Les données des Humanités numériques : des corpus d'images et du texte

Gaël Lejeune (gael.lejeune@sorbonne-universite.fr)

16 octobre 2025

UFR de Sociologie et d'Informatique pour les Sciences Humaines
STIH, Sorbonne Université
CERES (Expérimentations en Méthodes numériques pour les SHS)

Là où nous nous sommes arrêtés

- Des données
- Représentatives
- Dans un format approprié
- En un mot : un **corpus** !

Qu'est-ce qu'un corpus ?

Définir le concept de corpus

- **Corpus** : ensemble structuré et délimité de documents choisis pour une analyse scientifique.
- Différentes destinnations :
  - Corpus linguistique
  - Corpus documentaire
  - Corpus numérique
- Objectif : construire un **objet empirique** adaptées à des recherches reproductibles

Les critères du corpus selon Bénédicte Pincemin

Référence
B. Pincemin, *≪ Qu'est-ce qu'un corpus ? ≫*, in *Corpus* [En ligne], 1 — 2002.

- **Pertinence** : adéquation entre le corpus et la problématique de recherche.
- **Cohérence** : homogénéité thématique, temporelle, générique ou linguistique.
- **Représentativité** : capacité du corpus à refléter un phénomène linguistique ou culturel plus large.
- **Complétude/clôture** : le corpus doit être défini par un ensemble fini et stable de textes.
- **Homogénéité** : le corpus n'est pas un sac de textes
- **Volume** : la taille doit être significative

Les critères du corpus selon Bénédicte Pincemin

Référence
B. Pincemin, *≪ Qu'est-ce qu'un corpus ? ≫*, in *Corpus* [En ligne], 1 — 2002.

- **Pertinence** : adéquation entre le corpus et la problématique de recherche.
- **Cohérence** : homogénéité thématique, temporelle, générique ou linguistique.
- **Représentativité** : capacité du corpus à refléter un phénomène linguistique ou culturel plus large.
- **Complétude/clôture** : le corpus doit être défini par un ensemble fini et stable de textes.
- **Homogénéité** : le corpus n'est pas un sac de textes
- **Volume** : la taille doit être significative

1. Pertinence

Définition (Pincemin, 2002)
La pertinence renvoie à l'adéquation du corpus à la problématique de recherche.

- Le corpus doit être **délibérément construit** pour répondre à une question scientifique précise (objectif ou tâche).
- Il ne s'agit pas d'un simple ensemble de textes « disponibles », mais d'un ensemble **signifiant** pour l'analyse envisagée.

1. Pertinence

Définition (Pincemin, 2002)
La pertinence renvoie à l'adéquation du corpus à la problématique de recherche.

- Le corpus doit être **délibérément construit** pour répondre à une question scientifique précise (objectif ou tâche).
- Il ne s'agit pas d'un simple ensemble de textes « disponibles », mais d'un ensemble **signifiant** pour l'analyse envisagée.

Exemple
Constituer un corpus de discours politiques pour étudier la rhétorique persuasive, pas pour l'analyse lexicale générale.

2. Cohérence

Définition
Un corpus doit présenter une unité interne suffisante pour être interprété comme un tout.

- Cohérence thématique, générique, linguistique ou temporelle.
- Les textes doivent partager un cadre ou un type de production commun.
- Une trop grande hétérogénéité empêche l'analyse comparative.

2. Cohérence

Définition
Un corpus doit présenter une unité interne suffisante pour être interprété comme un tout.

- Cohérence thématique, générique, linguistique ou temporelle.
- Les textes doivent partager un cadre ou un type de production commun.
- Une trop grande hétérogénéité empêche l'analyse comparative.

Exemple
Corpus de lettres de poilus (1914–1918) : cohérence temporelle et typologique.

3. Représentativité

Définition
Le corpus doit permettre de décrire un phénomène linguistique ou discursif au-delà de ses cas particuliers.

- Capacité du corpus à **représenter une variété ou un usage** de manière équilibrée.
- Peut impliquer un échantillonnage raisonné (par auteur, époque, genre, etc.).
- Éviter les biais : surreprésentation d'un auteur, d'une époque, d'un support.

3. Représentativité

Définition
Le corpus doit permettre de décrire un phénomène linguistique ou discursif au-delà de ses cas particuliers.

- Capacité du corpus à **représenter une variété ou un usage** de manière équilibrée.
- Peut impliquer un échantillonnage raisonné (par auteur, époque, genre, etc.).
- Éviter les biais : surreprésentation d'un auteur, d'une époque, d'un support.

Exemple
Un corpus sur l'impact des JO à Paris avec des sources complémentaires et contradictoires.

4. Régularité

Définition
Le corpus doit présenter une certaine constance dans sa composition et son traitement.

- Régularité de sélection : même méthode pour tous les textes.
- Régularité de traitement : mêmes niveaux de nettoyage, d'encodage et de métadonnées.
- Assure la comparabilité entre sous-ensembles.

4. Régularité

Définition
Le corpus doit présenter une certaine constance dans sa composition et son traitement.

- Régularité de sélection : même méthode pour tous les textes.
- Régularité de traitement : mêmes niveaux de nettoyage, d'encodage et de métadonnées.
- Assure la comparabilité entre sous-ensembles.

Exemple
Tous les fichiers encodés en UTF-8, annotés avec le même schéma TEI.

5. Complétude

Définition
Le corpus doit être **suffisamment complet** pour permettre des analyses valides.

- Dans un corpus évolutif, la version analysée doit être figée et documentée.
- Complétude = fermeture temporaire du corpus à un moment de l'analyse.

5. Complétude

Définition
Le corpus doit être **suffisamment complet** pour permettre des analyses valides.

- Dans un corpus évolutif, la version analysée doit être figée et documentée.
- Complétude = fermeture temporaire du corpus à un moment de l'analyse.

Contre-Exemple
Un corpus de presse obtenu de manière incrémentale où l'on compare des analyses réalisées sur différentes périodes

6. Homogénéité

Définition
Le corpus doit être homogène dans ses caractéristiques formelles et linguistiques, sauf justification contraire.

- Homogénéité du support (texte imprimé, transcription, OCR, etc.).
- Homogénéité des niveaux d'annotation ou des métadonnées.
- Une hétérogénéité contrôlée est possible (ex. corpus contrastif bilingue).

6. Homogénéité

Définition
Le corpus doit être homogène dans ses caractéristiques formelles et linguistiques, sauf justification contraire.

- Homogénéité du support (texte imprimé, transcription, OCR, etc.).
- Homogénéité des niveaux d'annotation ou des métadonnées.
- Une hétérogénéité contrôlée est possible (ex. corpus contrastif bilingue).

Exemple
Un corpus où l'on contraste un auteur avec un autre auteur de la même période et opérant dans le même champ

7. Volume

Définition
Le corpus doit avoir un volume suffisant pour permettre une analyse significative.

- Dépend du type d'étude :
- Taille et granularité doivent être justifiées dans la méthodologie.
- Pincemin insiste : « la quantité ne remplace pas la pertinence ».

Contre-Exemple
Un corpus censé illustrer des fautes de langue mais limité à 5 posts récupérés sur des réseaux sociaux

Synthèse des critères de Pincemin

Critères de signifiance

- Pertinence
- Cohérence

Critères d'acceptabilité

- Représentativité
- Régularité
- Complétude

Critères d'exploitabilité

- Homogénéité
- Volume

Et en Humanités Numériques ?

Corpus et Humanités Numériques

- Passage du corpus matériel au corpus numérique.
- Enjeux :
  - Structuration et enrichissement par métadonnées.
  - Interopérabilité (formats ouverts : TEI/XML, JSON, CSV...).
- Exemples de corpus numériques :
  - *Frantext*, *Eltec*

Questions éthiques et méthodologiques

- Droit d'auteur et diffusion des données.
- Représentativité et biais de sélection.
- Importance de la documentation du corpus :
  - Fiches descriptives
  - Historique des versions
  - Provenance et reproductibilité

De la donnée brute au texte exploitable : l'OCR

Qu'est-ce que l'OCR ?

Définition
OCR = *Optical Character Recognition* : technologie permettant de convertir des images contenant du texte en texte exploitable par machine.

- Elle identifie les formes de caractères dans une image (scan, photo, PDF).
- Produit en sortie un fichier texte ou XML indexable et analysable.

Historique et usages

- Débuts dans les années 1950 (lecture automatique de caractères imprimés).
- Années 2000 : intégration dans les processus de numérisation patrimoniale.
- Aujourd'hui :
  - Traitement de grandes masses de documents.
  - Numérisation d'archives et bibliothèques.
  - Extraction de données pour la recherche (analyse lexicale, stylométrie, traduction).

Les étapes du processus OCR

1. **Numérisation** : création d'une image (scan, PDF).
2. **Préparation** : amélioration de l'image (binarisation, redressement, suppression du bruit).
3. **Segmentation** : découpage en lignes, mots, caractères.
4. **Reconnaissance** : identification des caractères à partir d'un modèle linguistique.
5. **Post-traitement** : correction automatique et indexation.

Typologie des technologies OCR

- **OCR classique** : reconnaissance de texte imprimé (Tesseract, ABBYY).
- **HTR** — *Handwritten Text Recognition* : pour manuscrits (Transkribus).
- **OLR** — *Optical Layout Recognition* : prise en compte de la structure du document.
- **OCR multilingue** : modèles spécifiques par langue (fra, eng, spa, deu...).

Est-ce qu'un ordinateur peut lire un texte ?

Pas vraiment.

Figure 1 – L'ordinateur préfère les QR code (Source : Wikipédia)

Mais les QR code c'est illisible non ?

Pas vraiment, il suffit de connaître ... le code

Mais les QR code c'est illisible non ?

Pas vraiment, il suffit de connaître ... le code

Comme si on lisait dans une autre langue

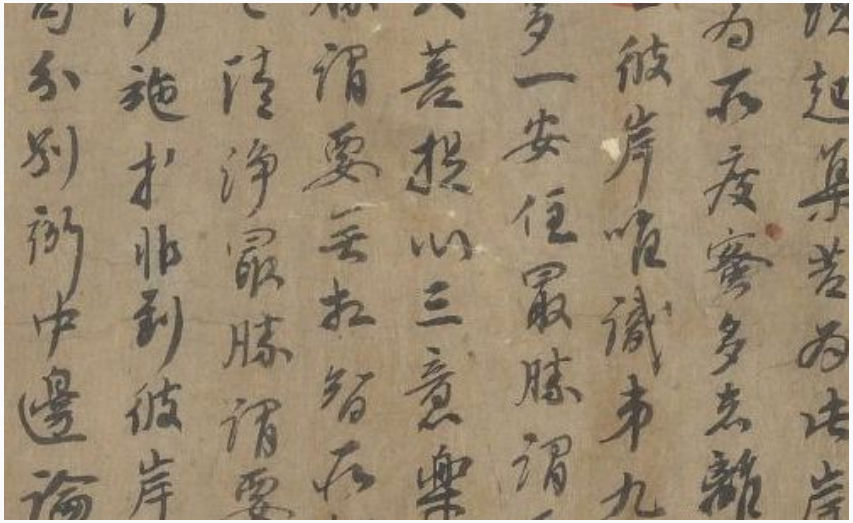
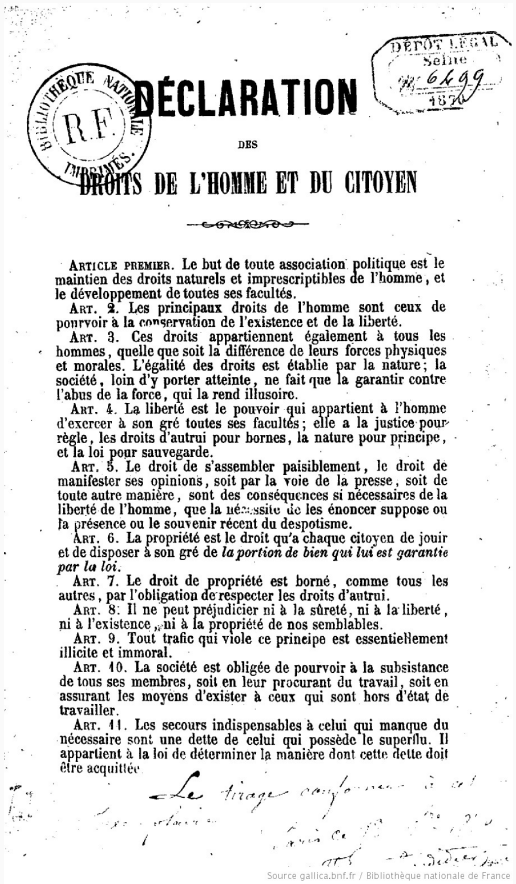


Figure 2 – Un texte en chinois (Source : Gallica)

Si j'arrive à lire, la machine le fait encore plus facilement ?

Pas vraiment.



DÉCLARATION

DES

DROITS DE L'HOMME ET DU CITOYEN

ARTICLE PREMIER. Le but de toute association politique est le maintien des droits naturels et imprescriptibles de l'homme, et le développement de toutes ses facultés.

ART. 2. Les principaux droits de l'homme sont ceux de pourvoir à la conservation de l'existence et de la liberté.

ART. 3. Ces droits appartiennent également à tous les hommes, quelle que soit la différence de leurs forces physiques et morales. L'égalité des droits est établie par la nature; la société, loin d'y porter atteinte, ne fait que la garantir contre l'abus de la force, qui la rend illusoire.

ART. 4. La liberté est le pouvoir qui appartient à l'homme d'exercer à son gré toutes ses facultés; elle a la justice pour règle, les droits d'autrui pour bornes, la nature pour principe, et la loi pour sauvegarde.

ART. 5. Le droit de s'assembler paisiblement, le droit de manifester ses opinions, soit par la voie de la presse, soit de toute autre manière, sont des conséquences si nécessaires de la liberté de l'homme, que la nécessité de les énoncer suppose ou la présence ou le souvenir récent du despotisme.

ART. 6. La propriété est le droit qu'a chaque citoyen de jouir et de disposer à son gré de *la portion de bien qui lui est garantie par la loi.*

ART. 7. Le droit de propriété est borné, comme tous les autres, par l'obligation de respecter les droits d'autrui.

ART. 8. Il ne peut préjudicier ni à la sûreté, ni à la liberté, ni à l'existence, ni à la propriété de nos semblables.

ART. 9. Tout trafic qui viole ce principe est essentiellement illicite et immoral.

ART. 10. La société est obligée de pourvoir à la subsistance de tous ses membres, soit en leur procurant du travail, soit en assurant les moyens d'exister à ceux qui sont hors d'état de travailler.

ART. 11. Les secours indispensables à celui qui manque du nécessaire sont une dette de celui qui possède le superflu. Il appartient à la loi de déterminer la manière dont cette dette doit être acquittée

Source gallica.bnf.fr / Bibliothèque nationale de France

Figure 3 – La déclaration des droits de l'Homme et du citoyen vue par l'humain (Source : Gallica)

ARIcs nsun. Le but de teute asseciation politique est le maintien des droits natarels et imprescriptibles de l'homme, et le developpement de toutes ses facult6s.

ia. i priiii ai d ros et eas 4 ponrvoir a la cnrseralian de l'existerce et de la lierte.

ARr. 3. Ces droits appertiennent egaslement teus les hommes, quelle qe soit la difference le leurs forces physiques et morales. Lgilit des droits est etblie par la iature; la societe, loin d'y porter atteinte, ne fait que la garantir contre 1'abus de la fore, qui la rend illusoiro.

AnT. i. L liberle est le pouoir qui appartient lhomme d'eercer son gre toutes ses facultes; elle a la justice pour regle, les droits d'autrui pour hornes, la nature peur prineipe, et la loi ponr sauvegarde.

. . ILs ani de assemnbler paisiblemnent, le droit de manifester ses opinions, sit par la oie de la presse, soit de oute autre maniere, sont des conseuences si necessaires de la liberte de l'homme, ue la ue:::ssile ie les enoncer suppose ou la presence ou le souenir recent du despotisme.

AR,. 6, La propriete est le droit qu'a chaque citoyen de jouir et de disposer a son gre de la portion de je qui lui est garantie par l loi.

ARr. T. ILe dreit de propriete est borne, comme tous les autres, par l'obligation derespecter les droits d'autrui,

ARI, 8: II ne peut prejudicier ni a la srele, ni a la liherte, ni b l'existece,,i a la propriete de nos semblables

Anr. 9. Tout trafic uii viole ce prineipe est essentiellement illicite et immoral.

Anr. 10. La societe est ohligee de pourvoir a la subsistanee de tous ses membres, soit en leur procurant du travail, soil en t les moyens d'exisuer a ceux qui sont hors d'etat de travailler.

Aar, 11. Les secours indispensables a celui qi manqe d necessaire sont une dette de celui qui possede le superilu. II appartient a la loi de determiner la maniere dant cette lette doil elre acquitlee

-eL-za4-

Figure 4 – La déclaration des droits de l'Homme et du citoyen vue par la machine

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie

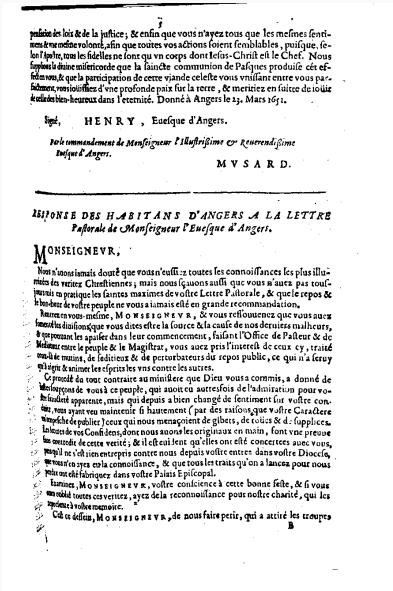
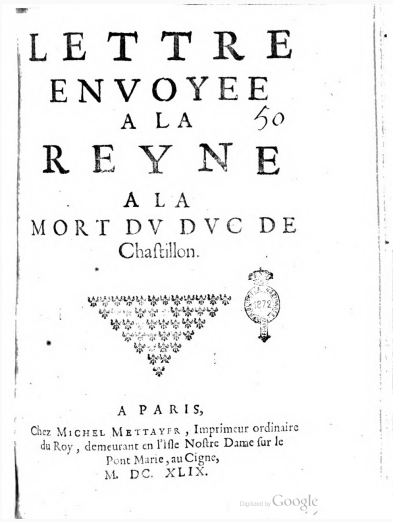


Figure 5 – Des lignes à redresser (source : ANTONOMAZ)

2. Segmentation : isoler les lignes et les caractères
3. Reconnaissance : à quoi le caractère segmenté ressemble
4. Post-traitement : correction d'éventuelles aberrations

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie


LETTRE
ENVOYEE
A LA
REYNE
A LA
MORT DV DVC DE
Chaſtillon.

A PARIS,
Chez MICHEL METTAYER, Imprimeur ordinaire du Roy, demeurant en l'iſle Noſtre Dame ſur le Pont Marie, au Cigne,
M. DC. XLIX.

Figure 5 – Augmenter le contraste (source : ANTONOMAZ)

2. Segmentation : isoler les lignes et les caractères
3. Reconnaissance : à quoi le caractère segmenté ressemble
4. Post-traitement : correction d'éventuelles aberrations

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie
2. Segmentation : isoler les lignes et les caractères

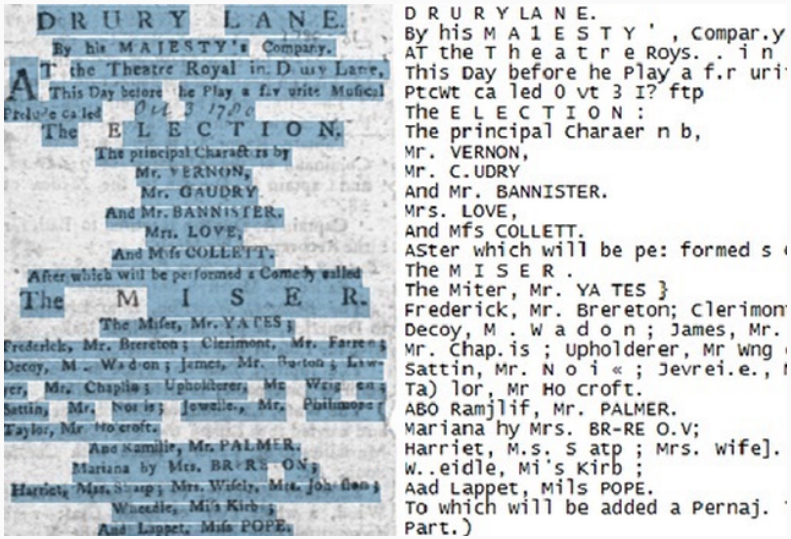


Figure 5 – Segmentation (source : EUROPEANA

3. Reconnaissance : à quoi le caractère segmenté ressemble
4. Post-traitement : correction d'éventuelles aberrations

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie
2. Segmentation : isoler les lignes et les caractères
3. Reconnaissance : à quoi le caractère segmenté ressemble

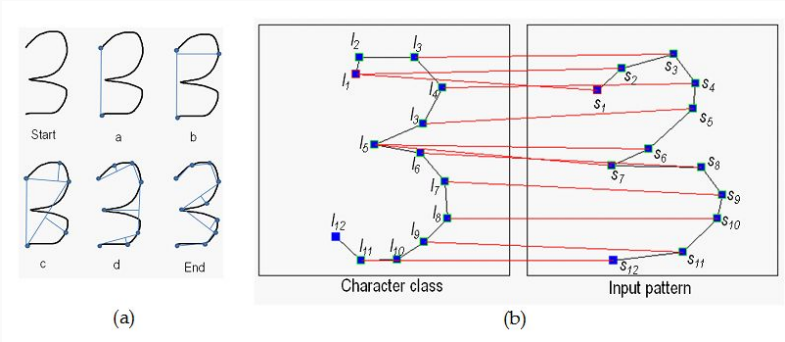


Figure 5 – Reconnaître un caractère (source : INTECHOPEN.COM)

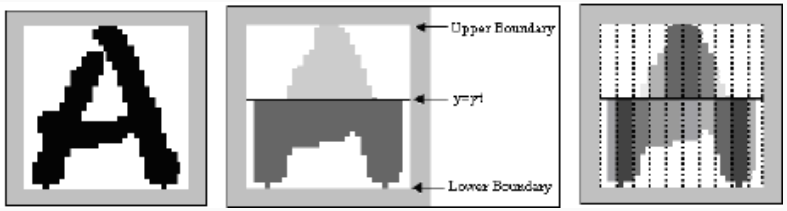


Figure 6 – Reconnaître un caractère II

4. Post-traitement : correction d'éventuelles aberrations

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie
2. Segmentation : isoler les lignes et les caractères
3. Reconnaissance : à quoi le caractère segmenté ressemble
4. Post-traitement : correction d'éventuelles aberrations

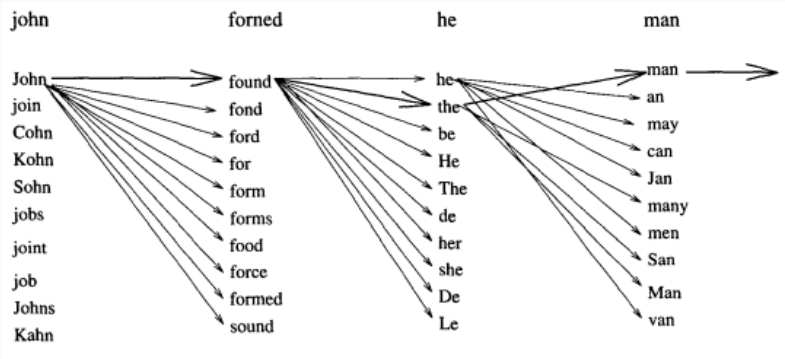


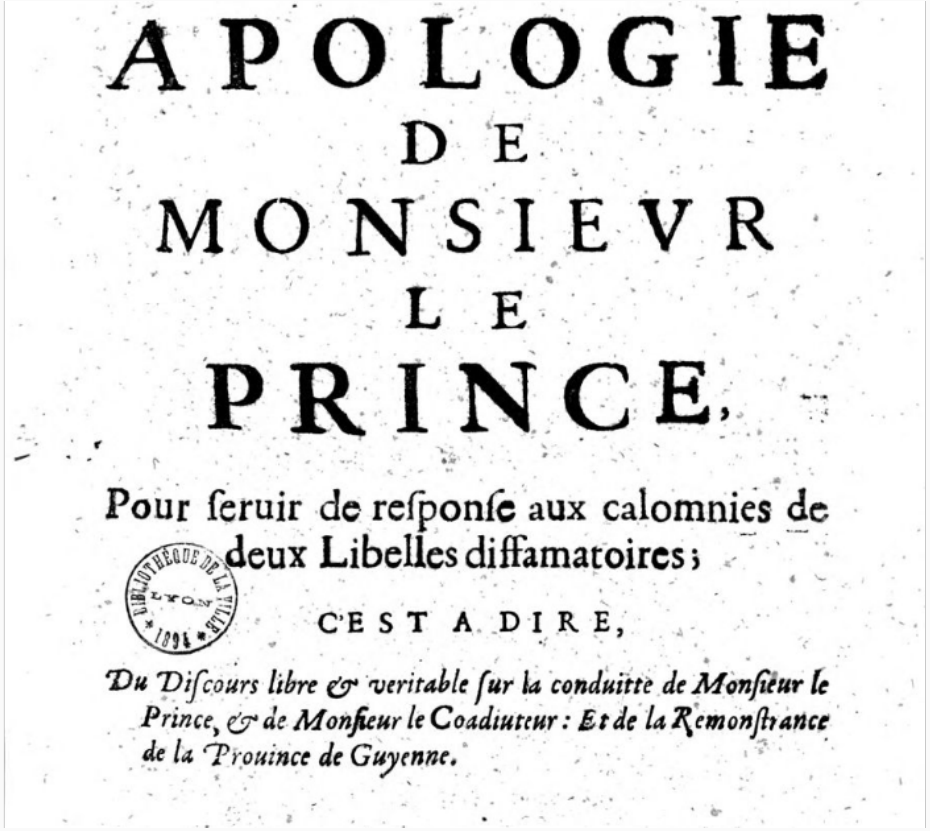
Figure 5 – Trouver la séquence de mots probable (Tong et Evans 1997)

Optical Character Recognition par étapes

1. Préparation : redresser l'image, simplifier la colorimétrie
2. Segmentation : isoler les lignes et les caractères
3. Reconnaissance : à quoi le caractère segmenté ressemble
4. Post-traitement : correction d'éventuelles aberrations

En sortie : une série de "chaînes de caractères" que l'on espère le plus proche possible du texte "réel"

Exemple de données Humanités Numériques



APOLOGIE
DE
MONSIEVR
LE
PRINCE,

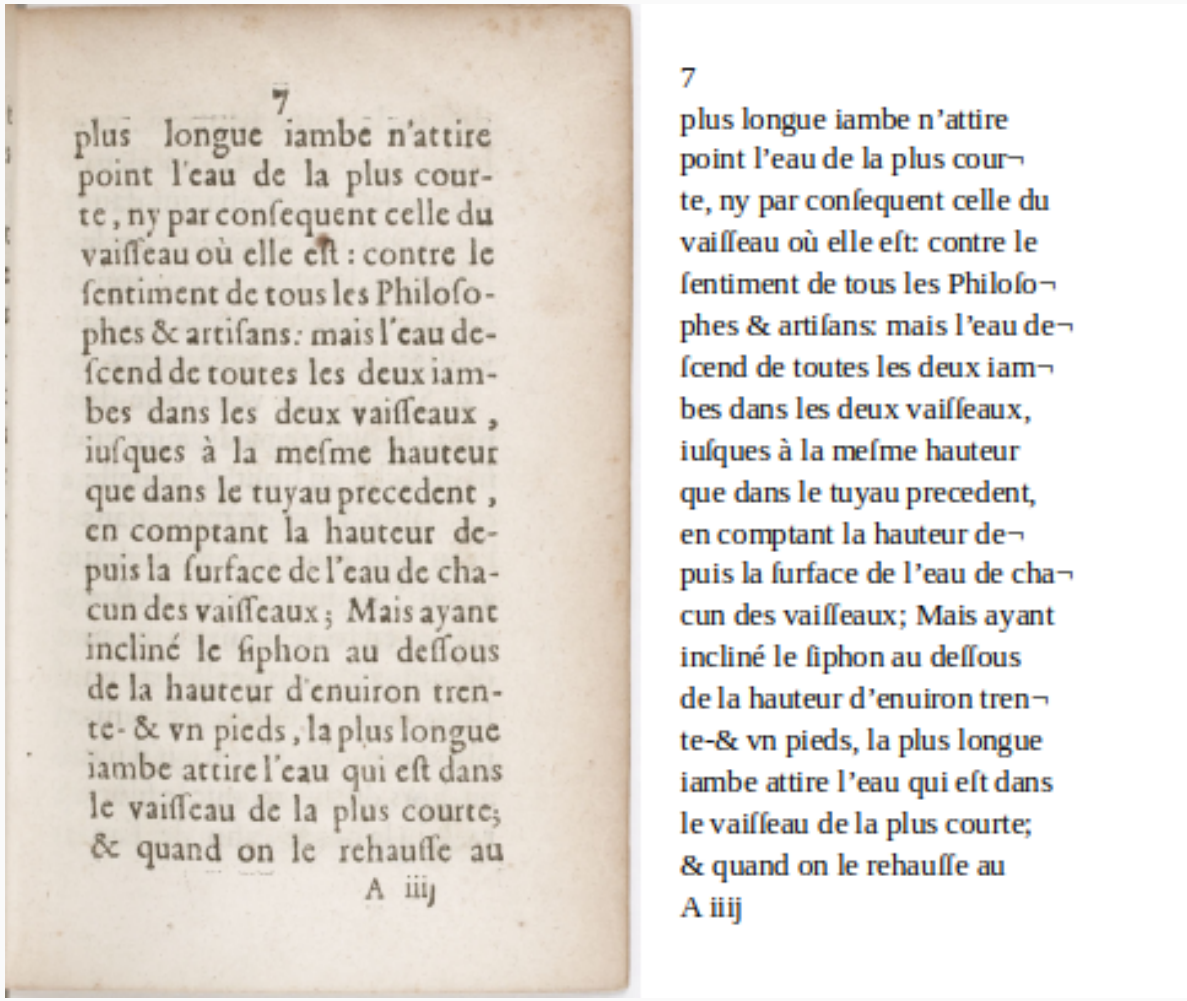
Pour ſeruir de reſponſe aux calomnies de deux Libelles diffamatoires;

C'EST A DIRE,

Du Diſcours libre & veritable ſur la conduitte de Monſieur le Prince, & de Monſieur le Coadiuteur: Et de la Remonſtrance de la Prouince de Guyenne.

Apologie de monsieur le Prince . . ., 1651 (Google Books)

Dans l'idéal on a (presque) le texte complet



7

plus longue iambe n'attire point l'eau de la plus courte, ny par conſequent celle du vaiſſeau où elle eſt : contre le ſentiment de tous les Philoſophes & artiſans: mais l'eau deſcend de toutes les deux iambes dans les deux vaiſſeaux, iuſques à la meſme hauteur que dans le tuyau precedent, en comptant la hauteur depuis la ſurface de l'eau de chacun des vaiſſeaux; Mais ayant incliné le ſiphon au deſſous de la hauteur d'enuiron trente- & vn pieds, la plus longue iambe attire l'eau qui eſt dans le vaiſſeau de la plus courte; & quand on le rehauſſe au

A iiij

7
plus longue iambe n'attire
point l'eau de la plus cour¬
te, ny par conſequent celle du
vaiſſeau où elle eſt: contre le
ſentiment de tous les Philoſo¬
phes & artiſans: mais l'eau de¬
ſcend de toutes les deux iam¬
bes dans les deux vaiſſeaux,
iuſques à la meſme hauteur
que dans le tuyau precedent,
en comptant la hauteur de¬
puis la ſurface de l'eau de cha¬
cun des vaiſſeaux; Mais ayant
incliné le ſiphon au deſſous
de la hauteur d'enuiron tren¬
te-& vn pieds, la plus longue
iambe attire l'eau qui eſt dans
le vaiſſeau de la plus courte;
& quand on le rehauſſe au
A iiij

Une page bien "formée", des caractères bien détectés (modèle OCR-17 [Gabay et al., 2023])

Dans les faits

1
l
[
1
3
e
aaaaa
A
pn eer
ees?
ee
e
aae?
ateu
ae
aeoe
ees¬
¬
dai
5
il
ee
aa
aeaaaaul
a
.
leu
:
eee
an)
IVLES a2a C dnsr eiller du y
l en les onseils, Abba des dbbayes de SsNab or enmorraine,
de SMda-d de Soissonis, et de SPiere de Corbie Cc.
DPaletar oneoret ecuuae-pmanl du Bex
, ¬
38

Une page difficile à traiter, du bruit

Avantages et limites de l'OCR

- **Avantages :**
  - Automatisation de la numérisation.
  - Extraction de données textuelles pour analyse.
  - Valorisation de corpus patrimoniaux.

Avantages et limites de l'OCR

- **Avantages :**
  - Automatisation de la numérisation.
  - Extraction de données textuelles pour analyse.
  - Valorisation de corpus patrimoniaux.
- **Limites :**
  - Dépendance à la qualité du scan et de la typographie.
  - Langues avec caractères spéciaux
  - Besoin de vérification et de correction manuelle.

Conclusion

Conclusion

- Du corpus conceptuel au corpus exploitable : un continuum.
- L'OCR comme étape de transformation et de standardisation.
- Importance :
  - de la reproductibilité
  - de la transparence
  - de la documentation
- Pour aller plus loin :
  - Post-correction automatique
  - Structuration en XML/TEI
  - Alignement multilingue

References i

Gabay, S., Clérice, T., and Reul, C. (2023).

OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more).

Journal of Data Mining and Digital Humanities, 2023.