Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique.

Caroline Koudoro-Parfait (1,2,3), Gaël Lejeune (2), Richy Buth (2)

TALN | RECITAL 2022 - Atelier TALN & HN 2022

27 Juin 2022

caroline.parfait@sorbonne-universite.fr
gael.lejeune@sorbonne-universite.fr
buth_richy@hotmail.com
(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France

Plan

Problématique(s) et Méthode

- "Impact du bruit des transcriptions OCR sur la REN ?"
  - Comment évaluer la qualité des résultats des REN sur des corpus OCR bruités ?
    - Précision, Rappel, F-score ?
    - Approche non supervisée ?
- Il est possible de récupérer des EN sur des textes bruités :
  - Comment associer les EN contaminées à leur forme standard ?

Constitution du/des Corpus

Données

- Corpus ELTeC [1] :
  - Littérature Française 19ème siècle,
  - 11 ouvrages, 3195 pages

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`

Constitution du/des Corpus

Données

- Corpus ELTeC [1] :
  - Littérature Française 19ème siècle,
  - 11 ouvrages, 3195 pages

Outils OCR

- Kraken (Modèle de base)
- Tesseract (Modèle français et de base)

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`

Constitution du/des Corpus

Données

- Corpus ELTeC [1] :
  - Littérature Française 19ème siècle,
  - 11 ouvrages, 3195 pages

Outils OCR

- Kraken (Modèle de base)
- Tesseract (Modèle français et de base)

Outils de REN

- Spacy (`small`, `medium` et `large`)
- Stanza
- SEM (`WiNER`)
- CasEN

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`

Constitution du/des Corpus

Données
- Corpus ELTeC [1] :
  - Littérature Française 19ème siècle,
  - 11 ouvrages, 3195 pages

Outils OCR
- Kraken (Modèle de base)
- Tesseract (Modèle français et de base)

Outils de REN
- Spacy (`small`, `medium` et `large`)
- Stanza
- SEM (`WiNER`)
- CasEN

Alignement & Éval.
- Nerval : évaluation de la REN sur du texte bruité

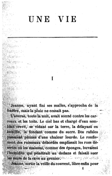
1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`

Corpus ELTeC - European Literary Text Collection

Book	Year	Page nb.
"Le château de Pinon, vol. I ", G. A. Dash, Comtesse	1844	332 p.
"Albert Savarus. Une fille d'Ève. ", Honoré de Balzac	1853	60 p.
"Les trappeurs de l'Arkansas ", Gustave Aimard	1858	450 p.
"Mon village",Juliette Adam (Lambert)	1860	200 p.
"Le petit chose", Alphonse Daudet	1868	292 p.
"L'Éducation sentimentale histoire d'un jeune homme", Gustave Flaubert	1880	520 p.
"Une vie", Guy de Maupassant	1883	337 p.
"La petite Jeanne", Zulma Carraud	1884	220 p.
"La Belle rivière", Gustave Aimard	1894	339 p.
"La nouvelle espérance", Anna de Noailles	1903	325 p.
"Marie-Claire",Marguerite Audoux	1925	120 p.

11 ouvrages, 3195 pages

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



UNE VIE

I

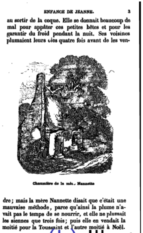
Jeanne, ayant fini ses malles, s'approcha de la fenêtre, mais la pluie ne cessait pas.

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) Du bruit dans l'image.



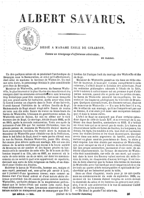
ENFANCE DE JEANNE. 3

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.

(b) Illustration et légende.



ALBERT SAVARUS.

DÉDIÉ A MADAME ÉMILE DE GIRARDIN.

(c) Texte sur deux colonnes.

Figure – Les difficultés de l'OCR [2]

2. a) *"Une vie"*, Guy de Maupassant. b) *"La petite Jeanne", Carraud.* c) *"Albert Savarus"*, Balzac.

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

Table – Transcriptions OCR d'un texte mis en page sur deux colonnes, *"Albert Savarus"*, *Balzac.*

Kraken	**Tess fr**
Un des quelques salons oh se produisait l'arehe_eque de [] lomene fut l'unique fruit du mariago des Wattoville et des Besangcon sous la Reslauralion, et celui qu'il affectionnait,] de Rupt. etait celui de madame la baronne do Watltev_lle. Un mot] Monsieur de Wotteville passait sn vie dans un riche	Un des quelques salons où se produisait l'archevêque de Besançon sous la Restauration, et celui qu'il affectionnait,[...] Les savans observateurs de la nature sociale ne manqueront pas de remarquer que Phi-[3]omène fut l'unique fruit du mariage des Watteville et des de Rupt[4].

3. Colonne de droite en vert
4. Colonne de gauche en noir

Problèmes d'entity linking et silence.

Version	Context	Spacy_lg	Stanza	SEM	CasEN
Ref.	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré	rue Saint-Honoré	rue Saint-Honoré
Kraken	*[...] la rue Saint-Honore ;*	rue Saint-Honore	rue Saint-Honore	rue Saint-Honore	rue Saint-Honore
Tess	*[...] larue Saint-Honoré ;*	_ Saint-Honoré	()	larue Saint-Honoré	_ Saint-Honoré
Tess fr	*[...] la rue Saint-Honoré ;*	rue Saint-Honoré	rue Saint-Honoré	rue Saint-Honoré	rue Saint-Honoré
Ref.	*les États [...] de Guadalajara*	Guadalajara	Guadalajara	Guadalajara	Guadalajara
Kraken	*les Etats [...] de Guadalazara*	Guadalazara	Guadalazara	Guadalazara	()
Tess	*les Etats [...] de Guadalaxara*	Guadalaxara	Guadalaxara	Guadalaxara	()
Tess fr	*les États [...] de Guadalaæw*a*	Guadalaæw*a	Guadalaæw*a	()	()

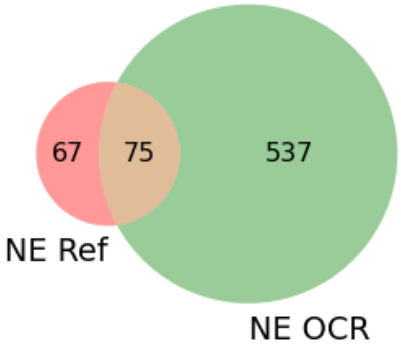
Table – Problèmes d'entity linking : Faux Positif ou Vrai Positif ?[1]

Version	spacy_sm	spacy_lg	Stanza	SEM	CasEN
Ref.	Grèce	Grèce **bleue (M)**	Grèce **bleue (M)**	Grèce bleue	Grèce
Kraken	Grece	Grece bleue	Grece **bleue (M)**	()	()
Tess	Gréce **(P)**	Gréce	Gréce **bleue**	()	()
Tess fr	Grèce	Grèce	Grèce **bleue (M)**	Grèce bleue	Grèce

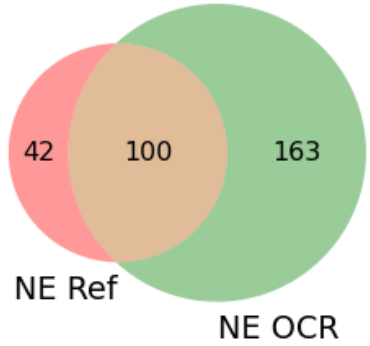
Table – Silence : Vrai négatif ou problème de label ?[2]

[1] *"Le chateau de Pinon"*, Dash , *"Les trappeurs de l'Arkansas"*, Aimard. [2] *"La nouvelle espérance"*, Noailles

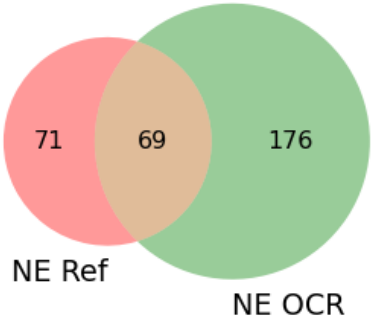
Corrélation entre la qualité de l'OCR et la REN ?



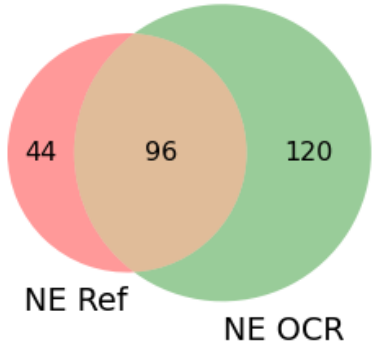
(a) `spacy_lg` - Kraken



(b) `spacy_lg` - Tess fr

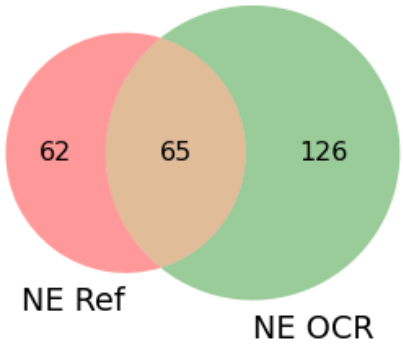


(c) `stanza` - Kraken

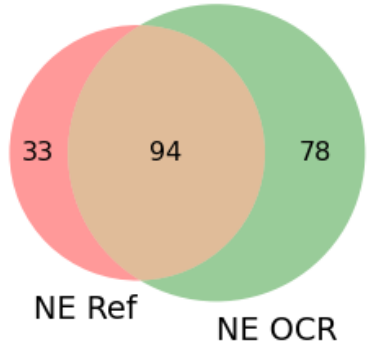


(d) `stanza` - Tess fr

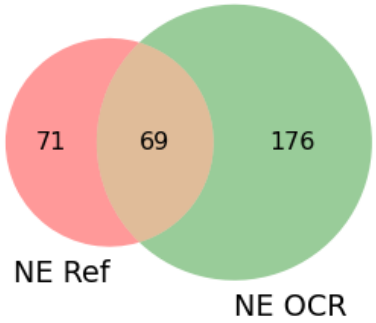
Corrélation entre la qualité de l'OCR et la REN ?



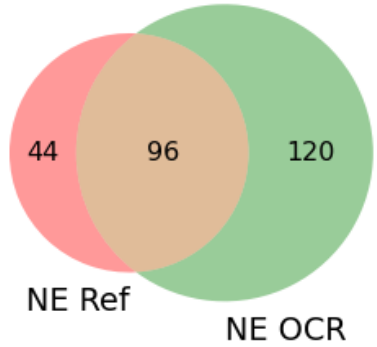
(a) SEM - Kraken



(b) SEM - Tess fr



(c) stanza - Kraken



(d) stanza - Tess fr

➜ **OCR de meilleure qualité** $\rightarrow$ **Moins de Faux Négatifs ?** [5]

5. *"Une vie"*, Maupassant, 1883.

Évaluation et alignement avec Nerval

Version	#Entités		Évaluation par NERVAL			
	Version OCR	Référence	Intersection	Précision	Rappel	F_1 mesure
Kraken	*1391*	965	576	*0.414*	*0.597*	0.489
Tess fr	980	965	713	0.728	0.739	**0.733**
Tess	1090	965	608	0.558	0.630	0.592

Table – Comparaison des résultats de la reconnaissance d'entités nommées avec `spacy_lg` sur différentes versions de *"Le petit chose"*, Daudet, 1868, après alignement avec NERVAL

Version	#Entités		Évaluation par NERVAL			
	Version OCR	Référence	Intersection	Précision	Rappel	F_1 mesure
Kraken	364	100	41	0.113	0.410	0.177
Tess fr	289	100	49	0.170	0.490	**0.252**
Tess	433	100	47	0.109	0.470	0.176

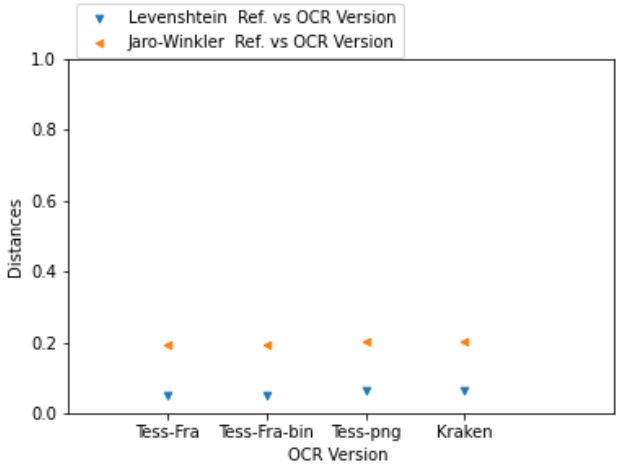
Table – Résultats de la reconnaissance d'entités nommées par système sur différentes versions de *"Marie-Claire"*, Audoux après alignement avec NERVAL

Liage des Entités contaminées avec NERVAL

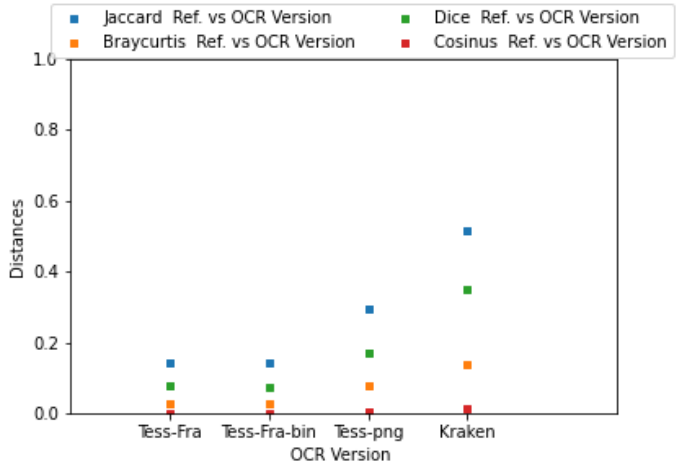
Versions	Entités Réf.[I]	Entités Align	Entités Réf.[II]	Entités Align
Kraken	l'Amérique	— merique	Paris	**eares**
Tess fr		**dé**Amérique		**-aran**
Tess		–Amérique		**aran**
Kraken	rio Gila	Rio Gila	Laon	Lao**u**
Tess fr		Bio Gz–		Laon
Tess		rio Gila		Laon

Table – Alignement des entités (`spacy_lg`) de la référence et des versions OCR avec NERVAL, [I] *"Les trappeurs de l'Arkansas"*, Aimard, 1858 et [II] *"Le château de Pinon, vol. I."*, *Dash, 1844*

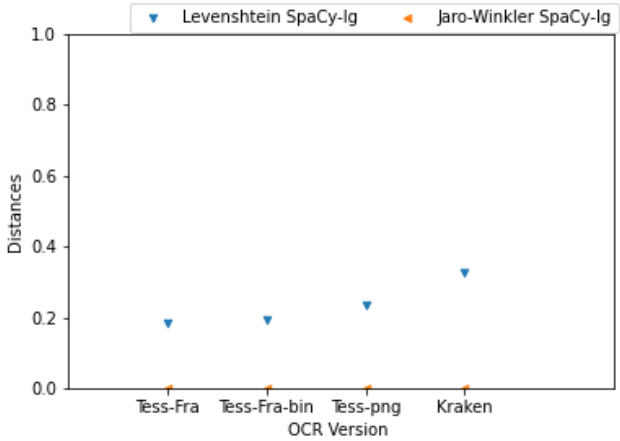
Importance des métriques : Levenshtein vs. Cosinus



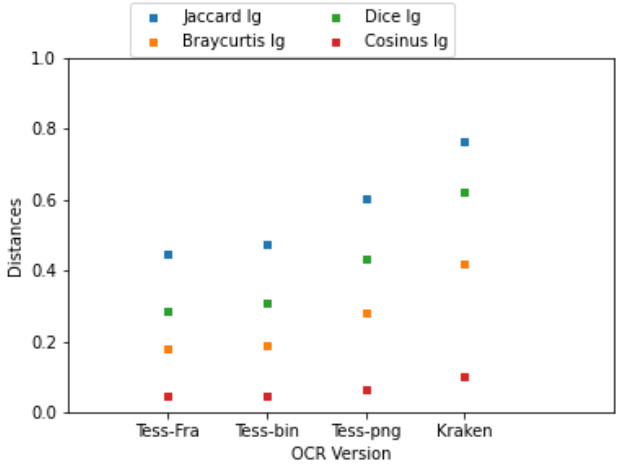
(a) `Réf. vs OCR`



(b) `Réf. vs OCR`



(c) `spacy_lg`



(d) `spacy_lg`

Figure – Distances (gauche) de Levenshtein et Jaro-Winkler entre les Entités Nommées "LOC" Réf. et OCR, *"Le petit chose"*, Daudet, 1868.

Des pistes pour le liage des Entités contaminées

Versions	Entités Réf.	Entités Align	Jaccard[6]	Cosinus[6]
Kraken	Morlincourt	Mlorlincourt	0.1428	0.0715
		Mlorlincourtl	0.1818	0.1210
Tess fr	Morlincourt	Morlinco'urt	0.1818	0.0762
		Morlin	**0.4761**	**0.2788**
Kraken	Saint-Brunelle	Bruncllе	**0.5925**	**0.3244**
		Brunelle	0.4583	0.2012
Tess fr	Saint-Brunelle	Saint—Brunelle	0.2222	0.0909
		Saint—anelle	0.4642	0.2183

Table – Résultats de la récupération automatique des entités contaminées récupérées par `spacy_lg` sur différentes versions de "Mon village", Adam, 1860.

6. Bigrammes de caractères

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
  ➜ Variations orthographiques d'un toponyme : comment les lier ?
- Le choix des métriques n'est pas trivial
  ➜ Distance Cosinus : évalue la différence de quantité entre deux groupes(?)
  ➜ Distance de Jaccard : sur-évalue les différences(?)
  ➜ Levenshtein vs. Cosinus : résultats proches, cos. moindre coût de calcul
- Les humains peuvent gérer le bruit, mais que faire s'il y a du silence ?

Perspectives

- Résolution des problèmes de liage (diachronie, formes fautives des mots)
- Désambiguïsation automatique : liage avec des bases de données ou embeddings ?