#dhnord2021

Réutiliser les corpus numériques constitués par OCR : bruit et silence dans les data papers

Caroline Koudoro-Parfait (1,2,3), Gaël Lejeune (2)
Jean-Baptiste Tanguy (1,2)

DHNord 2021

17 Novembre 2021

caroline.parfait@sorbonne-universite.fr,
jean-baptiste.tanguy@sorbonne-universite.fr
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

Construction du Corpus à partir de numérisations en ligne

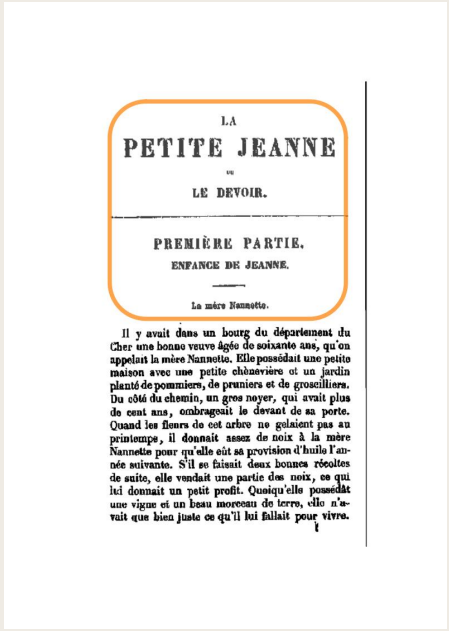
Données 17/19ème
- Corpus ANTONOMAZ[a]
- Corpus ELTEC[b]

Outils OCR
- Kraken (Modèle par défaut et modèle du 17ème)
- Tesseract (Modèle par défaut et modèle du français)

Outils REN[c]
- Spacy (modèles `large`, `medium` et `small` pour le français)
- Stanza (modèle du français)

Data Paper
- Observer les sorties d'OCR et de REN manuellement
- Évaluer le bruit et le silence

dans les sorties :
- Intersections
- Taux de confiance
- Distances



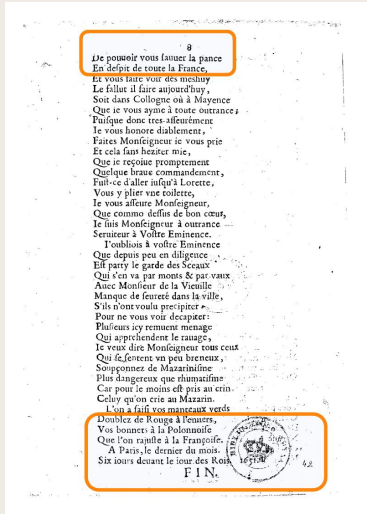
LA
PETITE JEANNE
OU
LE DEVOIR.

PREMIÈRE PARTIE.
ENFANCE DE JEANNE.

La mère Nannette.

Il y avait dans un bourg du département du Cher une bonne veuve âgée de soixante ans, qu'on appelait la mère Nannette. Elle possédait une petite maison avec une petite chènevière et un jardin planté de pommiers, de pruniers et de groseilliers. Du côté du chemin, un gros noyer, qui avait plus de cent ans, ombrageait le devant de sa porte. Quand les fleurs de cet arbre ne gelaient pas au printemps, il donnait assez de noix à la mère Nannette pour qu'elle eût sa provision d'huile l'année suivante. S'il se faisait deux bonnes récoltes de suite, elle vendait une partie des noix, ce qui lui donnait un petit profit. Quoiqu'elle possédât une vigne et un beau morceau de terre, elle n'avait que bien juste ce qu'il lui fallait pour vivre.

(a) Titre, sous titre, titre de partie...*



8

De pouvoir vous fauuer la pance
En defpit de toute la France,
Et vous faire voir des meshuy
Le fallut il faire aujourd'huy,
Soit dans Collogne où à Mayence
Que ie vous ayme à toute outrance;
Puifque donc tres-aſſeurément
Ie vous honore diablement,
Faites Monſeigneur ie vous prie
Et cela fans heziter mie,
Que ie reçoiue promptement
Quelque braue commandement,
Fuſt-ce d'aller iuſqu'à Lorette,
Vous y plier vne toilette,
Ie vous aſſeure Monſeigneur,
Que comme deſſus de bon cœur,
Ie ſuis Monſeigneur à outrance
Seruiteur à Voſtre Eminence.
I'oubliois à voſtre Eminence
Que depuis peu en diligence
Eſt party le garde des Sceaux
Qui s'en va par monts & par vaux
Auec Monſieur de la Vieuſille
Manque de feureté dans la ville,
S'ils n'ont voulu precipiter
Pour ne vous voir decapiter:
Pluſieurs icy remuent menage
Qui apprehendent le rauage,
Ie veux dire Monſeigneur tous ceux
Qui ſe ſentent vn peu breneux,
Soupçonnez de Mazarinifme
Plus dangereux que rhumatiſme
Car pour le moins eſt pis au crin
Celuy qu'on crie au Mazarin.
L'on a faiſi vos manteaux verds
Doublez de Rouge à l'eſpagnol,
Vos bonnets à la Polonnoiſe
Que l'on rajuſte à la Françoiſe.
A Paris, le dernier du mois.
Six iours deuant le iour des Rois.
FIN.

(b) Numéro de la page, tampon, (illustrations)* *

Figure – Difficultés de l'OCR pour la récupération des métadonnées et des données .

[a] ANalyse auTOmatique et NumérisatiOn des MAZarinades, 2700 imprimés, 20 000 pages.
[b] European Literary Text Collection , 11 oeuvres, 3 195 pages.
[c] Reconnaissance d'entités nommées.

* *"La petite Jeanne"*, Zulma Carraud,1884, ELTEC.
* * *Les estrennes burlesques de M. Scaron envoyées à Mazarin*, 1652, ANTONOMAZ.

Récupération de données à partir de transcriptions automatiques

→ Choix de l'OCR ? Le data Paper propose :

➢ Tableau d'observation manuelle

Kraken	Tess fr
__ P_TITB JLANN_ IL DRV0IR. PHEIIIkI PAIII_. EM_ANCE DE JEANME, _a mbre Nannette.	__ PETITE JEANNE LE DEVOIR. _— PREMIÈRE PARTIE. ENFANCE DE JEANNE. La mère Nannette.

Table – 2 sorties OCR sur le texte * ELTEC.

Kraken 17
8 .2 S De pouuoir vous lauuver la pance En deſpit de toute la France, [...] Vos bonnets a la Polonnoiſe lſe¬ Que l'on rajuſte a la Francoiſe ;l , ile 4 A Par _s, le dermier du mois , je Six io nrs deuant le iour des loils _ 5esifi a2 4 . FIN. ¬¬

Table – 1 sortie OCR sur le texte ** ANTONOMAZ.

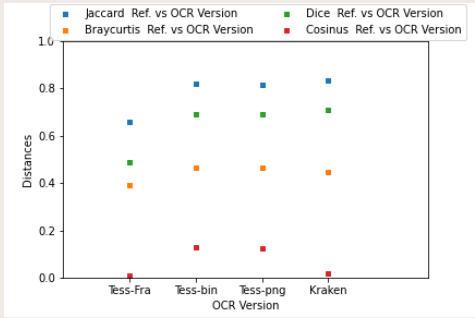
Bruit et silence OCR :

➢ Rouge : substitutions de caractères.

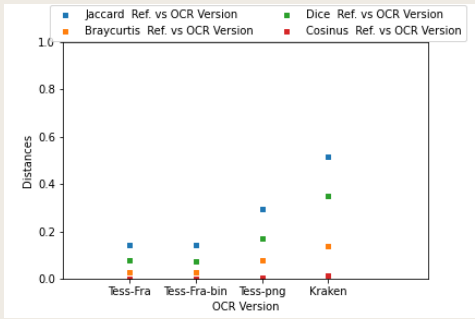
➢ Bleu : transcription d'un graphisme.

➢ "_" : caractère manquant.

➢ Evaluation automatique des sorties OCR

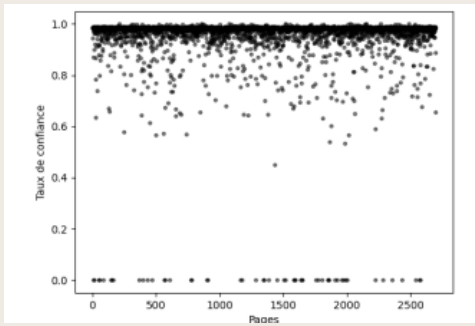


(a) *"Albert Savarus"*, Balzac, 1853.

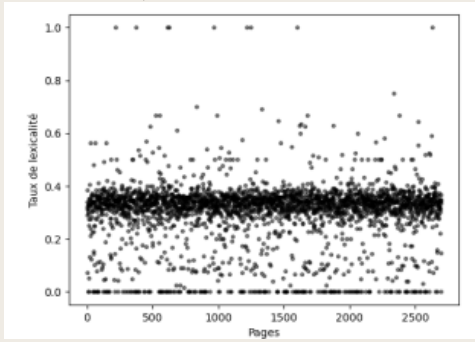


(b) *"Le petit chose"*, Daudet, 1868.

Figure – Mesurer la qualité de l'OCR, métriques : Jaccard, Cosinus, Dice et Braycurtis (1= Différent, 0= Similaire).



(a) Taux de confiance.



(b) Taux de Lexicalité.

Figure – Mesurer la qualité de l'OCR pour tout le corpus Antonomaz (1= Qualité correcte).

REN sur des corpus bruités : Observation manuelle des données

REN Réf.	REN OCR	Verdict 1	Détails	Verdict 2
Oui	Oui	VP	Entité nommée réelle	Vrai VP
Oui	Oui	VP	Erreurs dans les sorties de REN pour la réf. et l'OCR	Faux VP
Non	Non	VN	Aucune entité dans aucune version	Vrai VN
Non	Non	VN	Entité manquante dans les deux versions	Faux VN
Oui	Non	FN	Entité manquante dans la version OCR	Vrai FN
Oui	Non	FN	Erreur d'entité dans la réf.	Faux FN
Non	Oui	FP	Erreur d'entité dans l'OCR	Vrai FP
Non	Oui	FP	Entité manquante dans la réf.	Faux FP (VP)
Non	Oui	FP	Problème de liaison entre entités (*entity linking*)	Faux FP (VP ?)

Table – Typologie des erreurs de REN.

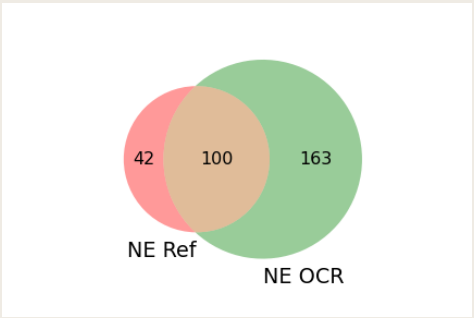
Version	Context	spacy_lg	stanza
Ref.	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré
Kraken	*[...] la rue Saint-Honore ;*	rue Saint-Honore	rue Saint-Honore
Tess	*[...] larue Saint-Honoré ;*	_ Saint-Honoré	()
Tess fr	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré
Ref.	*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.	*La Rochelle.[...] Thouloufe.*	La Rochelle, () *	La Rochelle, Thouloufe
Kraken 17	*La Rocdhelle .[...] Thouloufe.*	Rocdhelle , () *	La Rocdhelle , Mer. Thouloufe
Ref.	*[...]Viennes en Auftriche.*	Viennes, Auftriche	Viennes, Auftriche
Kraken 17	*[...]Viennes en Auftriche.*	Viennes, Auftriche	Viennes, Auftriche

Table – Problèmes d'entity linking et silence.

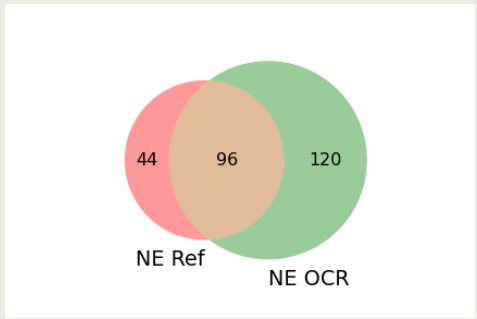
*Silence dû au s long non utilisé dans la graphie actuelle ? cf. exemple suivant.

Évaluations des sorties de REN dans le data paper

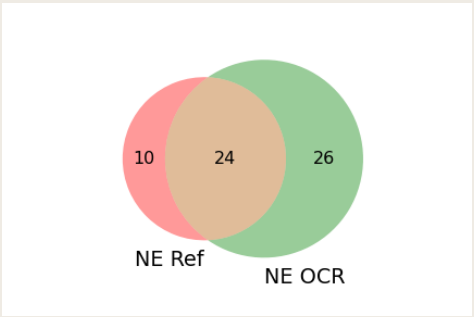
➢ Intersections :



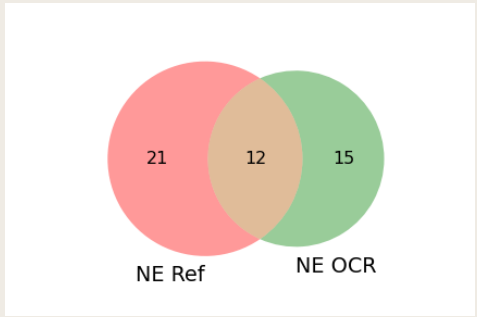
(a) spacy_lg - Tess fr



(b) stanza - Tess fr



(c) spacy_lg - Kraken 17



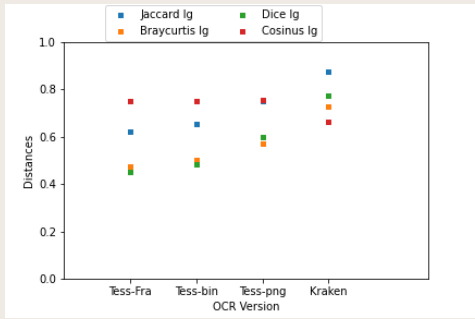
(d) stanza - Kraken 17

- NE OCR : Problèmes d'*entity linking* et bruit corrigeables.
- NE REF : Vrai positif (potentiel silence dans la référence ?).
- Intersection : Vrai positif et erreurs Réf./OCR de l'outil REN.

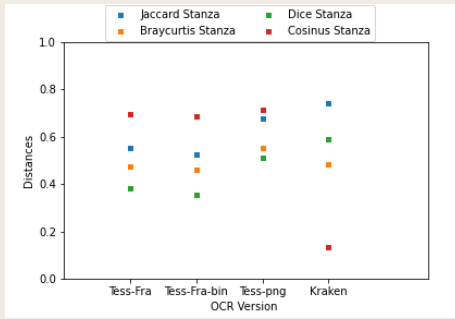
a b *"Une vie"*, Maupassant, 1883.

c d *Adieu (l') de Mazarin, burlesques*, 1649.

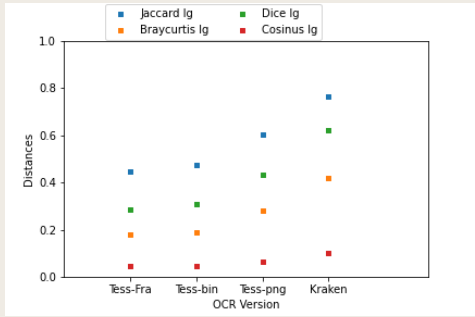
➢ Distances et Similarités :



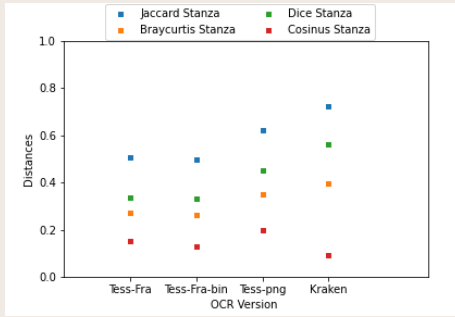
(a) spacy_lg



(b) stanza



(c) spacy_lg



(d) stanza

- Distance Cosinus : sous-évalue les différences(?)
- Distance de Jaccard : sur-évalue les différences(?)
- Proximité distances Cosinus et Jaccard : Problème ?

a b *"Une vie"*, Maupassant, 1883.

c d *"Le petit chose"*, Daudet,1868.