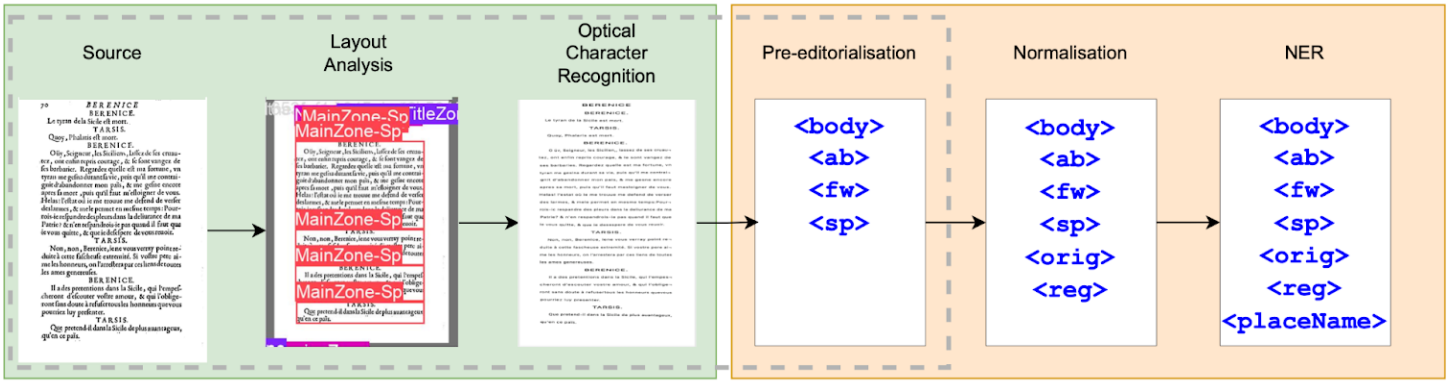
Enchaîner les IA

Vom Bild zum kodierten Dokument

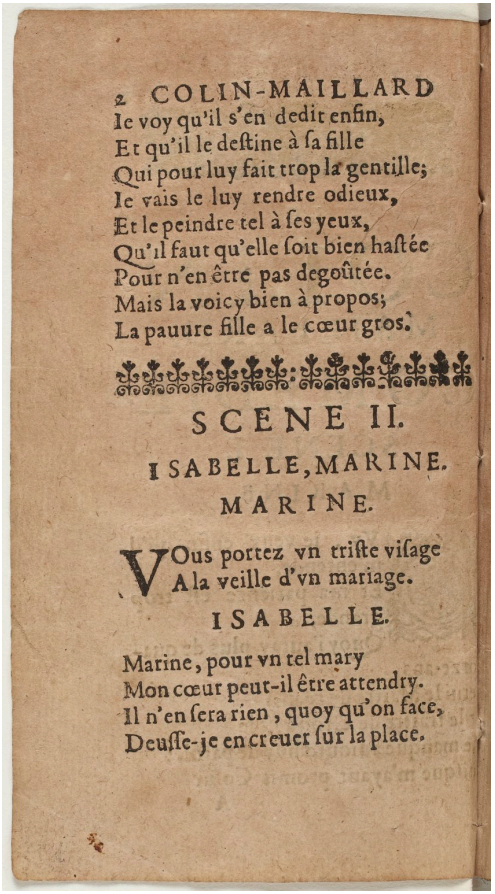
Simon Gabay [1]

[1]Université de Genève
prenom.nom@unige.ch

- Texterkennung (optical character recognition, OCR) ist nicht die einzige wichtige Technologie für Bibliotheken.
- Viele Aufgaben (*tasks*) sind auch wichtig : Layoutanalyse, Textnormalisierung, Eigennamenerkennung, etc.
- Diese Aufgaben sind nicht trivial, sondern erfordern erhebliche philologische Arbeit.



Texterkennung



2 COLIN-MAILLARD
Ie voy qu'il s'en dedit enfin,
Et qu'il le deſtine à ſa fille
Qui pour luy fait trop la gentille;
Ie vais le luy rendre odieux,
Et le peindre tel à ſes yeux,
Qu'il faut qu'elle ſoit bien haſtée
Pour n'en être pas degoûtée.
Mais la voicy bien à propos;
La pauure fille a le cœur gros.

SCENE II.

ISABELLE, MARINE.

MARINE.

VOus portez vn triſte viſage
A la veille d'vn mariage.

ISABELLE.

Marine, pour vn tel mary
Mon cœur peut-il être attendry.
Il n'en ſera rien, quoy qu'on face,
Deuſſe-je en creuer ſur la place.

Figure – Samuel Chappuzeau, *Le colin-maillard*, Paris : Jean-Baptiste Loyson, 1662
Source : Gallica

2 COLIN-MAILLARD
Ie voy qu'il s'en dedit enfin,
Et qu'il le destine à sa fille
Qui pour luy fait trop la gentille ;
Ie vais le luy rendre odieux,
Et le peindre tel à ses yeux,
Qu'il faut qu'elle soit bien hastée
Pour n'en être pas degoûtée.
Mais la voicy bien à propos ;
La pauure fille a le cœur gros.
SCENE II
ISABELLE, MARINE
MARINE
VOus portez vn triste visage
A la veille d'vn mariage.
ISABELLE.
Marine, pour vn tel mary
Mon cœur peut-il être attendry.
Il n'en sera rien, quoy qu'on face,
Deusse-je encreuer sur la place.

Texterkennung

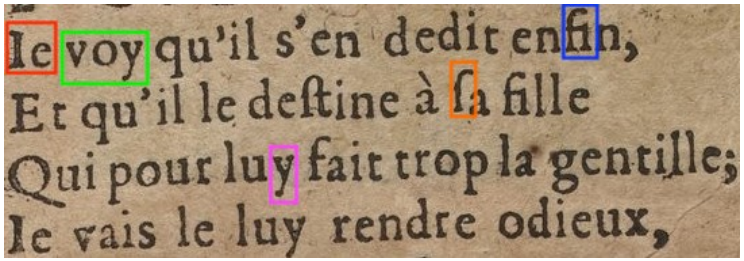
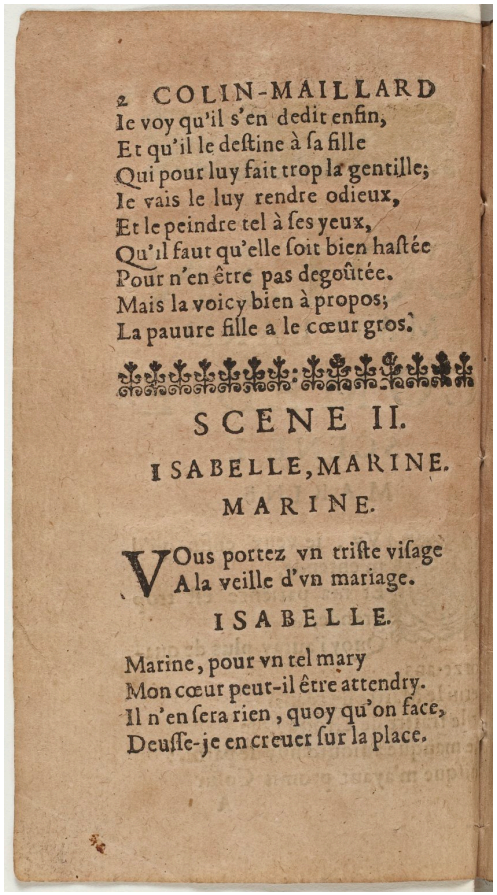


Figure – Samuel Chappuzeau, *Le colin-maillard*, Paris : Jean-Baptiste Loyson, 1662
Source : Gallica

- **Entwicklungen des Alphabets** (Unterscheidung *i*/*j* und *u*/*v*, sogenannte „Ramistische Buchstaben“)
- **Alte Konjugationen** (Erste Person ohne *s* vor der analogen Reparatur)
- **Alte Buchstaben** (Formvariante für den Buchstaben *s*)
- **Antike Verwendungen** (*y* kalligraphisch)

Layoutanalyse



2 COLIN-MAILLARD

Ie voy qu'il s'en dedit enfin,
Et qu'il le deſtine à ſa fille
Qui pour luy fait trop la gentille;
Ie vais le luy rendre odieux,
Et le peindre tel à ſes yeux,
Qu'il faut qu'elle ſoit bien haſtée
Pour n'en être pas degoûtée.
Mais la voicy bien à propos;
La pauure fille a le cœur gros.

SCENE II.

ISABELLE, MARINE.

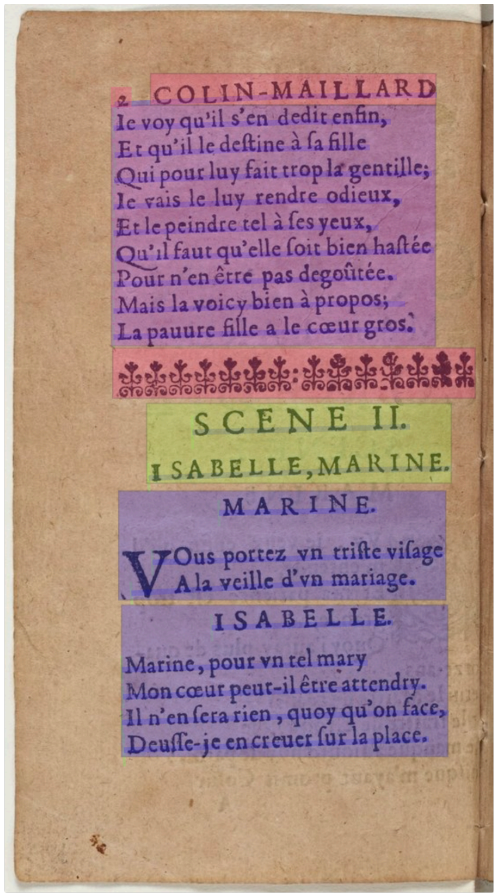
MARINE.

VOus portez vn triſte viſage
A la veille d'vn mariage.

ISABELLE.

Marine, pour vn tel mary
Mon cœur peut-il être attendry.
Il n'en ſera rien, quoy qu'on face,
Deuſſe-je en creuer ſur la place.

Figure – Samuel Chappuzeau, *Le colin-maillard*, Paris : Jean-Baptiste Loyson, 1662
Source : Gallica



2 COLIN-MAILLARD

Ie voy qu'il s'en dedit enfin,
Et qu'il le deſtine à ſa fille
Qui pour luy fait trop la gentille;
Ie vais le luy rendre odieux,
Et le peindre tel à ſes yeux,
Qu'il faut qu'elle ſoit bien haſtée
Pour n'en être pas degoûtée.
Mais la voicy bien à propos;
La pauure fille a le cœur gros.

SCENE II.

ISABELLE, MARINE.

MARINE.

VOus portez vn triſte viſage
A la veille d'vn mariage.

ISABELLE.

Marine, pour vn tel mary
Mon cœur peut-il être attendry.
Il n'en ſera rien, quoy qu'on face,
Deuſſe-je en creuer ſur la place.

Figure – Automatische Erkennung von Zonen mit SegmOnto/LaDaS-gesteuertem Vokabular.

Layoutanalyse

```
    <fw>2</fw>
    <fw>COLIN-MAILLARD</fw>
  <lb/>Ie voy qu'il s'en dedit enfin,
  <lb/>Et qu'il le destine à sa fille
  <lb/>Qui pour luy fait trop la gentille;
  <lb/>Ie vais le luy rendre odieux,
  <lb/>Et le peindre tel à ses yeux,
  <lb/>Qu'il faut qu'elle soit bien hastée
  <lb/>Pour n'en être pas degoûtée.
  <lb/>Mais la voicy bien à propos;
  <lb/>La pauure fille a le cœur gros.
</sp>
<head>SCENE II.
  <lb/>ISABELLE, MARINE</head>
<sp>MARINE
  <lb/>VOus porterez vn triste visage
  <lb/>A la veille d'vn mariage
</sp>
<sp>ISABELLE
  <lb/>Marine, pour vn tel mary
  <lb/>Mon cœur peut-il être attendry.
  <lb/>Il n'en sera rien, quoy qu'on face,
  <lb/>Deusse-je encreuer sur la place.
</sp>
```

Figure – Beispiel für eine automatische TEI-Kodierung (Proto-Edition).

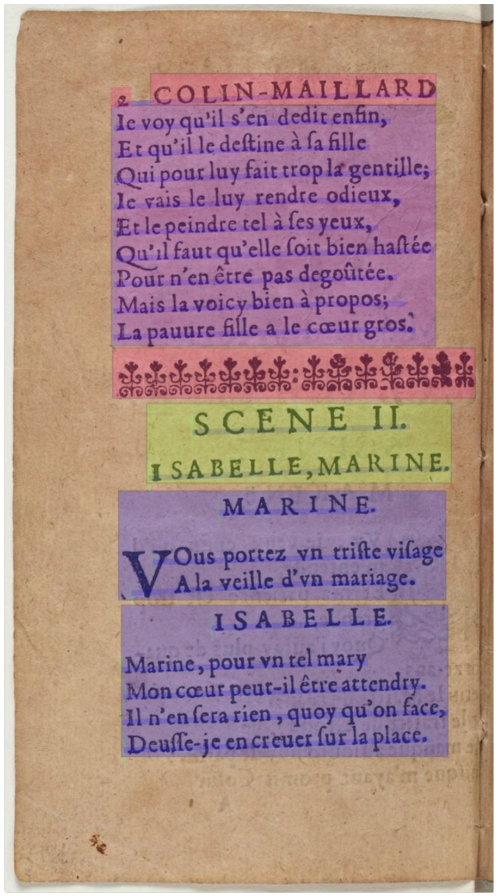


Figure – Automatische Erkennung von Zonen mit SegmOnto/LaDaS-gesteuertem Vokabular.

Textnormalisierung

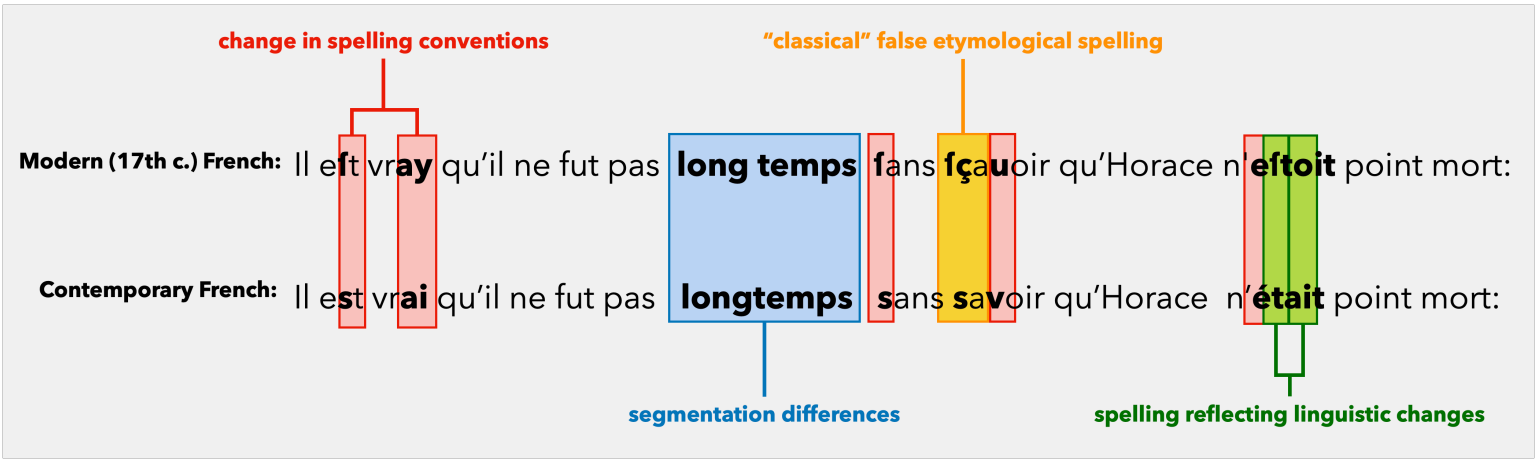


Figure – Beispiel für linguistische Normalisierung

- Für die automatische Transkription spricht man von eine
  - allographetische Transkription
  - graphematische Transkription
- Für der Transkription für eine Ausgabe spricht man von eine
  - (semi-)diplomatische Transkription
  - interpretierende Transkription

Textnormalisierung

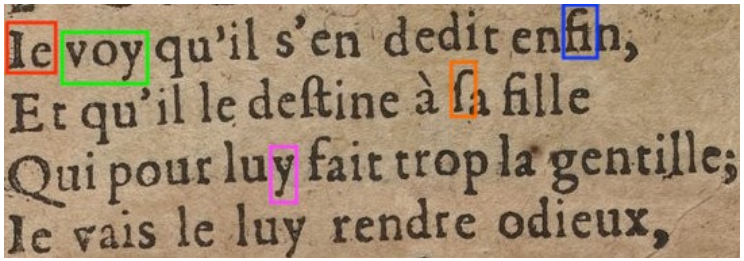


Figure – Samuel Chappuzeau, *Le colin-maillard*, Paris : Jean-Baptiste Loyson, 1662
Source : Gallica

Ie voy qu'il s'en dedit enfin,
Et qu'il le deſtine à ſa fille
Qui pour luy fait trop la gentille ;
Ie vais le luy rendre odieux,

Ie voy qu'il s'en dedit enfin,
Et qu'il le destine à sa fille
Qui pour luy fait trop la gentille ;
Ie vais le luy rendre odieux,

Je voy qu'il s'en dedit enfin,
Et qu'il le destine à sa fille
Qui pour lui fait trop la gentille ;
Je vais le lui rendre odieux,

Je vois qu'il s'en dédit enfin,
Et qu'il le destine à sa fille
Qui pour lui fait trop la gentille ;
Je vais le lui rendre odieux,

XML-TEI

Es muss nicht nur das Ergebnis der Normalisierung erhalten bleiben, sondern *alle* Zustände des Textes.

```
<sp>
   <ab>
      <seg>
         <orig>SGANARELLE.</orig>
         <reg>SGANARELLE.</reg>
      </seg>
      <seg>
         <orig>Promettez-moy donc, Seigneur Geronimo, de me parler avec toute
         ſorte de franchiſe.</orig>
         <reg>Promettez-moi donc, Seigneur Geronimo, de me parler avec toute
         sorte de franchise.</reg>
      </seg>
   </ab>
</sp>
<sp>
   <ab>
      <seg>
         <orig>GERONIMO.</orig>
         <reg>GERONIMO.</reg>
      </seg>
      <seg>
         <orig>Ie vous le promets.</orig>
         <reg>Je vous le promets.</reg>
      </seg>
   </ab>
</sp>
```

Figure – Beispiel einer XML-TEI-Kodierung mit Normalisierung.

Eigennamenerkennung

Token	Lemma	POS	COARSE	FINE	FINE-COMP	NESTED	Wikidata ID
Les	le	Da	O	O	O	O	_
allemands	allemand	Nc	O	O	O	O	_
élurent	élire	Vvc	O	O	O	O	_
pour	pour	S	O	O	O	O	_
empereur	empereur	Nc	B-pers	B-pers.ind	B-comp.title	O	Q438435
Rodolphe	Rodolphe	Np	I-pers	I-pers.ind	B-comp.name	O	Q438435
duc	duc	Nc	I-pers	I-pers.ind	B-comp.title	O	Q438435
de	de	S	I-pers	I-pers.ind	I-comp.title	O	Q438435
Suabe	Souabe	Np	I-pers	I-pers.ind	I-comp.title	B-loc.adm.reg	Q438435

Table – Erkennung und Begriffsklärung benannter Entitäten

[Pedro Ortiz Suarez, Simon Gabay. A Data-driven Approach to Named Entity Recognition for Early Modern French. *Computational Linguistics*, Oct 2022, Gyeongju, South Korea. pp.3722-3730. hal-04110765]

Pipeline

Hier stellen wir nur einige Beispiele vor :

1. Automatische Erstellung von Metadaten
2. Textanreicherung
3. etc.