Reconnaissance des Entités Nommées spatiales dans un corpus littéraire bruité : Des entités à la carte

Caroline Koudoro-Parfait (1,2,3)

Séminaire STIH

10 Mars 2022

caroline.parfait@outlook.fr
(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

Comment rendre les systèmes de NER plus robustes face aux variations dans les données qui leur sont soumises par les utilisateurs ?

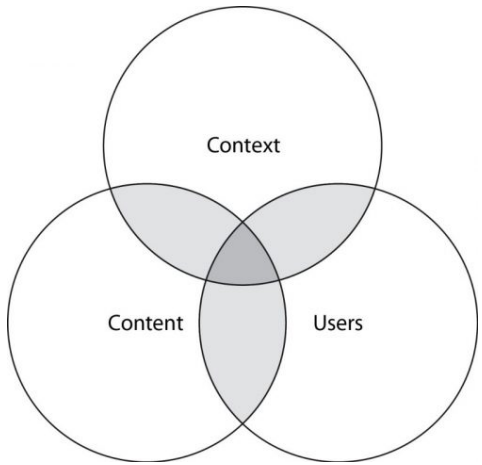


Figure – Peter Morville, *"Three Circles of Information Architecture"*, 2004

- **Utilisateurs** : chercheurs SHS, HN, TAL ... Interdisciplinarité
- **Données** : Corpus ELTeC (Romans Français 19/20ème s.)
- **Variabilités** : diachronie, diatopie, qualité de la transcription OCR ?

Problématique(s) et Méthode

- Enquêtes **Utilisateurs** quantitative et qualitative et Workshop avec J-B. Tanguy
- **Évaluation des systèmes** Spacy, Stanza, SEM et CasEN
- **Angle Expé.** : "Impact du bruit des transcriptions OCR sur la REN ?"
  - Comment évaluer la qualité des résultats des NER sur des corpus OCR bruités ?
    - Précision, Rappel, F-score ?
    - Approche non supervisée ?

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

Constitution du/des Corpus

Données

- Littérature, ELTeC [1] Corpus
- Version ELTeC (réf.) and Version OCR
- Comment est constituée la version OCR du corpus ?
  - Récupération des images sur le site de Gallica [2]

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`
2. `https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop`

Constitution du/des Corpus

Données

- Littérature, ELTeC [1] Corpus
- Version ELTeC (réf.) and Version OCR
- Comment est constituée la version OCR du corpus ?
  - Récupération des images sur le site de Gallica [2]

Outils OCR

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

Outils de REN

- Spacy (Modèles français : petit(sm), grand(lg), moyen(md))
- Stanza (Modèle français)
- SEM (Modèle WiNER français)
- CasEN (Modèle de base)

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`
2. `https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop`

Constitution du/des Corpus

Données

- Littérature, ELTeC [1] Corpus
- Version ELTeC (réf.) and Version OCR
- Comment est constituée la version OCR du corpus ?
  - Récupération des images sur le site de Gallica [2]

Outils OCR

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

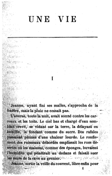
Outils de REN

- Spacy (Modèles français : petit(sm), grand(lg), moyen(md))
- Stanza (Modèle français)
- SEM (Modèle WiNER français)
- CasEN (Modèle de base)

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`
2. `https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop`

1. Problématique et enjeux
2. Constitution du/des corpus
3. Mesurer l'impact du bruit de L'OCR sur la REN ?
   - Corrélation entre la qualité de l'image et la sortie OCR ?
   - Comparaisons manuelles
   - Comparaison automatique des résultats des REN
     - Problèmes de mesure : Précision, Rappel, F-score
     - Intersections
     - Distances et similarités : Importance du choix des métriques
4. Cartographie
   - Évaluations Cartographiques
   - Projet NER&MAP
5. Et après ?

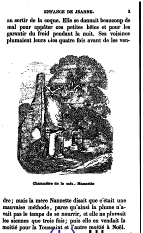
Corrélation entre la qualité de l'image et la sortie OCR ?


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

(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.

DEDIÉ A MADAME EMILE DE GIRARDIN.

(c) Texte sur deux colonnes.

Figure – Les difficultés de l'OCR [3]

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

Corrélation entre la qualité de l'image et la sortie OCR ?

Table – Transcriptions OCR d'une image bruitée, première page de *"Une vie"*, Guy de Maupassant.

Kraken-Base	**Tesseract Français**
Texte non transcrit	**Texte non transcrit**

Corrélation entre la qualité de l'image et la sortie OCR ?

Table – Transcriptions OCR d'une illustration et de sa légende, *"La petite Jeanne"*, Carraud.

Kraken	**Tess fr**
Ses voisines plumaient leurs oioes quatre fois avant de les ven- LL I II II I I IIF M ii I I II E E g[4] Chamnnlhrs de ta mn Mamn-netta[5] dre ; mais la mere Nan-nette disait que eetait une mau-vaise m6thode, paree qu'ainsi la plume ...	Ses voisines plumaient leurs vies quatre fois avant de les ven- Chaumière de la mè1. Nannette[6] dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi l2 plume ...

4. Illustration
5. Légende de l'illustration
6. Légende de l'illustration

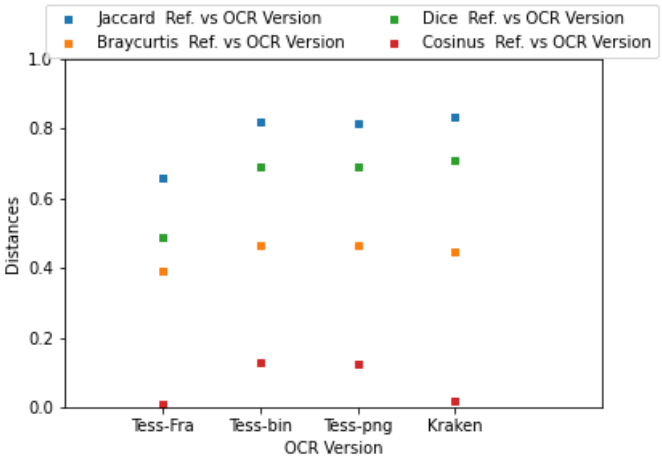
Correlation between input quality and output quality ?

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

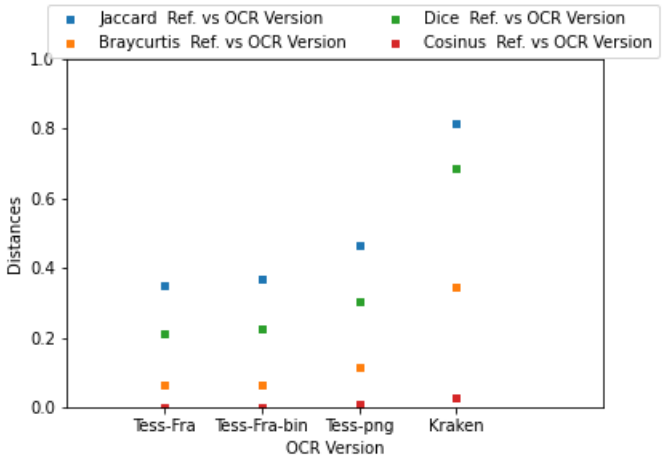
Kraken	**Tess fr**
Un des quelques salons oh se produisait l'arehe_eque de [] lomene fut l'unique fruit du mariago des Wattoville et des Besangcon sous la Reslauralion, et celui qu'il affectionnait,] de Rupt. etait celui de madame la baronne do Watltev_lle. Un mot] Monsieur de Wotteville passait sn vie 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- [7] omène fut l'unique fruit du mariage des Watteville et des de Rupt [8].

7. Colonne de droite en vert
8. Colonne de gauche en noir

OCR le plus performant ?



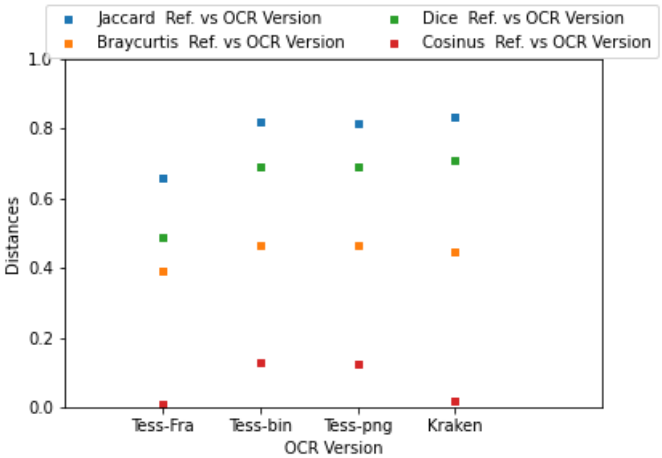
(a) *"Albert Savarus"*, Balzac.



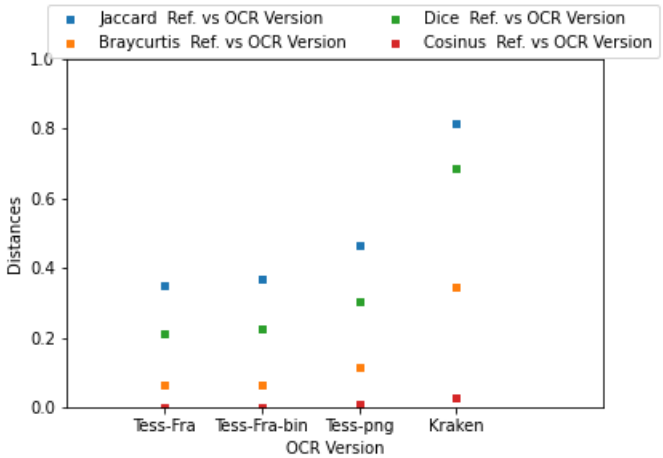
(b) *"La petite Jeanne"*, Carraud.

Figure – Distances entre les versions de Référence et les versions OCR

OCR le plus performant ?



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



(b) *"La petite Jeanne"*, Carraud.

Figure – Distances entre les versions de Référence et les versions OCR

➜ Tesseract semble meilleur que Kraken

Impact des erreurs orthographique sur la RENS

Version	contexte	spacy_sm	stanza	SEM	CasEN
Réf.[1]	*en faïence de Hollande.*	Hollande	Hollande	()*	Hollande
Kraken	*en faience _e kollande*	()	()	()	()
Tess	*en faience ___ Hollande.. 7 _*	Hollande	Hollande	()*	Hollande
Tess fr	*en faïence ___ Hollande.*	Hollande	Hollande	()*	Hollande
Réf.[2]	*prendre la diligence de Châlons*	Châlons	Châlons	Châlons	()
Kraken	*prendre la diligence de Ch_lons*	Ch_lons	()	Ch_lons	()
Tess	*prendre la diligence de Chalons*	Chalons	Chalons	Chalons	()
Tess fr	*prendre la diligence de Châlons*	Châlons	Châlons	Châlons	()
Réf.[3]	*on se voit forcé d'aller à Morlincourt.*	Morlincourt	Morlincourt	Morlincourt	()*
Kraken	*on se voit force d'aller _ Mlorlincourt.*	Mlorlincourt	Mlorlincourt	Mlorlincourt	()
Tess	*on se v01t force (1 al- '- fi.' . . :3 ler £1 Morlincourt.*	()	()	Morlincourt	()*
Tess fr	*"% ler à Morlincourt.*	Morlincourt	Morlincourt	Morlincourt	()*

Table – Substitution ou absence d'un caractère,

[1] *"Une vie"*, Maupassant.

[2] *"L'éducation sentimentale"*, Flaubert.

[3] *"Mon village"*, Adam.

Impact des erreurs orthographique sur la NER

Version	NER model	Entitiy	# Miss.
Ref.	spacy_lg	Morlincourt : 18	N/A
	Stanza	Morlincourt : 16	N/A
	SEM	Morlincourt : 18	N/A
	CasEN	Morlincourt : 0*	N/A
Kraken	spacy_lg	Morlincourt : 8 Mlorlincourt : 1 Mlorlincourtl : 1	8
	Stanza	Morlincourt : 6 Mlorlincourt : 3 Mlorlincourtl : 1	6
	SEM	Morlincourt : 8 Mlorlincourt : 3	7
	CasEN	Mlorlincourt : 1	N/A
Tess	spacy_lg	Morlincourt : 11 Morlinco‘urt : 1	6
	Stanza	Morlincourt : 9 Morlinco‘urt : 1	6
	SEM	Morlincourt : 11	7
	CasEN	Morlincourt : 0*	N/A
Tess fr	spacy_lg	Morlincourt : 9 Morlin— : 1	8
	Stanza	Morlincourt : 7 Morlinco’urt : 1	8
	SEM	Morlincourt : 11 Morlinco‘urt : 1	6
	CasEN	Morlincourt : 0*	N/A

Table – Variations orthographiques de l’EN ”Morlincourt”

1. Problématique et enjeux
2. Constitution du/des corpus
3. Mesurer l'impact du bruit de L'OCR sur la REN ?
   - Corrélation entre la qualité de l'image et la sortie OCR ?
   - Comparaisons manuelles
   - Comparaison automatique des résultats des REN
     - Problèmes de mesure : Précision, Rappel, F-score
     - Intersections
     - Distances et similarités : Importance du choix des métriques
4. Cartographie
   - Évaluations Cartographiques
   - Projet NER&MAP
5. Et après ?

Précision, Rappel, F-score : problèmes d'alignement

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.

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]

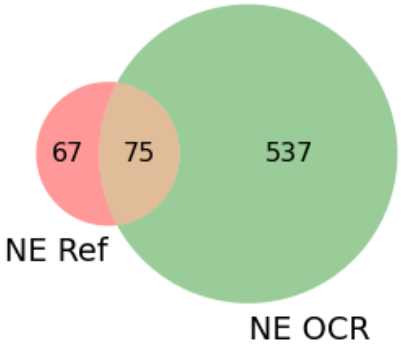
[1] *"Le chateau de Pinon"*, Dash , *"Les trappeurs de l'Arkansas"*, Aimard. [2] *"La nouvelle espérance"*, Noailles

Comparaisons Manuelles Vrai Positif/Faux Positif

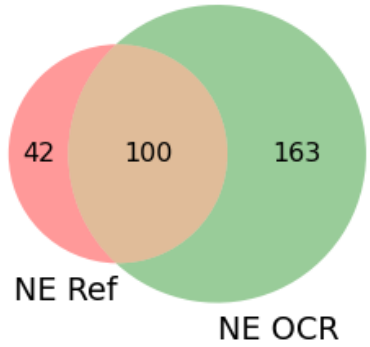
	REF			Kraken		
	SEM	stanza	spacy_lg	SEM	stanza	spacy_lg
TOTAL EN	593	320	273	939	421	946
VP Total	173	179	174	120	137	131
FP Total	415	122	80	811	271	785
AMB Total	5	19	19	8	13	30
Fréq. Total	128	140	142	191	245	612
Répét.	55	40	39	76	48	67
Hapax	73	100	103	115	197	545
VP	67	79	79	49	71	75
FP	58	50	49	135	162	512

Table – *"Une vie"*, Maupassant, 1883.

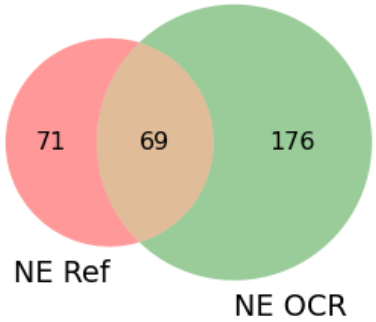
Corrélation entre la qualité de l'OCR et la REN ?



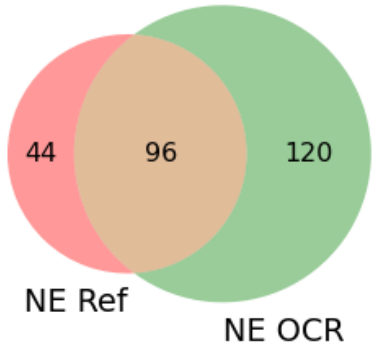
(a) `spacy_lg` - Kraken



(b) `spacy_lg` - Tess fr



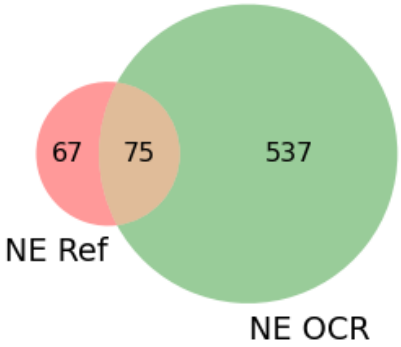
(c) `stanza` - Kraken



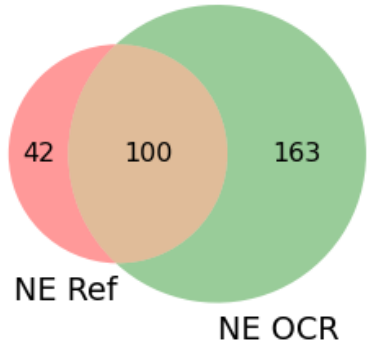
(d) `stanza` - Tess fr

9. *"Une vie"*, Maupassant

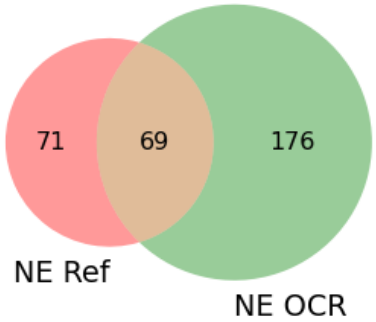
Corrélation entre la qualité de l'OCR et la REN ?



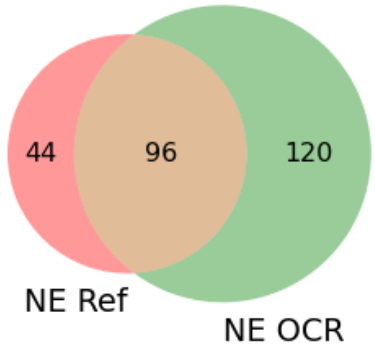
(a) `spacy_lg` - Kraken



(b) `spacy_lg` - Tess fr



(c) `stanza` - Kraken

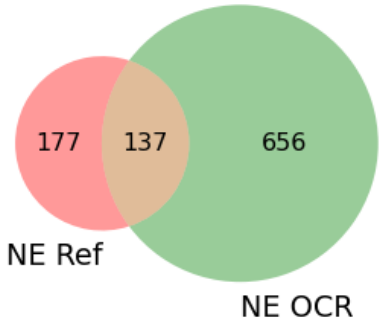


(d) `stanza` - Tess fr

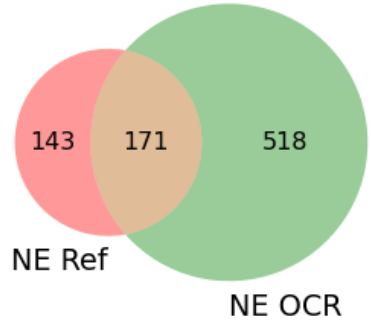
➜ **OCR de meilleure qualité** $\rightarrow$ **Moins de Faux Négatifs ?** [9]

9. *"Une vie"*, Maupassant

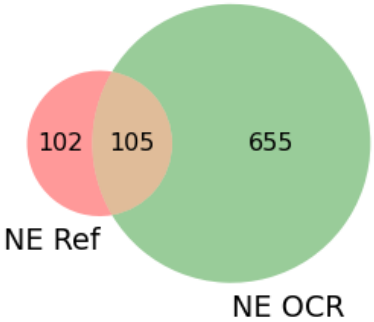
Impact du modèle de NER selon les versions OCR



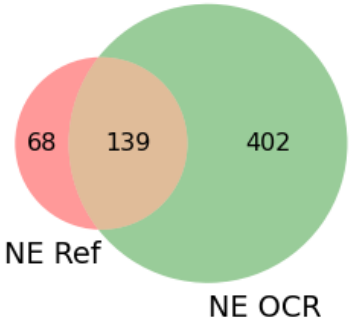
(a) `spacy_sm` - Kraken



(b) `spacy_sm` - Tess fr



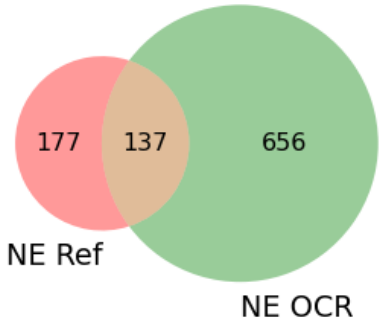
(c) `spacy_lg` - Kraken



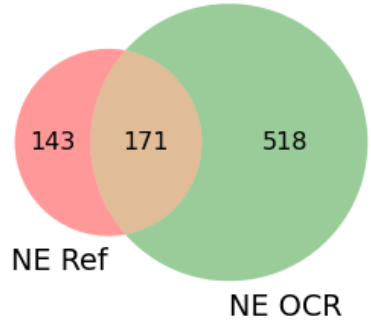
(d) `spacy_lg` - Tess fr

10. *"Albert Savarus"*, Balzac

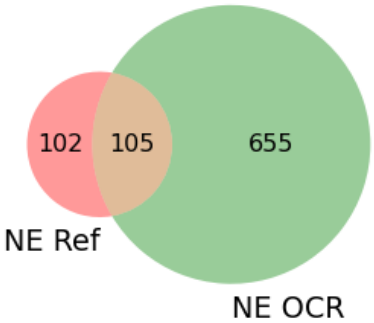
Impact du modèle de NER selon les versions OCR



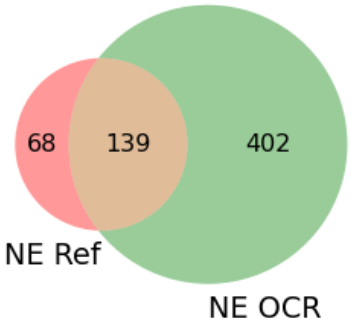
(a) `spacy_sm` - Kraken



(b) `spacy_sm` - Tess fr



(c) `spacy_lg` - Kraken

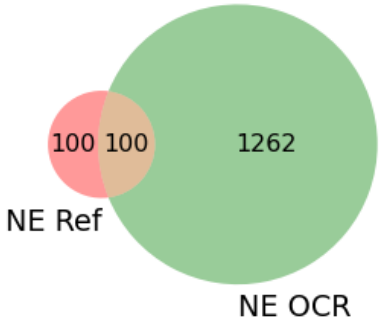


(d) `spacy_lg` - Tess fr

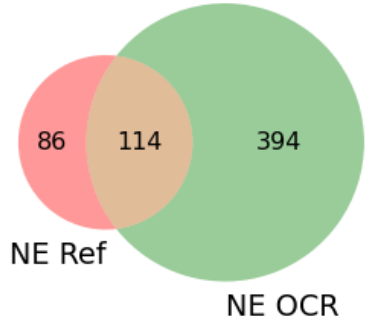
➜ **NER, + de données d'entraînement → Moins de FN ?** [10]

10. *"Albert Savarus"*, Balzac

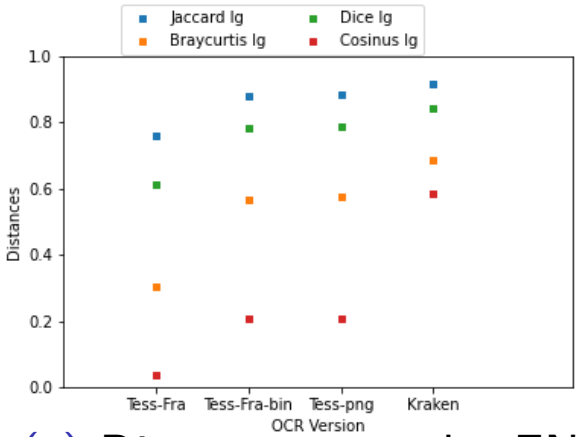
Corrélation entre la qualité de l'OCR et la REN ?



(a) `spacy_lg` - Kraken



(b) `spacy_lg` - Tess fr

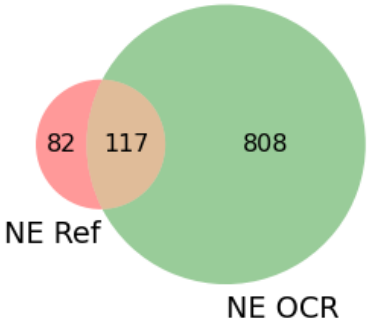


(c) Distances entre les ENS Réf. et les ENS OCR - `spacy_lg`

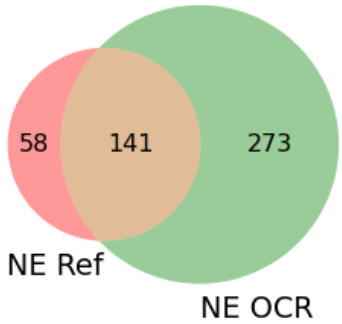
11

11. *"La Belle rivière"*, Aimard

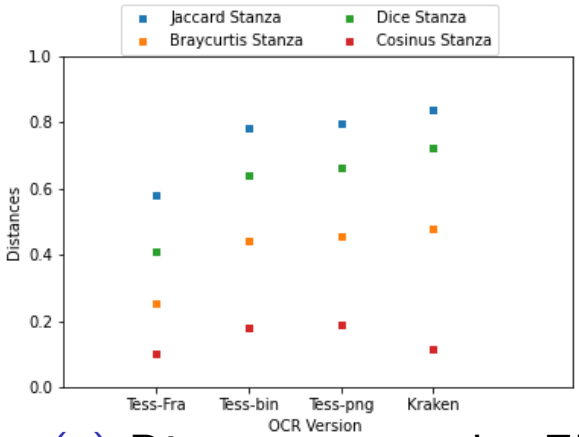
Corrélation entre la qualité de l'OCR et la REN ?



(a) Stanza - Kraken



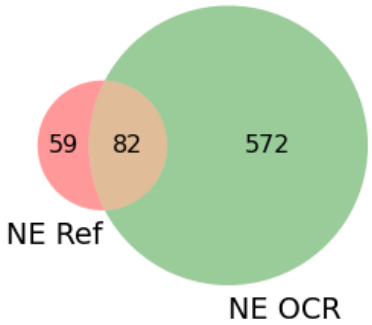
(b) Stanza - Tess fr



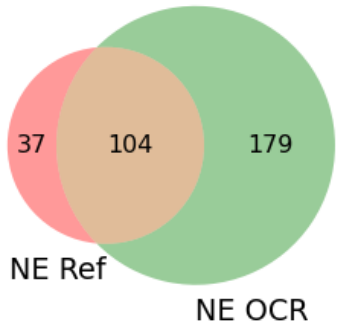
(c) Distances entre les ENS Réf. et les ENS OCR - Stanza [12]

12. *"La Belle rivière"*, Aimard

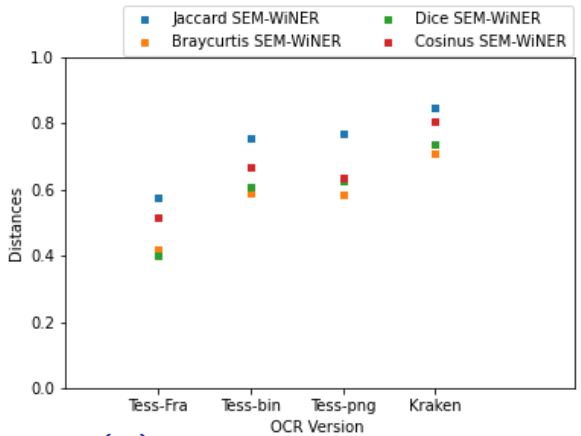
Corrélation entre la qualité de l'OCR et la REN ?



(a) SEM - Kraken



(b) SEM - Tess fr

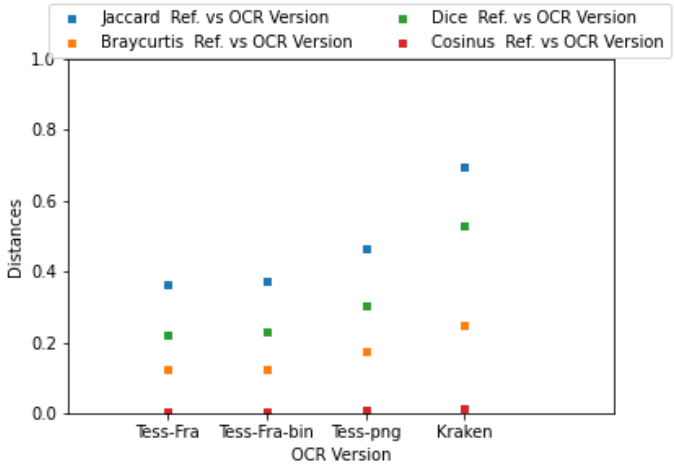


(c) Distances entre les ENS Réf. et les ENS OCR - SEM

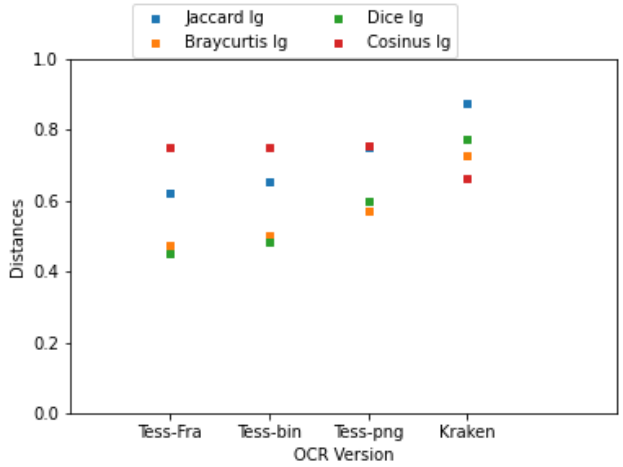
➜ **Comparaison des métriques : à propos du bruit et du silence ?** [13]

13. *"La Belle rivière"*, Aimard

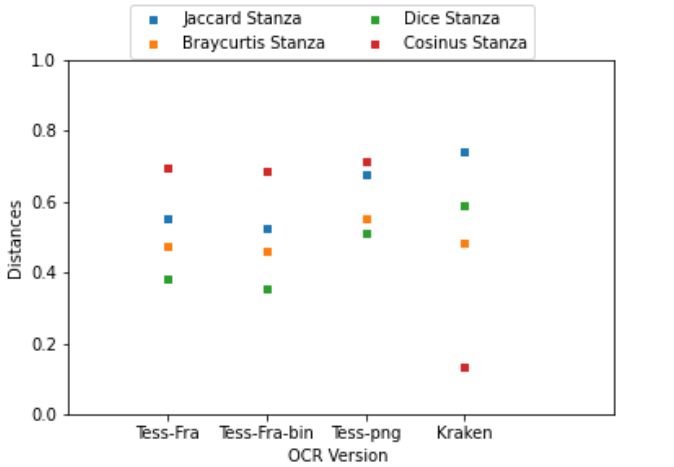
Évaluations des métriques pour les sorties de REN



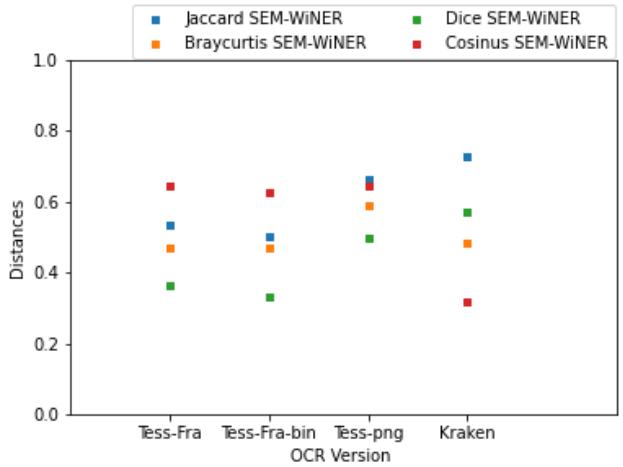
(a) `Texte Réf. vs. OCR`



(b) `spacy_lg`



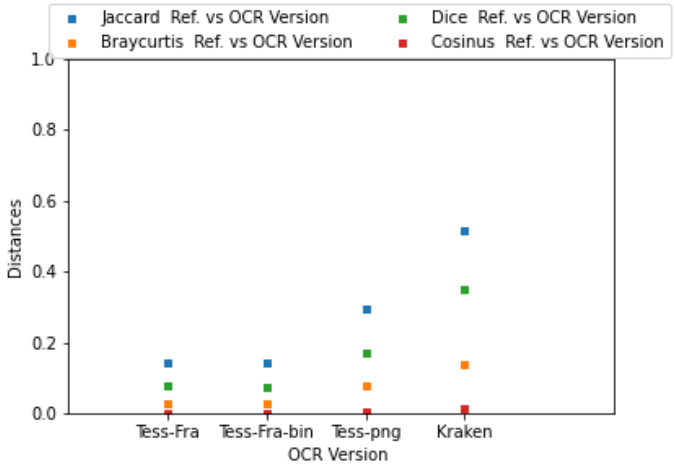
(c) `stanza`



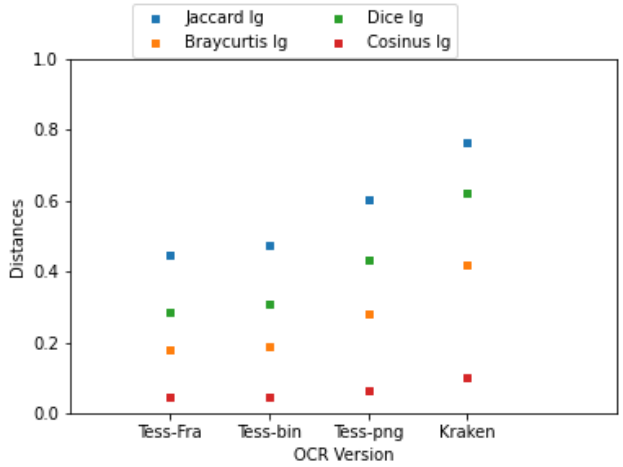
(d) `SEM`

"Une vie", Maupassant, 1883.

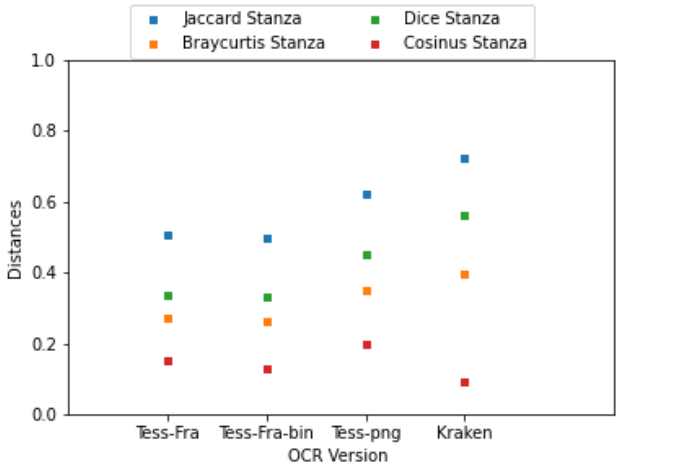
Évaluations des métriques pour les sorties de REN



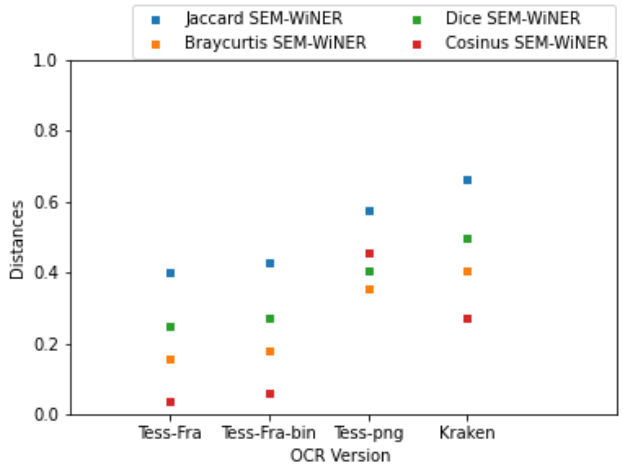
(a) `Texte Réf. vs. OCR`



(b) `spacy_lg`



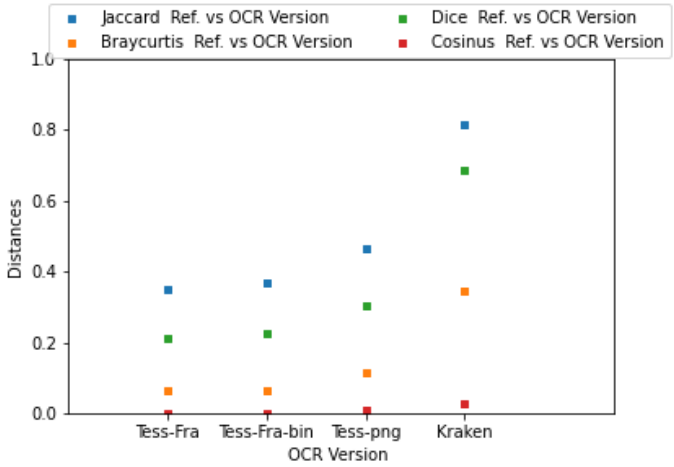
(c) `stanza`



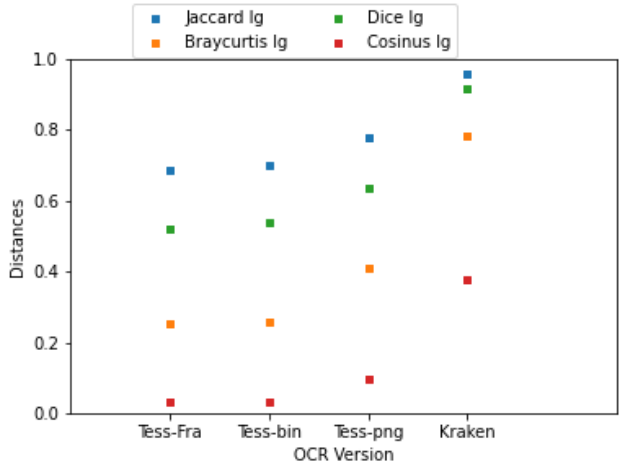
(d) `SEM`

"Le petit chose", Daudet,1868.

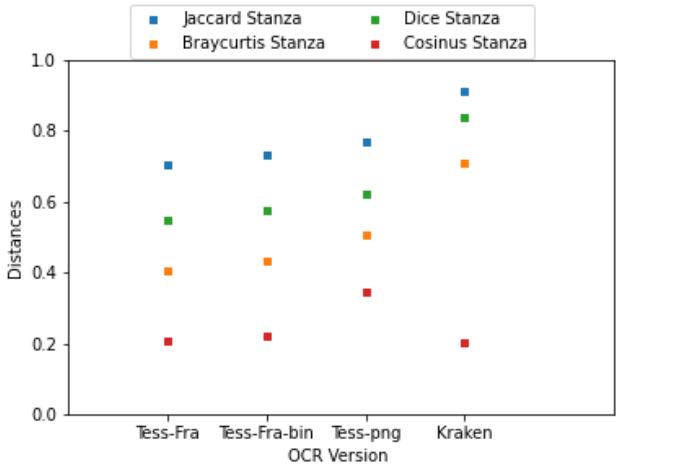
Évaluations des métriques pour les sorties de REN



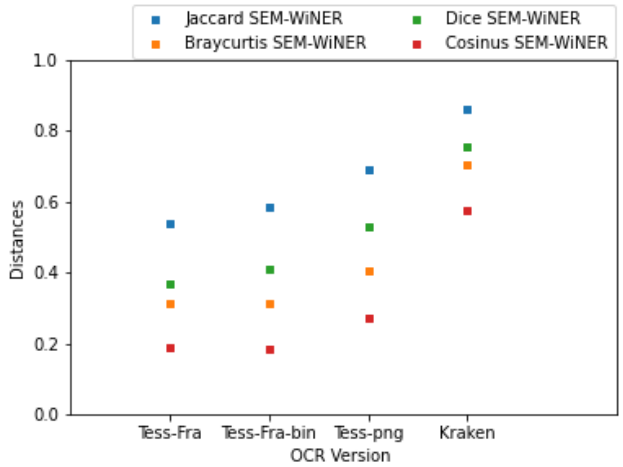
(a) `Texte Réf. vs. OCR`



(b) `spacy_lg`



(c) `stanza`



(d) `SEM`

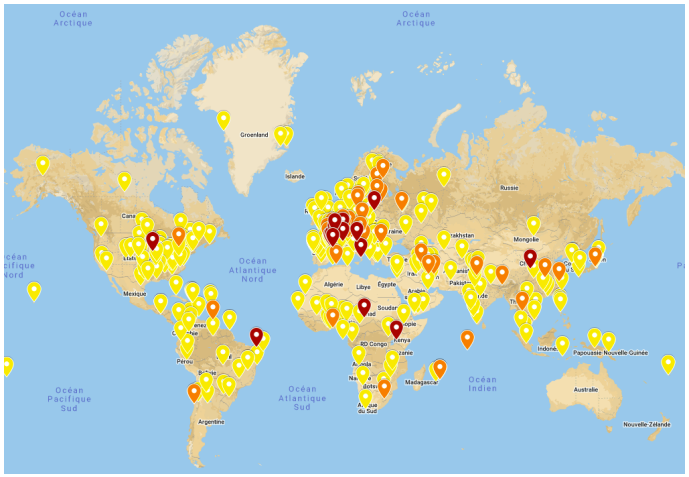
"La petite Jeanne", Carraud, 1853.

1. Problématique et enjeux
2. Constitution du/des corpus
3. Mesurer l'impact du bruit de L'OCR sur la REN ?
   - Corrélation entre la qualité de l'image et la sortie OCR ?
   - Comparaisons manuelles
   - Comparaison automatique des résultats des REN
     - Problèmes de mesure : Précision, Rappel, F-score
     - Intersections
     - Distances et similarités : Importance du choix des métriques
4. Cartographie
   - Évaluations Cartographiques
   - Projet NER&MAP
5. Et après ?

Cartographier les résultats OCR basse qualité

(a) `Réf. automatique`

(b) `Réf. corrections sorties REN`



(c) `Kraken automatique`

(d) `Kraken corrections sorties REN`

"La petite Jeanne", Carraud, 1853.

Cartographier les résultats OCR bonne qualité

(a) `Réf. automatique`

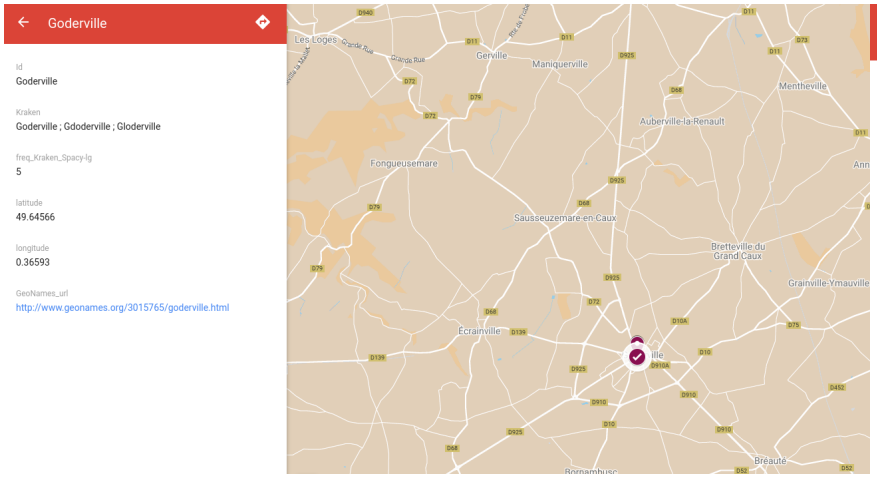
(b) `Réf. corrections sorties REN`

(c) `Kraken automatique`

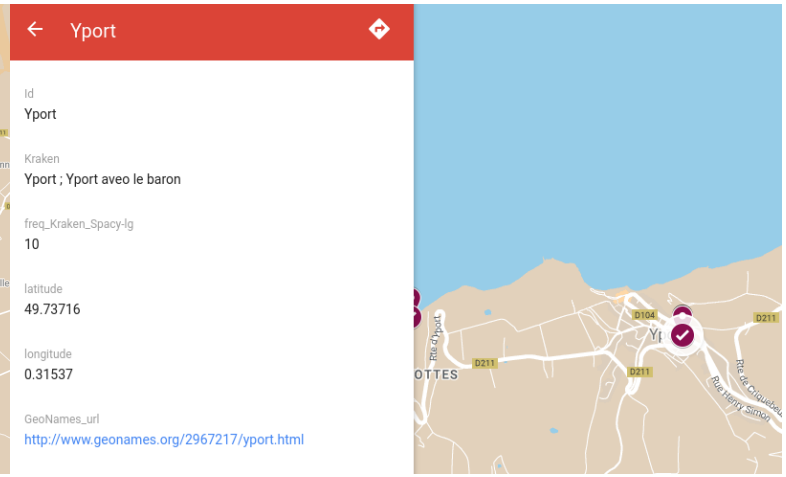
(d) `Kraken corrections sorties REN`

"Le petit chose", Daudet,1868.

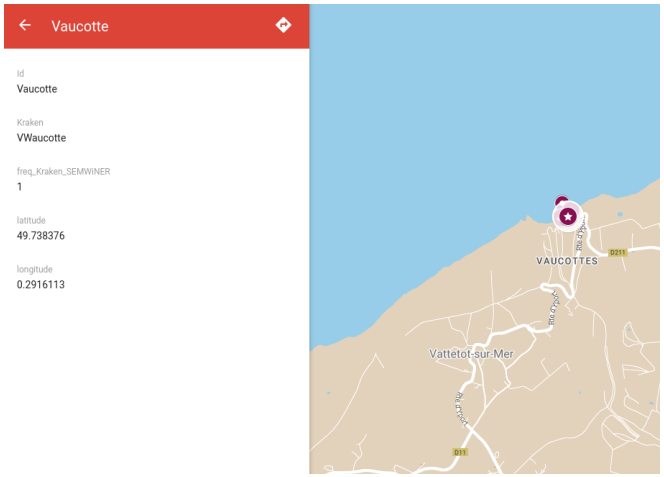
Cartographier des données imparfaites



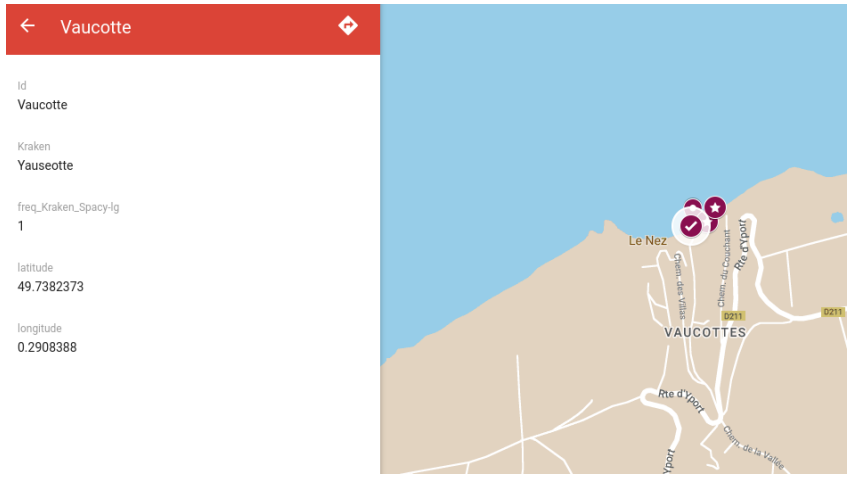
(a) Variations du terme `Goderville`



(b) Variations du terme `Yport`



(c) Variations du terme `Vaucotte`



(d) Variations du terme `Vaucotte`

"Une vie", Maupassant, 1883.

Évaluer visuellement les résultats

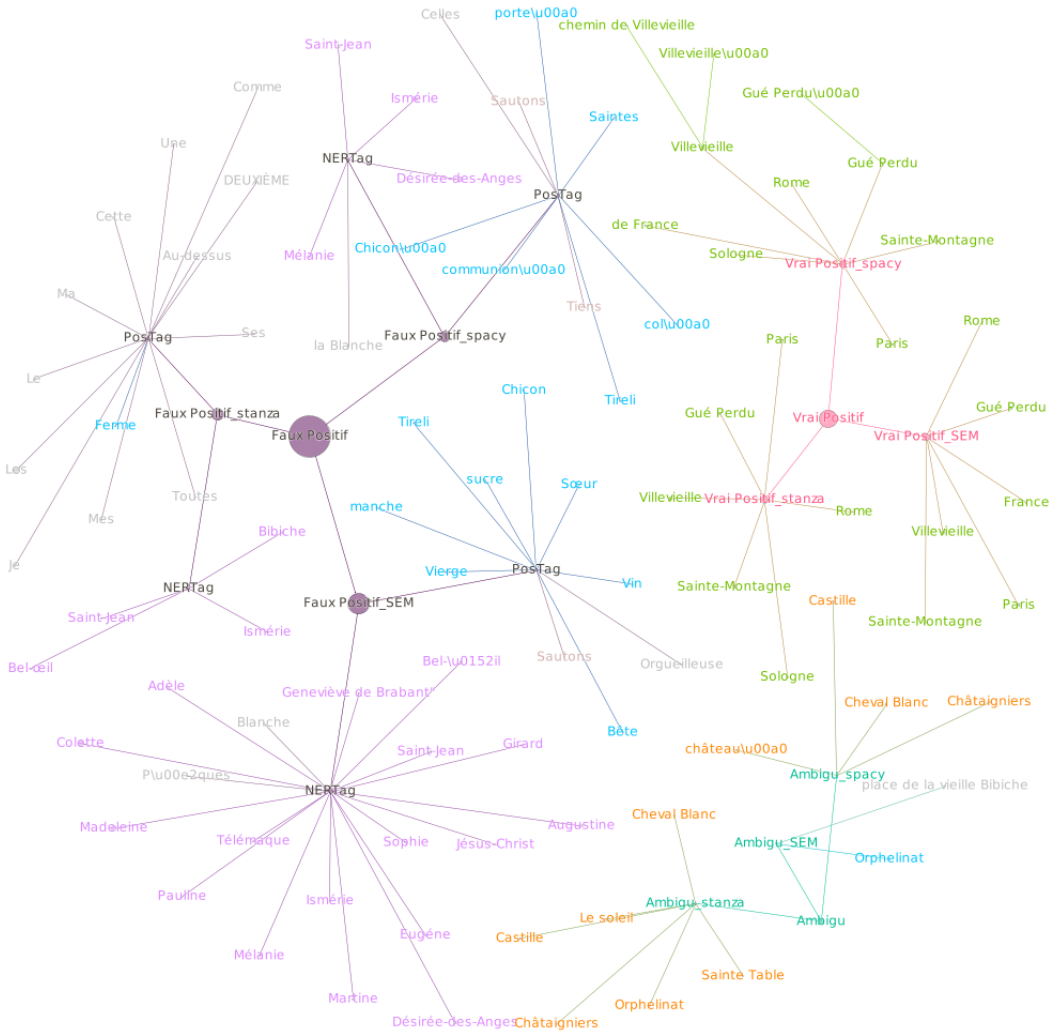
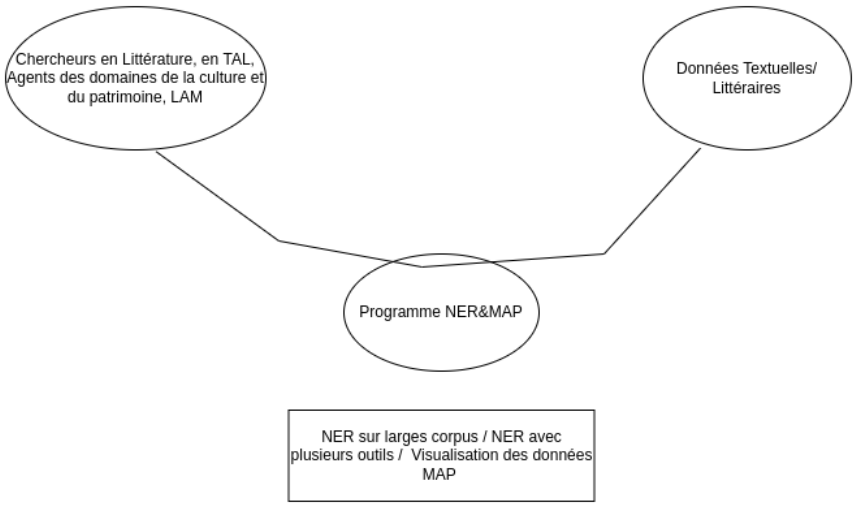
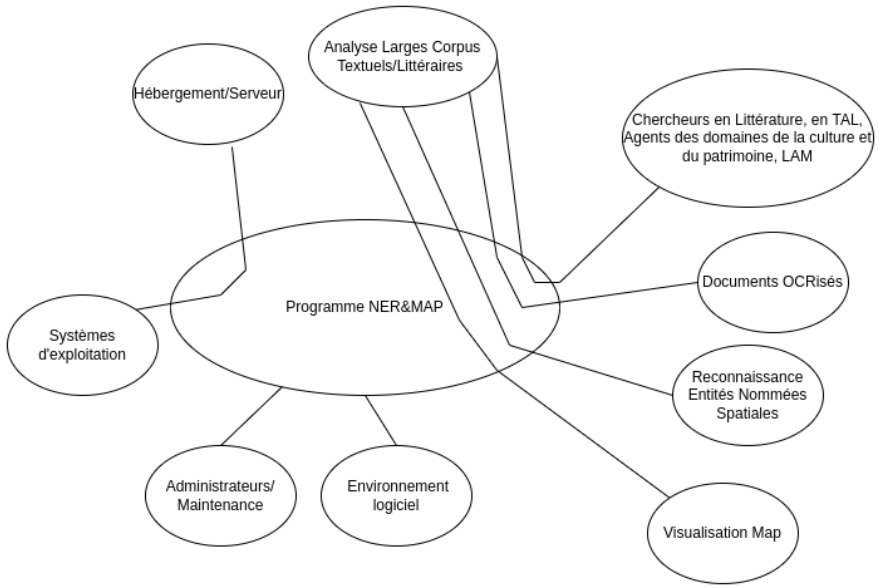


Figure – Visualisation des Vrais Positifs et Faux Positifs Spacy, Stanza et SEM

Projet NER&MAP



(a) Bête à corne - Spécifications



(b) Pieuvre - Spécifications

Cas Usage NER&MAP

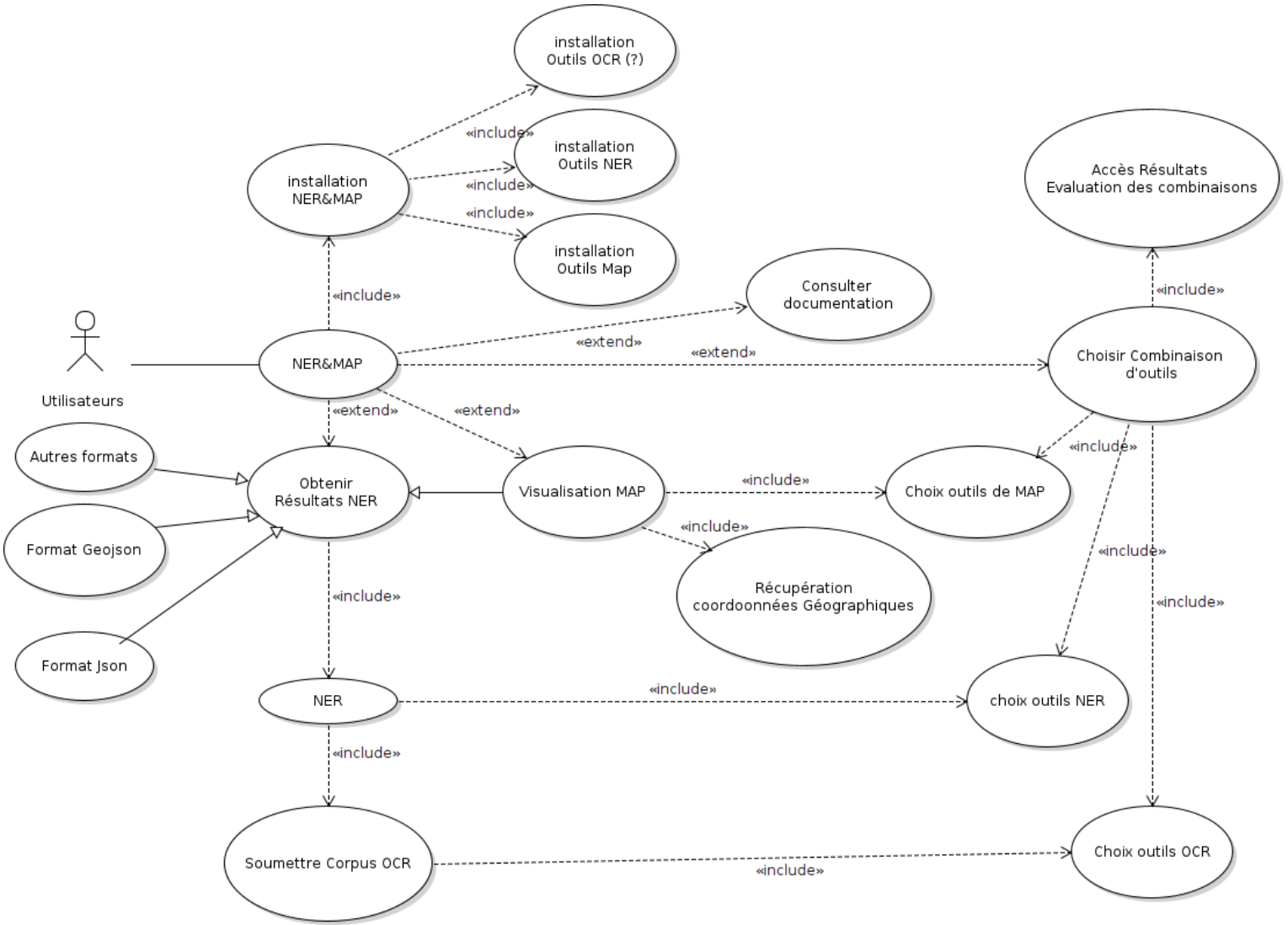
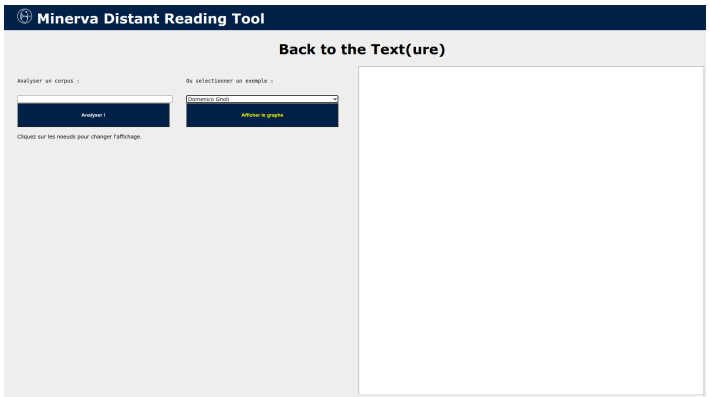
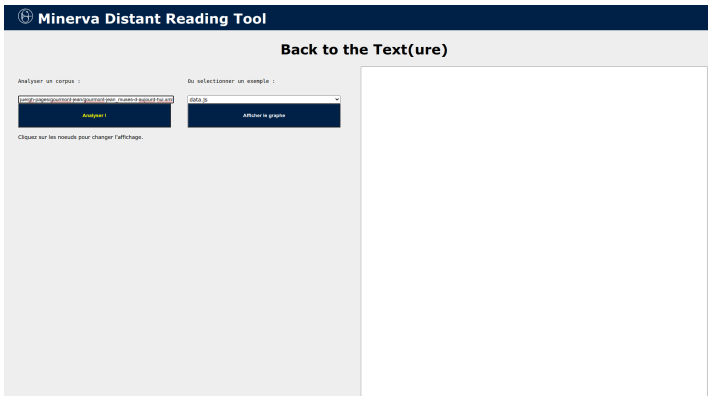


Figure – Cas d'usage

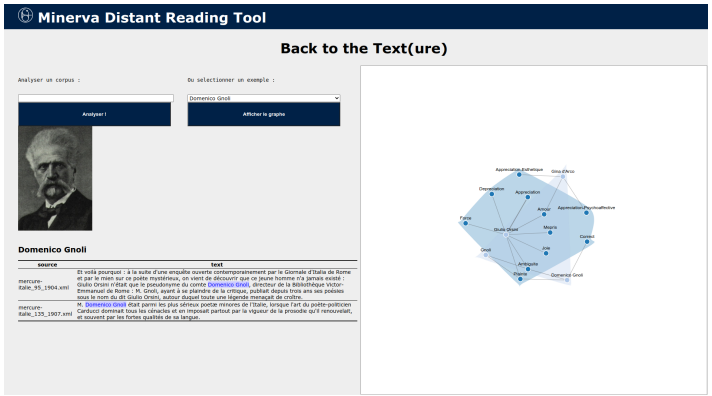
Modèle Application



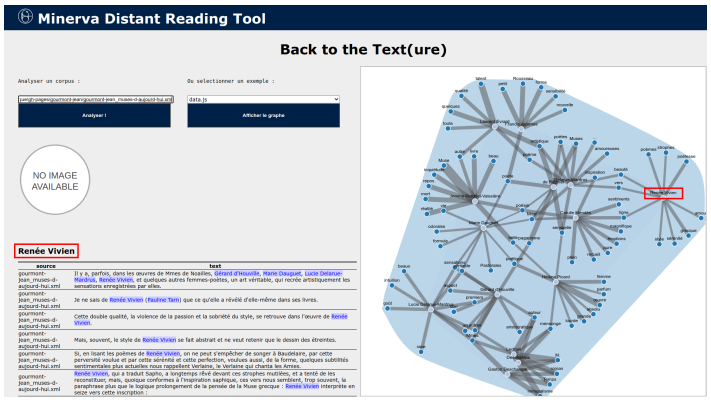
(a) Utiliser l'url d'un fichier xml



(b) Utiliser les données pré-enregistrées



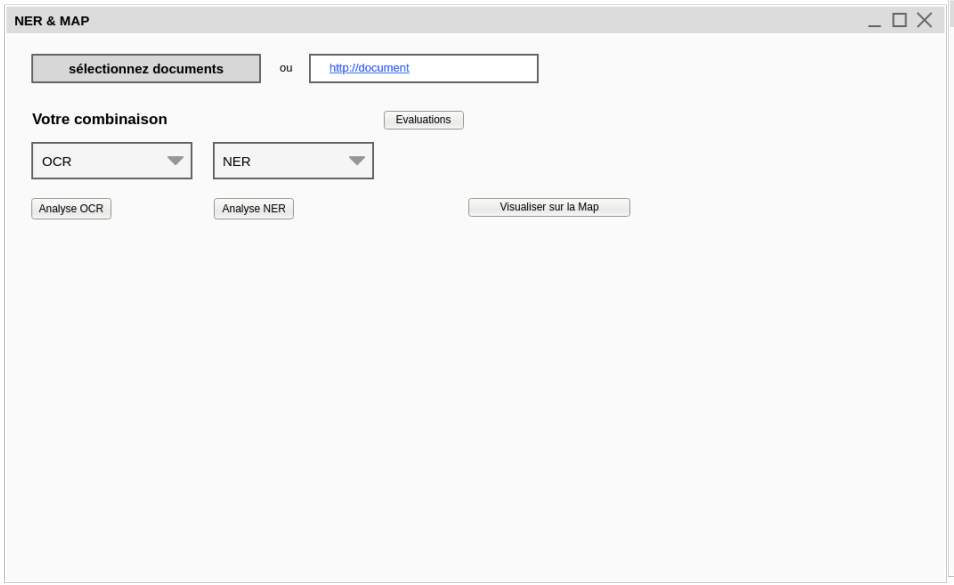
(a) Affichage Graph et métadonnées



(b) Cliquer sur un noeud pour afficher les métadonnées

Figure – Minerva, Y.Dupont

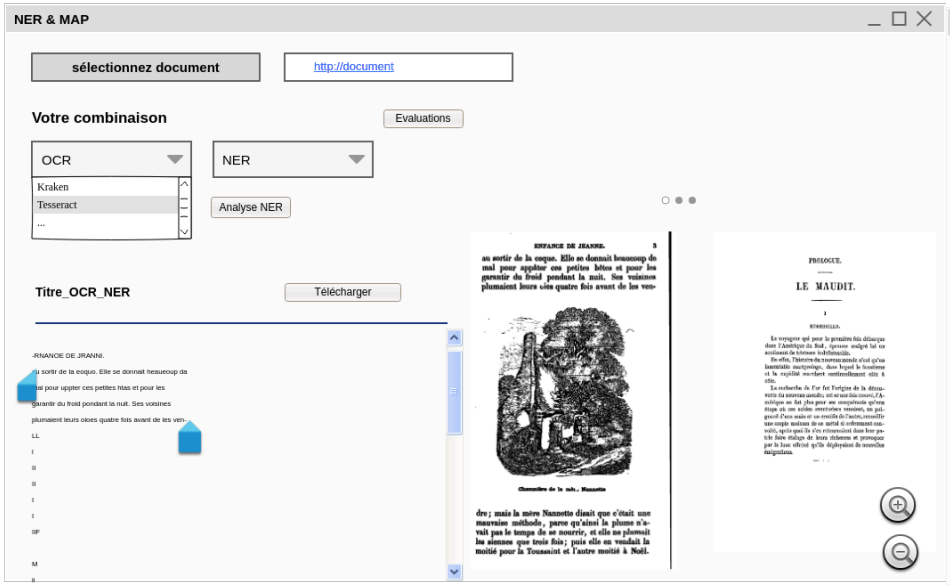
Wireframe NER&MAP



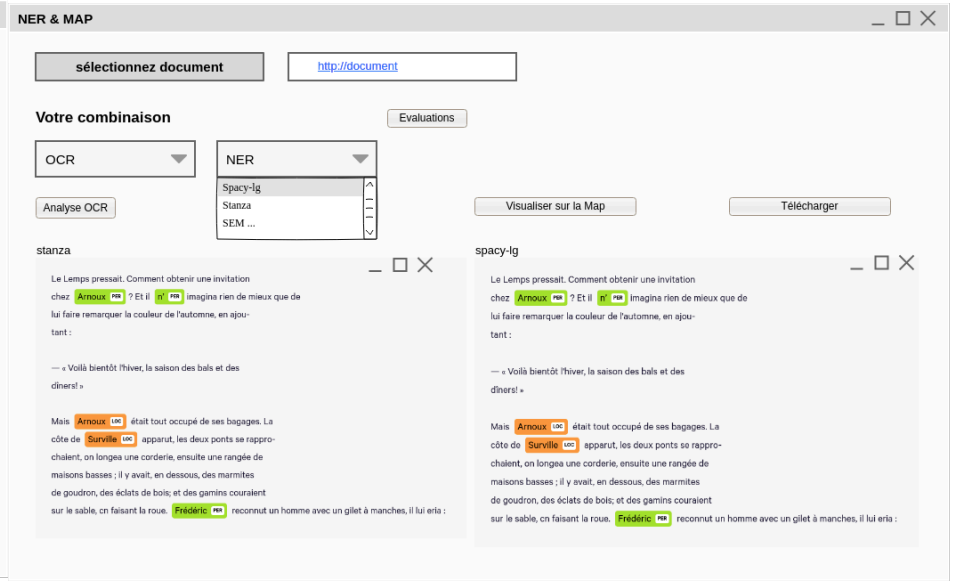
(a) Accueil

(b) Exemples évaluations

Wireframe NER&MAP



(a) Paramètre OCR



(b) Paramètre NER

Wireframe NER&MAP

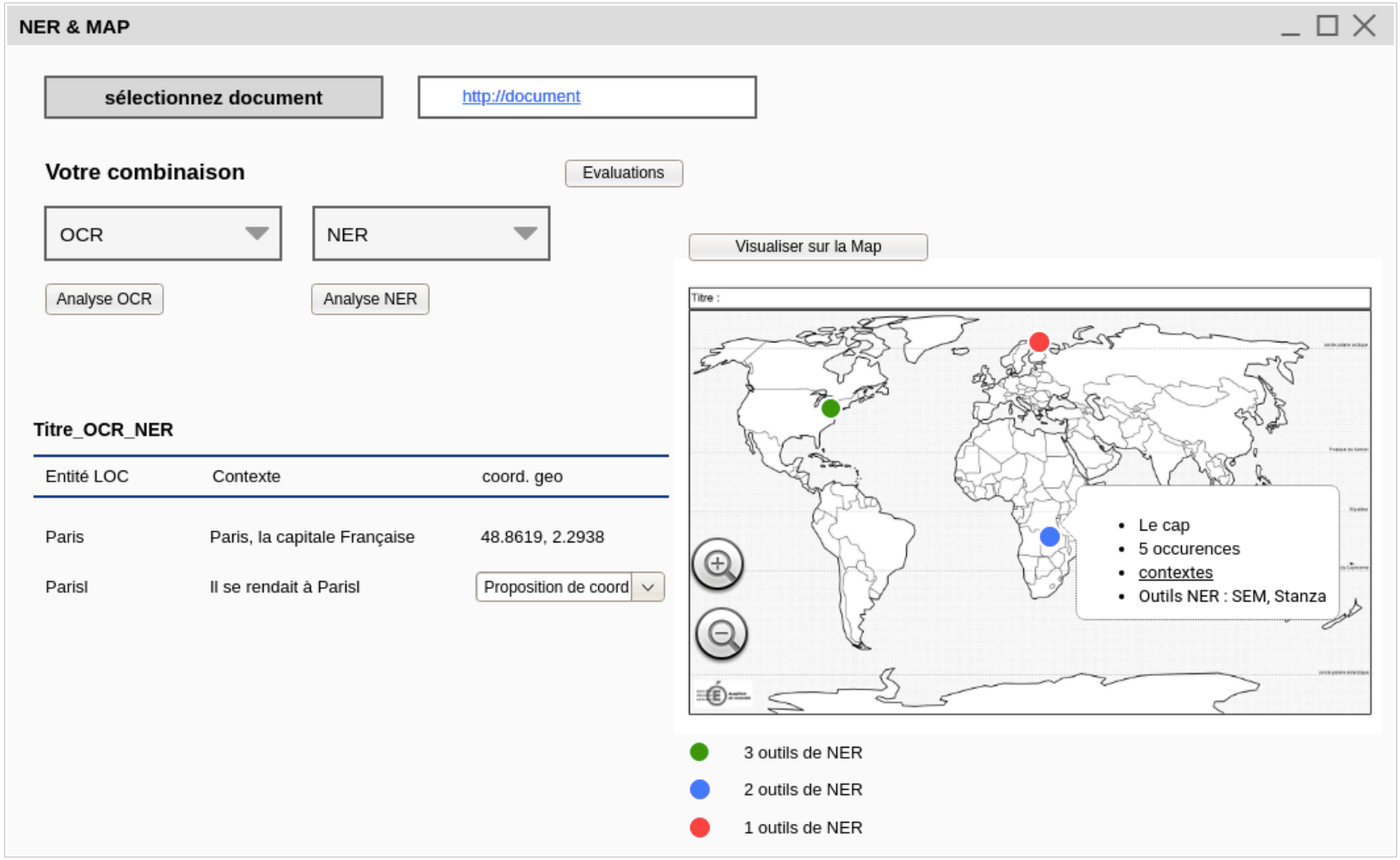


Figure – Visualisation Cartes

Outils de Cartographie

	GGMyMAP	**Leaflet**	**Qgis**
Avantages	Facile d'utilisation, customisation, intégration web facile	Customisation, nombreux tuto ; plusieurs formats d'entrée (Geojson, csv) ; intégration web	Plusieurs couches de cartes
Inconvénients	Les données sont sur Google	R et Rstudio windows	Coût d'entrée élevé

Table – Premiers retours sur différents outils de Visualisation

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 ?
  ➜ comment les cartographier ?
- Le choix des métriques n'est pas trivial
  ➜ Distance Cosinus : sous-évalue les différences(?)
  ➜ Distance de Jaccard : sur-évalue les différences(?)
  ➜ Proximité distances Cosinus et Jaccard : Problème ?
- La représentation cartographique permet de visualiser :
  ➜ EN récupérée par +ieurs outils est potentiellement un VP ?
  ➜ Une géographie littéraire du 19ème siècle
- Les humains peuvent gérer le bruit, mais que faire s'il y a du silence ?

Perspectives

- Résolution des problèmes d'entity linking (diachronie, formes fautives des mots)
- Améliorer la Géolocalisation automatique
- Annotation et entrainement d'un modèle sur des données bruitées
  ➜ Pour quel résultat ? Meilleur, moins bon, équivalent ?