont été incluses comme citations (*quote*). L'élément <choice> nous a permis d'encoder, côte-à-côte, une abréviation, une graphie ou un caractère ancien, et leur résolution ou modernisation graphique. Les informations lexicales et morpho-syntaxiques ont été ajoutées comme attributs des éléments <w> encapsulant chaque token, grâce au logiciel *AnaLog* (Lay & Pincemin 2010) et au jeu d'étiquettes *PRESTO*. Cinq scripts en langage Python nous ont ensuite permis de résoudre une partie conséquente des ambiguïtés de lemmatisation.

[59] En plus d'accélérer considérablement certaines tâches, les outils numériques nous permettent donc d'affranchir notre encodage de la forme initiale et de nous concentrer sur l'information pure. Nous sommes ainsi d'autant plus libres pour choisir une nouvelle forme : étant basée sur l'encodage plutôt que d'en être le centre, elle est malléable et adaptable. Loin d'une transcription numérique ou d'une remise-en-forme de sources textuelles, l'établissement de la base de données *ConDÉ* est ainsi le fruit d'une réflexion sur la nature des témoins et l'accès à leur structure par leur format originel. En considérant que la forme constitue l'accès à la nature des éléments de textes, on se donne les moyens de la transcender pour pouvoir la restituer dans une autre forme.

Abréviations et références bibliographiques

AnaLog = Marie-Hélène Lay 2008. Logiciel *AnaLog*.

Basnage 1678 = Henri Basnage 1678. *La coutume reformée du païs et duché de Normandie, anciens ressorts et enclaves d'iceluy*. Rouen : C. et J. Lucas. https://numelyo.bm-lyon.fr/f_view/BML:BML_00GOO0100137001101309818/IMG00000005, https://books.google.fr/books?id=nyiX4EyFTJkC.

Benoist 1995 = Jocelyn Benoist 1995. Qu'est-ce qu'un livre ? Création, droit et histoire. Emmanuel Kant 1995. *Qu'est-ce qu'un livre ? Textes de Kant et de Fichte traduits et présentés par Jocelyn Benoist*. Paris : Presses universitaires de France, 11-118.

Bérault 1614 = Josias Bérault 1614. *La coustume reformée du pays et duché de Normandie, anciens ressorts et enclaves d'iceluy*. 2e édition. Rouen : Raphaël du Petit Val. https://gallica.bnf.fr/ark:/12148/bpt6k3043196v.

BFM = École normale supérieure de Lyon (éd.) 1989-2022. *Base de français médiéval*. http://txm.ish-lyon.cnrs.fr/bfm.

Cazals, à paraître = Géraldine Cazals, à paraître. Coutume et jurisprudence. Point historiographique. *Cahiers historiques des Annales de droit* 5.

ConDÉ = Pierre Larrivée, Mathieu Goux (éds.) 2022. Projet *Constitution d'un droit européen : six siècles de coutumiers normands*. https://www.unicaen.fr/coutumiers/conde/accueil.html, https://github.com/RIN-ConDE.

CSV = *Comma separated values*.

EPELE = Pierre Larrivée (éd.) 2018. Projet *Écriture des peu lettrés*. https://www.unicaen.fr/epele.

Frantext = Laboratoire Analyse et traitement informatique de la langue française (ATILF) (éd.) 1998-2022. *Base textuelle Frantext*. http://www.frantext.fr.

Galleron & Idmhand 2020 = Ioana Galleron, Fatiha Idmhand 2020. De l'interopérabilité à la réutilisabilité des éditions électroniques. *Humanités numériques* 1. https://journals.openedition.org/revuehn/350.

Gallica = Bibliothèque nationale de France (éd.) 2022. Bibliothèque numérique *Gallica*. https://gallica.bnf.fr.

GC = *Grand Coutumier de Normandie* (Harvard Law School Library, ms 91, env. 1300). https://iiif.lib.harvard.edu/manifests/view/drs:11589675$1i.

Guillot-Barbance, Heiden & Lavrentiev 2017 = Céline Guillot-Barbance, Serge Heiden, Alexei Lavrentiev 2017. Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques* 7, 168-184.

HTR = *Handwritten text recognition*.

HTR-United = Alix Chagué, Thibault Clérice 2022. *HTR-United*. https://htr-united.github.io.

Instrucions et ensaignemens = *Instrucions et ensaignemens* (1386-1390). Georges Besnier, Robert Génestal (éds.) 1912. *Instrucions et ensaignemens : style de procéder d'une justice seigneuriale normande (1386-90)*. Caen : Jouan.

Junicode = Peter S. Baker 2006. Caractères *Junicode*. https://junicode.sourceforge.io.

Kraken = École pratique des hautes études 2015. Logiciel *Kraken*. http://kraken.re.

Lay & Pincemin 2010 = Marie-Hélène Lay, Bénédicte Pincemin 2010. Pour une exploration humaniste des testes : AnaLog. Sergio Bolasco, Isabella Chiari, Luca Giuliano (éds.). *Jadt 2010. Statistical analysis of textual data. Proceedings of the 10th international conference. 9-11 June 2010. Sapienza University of Rome*. http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1045-1056_106-Lay.pdf.

Le Rouillé 1539 = Guillaume Le Rouillé 1539. *Le grand coustumier du pays et duché de Normendie*. Rouen : Le Roux. https://gallica.bnf.fr/ark:/12148/bpt6k117438n.

MaX = Pôle du document numérique 2022. Logiciel *MaX*. http://www.unicaen.fr/recherche/mrsh/document_numerique/outils/max.

MCVF = France Martineau (éd.) 2010. *Modéliser le changement : les voies du français*. http://www.voies.uottawa.ca/index.html.
Mellot 2005 = Jean-Dominique Mellot (éd.) 2005. Production et usages de l'écrit juridique en France du Moyen Âge à nos jours. *Histoire et civilisation du livre* 1. https://revues.droz.org/index.php/HCL/issue/view/131.
Merville 1731 = Pierre Biarnoy de Merville 1731. *Décisions sur chaque article de la coutume de Normandie*. Paris : Valleyre. https://droit-normand.nakalona.fr/items/show/221.
Morisse *Notes* = Charles Morisse 15e siècle. *Notes sur le Grand Coutumier* (Bibliothèque municipale de Rouen, ms Y194a).
NCA = Achim Stein, Pierre Kunstmann, Martin-Dietrich Glessgen (éds) 2006. *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca. 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-Dietrich Glessgen*. https://sites.google.com/site/achimstein/research/resources/nca#nca.
Numelyo = Bibliothèque municipale de Lyon (éd.) 2022. Bibliothèque numérique *Numelyo*. https://numelyo.bm-lyon.fr.
OCR = *Optical character recognition*.
Pannier 1856 = Victor Pannier 1856. *Les ruines de la coutume de Normandie, ou Petit dictionnaire du droit normand restant en vigueur pour les droits acquis*. 2e édition. Rouen : Le Brument.
PDN = Pôle du document numérique. https://www.unicaen.fr/recherche/mrsh/document_numerique.
Pesnelle 1771 = Pesnelle 1771. *Coutume de Normandie expliquée par M. Pesnelle avec les observations de M. Roupnel de Chenilly*. 4e édition. Rouen : Lallemant. https://gallica.bnf.fr/ark:/12148/bpt6k9684477n, https://gallica.bnf.fr/ark:/12148/bpt6k9684468p.
POS = *Part of speech*.
PRESTO = Denis Vigier, Peter Blumenthal (éds.) 2013-2017. Projet *Évolution du système prépositionnel du français*. http://presto.ens-lyon.fr.
Python = Python Software Foundation 2001-2022. Langage de programmation Python. https://www.python.org.
RIN = Réseau d'intérêts normands.
SRCMF = Sophie Prévost, Achim Stein (éds.) 2013. *Syntactic reference corpus of medieval French*. http://srcmf.org.
TAC = *Très Ancien Coutumier de Normandie* (Bibliothèque Sainte-Geneviève, ms 1743, env. 1250). Retranscrit dans Ange-Ignace Marnier 1839. *Établissements et coutumes, assises et arrêts de l'échiquier de Normandie (de 1207 à 1245)*. Paris : De Stahl. https://droit-normand.nakalona.fr/items/show/319.
TEI = Text encoding initiative.
TEI 2022 = TEI 2022. *P5 : Recommandations pour l'encodage et l'échange de textes électroniques*. https://tei-c.org/release/doc/tei-p5-doc/fr/html/index.html.
Terrien 1578 = Guillaume Terrien 1578. *Commentaires du droict civil tant public que privé observé au pays et duché de Normandie*. Paris : Du Puys. https://gallica.bnf.fr/ark:/12148/bpt6k9107304x.
Transkribus = READ-COOP 2016-2022. Logiciel *Transkribus*. https://readcoop.eu/transkribus.
TXM = Serge Heiden et. al. 2007-2022. Logiciel *TXM*. Lyon : École normale supérieure. https://www.textometrie.org.
XML = *Extensible markup language*.
XPath = *XML XPath language*.
XSLT = *Extensible stylesheet language transformation*.