Correction automatique des interférences OCR dans la reconnaissance d'entités nommées spatiales : réel gain ou perte de l'information ?

Caroline Koudoro-Parfait[1,2,3], Ljudmila Petkovic[1], Gaël Lejeune[2]

Journée d'étude « Bruit de fond ou valeur ajoutée ? »

Grenoble, 28 avril 2023

(1) caroline.parfait@sorbonne-universite.fr
(2) ljudmila.petkovic@sorbonne-universite.fr
(3) gael.lejeune@sorbonne-universite.fr

(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris, France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France

Sommaire

Contexte et enjeux

Question Faut-il corriger les erreurs d'OCR ou tenter de rendre les systèmes de REN plus robustes face aux variations dans les données qui leur sont soumises par les utilisateurs ?

Angles Expé.
- « Impact du bruit des transcriptions OCR sur la REN ? » Comment évaluer la qualité des résultats de REN sur des corpus OCR bruités ?
- « Impact de la correction des transcriptions OCR sur la REN ? »

Données **Litt. Française 19e siècle**,
- Corpus ELTeC (Réf.) [1] : 10 ouvrages, 3195 pages.
- Très grande bibliothèque (TGB) [2] : 10 ouvrages, 1157 pages.

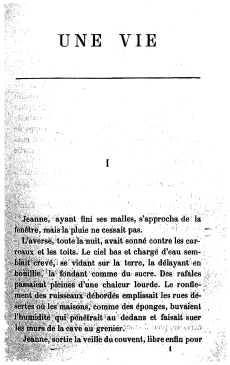
OCR Kraken, Tesseract (Modèle français et de base)

Modèles REN fra. Spacy, Stanza, SEM, CasEN

Corr. auto
- Jamspell [3]
- Un autre ?

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`
2. TGB, `http://obvil.lip6.fr/tgb/`
3. `https://github.com/bakwc/JamSpell`

Difficultés pour l'OCR



UNE VIE

I

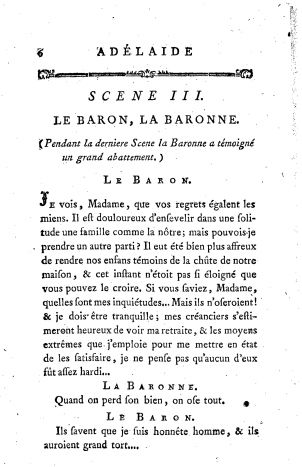
Jeanne, ayant fini ses malles, s'approcha de la fenêtre, mais la pluie ne cessait pas.

L'averse, toute la nuit, avait sonné contre les carreaux, et les toits. Le ciel bas et chargé d'eau semblait crevé, se vidant sur la terre, la délayant en bouillie, la fondant comme du sucre. Des rafales passaient pleines d'une chaleur lourde. Le ronflement des ruisseaux débordés emplissait les rues désertes où les maisons, comme des éponges, buvaient l'humidité qui pénétrait au dedans et faisait suer les murs de la cave au grenier.

Jeanne, sortie la veille du couvent, libre enfin pour

1

(a) Bruit dans l'image



6 ADÉLAIDE

SCENE III.

LE BARON, LA BARONNE.

(*Pendant la derniere Scene la Baronne a témoigné un grand abattement.*)

LE BARON.

Je vois, Madame, que vos regrets égalent les miens. Il eſt douloureux d'enſevelir dans une ſolitude une famille comme la nôtre; mais pouvois-je prendre un autre parti? Il eut été bien plus affreux de rendre nos enfans témoins de la chûte de notre maiſon, & cet inſtant n'étoit pas ſi éloigné que vous pouvez le croire. Si vous ſaviez, Madame, quelles ſont mes inquiétudes... Mais ils n'oſeroient! & je dois être tranquille; mes créanciers s'eſtimeront heureux de voir ma retraite, & les moyens extrêmes que j'emploie pour me mettre en état de les ſatisfaire, je ne penſe pas qu'aucun d'eux fût aſſez hardi...

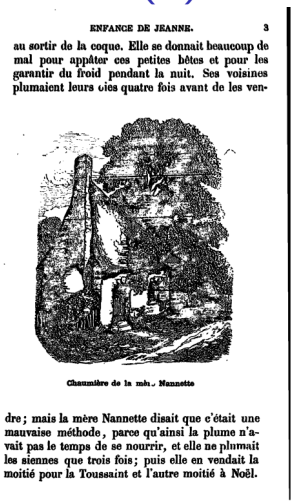
LA BARONNE.

Quand on perd ſon bien, on oſe tout.

LE BARON.

Ils ſavent que je ſuis honnête homme, & ils auroient grand tort....

(b) Décoration, Capitalisation



ENFANCE DE JEANNE. 3

au sortir de la coque. Elle se donnait beaucoup de mal pour appâter ces petites bêtes et pour les garantir du froid pendant la nuit. Ses voisines plumaient leurs oies quatre fois avant de les ven-

Chaumière de la mère Nannette

dre; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi la plume n'avait pas le temps de se nourrir, et elle ne plumait les siennes que trois fois; puis elle en vendait la moitié pour la Toussaint et l'autre moitié à Noël.

(c) Illustration + légende

ALBERT SAVARUS.

DÉDIÉ A MADAME ÉMILE DE GIRARDIN.

Comme un témoignage d'affectueuse admiration.

DE BALZAC.

(d) Texte en colonnes

Figure – a) *"Une vie"*, Maupassant. b) *Drame en cinq actes*, Mariendal. c) *"La petite Jeanne"*, Carraud. d) *"Albert Savarus"*, Balzac.

Influence de la qualité de l'image sur la sortie OCR ?

Table – Transcriptions OCR d'une page de texte théâtral, avec décoration, *"Drame en cinq actes"*, Inconnu.

Kraken	Tess fr
A D E L A I D E v reee= e,E=eeee=ees,g s c E N E I I I. a E LE BARON, LA B ARONNE. r (Pendant l_ derniere Scene la Baronne a temoigne te grand aattement.) L s B A R o N. _E vois, Madame, que vos regrets egalent les miens. [...] LA B A R o N_ _ . Quand on perd fon bien, on ofe tout.e L B B A R o N. Ils favent que je fuis honnete homme, & _ _ auroient grand tort...	6 'ADÉLAIDE SCENE III. LE BARON, LA BARONNE. (Pendant la derniere Scene la Baronne a témoigné un grand abattement,) Le BARON. $ E vois, Madame, que vos regrets égalent les miens.[...] LA BARONNE. Quand on perd fon bien, on ofe tout, LE Baron. Ils favent que je fuis honnête homme, & ils auroient grand tort... -°

➜Difficultés de l'OCR : Bleu, Décorations ; Vert, Capitalisation ; Rouge, bruits divers. Bruits = ajout/transformation/suppression

Alignement manuel d'EN contaminées, OCR non corrigés

Version	Context	spaCy_lg	stanza
Ref.*	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré
Kraken	*[...] la rue Saint-Honore ;*	rue Saint-Honore	rue Saint-Honore
Tess	*[...] larue Saint-Honoré ;*	_ Saint-Honoré	()
Tess fr	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré
Ref.*	*les États [...] de Guadalajara*	Guadalajara	Guadalajara
Kraken	*les Etats [...] de Guadalazara*	Guadalazara	Guadalazara
Tess	*les Etats [...] de Guadalaxara*	Guadalaxara	Guadalaxara
Tess fr	*les États [...] de Guadalaæw*a*	Guadalaæw*a	Guadalaæw*a
Ref.**	*[...] rue de Sèvres ;*	rue de Sèvres	rue de Sèvres
Kraken	*[...] ruc de Sbvrcs*	Sbvrcs	Sbvrcs
Tess	*[...] rue de Sévres*	Sévres	rue de Sérves
Tess fr	*[...] rue de Sèvres ;*	rue de Sèvres	rue de Sèvres
Ref.**	*le faubourg Saint-Germain*	faubourg Saint-Germain	faubourg Saint-Germain
Kraken	*lc faubourg Saint-Gcrmain*	Saint-Gcrmain	faubourg Saint-Gcrmain
Tess	*le faubourg Saint-Germain*	faubourg Saint-Germain	faubourg Saint-Germain
Tess fr	*le faubourg Saint-Germain*	faubourg Saint-Germain	faubourg Saint-Germain

Table – Alignement : Faux Positif ou Vrai Positif ?[1]

* *"Le chateau de Pinon"*, Dash , *"Les trappeurs de l'Arkansas"*, Aimard.** *"Souvenirs d'un vieux mélomane"*, Pontmartin.

Alignement manuel d'EN contaminées, OCR corrigés

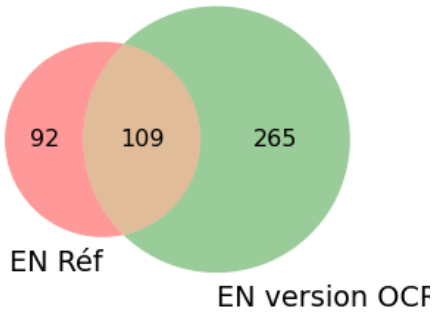
	ELTeC Français[1]			
	OCR(Kraken)	Corr.	REN	REN corr.
MOBC	"[. . .] Parisl"	"[. . .] Paris"	Parisl Ici	Paris
MOMC	"[. . .] Grand-hlail"	"[. . .] Grand-Hail"	Grand-hlail	X
MOI	"[. . .] Mlorlincourtl"	"[. . .] Mlorlincourtl"	Mlorlincourtl	Mlorlincourtl
BOIC	"[. . .] Morlincourt "	"[. . .] Martincourt"	Morlincourt	Martincourt

	TGB[2]			
	OCR	Corr.	REN	REN corr.
MOBC	"[. . .] Sainl-Cyr"	"[. . .] Saint-Cyr"	Sainl-Cyr	Saint-Cyr
MOMC	"[. . .] Nîme"	"[. . .] Mîme"	Nîme	Mîme
MOI	"[. . .] SaintAntoine"	"[. . .] SaintAntoine"	SaintAntoine	SaintAntoine
BOIC	"[. . .] Pinde"	"[. . .] Inde"	Pinde	Et du sommet du Inde

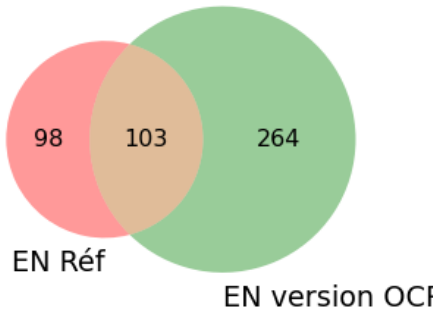
Table – Résultats de REN spaCy-lg. Types de formes : MOBC - mal océrisées bien corrigées ; MOMC - mal océrisées mal corrigées ; MOI - mal océrisées ignorées ; BOIC - bien océrisées indûment corrigées. Corr. - correction avec jamspell, REN corr. - REN après la correction avec Jamspell.

[1] *Mon Village*, Adam. *"Le petit chose"*, Daudet.
[2] *"Oeuvres poétiques de Boileau"*, Boileau. *"Poésies de Valentin"*, Bourette. *"Histoire des sociétés secrètes, politiques et religieuses"*, Zaccone.

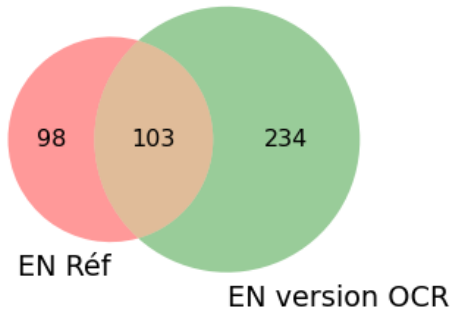
Évaluer l'impact de la correction OCR sur la REN



(a) non corrigé



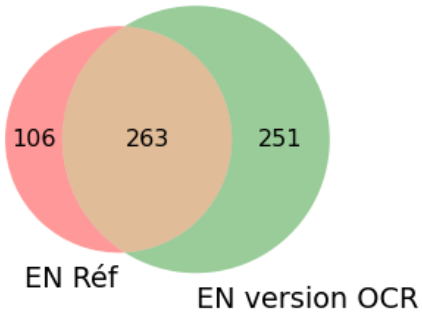
(b) jspl-pretrain



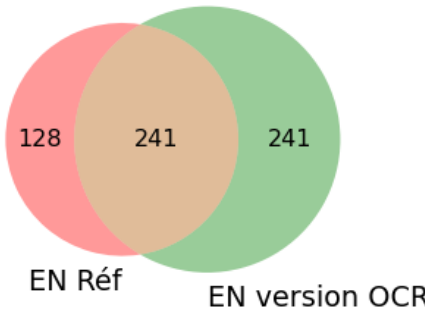
(c) jspl-ELTeCmodel

Figure – a), b), c) *"Le petit chose"*, Daudet. Configuration : Kraken, spaCy 3.5.1, modèle lg

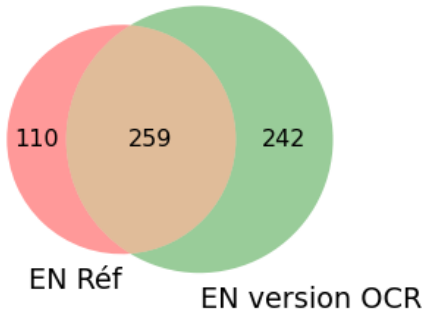
Évaluer l'impact de la correction OCR sur la REN



(a) non corrigé



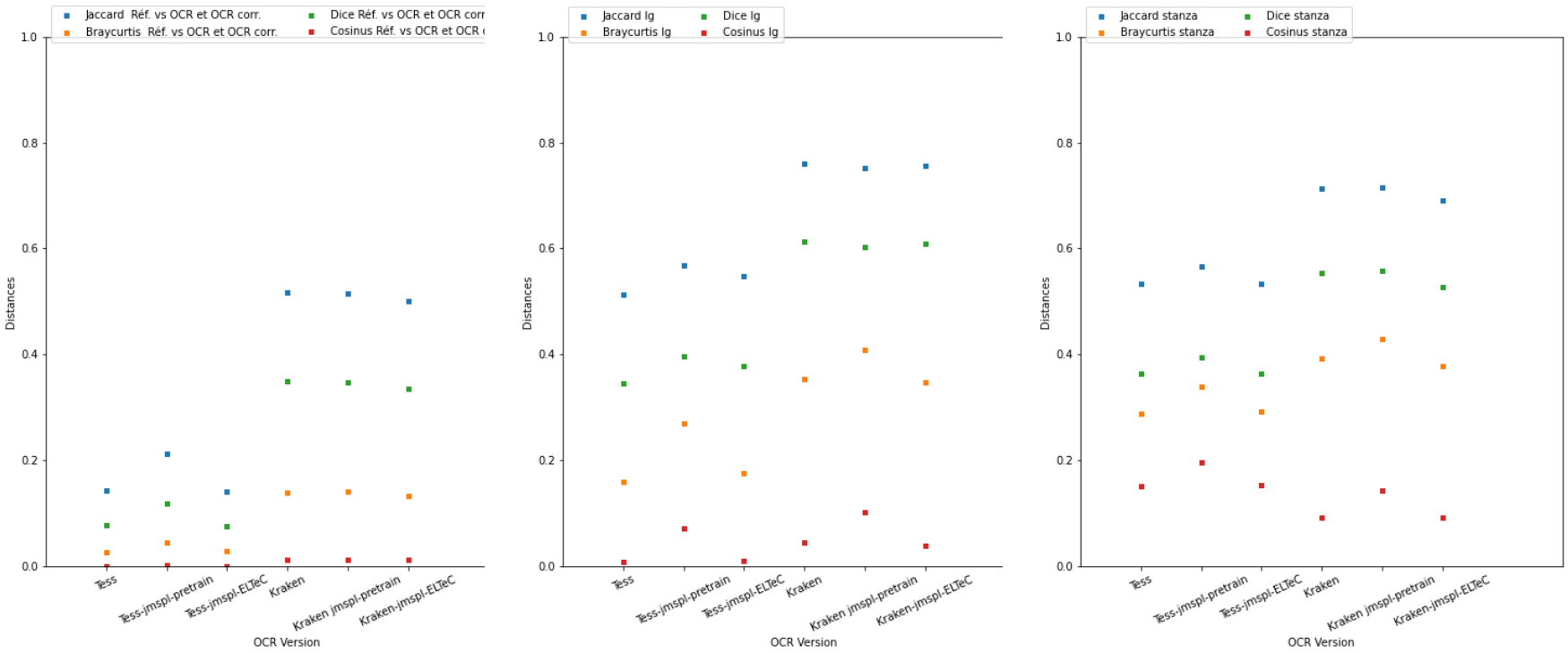
(b) jspl-pretrain



(c) jspl-ELTeCmodel

Figure – a), b), c) *"Le petit chose"*, Daudet. Configuration : TesseractFra-Png, spaCy 3.5.1, modèle lg

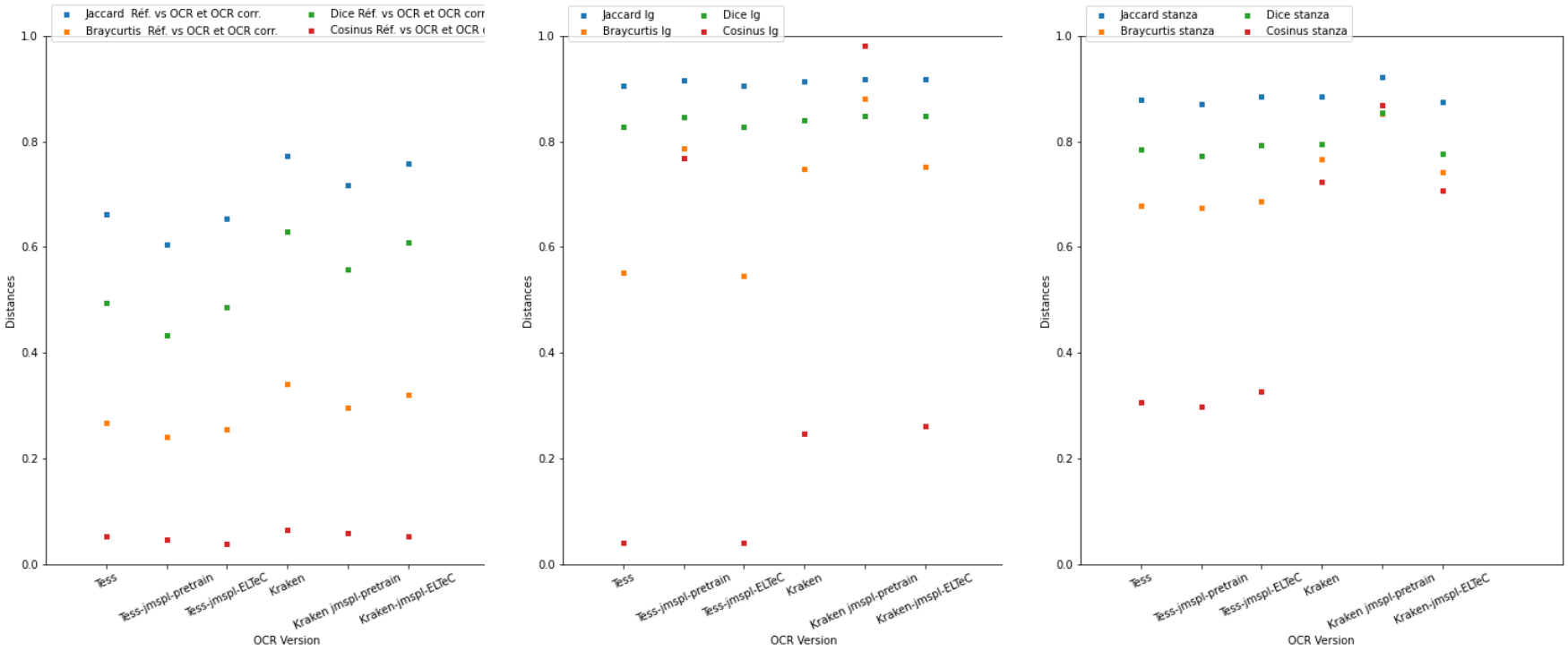
Évaluer l'impact de la correction OCR sur la REN



(a) `texte` (b) `spaCy_lg` (c) `stanza`

Figure – Distances Jaccard, Dice, Bray-Curtis, Cosinus, a), b), c) *"Le petit chose"*, Daudet.

Évaluer l'impact de la correction OCR sur la REN



(a) `texte` (b) `spaCy_lg` (c) `stanza`

Figure – Distances Jaccard, Dice, Bray-Curtis, Cosinus, a), b), c) *"Drame en 5 actes"*, Inconnu.

➜ **OCR de meilleure qualité → Meilleurs résultats de REN ? → Moins de Faux Négatif/silence**
➜ **Correction de meilleure qualité → ? Amélioration de la REN ?**

Nerval : un outil pour l'alignement des EN contaminées

	#Entités		Évaluation par Nerval			
Version	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken	1122	944	566	0.504	0.761	*0.607*
Tess fr	860	944	646	0.751	0.868	**0.805**
Tess	920	944	597	0.649	0.802	*0.718*

	#Entités		Évaluation par Nerval			
Version	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken + Jspl-ELTeC	1055	944	548	0.519	0.737	*0.609* ⇑
Tess fr + Jspl-ELTeC	838	944	621	0.741	0.835	*0.785* ⇓
Tess + Jspl-ELTeC	927	944	576	0.621	0.774	*0.689* ⇓

	#Entités		Évaluation par Nerval			
Version	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken + Jspl-pt	1027	944	471	0.459	0.633	**0.532** ⇓
Tess fr + Jspl-pt	794	944	532	0.67	0.715	***0.692*** ⇓
Tess + Jspl-pt	846	944	503	0.595	0.676	*0.633* ⇓

Table – Alignement des entités (`spaCy_lg format IOB`) de la référence et des versions OCR corrigées avec le modèle français préentrainé de Jamspell et un modèle entrainé sur une partie du corpus ELTeC, Nerval [4], *"Le petit chose"*, Daudet, 1868.

4. `https://gitlab.com/teklia/ner/nerval`

Contributions

- **Toutes** les **erreurs d'OCR** ne se **valent pas**
  ➜ fautes d'orthographe $\neq$ mots collés ensemble
- Impact du bruit OCR sur les sorties de NER
  ➜ **EN contaminées** = **Variations orthographiques** d'un toponyme
- Usages de **différentes métriques** : **évaluation fine**/nuancée
- Méthodes d'alignements et évaluation : évaluation Nerval

Contributions

- **Toutes** les **erreurs d'OCR** ne se **valent pas**
  ➜ fautes d'orthographe $\neq$ mots collés ensemble
- Impact du bruit OCR sur les sorties de NER
  ➜ **EN contaminées** = **Variations orthographiques** d'un toponyme
- Usages de **différentes métriques** : **évaluation fine**/nuancée
- Méthodes d'alignements et évaluation : évaluation Nerval
- **La correction automatique n'est pas automatiquement un gain** :
  - l'outil produit un nombre conséquent des sur-corrections
  - l'outil ne corrige pas toutes les contaminations/interférences OCR
  - La correction des EN a moins d'impact que la correction du contexte(?)
- **Typologie des erreurs de corrections automatiques** d'OCR