Vers une annotation automatique en TEI via l'analyse de mise en page

Ariane Pinche[1] Juliette Janes[1] Claire Jahan[1] Simon Gabay [2]

École nationale des Chartes|PSL, 65 Rue de Richelieu, 75002 Paris, France
{prenom.nom}@chartes.psl.eu

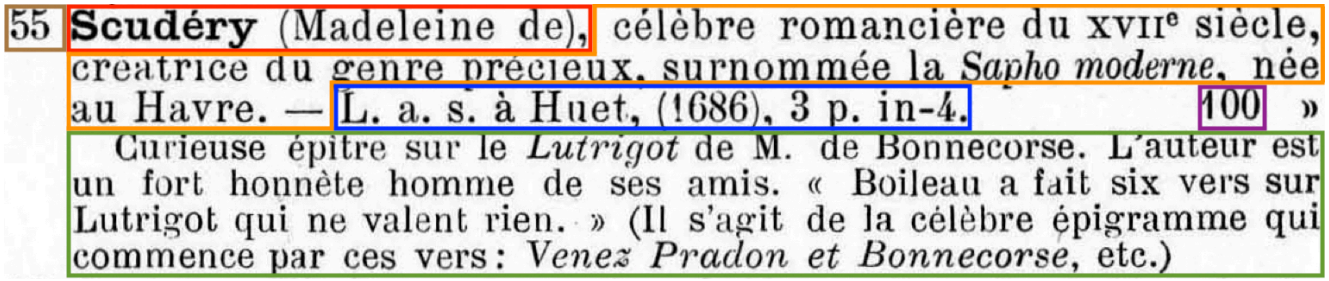
Université de Genève, Rue des Battoirs 7, 1205 Genève, Switzerland
{prenom.nom}@unige.ch

9 décembre 2021

Sommaire

Des documents complexes

Depuis des années, nous sommes capables d'extraire de l'information de documents très complexes comme les dictionnaires, les catalogues, les annuaires... (Gabay, Rondeau & Khemakhem 2020)



55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

Figure: *RDA*, n°67 (March 1881), lot N°55.

[S. Gabay, L. Rondeau Du Noyer, M. Khemakhem. "Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes", *IX Convegno AIUCD*, Jan. 2020, Milan.]

De l'extraction à l'encodage

L'extraction d'information n'est pas suffisante. Nous devons:

- Publier les données
- Partager les données
- Pérenniser les données

Pour tout cela, l'encodage en TEI est la meilleure solution.

```
<item n="55" xml:id="CAT_000037_e55">
 <num>55</num>
 <name type="author">Scudéry (Madeleine de),</name>
 <trait>
  <p>célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la Sapho
  moderne, née au Havre.</p>
 </trait>
 <desc>L.a.s. à Huet, (1686), 3 p. in-4. 100 »</desc>
 <note>Curieuse épître sur le Lutrigot de M. de Bonnecorse. L'auteur est un fort honnête
 homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien.» (Il
 s'agit de la célèbre épigramme qui commence par ces vers: Venez Pradon et Bonnecorse,
 etc.)</note>
</item>
```

Aller plus loin?

Figure: BNF, Fr. 412, f. 10^{r}.

S'il est possible d'extraire de l'information riche et complexe, il est important de pouvoir traiter des documents (en apparence) plus simple pour constituer des éditions numériques de qualité (*e.g.* Pinche 2021.)

→ **Notre objectif est d'encoder et structurer de l'information depuis des scans de ces documents**

[A. Pinche, *Édition nativement numérique du recueil hagiographique "Li Seint Confessor" de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France*, Université de Lyon, 2021]

Sommaire

Modéliser la mise en page

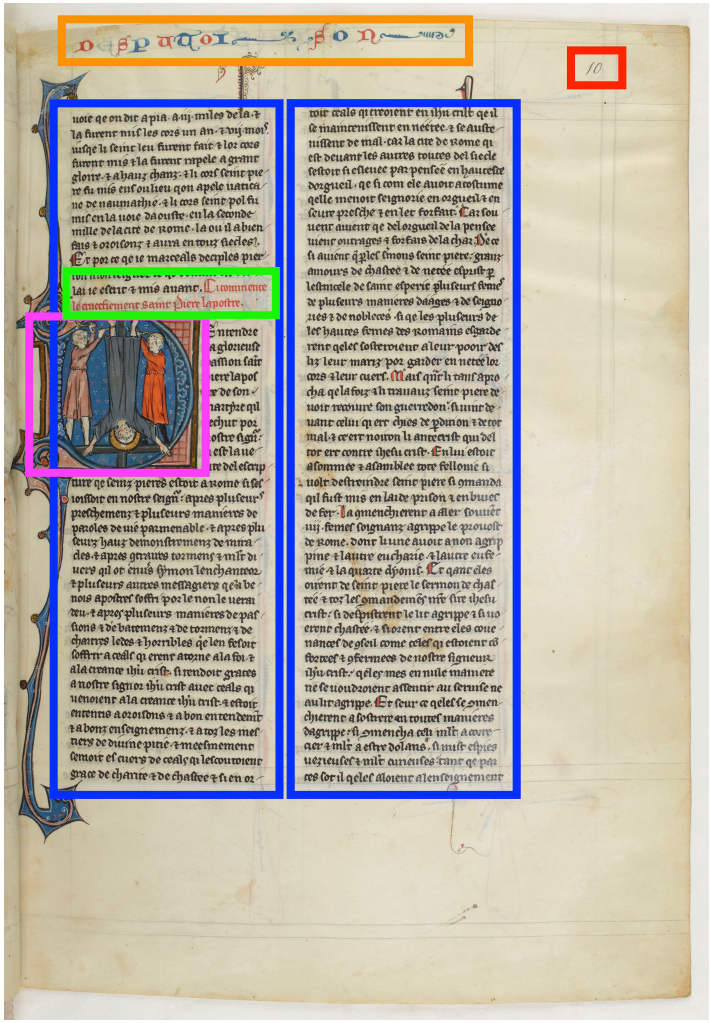


Figure: BNF, Fr. 412, f. 10^{r}.

- Analyse générale de la mise en page:
- Titre courant, **pagination**, corps du texte, **initiale**, **rubrique**.
- Une solution pour l'encodage de documents simples: la description standardisée de la page

Convergences: manuscrits et imprimés

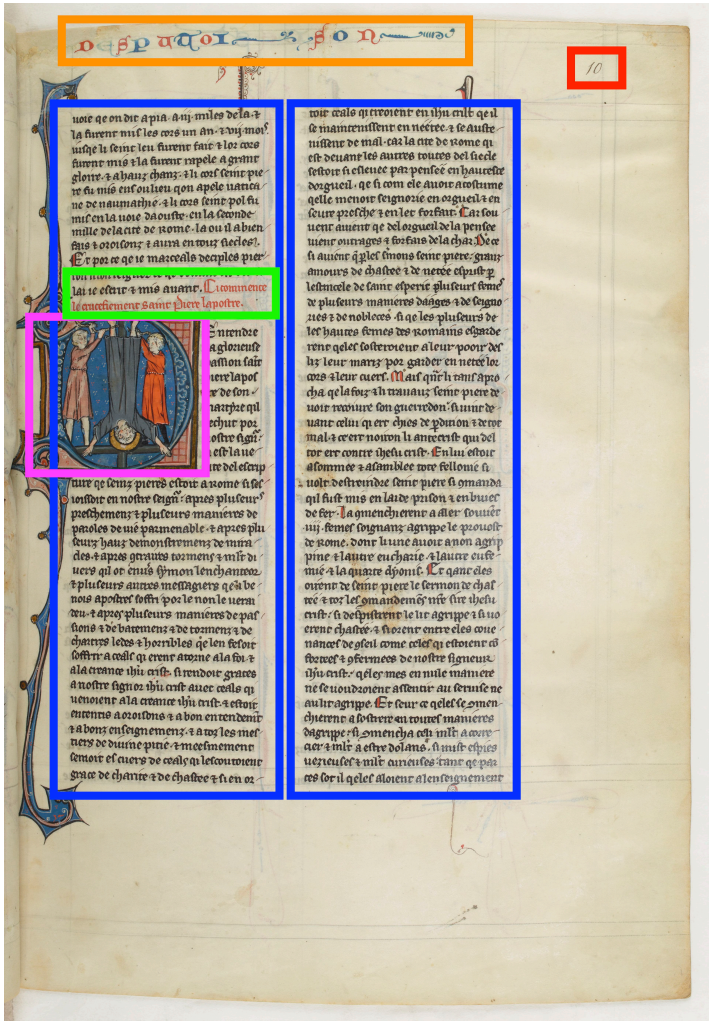


Figure: BNF, Fr. 412, f. 10^{r}.

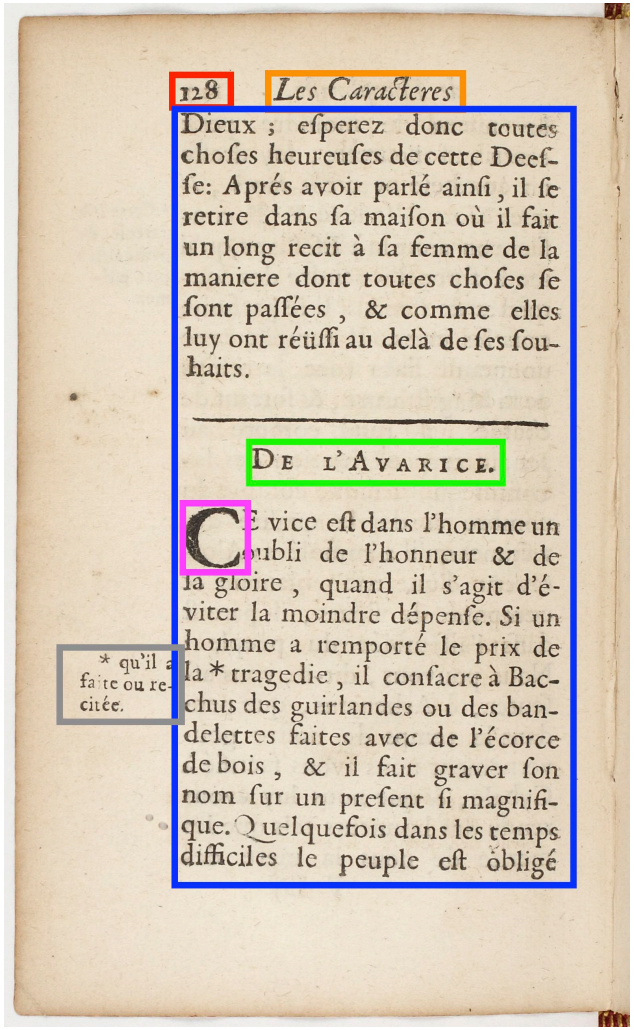


Figure: *Les Caractères*, 1688, p. 128.

Un vocabulaire contrôlé de description de la mise en page: SegmOnto

Développer un vocabulaire contrôlé: SegmOnto

- Non spécialisé
- Partage des données en amont: augmentation des jeux de données d'entraînement
- Partage des données en aval: mise en commun du post-traitement des données

https://github.com/SegmOnto/Guidelines

Listes des éléments SegmOnto

Zones	Lignes
CustomZone	CustomLine
DamageZone	DefaultLine
GraphicZone	DropCapitalLine
DigitizationArtefactZone	InterlinearLine
DropCapitalZone	MusicLine
GraphicZone	RubricLine
MainZone	
MarginTextZone	
MusicZone	
NumberingZone	
QuireMarksZone	
RunningTitleZone	
SealZone	
StampZone	
TableZone	
TitlePageZone	

Une chaîne de traitement

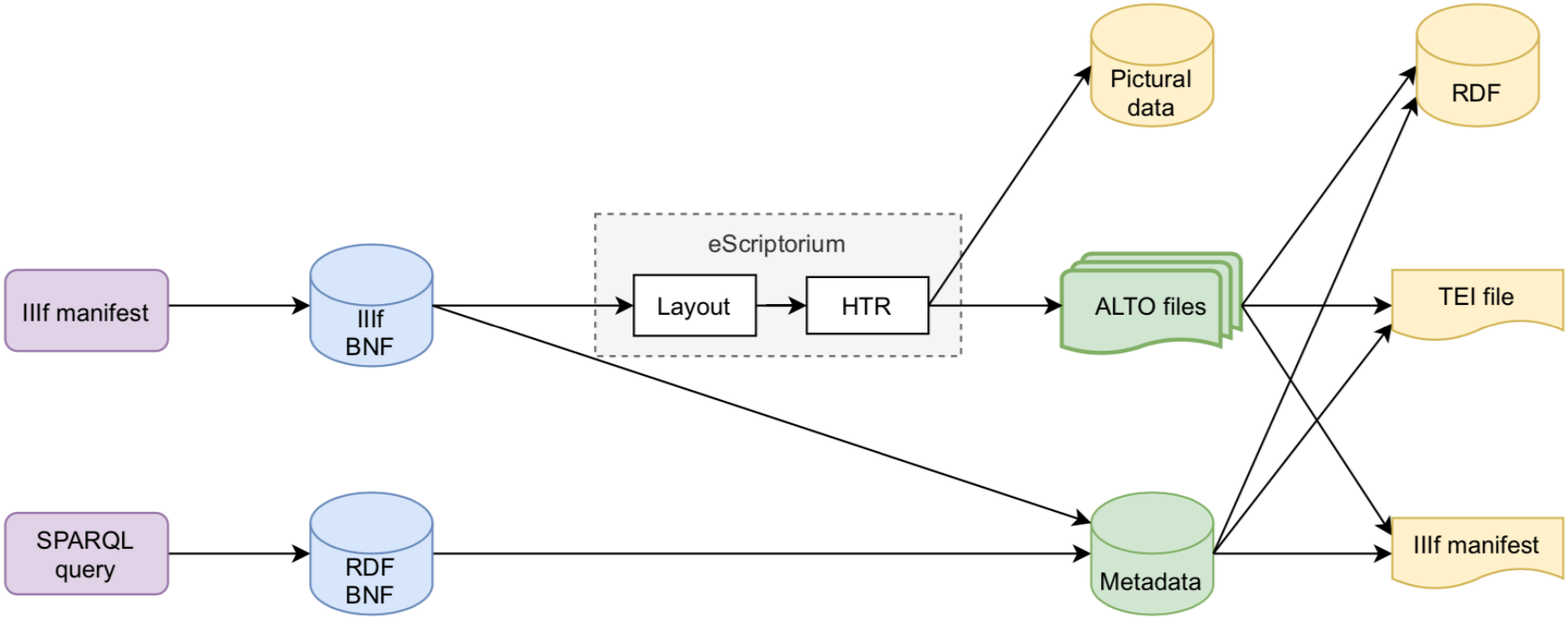


Figure: Chaîne de traitement Galli(corpor)a

Correspondance des éléments SegmOnto et XML-TEI

Zone SegmOnto	élément TEI
DamageZone	<damage>
GraphicZone	<figure>
DropCapitalZone	<hi>
MainZone	<p>
MarginTextZone	<note>
MusicZone	<figure>
NumberingZone	<fw type="Numbering">
RunningTitleZone	<fw type="RunningTitle">
SealZone	<figure>
QuireMarksZone	<fw type="Signatures">
StampZone	<figure>
TableZone	<table>
TitlePageZone	<p>

Proposition d'encodage TEI

avec <sourceDoc>

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<!-- Métadonnées -->
</teiHeader>
 <sourceDoc>
 <!-- Une page -->
  <surface>
   <!-- Une zone -->
   <zone type="ZoneSegmOnto" points="coordonnées">
    <!-- Une ligne -->
    <line type="LineSegmOnto" subtype="MySubtype"
    ↪    points="coordonnées">
     <!-- Une baseline -->
     <path points="coordonnées"/>
    </line>
    <!-- lignes suivantes -->
   </zone>
   <!-- zones suivantes -->
  </surface>
 </sourceDoc>
 <text>
   <body>
     <p> <!-- Texte structuré --> </p>
   </body>
 </text>
</TEI>
```

- Programme de conversion en python ALTO →XML-TEI
- Requête dans les données RDF de la BnF pour récupérer les métadonnées (absentes d'ALTO)
- Balise <sourceDoc>: conservation des données ALTO en TEI et stockage de la transcription automatique (pour rétroconversion vers ALTO)
- Texte transcrit: Structuration dans <body>

Sommaire

Préparer les données

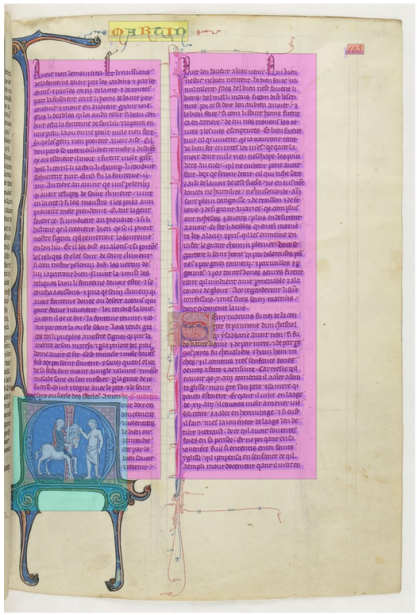
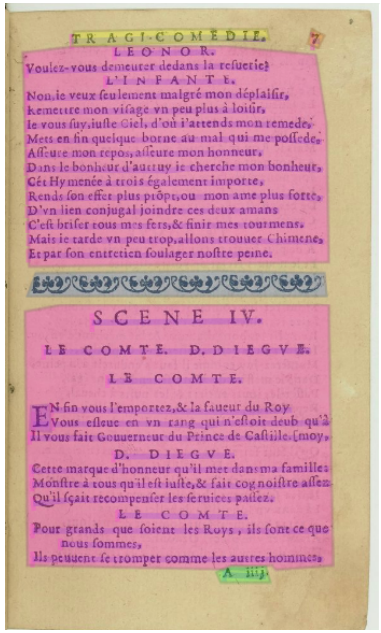


Figure: BNF, Fr. 412, f.1v



TRAGI-COMEDIE. 7

LEONOR.

Voulez-vous demeurer dedans la retraite?

L'INFANTE.

Non, ie veux seulement malgré mon déplaisir,
Remettre mon visage vn peu plus à loisir,
Ie vous fuy, iuste Ciel, d'où i'attends mon remede;
Mets en fin quelque borne au mal qui me possede.
Asseure mon repos, asseure mon honneur,
Dans le bonheur d'autruy ie cherche mon bonheur,
Cét Hymen à trois également importe,
Rends son effet plus prompt, ou mon ame plus forte.
D'vn lien conjugal joindre ces deux amans
C'est briser tous mes fers, & finir mes tourmens.
Mais ie tarde vn peu trop, allons trouuer Chimene,
Et par son entretien soulager nostre peine.

SCENE IV.

LE COMTE. D. DIEGVE.

LE COMTE.

ENfin vous l'emportez, & la faueur du Roy
Vous esleue en vn rang qui n'est soit deub qu'à
Il vous fait Gouuerneur du Prince de Castille, (moy,
D. DIEGVE.

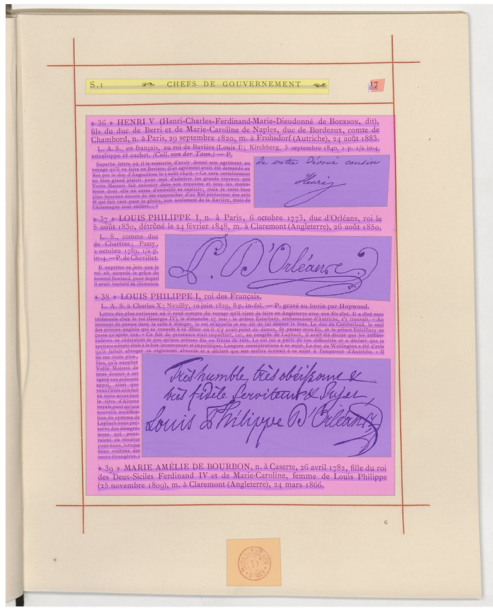
Cette marque d'honneur qu'il met dans ma famille
Monstre à tous qu'il est iuste, & fait cognoistre assez
Qu'il sçait recompenser les seruices passez.

LE COMTE.

Pour grands que soient les Roys, ils sont ce que
nous sommes,
Ils peuuent se tromper comme les autres hommes.

A iij

Figure: Le Cid, 1642, p.7



S.1 CHEFS DE GOUVERNEMENT 17

Figure: *Catalogue de manuscrits*, Bovet, p.81

Légende

MainZone, RunningTitleZone, GraphicZone:ornamentation, **GraphicZone:illustration**, **NumberingZone**, **QuireMarksZone**, StampZone, DropCapitalZone

Jeux de données et production de modèles

Datasets		Dates	Pages	Zones spécifiques
Romans/Pièces		XVII	569	Title, Running Title, Damage, Rubric
Manuscrits		XII-XIII	163	Running Title, DropCapital, 2 Main
Doc. structurés	Cat. expositions	XIX-XX	130	2 Main, Running Title, Title
	Cat. manuscrits	XIX-XX	102	Title, Stamp, Figure, Running Title
	Annuaires	XIX	50	2 Main

Table: Répartition des trois jeux

- Variation des paramètres
- Production de 10 modèles

Structure neuronale	Datasets
VGSL 1200	1 par jeu de données
VGSL 1800	1 XVII + Manuscrits
	1 XVII + Doc. Struct.

Table: Paramètres d'entraînement

Résultats

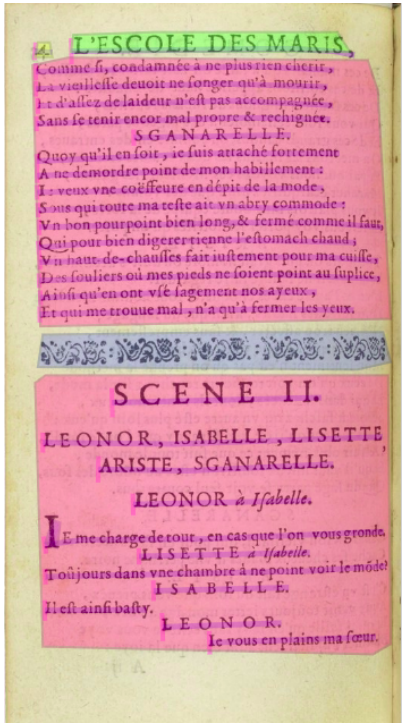

L'ESCOLE DES MARIS,
Comme ſi, condamnée à ne plus rien chérir,
La vieilleſſe deuoit ne ſonger qu'à mourir,
Et d'aſſez de laideur n'eſt pas accompagnée,
Sans ſe tenir encor mal propre & rechignée.
SGANARELLE.
Quoy qu'il en ſoit, ie ſuis attaché fortement
A ne demordre point de mon habillement:
Ie veux vne coëffeure en dépit de la mode,
Sous qui toute ma teſte ait vn abry commode;
Vn bon pourpoint bien long, & fermé comme il faut,
Qui pour bien digerer tienne l'eſtomach chaud;
Vn haut-de-chauſſes fait iuſtement pour ma cuiſſe,
Des ſouliers où mes pieds ne ſoient point au ſuplice,
Ainſi qu'en ont vſé ſagement nos ayeux,
Et qui me trouue mal, n'a qu'à fermer les yeux.
SCENE II.
LEONOR, ISABELLE, LISETTE, ARISTE, SGANARELLE.
LEONOR à Iſabelle.
Ie me charge de tout, en cas que l'on vous gronde.
LISETTE à Iſabelle.
Toûjours dans vne chambre à ne point voir le mõde?
ISABELLE.
Il eſt ainſi baſty.
LEONOR.
Ie vous en plains ma ſœur.

Figure: L'Escole des Maris, 1661, p.4

	Medieval data	17th c.	19th c.	Combination 1	Combination 2
MIoU	0.6017	0.1965	0.2092	0.4883	0.1964
FWIoU	0.7445	0.7794	0.7157	0.7537	0.7322
MAcc	0.9805	0.9846	0.9613	0.9886	0.9834
Acc	0.9805	0.9846	0.9613	0.9886	0.9834

Table: Mesures des Modèles Architecture 1

	Medieval data	17th c.	19th c.	Combination 1	Combination 2
MIoU	0.6127	0.1685	0.3662	0.5386	0.3269
FWIoU	0.7754	0.8002	0.7244	0.7845	0.7821
MAcc	0.9905	0.9886	0.9651	0.9917	0.9869
Acc	0.9905	0.9886	0.9651	0.9917	0.9869

Table: Mesures des Modèles Architecture 2

Légende

MainZone, **RunningTitleZone**,
GraphicZone:ornamentation, NumberingZone

(Mean Intersection-Over-Union (MIoU), frequency-weighted intersection over union(FWIoU), Mean accuracy (MAcc) and accuracy (Acc))

Sommaire

Conclusions

Cette chaîne de traitement devrait permettre de répondre aux besoins des chercheurs. Des améliorations sont cependant à prévoir:

- Le maillon faible reste la quantité de données, qu'il va falloir augmenter pour rendre le système efficace
- Prolonger l'annotation (en stand off?): lemmatisation, normalisation linguistique, entités nommées...

La prochaine étape sera la constitution automatisée d'un corpus de grande taille, en cours de préparation *via* un projet DataLab!

Remerciements

Merci à:

- Jean-Baptiste Camps
- Alix Chagué
- Hugo Scheithauer
- Laurent Romary