

OCR, bruitage, hétérogénéité : exploiter des données textuelles hors "conditions de laboratoire"

Gaël Lejeune, gael.lejeune@sorbonne-universite.fr

9 octobre 2026

Sorbonne Université

Pourquoi les données textuelles ?

Pourquoi les données textuelles ?

Par ce que je fais du Traitement Automatique des Langues ?

Pourquoi les données textuelles ?

Par ce que je fais du Traitement Automatique des Langues ?

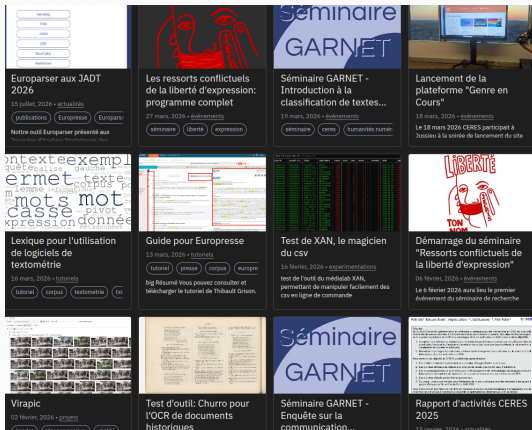


Figure 1 – Extrait des activités de CERES en 2026

Que fait-on des données textuelles des SHS ?

Que veulent faire les collègues en SHS ?

Que fait-on des données textuelles des SHS ?

Que veulent faire les collègues en SHS ?



Figure 2 – "C'est peut être pas votre question mais c'est ma réponse"

Que peut-on faire des données textuelles des SHS ?

Les données sont-elles *Machine Readable* ?

Que peut-on faire des données textuelles des SHS ?

Les données sont-elles *Machine Readable* ?

Sinon ce sont des données **non-structurées** !

Que peut-on faire des données textuelles des SHS ?

Les données sont-elles *Machine Readable* ?

Sinon ce sont des données **non-structurées** !

Ou **Computationnellement opaques** [de Busser and Moens, 2006]

Que peut-on faire des données textuelles des SHS ?

Les données sont-elles *Machine Readable* ?

Sinon ce sont des données **non-structurées** !

Ou **Computationnellement opaques** [de Busser and Moens, 2006]



Que peut-on faire des données textuelles des SHS ?

Les données sont-elles *Machine Readable* ?

Sinon ce sont des données **non-structurées** !

Ou **Computationnellement opaques** [de Busser and Moens, 2006]



D'ailleurs pourquoi "nettoyage" ou pré-traitement ? [Millour, 2020]

Extrait du résumé de la journée d'aujourd'hui

Les données des LSHS ne constituent pas seulement des **objets d'application pour l'informatique**,

Que faire de données textuelles

Extrait du résumé de la journée d'aujourd'hui

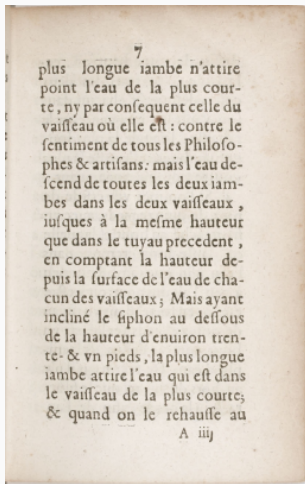
Les données des LSHS ne constituent pas seulement des **objets d'application pour l'informatique**, mais qu'elles sont aussi productrices de **questions théoriques et techniques**

Extrait du résumé de la journée d'aujourd'hui

Les données des LSHS ne constituent pas seulement des **objets d'application pour l'informatique**, mais qu'elles sont aussi productrices de **questions théoriques et techniques** susceptibles de **stimuler la recherche en informatique**.

Qualité réelle et qualité ressentie

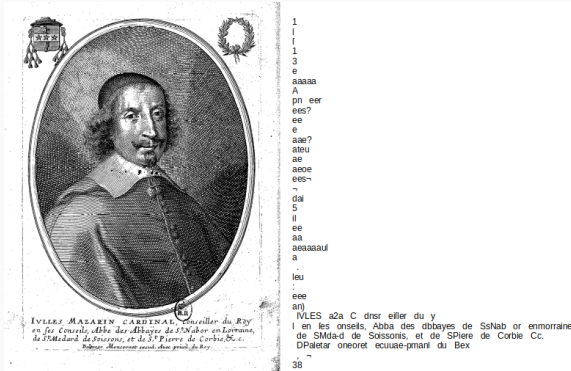
Dans l'idéal on a (presque) le texte complet



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'environ 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 iij

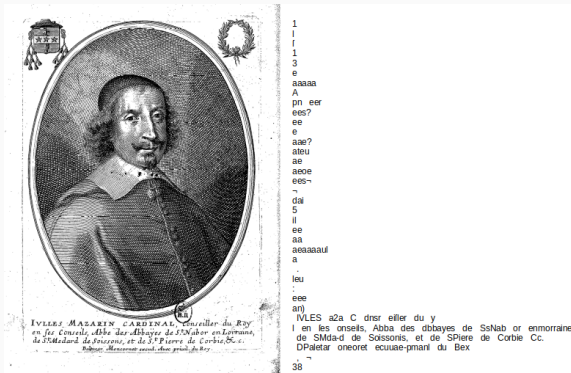
Une page bien "formée", des caractères bien détectés (modèle OCR-17
[Gabay et al., 2020])

Dans les faits



Une page difficile à traiter, du bruit

Dans les faits



Une page difficile à traiter, du bruit

Que faire avec ces données ?

Quelle qualité en entrée et sortie d'un OCR ?

Comment se manifeste le bruit ?

-
1. Expériences sur le corpus ELTEC

Quelle qualité en entrée et sortie d'un OCR ?

Comment se manifeste le bruit ?

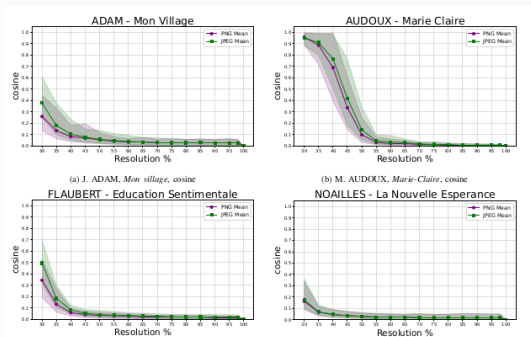


Figure 3 – Une légende (en bonne partie)

Nul besoin d'avoir des TIFF haute résolution pour avoir un *output* exploitable [Boubehziz et al., 2025] ¹

1. Expériences sur le corpus ELTEC

Quel bruit, quelle fréquence ?

Distance de l'input avec la référence

Quel bruit, quelle fréquence ?

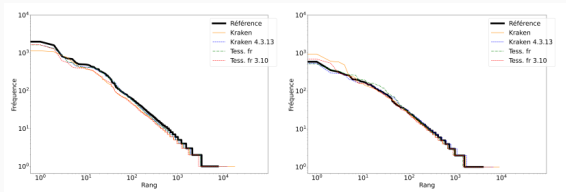


Figure 4 – Les textes bruités suivent aussi une courbe de Zipf

Seules les basses fréquences sont véritablement affectées[Parfait, 2025]

Le bruit et les tâches en aval

Que faire avec des textes ?

On a des données écrites pour être lues par un humain

On voudrait pouvoir les consulter avec la même acuité qu'un locuteur

Que faire avec des textes ?

On a des données écrites pour être lues par un humain

On voudrait pouvoir les consulter avec la même acuité qu'un locuteur

Pour ça, une manière de faire c'est "enrichir/expliciter" les données :

- raciniser : voyageur, voyage $\simeq$ voyage_
- lemmatiser : fassiez, faites $\simeq$ faire

Que faire avec des textes ?

On a des données écrites pour être lues par un humain

On voudrait pouvoir les consulter avec la même acuité qu'un locuteur

Pour ça, une manière de faire c'est "enrichir/expliciter" les données :

- raciniser : voyageur, voyage $\simeq$ voyage_
- lemmatiser : fassiez, faites $\simeq$ faire
- rapprocher : lièvre VS lapin et lièvre VS cochon d'inde
- désambigüiser : la porte vs la porte

Que faire avec des textes ?

On a des données écrites pour être lues par un humain

On voudrait pouvoir les consulter avec la même acuité qu'un locuteur

Pour ça, une manière de faire c'est "enrichir/expliciter" les données :

- raciniser : voyageur, voyage $\simeq$ voyage_
- lemmatiser : fassiez, faites $\simeq$ faire
- rapprocher : lièvre VS lapin et lièvre VS cochon d'inde
- désambigüiser : la porte vs la porte
- **reconnaître/compter des expressions**

Que faire avec des textes ?

Ce mardi 6 octobre est un grand jour pour les amateurs de football, fans du FC Metz ORG . C'est ce mardi que Robert Pirès PER organise son Jubilé MISC au stade Saint-Symphorien LOC . Il a dévoilé fin septembre, les contours de cette soirée. À partir de 20 h 45, trois équipes se retrouveront sur le terrain : FC Metz Légendes PER , Pirès All Stars MISC et Arsenal Légendes PER . Lilian Thuram PER , Thierry Henry PER , Youri Djorkaeff PER ou encore Didier Drogba PER : voici quelques noms des joueurs qui seront présents.

Ce mardi 6 octobre est un grand jour pour les amateurs de football, fans du FC Metz ORG . C'est ce mardi que Robert Pires PER organise son Jubilé MISC au stade Saint Symphorien LOC . Il a dévoilé fin septembre, les contours de cette soirée. À partir de 20 h 45, trois équipes se retrouveront sur le terrain : FC metz légendes ORG , Pirès All Stars MISC et Arsenal Légendes PER . Lilian Thuram PER , Gaël Lejeune PER , Youri Djorkaef PER ou encore Didier Drogba PER : voici quelques noms des joueurs qui seront présents.

Mais quand même si on corrigeait ?

On me dit que la reconnaissance d'entités nommées marcherait mieux si mes données étaient propres ?

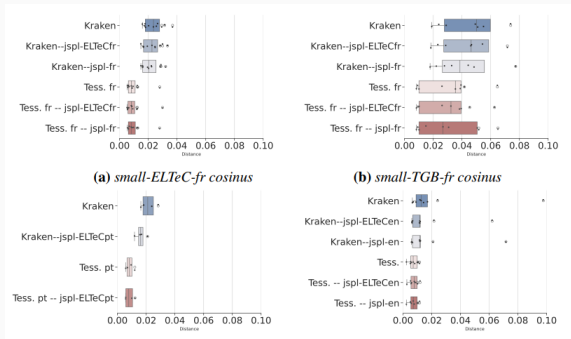


Figure 5 – Influence de la correction automatique sur la qualité des entités extraites

La correction "traditionnelle" n'améliore que marginalement les résultats[Koudoro-Parfait et al., 2024]

Biais de lecture des résultats de REN

Gold et Silver standard

Biais de lecture des résultats de REN

GOLD (référence manuelle)	SILVER (données "parfaites")	OCR
"Languedoc" "Marseille"	"Languedoc" "Marseille"	"Languedoc" "Marseille"
"département du Cher" "Rhône"	"département du Cher" "Rhône"	"Cher" "Rh6ne/Rhone"
"La grande route"	—	—
N/A "Lyon" N/A	"Eysette" — "Annou"	"Eysette" "Lyon" —

Conclusion

1. Les systèmes NLP/CV sont imparfaits . . . même sur données parfaites

1. Les systèmes NLP/CV sont imparfaits . . . même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose

Quatre enseignements

1. Les systèmes NLP/CV sont imparfaits . . . même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose
2. Avec un bruit raisonnable, les tendances statistiques restent
 - On peut donc travailler hors "**conditions de laboratoire**"

Quatre enseignements

1. Les systèmes NLP/CV sont imparfaits ... même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose
2. Avec un bruit raisonnable, les tendances statistiques restent
 - On peut donc travailler hors "**conditions de laboratoire**"
3. On peut accepter d'aménager sa question de recherche
 - ... mais si elle a été trop tordue, **se méfier**

1. Les systèmes NLP/CV sont imparfaits ... même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose
2. Avec un bruit raisonnable, les tendances statistiques restent
 - On peut donc travailler hors "**conditions de laboratoire**"
3. On peut accepter d'aménager sa question de recherche
 - ... mais si elle a été trop tordue, **se méfier**
4. Le volet numérique n'est pas une fin en soi
 - ... même si c'est plus vendeur ...

Quatre enseignements

1. Les systèmes NLP/CV sont imparfaits ... même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose
2. Avec un bruit raisonnable, les tendances statistiques restent
 - On peut donc travailler hors "**conditions de laboratoire**"
3. On peut accepter d'aménager sa question de recherche
 - ... mais si elle a été trop tordue, **se méfier**
4. Le volet numérique n'est pas une fin en soi
 - ... même si c'est plus vendeur ...

Question subsidiaire : faut-il tout passer dans des LLM/VLM ?

Quatre enseignements

1. Les systèmes NLP/CV sont imparfaits ... même sur données parfaites
 - **Corriger les données** rassure plus qu'autre chose
2. Avec un bruit raisonnable, les tendances statistiques restent
 - On peut donc travailler hors "**conditions de laboratoire**"
3. On peut accepter d'aménager sa question de recherche
 - ... mais si elle a été trop tordue, **se méfier**
4. Le volet numérique n'est pas une fin en soi
 - ... même si c'est plus vendeur ...

Question subsidiaire : faut-il tout passer dans des LLM/VLM ?

Et l'hétérogénéité alors ?



Boros, E. and Ehrmann, M. (2025).

Investigating ocr-sensitive neurons to improve entity recognition in historical documents.

In Oliver, G. e. a., editor, *Sustainability and Empowerment in the Context of Digital Libraries*, pages 54–66, Singapore.



Boubehziz, T., Koudoro-Parfait, C., and Lejeune, G. (2025).

Assessing the Impact of Image Resolution on OCR Transcription Accuracy.

In *3rd International Workshop on Digital Language Archives @JCDL 2025 (CORE A*)*, pages 25–29, Online. .



de Busser, R. and Moens, M.-F. (2006).

Information extraction and information technology, pages 1–22.

Springer.



Gabay, S., Bartz, A., and Deguin, Y. (2020).

Corpus17 : a philological french corpus for 17th century.

In *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, pages 1–7.



Koudoro-Parfait, C., Petkovic, L., and Roe, G. (2024).

Analyse multilingue de l'impact de la correction automatique de la ROC sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires.

Traitement Automatique des Langues, 64(2) :43–67.



Millour, A. (2020).

Myriadisation de ressources linguistiques pour le TA de langues non standardisées.

PhD thesis, Sorbonne Université, France.



Parfait, C. (2025).

**Des IA au service de l'espace littéraire du XIXe siècle :
évaluation et analyse des outils de reconnaissance d'entités
nommées spatiales.**

Theses, Sorbonne Université.