Explorer la collection Française d'ELTeC avec l'outil de cartographie textuelle Épiméthée

Caroline Koudoro-Parfait
soutenance : 6 janvier 2025 (4 ans)
caroline.parfait@sorbonne-universite.fr
https://hal.science/tel-05042915

SAGEO2025, Ateliers MAGIS, Humanités numériques spatialisées, 21 mai 2025

Observatoire des Textes des Idées et des Corpus - Obtic,
Sorbonne Center for Artificial Intelligence - SCAI,
Sens Textes Informatique Histoire - STIH EA 4509, Sorbonne Université

Humanités Numériques : espace de ma recherche

Le numérique au service d'utilisateur·ice·s non expert·e·s

Doctorat, Littératures françaises et comparées, HN, 2020 - 2024

Combiner pratiques plastiques et numériques

Diplôme national supérieure d'expression Plastique, ENSAPC, 2017

À l'interface : utilisatrice hier, conceptrice demain

Master IDCP (HN), CESR, 2020

Nouer philosophie de l'art et des sciences

Maîtrise de Philosophie, Sorbonne Université, 2011.

Plan de la présentation

Enjeux, usages et verrous à surmonter

Impact des contaminations OCR sur la REN

Stratégies pour aider les utilisateur·ice·s

Explorer l’espace littéraire européen

Et après ?

Enjeux, usages et verrous à surmonter

Appliquer des outils de TAL sur des données réelles

Utilisateurs et **interdisciplinarité** :

- Sciences Humaines et Sociales (SHS) et Lettres ;
- Traitement automatique des langues (TAL) ;

➥ **Acquisition** des **corpus** par OCR : **Variabilités** ;

➥ **Variabilité** dans le contexte d'usage

→ Les interférences OCR un défi à relever ;

→ Des **enjeux** qui concernent de nombreux·ses chercheur·euse·s dans **différents domaines** ;

→ Séminaire **NER for OCR'ed historical documents** [1]

1. `https://ner-for-historical-docs.github.io/`

Angles Expérimentaux

→ **Impact de la qualité de la transcription OCR sur la REN** ;

- Comment **évaluer la qualité des résultats de REN** sur des corpus contaminés par les interférences OCR ?
- Quelles stratégies pour dépasser les contaminations OCR ?

Des corpus pour une évaluation multilingue

✱ European Literary Text Collection (ELTeC)[2] [Schöch et al., 2021] :

- 22 langues, ≈ 100 romans par langue ;
- période : 1840 à 1920 ;

✂ Corpus constitués

Corpus	Books	Pages	Words	# Named Entities : LOC		# Évaluations		
				SPACY_LG	FLAIR	qti. Silver	qli Gold – Annot.	Cluster
small-ELTeC-fra	11	3 195	829 604	5 765	4 814	✓	**51 000 tok.**	✓
small-ELTeC-eng	9	5 281	2 063 246	5 551	8 867	✓	✓	✓
small-ELTeC-por	4	1 795	421 915	7 590	N/A	✓	✓	✗
Total	24	10 271	**3 314 765**	13 906	13 681	-	-	-

Table 1 – Statistiques sur les corpus dans la version de référence

2. https://www.distant-reading.net/eltec/

Privilégier le *Silver* standard

➜ *Silver* standard :

→ annotation automatique sur textes de réf. ;

→ évaluation à grande échelle ;

➜ Annotation pour *Gold* standard est coûteuse ;

corpus	*Silver Small*-ELTeC	*Gold* Cle HIPE2020	*Gold* TENs [3]
nb. tokens	**3 314 765**	444 596	*51 000*

3. TAL-ENS et TAL-ENS2, 5000 - 6 000 tokens, trois textes dans trois versions OCR de *small*-ELTeC-fr

Systèmes OCR utilisés

✏ 2 campagnes OCR 2021 et 2024 :

Année	Paramètres	Kraken	Tesseract
2021	Version	3.0	4.1.1 (PyTess. 0.3.6)
	Modèle	Modèle de base	`fr, en, pt`
2024	Version	4.3.13	5 (PyTess. 0.3.10)
	Modèle	Modèle de base, Lectaurep [4]	`fr`

4. `https://lectaurep.hypotheses.org/`

Système de REN évalués

✏ Modèles de REN par langues :

	fr	en	pt
spaCy	✓	✓	✓
Bert	✓	✓	✓
Stanza	✓	✓	✗
Flair	✓	✓	✓
SEM	✓	✗	✗
CasEN	✓	✗	✗

Méthodes explorées et leurs performances

Rechercher des solutions appropriées pour rendre la REN utilisable sur des données bruitées :

- **Désambiguïsation**, en utilisant des **métriques de similarité** et des ***clusters*** ✓ [Koudoro-Parfait et al., 2022]
- **Combinaison**, plusieurs modèles de REN pour filtrer les résultats ✓ [Petkovic et al., 2025]
- **Cartes** Une vue d'ensemble des entités ✓ [Koudoro-Parfait and Lejeune, 2024]
- **Correction automatique**, problèmes de sur-correction ✗ [Koudoro-Parfait et al., 2024]
- **Linking**, Lier des entités contaminées à leur forme standard✗

Performances : ✓ très convaincantes ; ✓ convaincante ; ✓ peu convaincantes.

Épiméthée, agir avant de réfléchir !

- ☞ OCR → REN – sans correction de la sortie OCR ;
- ☞ Intégration stratégies de filtrages auto. des FP de REN :
  - Combiner des systèmes de REN ;
  - Clustering ;
  - Géolocalisation ;

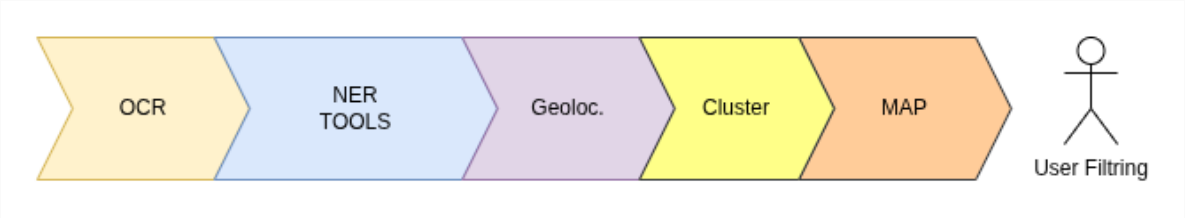


Figure 2 – Chaîne de traitement Épiméthée,
`https://github.com/These-SCAI2023/EPIMETHEE`

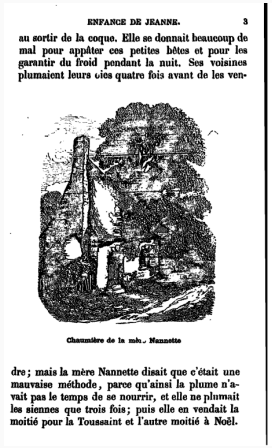
- ☞ Interface : filtre manuel par l'utilisateur·ice ;
- ☞ Récupération (CSV), réemploi et analyse littéraire.

Impact des contaminations OCR sur la REN

✏ Contexte des évaluations

- Corpus multilingue ;
- Multiples versions OCR ;
- Multiples configurations → version OCR + version REN ;
- *Silver* standard → grande échelle ;

Difficultés rencontrées par les modèles d'OCR



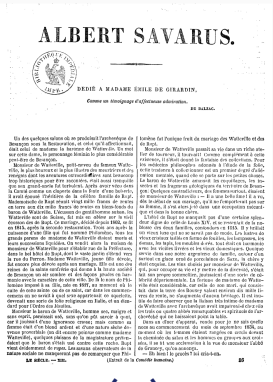
ENFANCE DE JEANNE. 3

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.

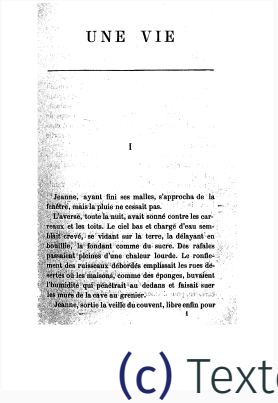
(a) Illustration + légende



ALBERT SAVARUS.

DÉDIÉ A MADAME ÉMILE DE GIRARDIN.

(b) Texte en colonnes



UNE VIE

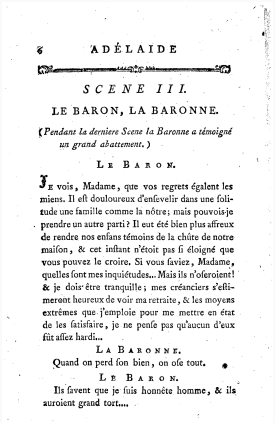
I

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 roulement des ruisseaux débordés emplissait les rues désertes où les maisons, comme des éponges, buvaient l'humidité qui pénétrait au dedans et suintait sur les murs de la cave au grenier.

Jeanne, sortie la veille du couvent, libre enfin pour

(c) Texte en filigrane



ADÉLAIDE

SCENE III.

LE BARON, LA BARONNE.

(*Pendant la dernière 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....

(d) Décoration, Capitalisation

Figure 3 – a) G. de Maupassant, *Une vie*, 1883. b) Inconnu, *Adélaïde de Mariendal, drame en cinq actes*, 1783. c) Z. Carraud, *La petite Jeanne*, 1884. (d) H. de Balzac, *Albert Savarus*, 1853.

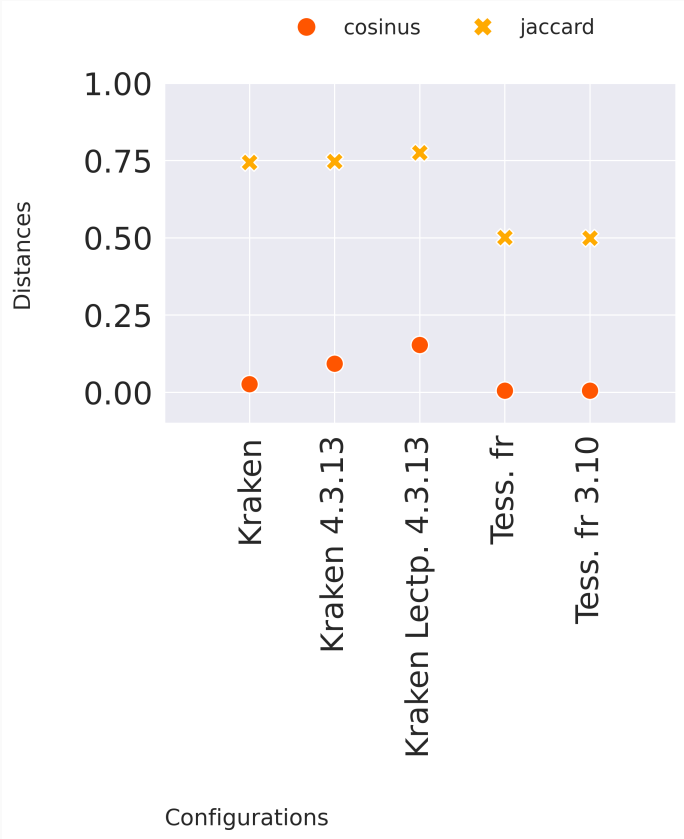
Influence de la qualité de l'image sur la sortie OCR ?

Kraken 3.0	Tess. fr 0.3.6
Ses voisines plumaient leurs	Ses voisines plumaient leurs
oioes quatre fois avant de les ven-	vies quatre fois avant de les ven-
LL l II II I I IIF M ii I I II E E g	
Chamnnlhrs de ta mn Mamnnetta	Chaumière de la mè1. Nannette
dre ; mais la mere Nannette	dre ; mais la mère Nannette
disait que eetait une	disait que c'était une
mauvaise m6thode,	mauvaise méthode,
paree qu'ainsi la plume …	parce qu'ainsi l2 plume …

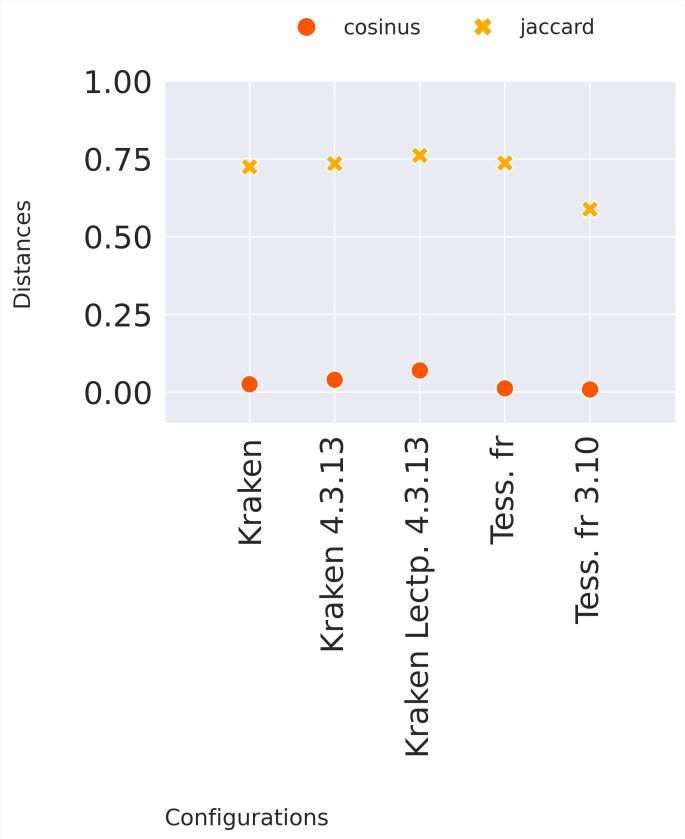
Table 2 – Transcriptions* OCR d'une illustration et de sa légende. 🟠 = illustration, 🔵 = légende, 🔴 = contaminations orthographiques.

* Z. Carraud, *La petite Jeanne*, 1884.

Importance du choix des métriques



(a) Illustration + légende.
CER Kraken 3.0 = 0.13,
Tess. fr 0.3.6 = 0.03,
Tess. fr 0.3.10 = 0.04



(b) Texte en colonnes.
CER Kraken 3.0 = 1.34,
Tess. fr 0.3.6 = 1.11,
Tess. fr 0.3.10 = 1.12

Figure 4 – Évaluation de la qualité des transcriptions OCR.

(a) Z. Carraud, *La petite Jeanne*, 1884. (b) H. de Balzac, *Albert Savarus*, 1853.

Les hapax : indice de la contamination de l'OCR

➔ Loi de Zipf, transcription* qualité ↘, CER : Kraken 3.0 = 0.1768, Tess. fr 0.3.6 = 0.0976.

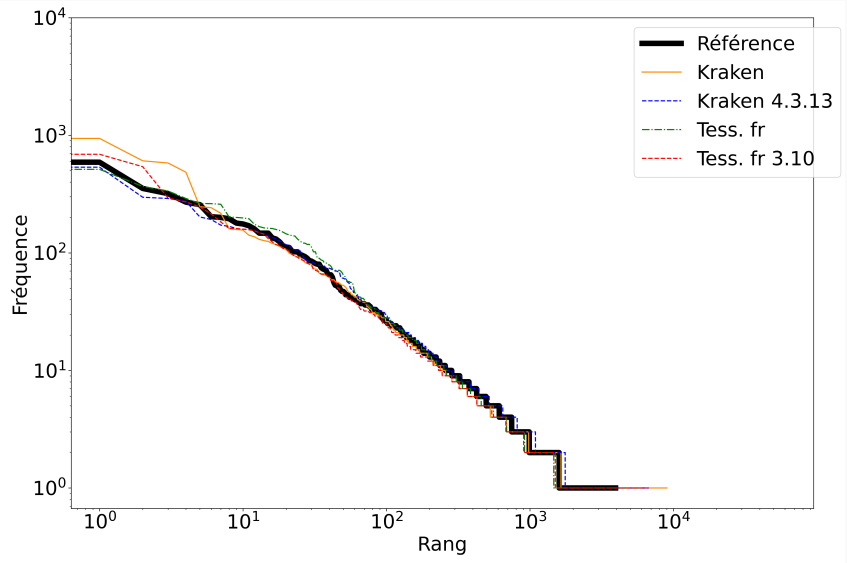


Figure 5 – Longue traîne sur les textes.

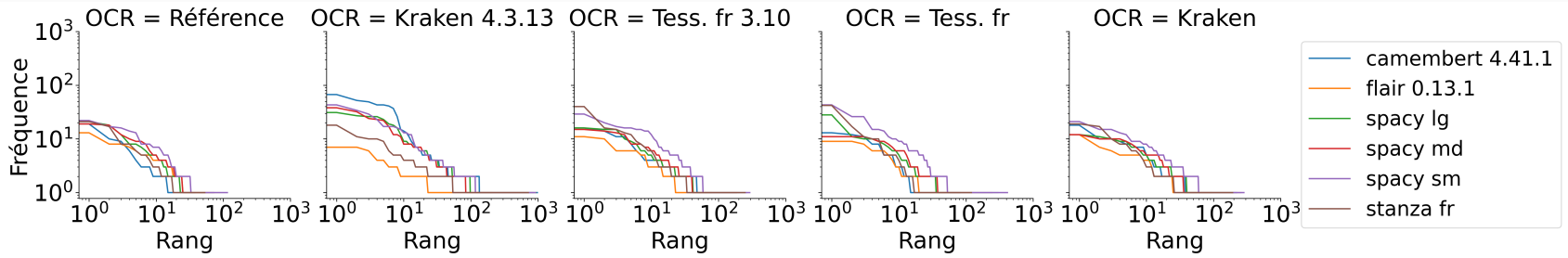


Figure 6 – Longue traîne sur les résultats de REN.

* J. Adam, *Mon village*,1860.

Difficultés d'entity matching et silence.

Entités contaminées → Faux Positif ou **Vrai Positif contaminé**

Version	Contexte	`sp_lg`	`stz`	CmBert	flair	SEM	CasEN
Réf.*	sur la **grand-route de Montivilliers**	✓	✓	✓	✗	✓	✓
Kraken 3.0	sur la **grand'route _ Yontivilliers**	✓	✓	✓	✗	✗	✓
Krak. 4.3	sur la **grandroute _ BMontivilliers**	Misc.	✗	✓	✗	N/A	N/A
Krak. Lrep	sur la **Crand route R Mntirittiere**	✓	✓	✓	✓	N/A	N/A
Tess fr 0.3.6	sur la **grand'route _ M_ntivilliers**	Misc.	✓	✓	✗	✗	✗
Tess. fr 0.3.10	sur la **grand'route _ Montivilliers**	✓	✓	✓	✗	N/A	N/A

✓ EN correctement reconnue ;

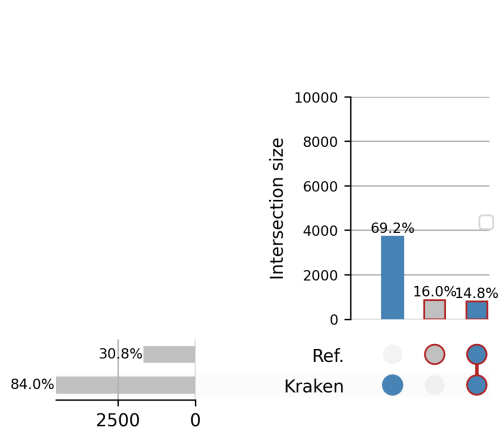
✓ erreur périmètre de l'EN ;

✗ EN non reconnue ;

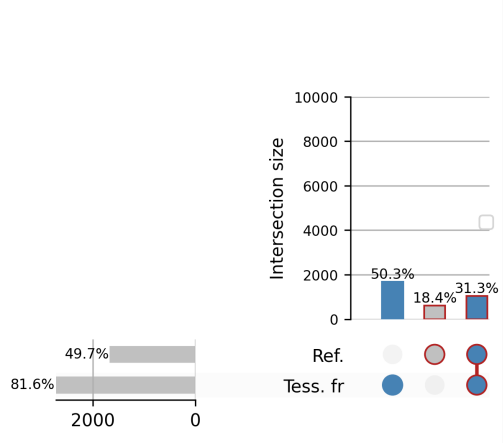
Abréviations Per., Misc., Org. : erreur d'étiquette ;

* G. de Maupassant, *Une vie*, 1883.

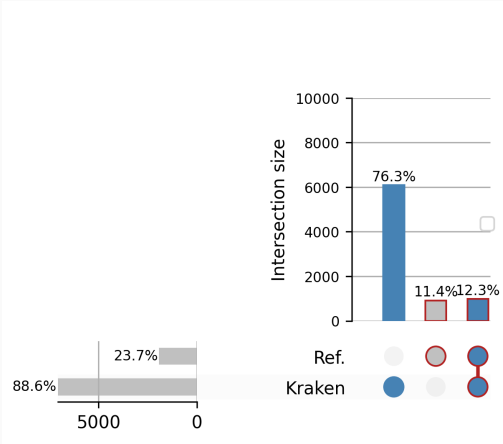
Évaluation stricte et problèmes d'alignements



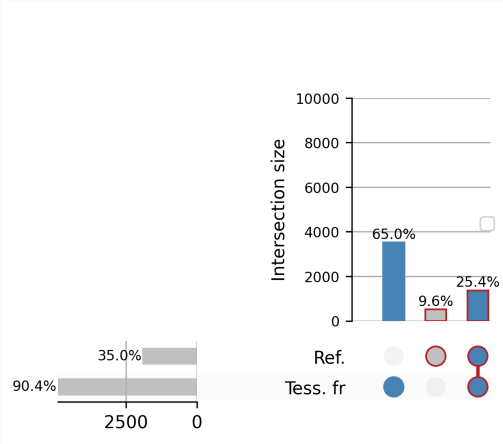
(a) KRAKEN – `CamemBert`



(b) TESS. FR – `CamemBert`



(c) KRAKEN – `spaCy_lg`



(d) TESS. FR – `spaCy_lg`

Figure 7 – Intersections pour les configurations avec Kraken 3.0 et Tess. fr 0.3.6 pour le corpus *small*-ELTeC-fra (global).

Évaluation et problèmes d'alignements

Version	#Entités		Évaluation			
	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken 3	1122	744	134	0.048	0.026	0.033
Kraken 4.3.13	3097	744	104	0.0	0.0	0.0
Tess. fr 0.3.6	860	744	182	0.063	0.052	**0.057**
Tess. fr 0.3.10	860	744	180	0.079	0.064	**0.070**

Table 3 – Résultats pour les EN de `spaCy_lg` sur différentes versions OCR, pour la catégorie **LOC**.

Version	#Entités		Évaluation avec `Nerval`			
	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken 3	1122	744	561	0.500	0.754	0.601
Kraken 4.3.13	3097	744	397	0.128	0.533	0.206
Tess. fr 0.3.6	860	744	640	0.744	0.860	**0.798**
Tess. fr 0.3.10	860	744	647	0.752	0.869	**0.806**

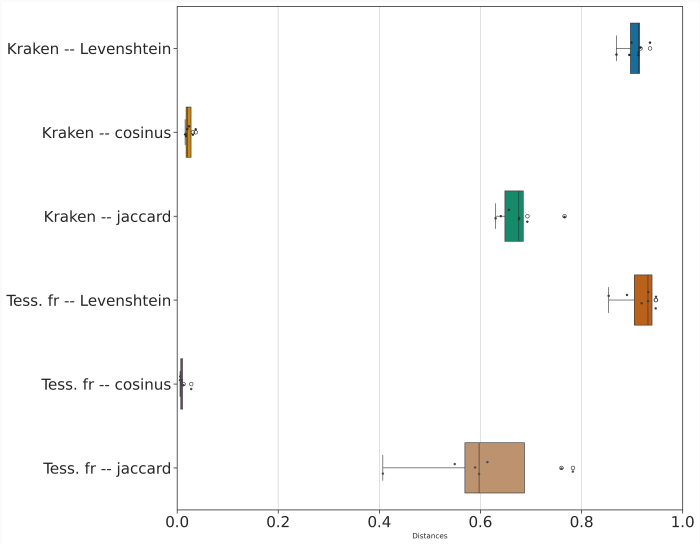
Table 4 – Résultats après alignement avec `Nerval` des EN de `spaCy_lg` sur différentes versions, pour la catégorie **LOC**.

A. Daudet, *Le petit chose*, 1868.

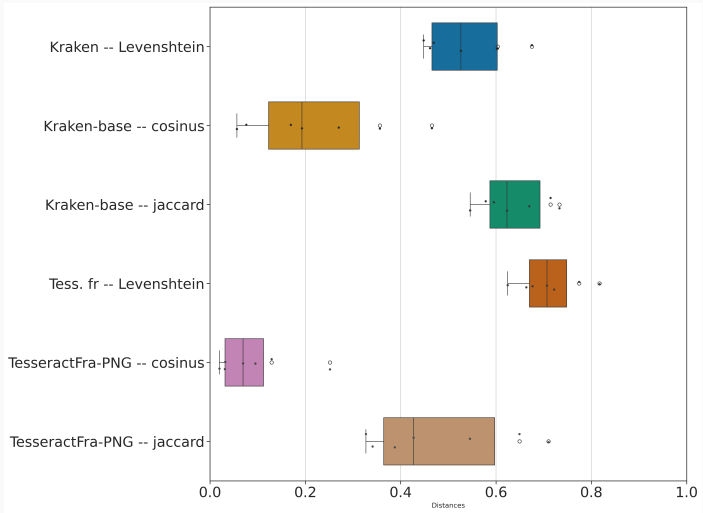
Évaluations, importance du choix des métriques

Distances cosinus, Jaccard et Levenshtein :

- ➢ différents temps de calcul ;
- ➢ éclairent différents phénomènes ;



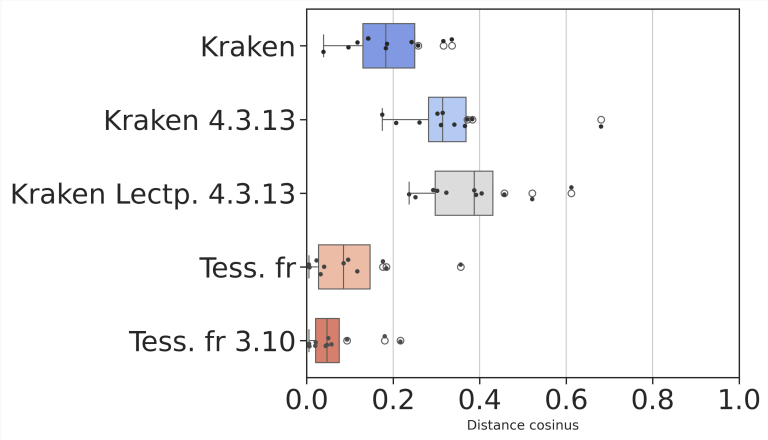
(a) Versions OCR



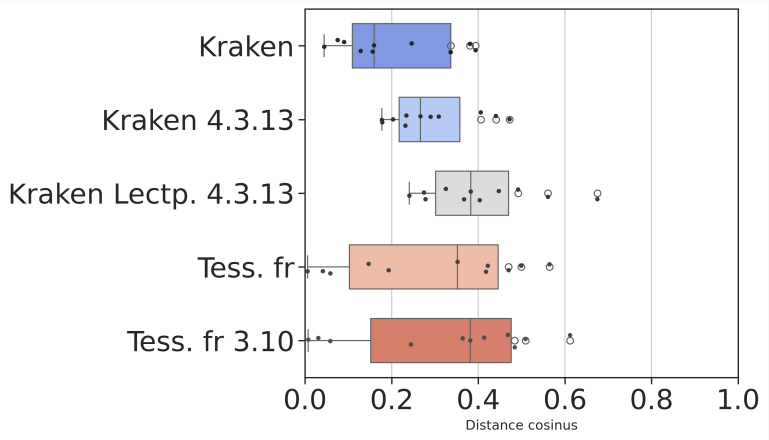
(b) Résultats de SPACY_LG

Figure 8 – Distances de Levenshtein, cosinus et Jaccard pour les configurations avec `spaCy_lg` de *small*-ELTeC-fra (global). Boîte à moustache proche de 0 = EN similaires, 1 = EN différentes.

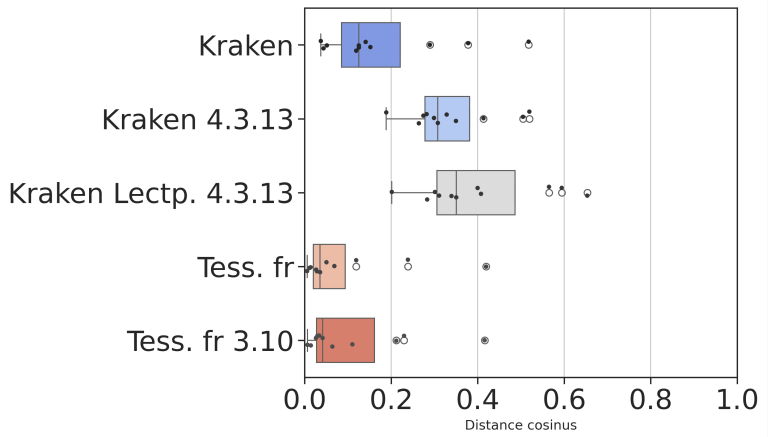
Meilleure configuration selon cosinus



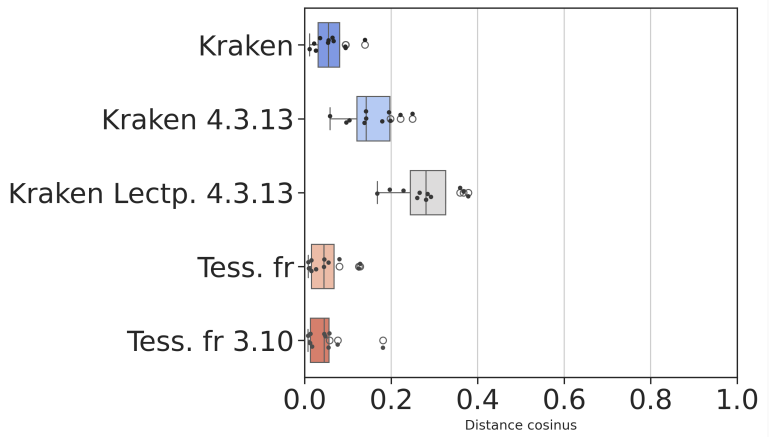
(a) SPACY_LG, cosinus



(b) STANZA, cosinus



(c) CAMEMBERT, cosinus



(d) FLAIR, cosinus.

Figure 9 – Distances cosinus pour la REN sur le corpus *small*-ELTeC-fra (global).

Points clés

- Toutes les erreurs d'OCR ne se valent pas ;
- Meilleures configurations : `Tesseract` – `spaCy_lg` ou `Tesseract` – `flair`
- Quantité d'**hapax** = indice de la contamination OCR ;

  → + d'**hapax** = **texte contaminé** ;
- L'évaluation demande **alignement** ➡ **verrou** ;
- Chaque **métrique** met en évidence un **phénomène particulier** ;

Évaluation qualitative : Cas attendus

Faux positif (FP) : EN détectées à tort dans la version OCR ;

Faux négatif (FN) : EN manquantes dans la version OCR ;

Vrais positifs (VP) : EN détectées dans les deux versions ;

Type	Version	Contexte	`spaCy-lg`
FP	Réf.[a]	*dans ce village, ordinairement désert et silencieux*	()
	Kraken	*dans ce village, ordinairement desert et silencieux*	silencieux
FN	Réf.[b]	*devenait une sorte de folie religieuse en Italie*	Italie
	Kraken	*devenait une sorle de folie religieuse en Ilalie*	()
VP	Réf.[c]	*filles de son âge venaient à leur maison à Paris*	Paris
	Kraken	*filles de son age venaient a leur maison a Paris*	Paris

[a] G. Aimard, *La Belle rivière*, 1894. [b] H. de Balzac, *Albert Savarus*, 1853. [c] A. de Noailles, *La nouvelle espérance*, 1903.

Évaluation qualitative : cas complémentaires

➔ Sous évaluation du bruit et du silence de la REN :

Faux vrais positifs (FVP) : EN détectées à tort dans les deux versions ;

Faux vrais négatifs (FVN) : EN manquantes dans les deux versions ;

Type	Version	Contexte	`spaCy-lg`
FVP	Réf.[a]	*[...] better than the milk-and-water lagrime*	lagrime
	Kraken	*[...] better than the _ilk- and-water lagrime*	lagrime
FVN	Réf.[b]	*[...] l'été dans leur **propriété des Peuples***	()
	Kraken	*[...] l'ete dans leur **pro- priete des Peuples***	()

[a] W. M. Thackerey, Vanity Fair, 1848. [b] G. de Maupassant, Une vie, 1883.

Évaluation qualitative : cas complémentaires

➔ sur évaluation du bruit et du silence de la REN

Faux faux positifs (FFP) : EN détectées **seulement dans la version OCR** :

- EN **manquantes** dans la référence$^{(i)}$;
- EN détectées dans l'OCR mais sous **forme contaminée**$^{(ii)}$;

Faux faux négatifs (FFN) : EN détectées à tort dans le texte de référence ;

FFP $^{(i)}$	Réf.^{a} Kraken	*[...] a sua entrada para o* ***colegio militar*** *[...] a s_a entrada para o colcgio militar*	() colcgio militar
FFP $^{(ii)}$	Réf.^{b} Kraken	*[...] e na vespera delle ir para Coimbra* *[...] e na vespera delle ir para Coimhra*	Coimbra Coimhra
FFN	Réf.^{c} Kraken	*[...] fleurs emblématiques que les Bachagas* *[...] fleurs emble- matiques que les Bach'agas*	Bachagas ()

^{a} ^{b} A. Castro Osorio, *Quattro Novelas*, 1908. ^{c} A. Daudet, *Le petit chose*, 1868.

Évaluation qualitative

✏ Évaluation manuelle à posteriori du *Silver standard*

- Les **FFP** qui sont des **VP** ➜
  - Kraken : 71+50 = 121 ;
  - Tess. fr : 100 + 17 = 117 ;

	Réf.[a]	Kraken	Tess. fr
Nb. types	207	454	294
VP	105	**71**	**100**
FP	102	283	105
FFP → **VP**		**50**	**17**
FVP → FP		50	72
FFN → FP		45	32

Table 5 – Nombre des VP et FP prenant en compte les cas complémentaires.

➜ La plupart des FP sont dus à la REN et non pas à l'OCR ;

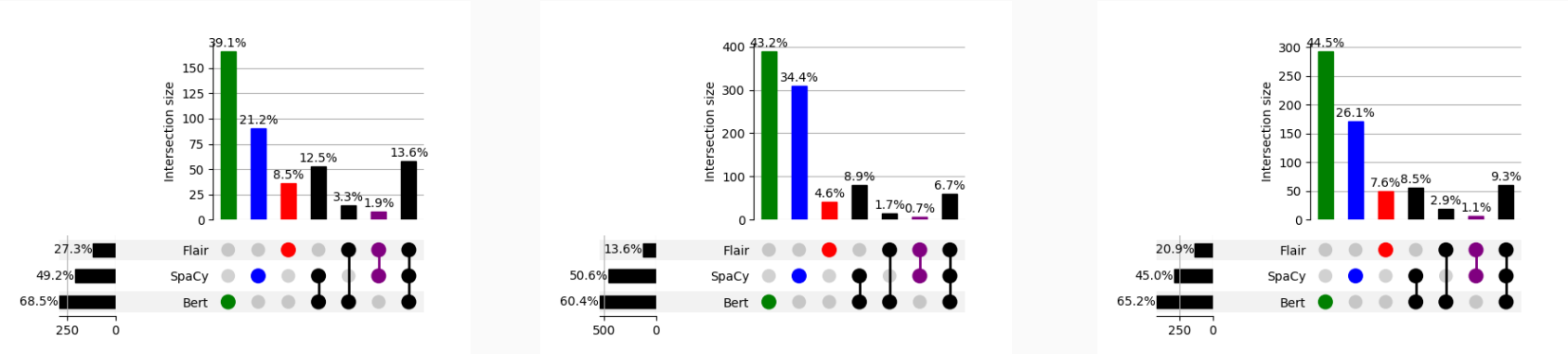
[a] A. Daudet, *Le petit chose*, 1868.

Synthèse

- **Typologie étendue** pour l'évaluation fine de l'impact des erreurs OCR sur la REN ;
- **Difficulté** éval. à **grande échelle** des **EN contaminées** ;
- La REN sur *données propres* n'est **pas parfaite** ;

Stratégies pour aider les utilisateur·ice·s

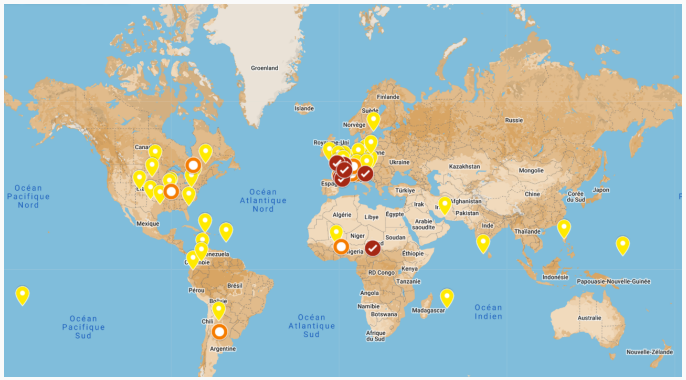
Combinaison de ≠ systèmes de REN



(a) Référence (b) KRAKEN (c) TESS_FR

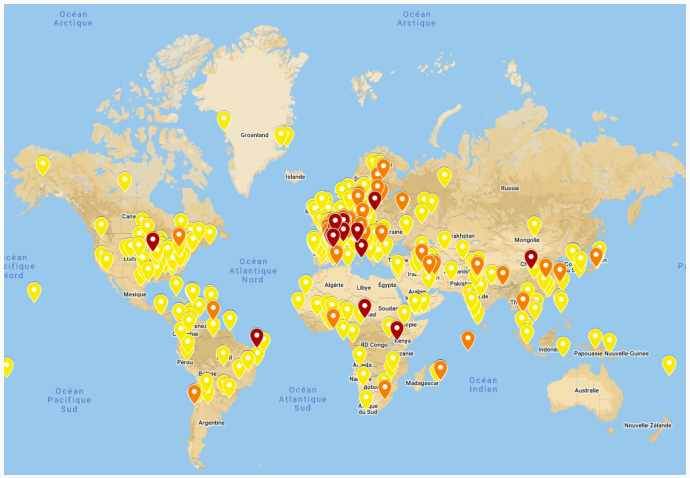
Figure 10 – Combinaison de trois outils de REN sur un OCR de haute qualité (TESS.FR : CER = 0,03, WER = 0,05 ; KRAKEN : CER = 0,05, WER = 0,18). Chaque colonne représente le pourcentage d'entités nommées trouvées par chaque combinaison exclusive d'outil(s). Chaque ligne indique la couverture individuelle et inclusive de chaque sous-ensemble. Daudet, 1868.

Cartographier les résultats OCR basse qualité (CER = 0.13)



(a) `Réf. automatique`

(b) `Réf. filtrage sorties REN`



(c) `Kraken automatique`

(d) `Kraken filtrage sorties REN`

Figure 11 – Légende : 🔴 3 REN ; 🟠 2 REN ; 🟡 1REN.

Z. Carraud, *La petite Jeanne*, 1853.

Similarités : des pistes pour le liage des entités contaminées

Versions	Entités Réf.	Entités Align	Jaccard	Cosinus
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 6 – Récupération automatique des formes contaminées par `spacy_lg` sur différentes versions de J. Adam, *Mon village*, 1860.

Cluster : des pistes pour le liage des entités contaminées ?

Figure 12 – Cluster cosinus bigramme de caractères sur les sorties Tesseract, spaCy_lg. A. Daudet, *Le petit chose*, 1868.

Cluster : des pistes pour le liage des entités contaminées ?

➜ Clusters intéressants, le centroïde et un VP.

Version	Centroid	Cluster members
Réf.[a]	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken	Montparnasse	Montparnasse, boulevard Montparnasse, theatre Montparnasse, Gymnase, Debarrassez, WWt3, rs5, ytP
Kraken[b]	Goderville	Gdoderville, Gloderville, Goderville, Barville, Fourville, OD0

➜ Mais parfois le centroïde est un FP.

Version	Centroid	Cluster members
Réf.[a]	PION	Lyon, Odéon, PION, Rio
Kraken [a]	Fougeroux	Broum, Fougeroux, Luxembourg, MY, Perou, Vaudoux, lesFougeroux

- erreur Clustering,
- erreur Clustering + interférence OCR,
- erreur Clustering + bruit REN,
- EN LOC.

[a] modèle REN `spaCy_lg`, *"Le petit chose"*, Daudet, 1868.

[b] modèle REN `spaCy_lg`, *"Une vie"*, G. de Maupassant, 1883

Évaluation quantitative de $\neq$ algo. de clustering

	small-ELTeC-fra	*small*-ELTeC-eng
Ref.	329	93
Kraken	828	356
Tess.	1035	116
Total	2192	565

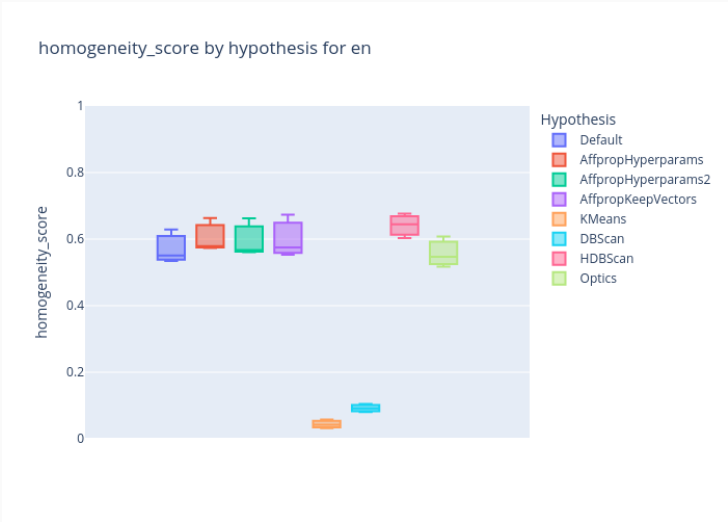
Table 7 – Nombre de tokens annotés pour les sous-corpus *small*-ELTeC-fra et *small*-ELTeC-eng.

Évaluation quantitative de $\neq$ algo. de clustering

Les algo. évalués + CountVectorizer, n-gram(min, max)

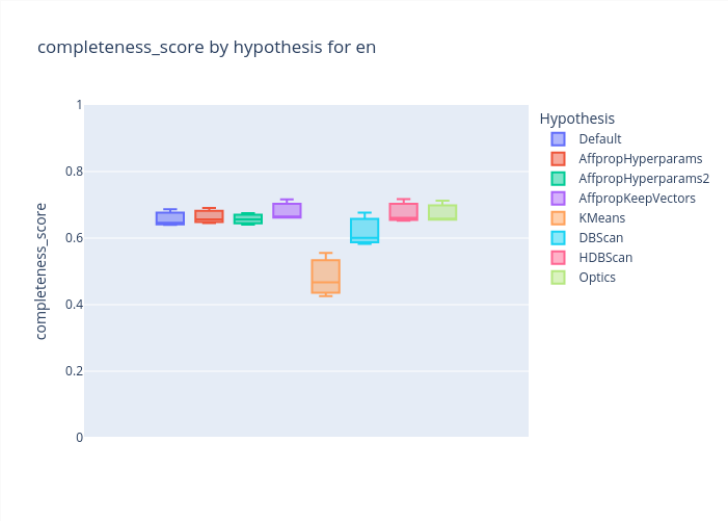
- Affinity Propagation :
  - Default (2,2) ;
  - Hyperparams (2,4) ;
  - Hyperparams2 (3,4) ;
  - KeepVectors (3,4) ;
- DBScan (2,4) ;
- HDBScan (2,4) ;
- Optics (2,4) ;

Évaluation quantitative de ≠ algo. de clustering

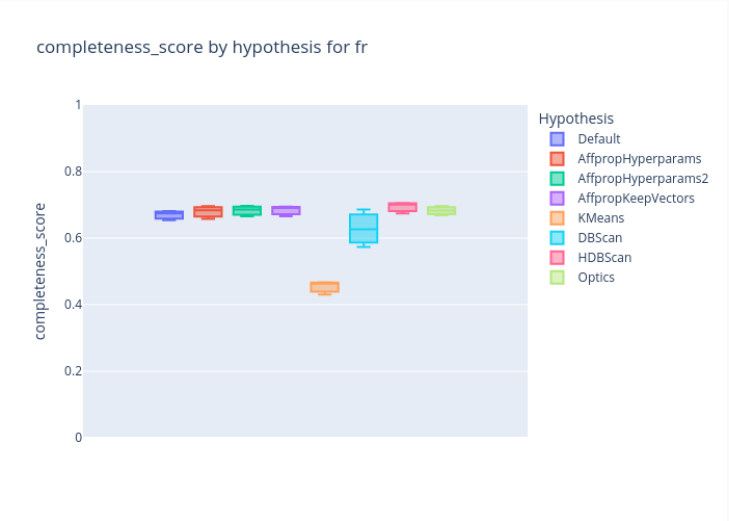


(a) homogeneity score, en.

(b) homogeneity score, fr.



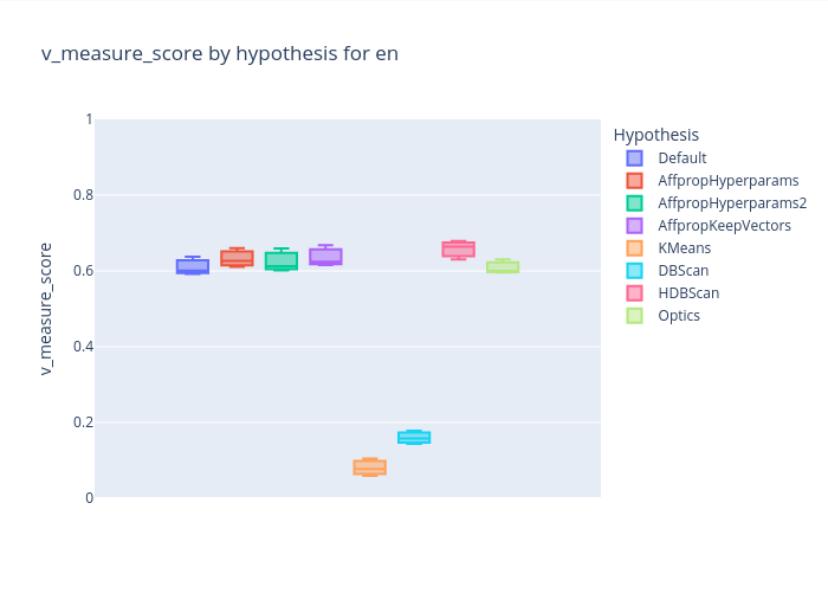
(c) completeness score, en.



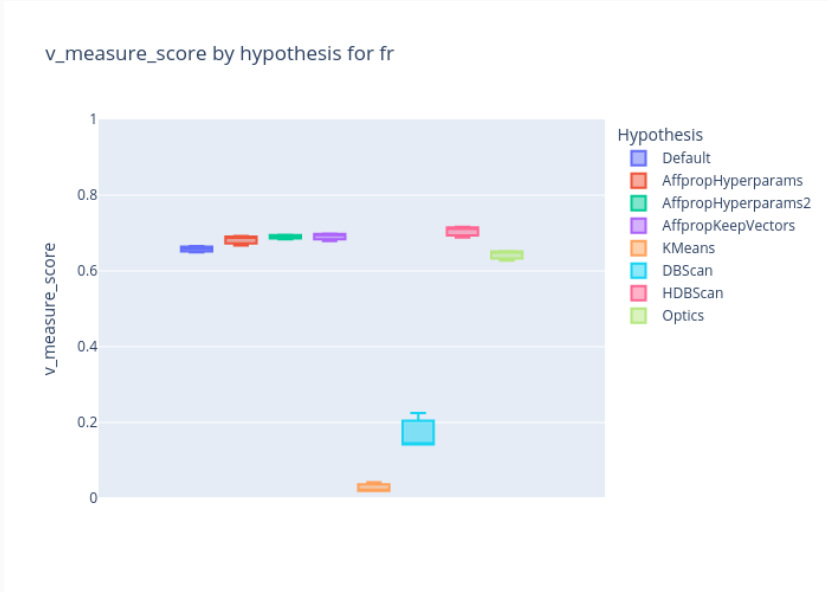
(d) completeness score, fr.

Figure 13 – Évaluation des algorithmes de clustering pour

Évaluation quantitative de ≠ algo. de clustering



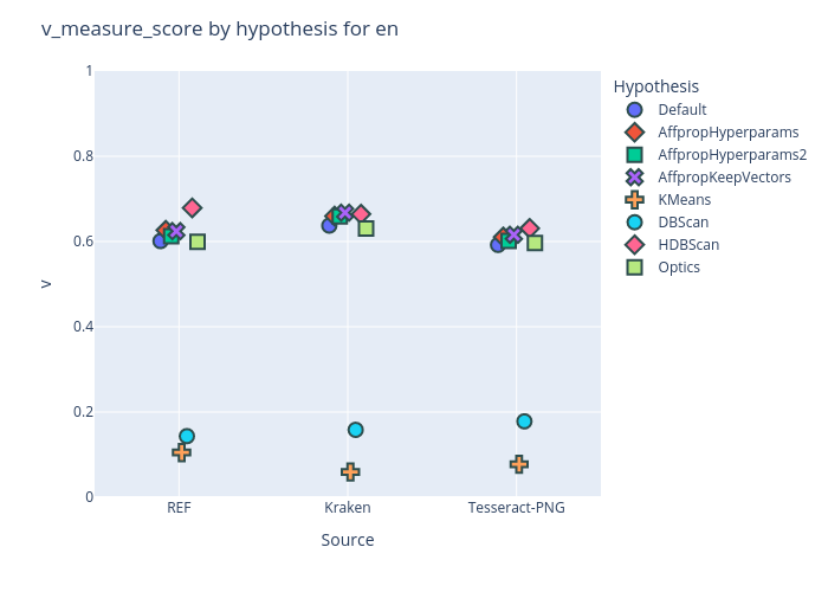
(a) v-measure, en.



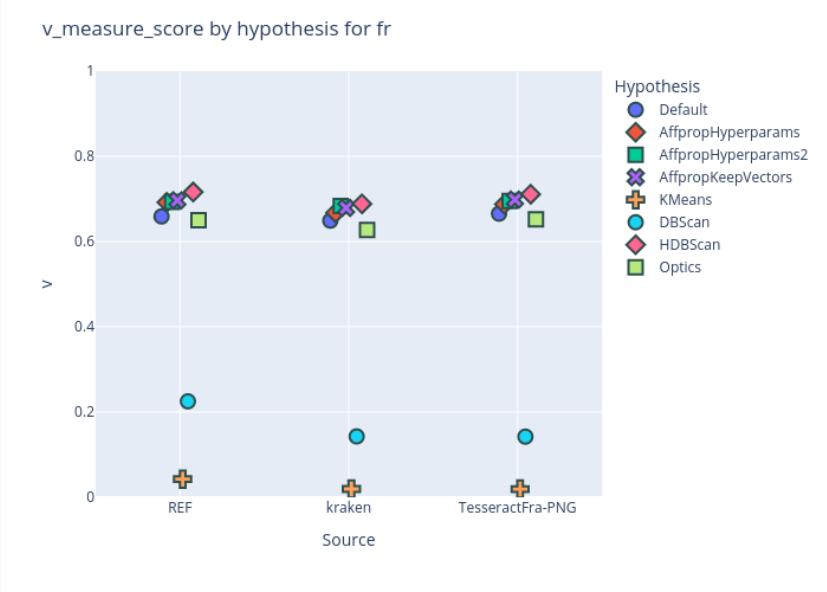
(b) v-measure, fr.

Figure 14 – Évaluation des algorithmes de clustering pour *small*-ELTeC-eng and *small*-ELTeC-fra.

Évaluation quantitative de ≠ algo. de clustering



(a) v-measure, en.



(b) v-measure, fr

Figure 15 – Évaluation de divers algorithmes de clustering sur différentes versions de texte (référence et OCR) en utilisant la *v-measure*, pour *small*-ELTeC-eng et *small*-ELTeC-fra.

Explorer l’espace littéraire européen

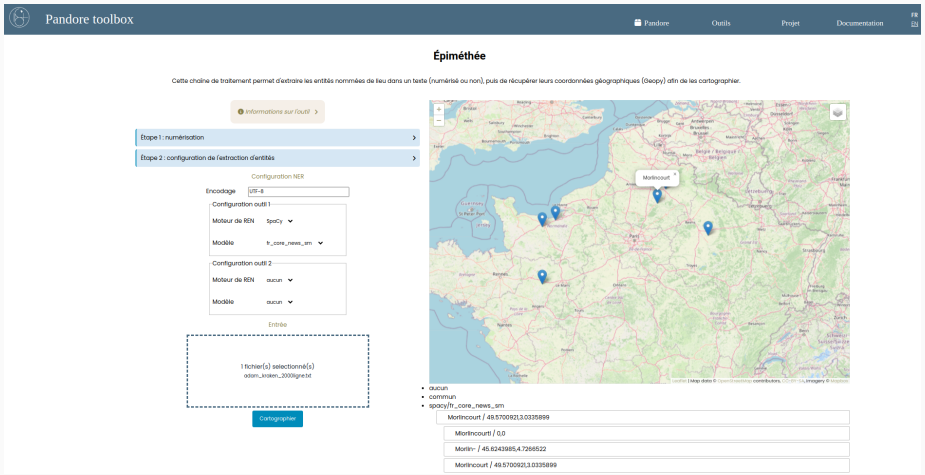
Systèmes d'OCR et de REN intégrés dans Épiméthée

- ✏ Modèles d'OCR : Tesseract Français, Anglais, Portugais ;
- ✏ Modèles de REN par langues :

	fr	en	pt
spaCy	✓	✓	✓
Flair	✓	✓	✓

Épiméthée : un système de bout-en-bout

- ☛ Conception du projet;
- ☛ Conduite du projet;
- ☛ Gestion équipe;



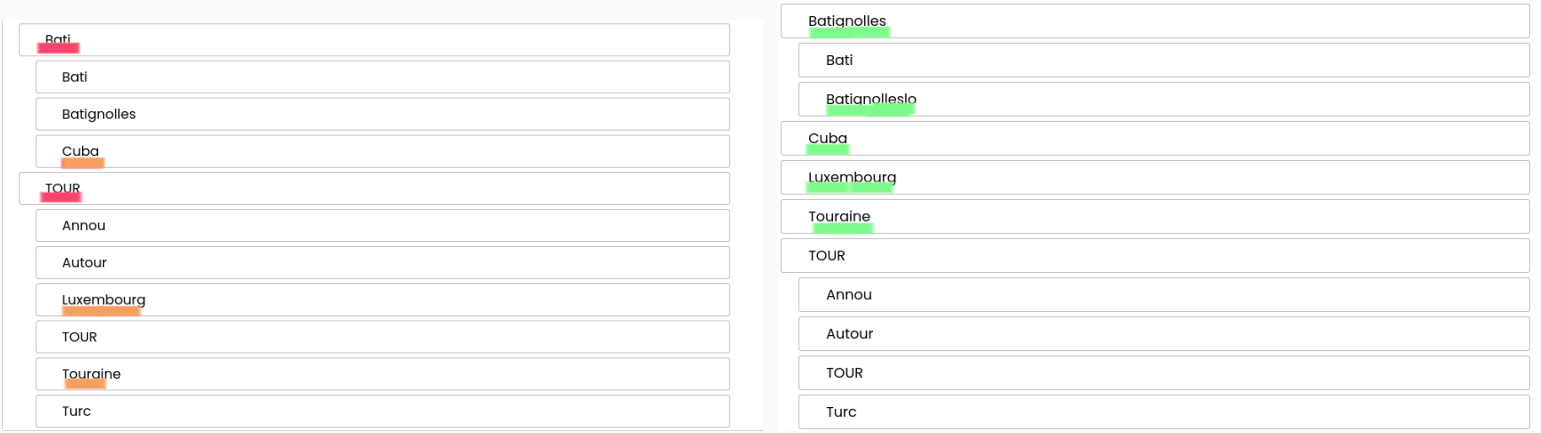
✍ Interface Épiméthée

✍ Github [a]

a. https://github.com/These-SCAI2023/EPIMETHEE

✍ Accès à l'interface

Épiméthée : un système de bout-en-bout



(a) Output from KRAKEN (b) Output from KRAKEN man. corr.

Figure 16 – Résultats de la chaîne de traitement Épiméthée avec les modèles `spaCy_lg` et `flair`, sur la version KRAKEN. Avant et après filtrage utilisateur *man. corr. = corrigé manuellement par l'utilisateur* ; ● erreur du centroïde par rapport aux candidats, ● erreur de candidat, ● filtrage manuel.

Limites de la chaîne de traitement

Le cas *"Arthurville"*, *Capitaine Cap*, A. Allais.

- `spaCy_lg` et `flair` = OK
- Épiméthée absent ⇝ qui est coupable :
  - Géoloc. ? ⇝ Actuel Municipalité de Saint Raphaël (Québec).
  - Clustering ? *Aff. Prop.* écarte des entités.

➜ Combien d'autres cas ?

➜ Quelles stratégies d'évaluations ?

Une géographie littéraire européanocentrée ?

➤ Étapes de production des cartes à partir de la chaîne de traitement Épiméthée :

- Filtrage depuis l'interface Épiméthée ;
- Récupération du fichier csv ;
- Correction manuelle des géoloc le cas échéant ;
- Enregistrement du csv (sép. « , »)
- Nouvelle visualisation (Google My map).

➤ Corpus utilisé, les textes sont issus de ELTeC-fra :

- Hector Malot, "Sans famille", 1878 ;
- Pierre Loti, "Mon frère Yves", 1883 ;
- Catulle Mendès, "Luscignole", 1892 ;
- Alphonse Allais, "Capitaine Cap", 1902.

Une géographie littéraire européanocentrée ?

Titre	nb tok.	nb LOC	Europe	Hs. Europe	France	Hs. France
Sans famille ✈	153 282	186	169	17	152	34
Mon frère Yves ✈	88 730	102	62	40	52	50
Capitaine Cap ✈	50 833	103	73	30	56	47
Luscignole ✈	36 667	25	17	8	19	6

Table 9 – Statistiques sur les lieux nommés dans les romans.

Et après ?

Contributions

➡ Éval. impact erreurs OCR sur REN → **tâche non triviale** :

- Éval. quali. : Typologie → cas attendus + cas additionnels ;
- Éval. quanti. : Stratégies d'évaluations → verrou alignement ;

➡ Épiméthée → *end-to-end* pour dépasser les contaminations OCR ;

➡ Évaluation d'algo. de clustering et paramètrages → amélioration des **clusters** ;

➡ Vers une géographie littéraire du XIX^{e} siècle :

✏ Une géographie européenne dans le roman français ?

Recommandations

➡ Un texte *propre* **ne garantit pas** de meilleurs résultats de REN :

- Correction auto. et **sur-correction** ;
- **Erreurs** des outils de REN même sur textes *propre* ;
- Importance choix **métrique** éval. ;

➡ Assistance utilisateur·ice·s :

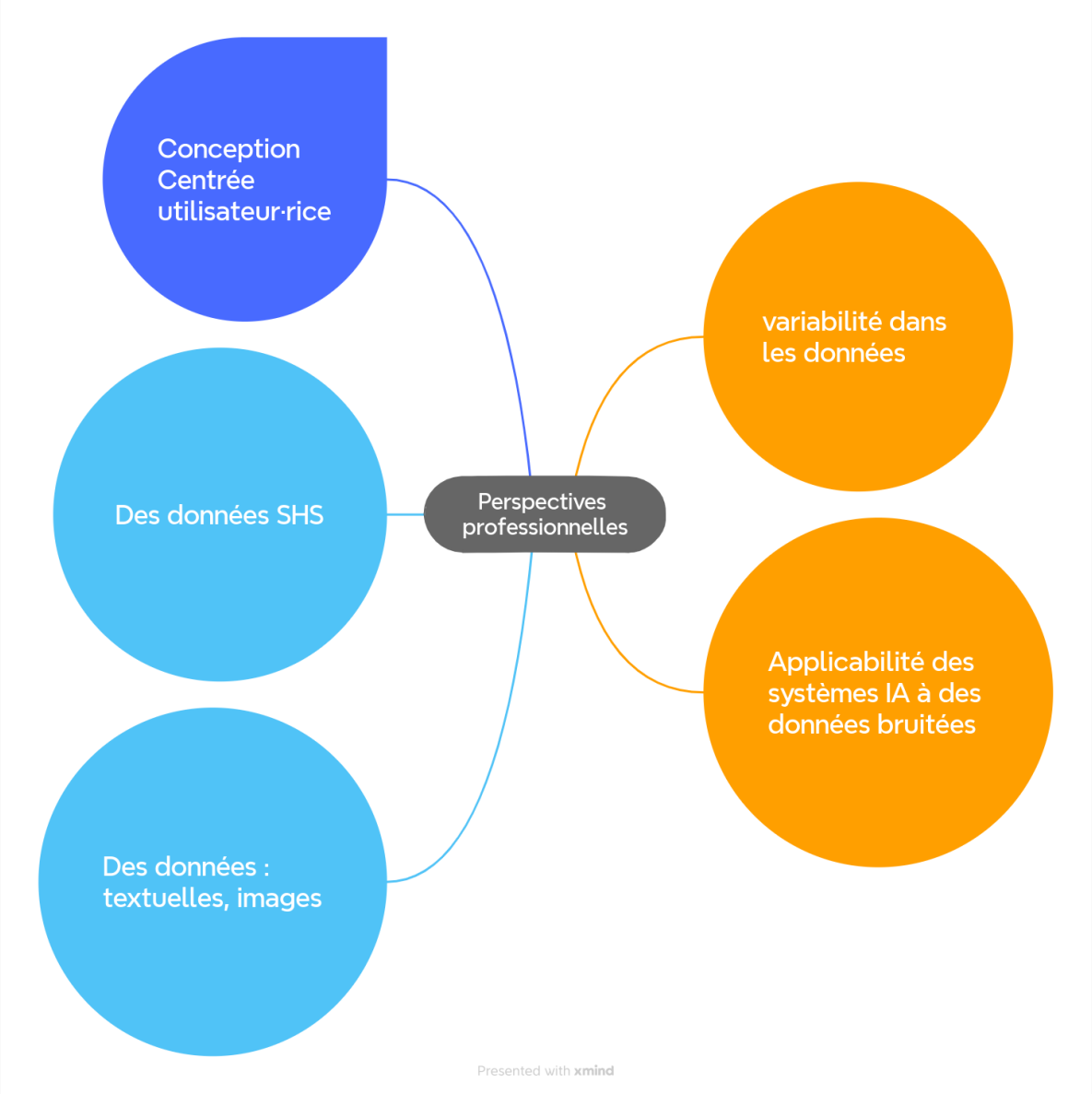
- **Combinaison** systèmes, EN extraite par +ieurs outils → VP ;
- **Clustering** : rapprocher des formes contaminées d'une EN ;

Perspectives

- Résolution **entity matching** → amélioration des **clusters**;
- Liage avec des bases de données de toponymes anciens;
- Améliorer la Géolocalisation auto.;
- Contribuer à une géographie littéraire du XIX^{e} siècle;

Merci de votre attention !

✈ https://carolinekoudoroparfait.github.io/



References i

Koudoro-Parfait, C. and Lejeune, G. (2024).
Reconnaissance des Entités Nommées spatiales sur un corpus littéraire bruité : des entités à la carte.
In *Séminaire des sources aux Systèmes d'Information Géographique*.

Koudoro-Parfait, C., Lejeune, G., and Buth, R. (2022).
Reconnaissance d'entités nommées sur des sorties ocr bruitées : des pistes pour la désambiguïsation morphologique automatique.
In *TAL-HN @ TALN(Traitement Automatique des Langues Naturelles) 2022*.

References ii

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.
Revue TAL.
À paraître.

Petkovic, L., Koudoro-Parfait, C., Desmarest, M.-S., and Lejeune, G. (2025).
Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ?
Corpus, (26).

References iii

Schöch, C., Patras, R., Erjavec, T., and Santos, D. (2021). **Creating the european literary text collection (eltec) : Challenges and perspectives.** *Modern Languages Open*, 0(1) :25.

Corpus ELTeC - European Literary Text Collection

Ouvrage	Auteur	Année	Pages	Mots	spaCy_lg	stanza
Mon village	J. Adam	1860	200	20938	213	152
La Belle rivière	G. Aimard	1894	339	137392	1004	959
Les trappeurs de l'Arkansas	G. Aimard	1858	450	91119	646	606
Marie-Claire	M. Audoux	1925	120	35780	101	108
Albert Savarus. Une fille d'Ève	H. de Balzac	1853	60	79924	682	684
La petite Jeanne	Z. Carraud	1884	220	53212	316	95
Le château de Pinon, vol. I	G. A. Dash	1844	332	44246	271	311
Le petit chose	A. Daudet	1868	292	86482	744	580
L'Éducation sentimentale	G. Flaubert	1880	520	150494	1304	1098
Une vie	G. de Maupassant	1883	337	75745	302	312
La nouvelle espérance	A. de Noailles	1903	325	54272	182	236

Table 10 – *small*-ELTeC-fra, 11 ouvrages, 3195 pages

Corpus ELTeC - European Literary Text Collection

Ouvrage	Auteur	Année	Pages	mots	spaCy_lg	stanza
Home influence	G. Aguillar	1847	628	171342	205	244
Auriol	W. H. Ainsworth	1844	246	46388	82	55
Wuthering Heights	E. Brontë	1847	764	94986	140	132
Coningsby	B. Disraeli	1844	983	101778	634	543
Mary Barton	E. Gaskell	1848	423	161568	290	281
The Mysteries of London	G. Reynolds	1844	840	810167	2019	2312
Modern Flirtations vol.1	C. Sinclair	1841	386	189057	502	248
Vanity Fair	W. M. Thackeray	1848	624	298568	1492	1164
The Life and Adventures of M. Armstrong	F. Trolloppe	1840	387	189392	187	207

Table 11 – *small*-ELTeC-eng, 9 ouvrages, 5281 pages

Corpus ELTeC - European Literary Text Collection

Ouvrage	Auteur	Année	Pages	Mots	spaCy_lg	stanza
Quattro Novelas	A. Castro Osorio	1908	272	50766	353	N/A
A illustre casa de Ramires	E. de Queirós	1900	543	107441	3881	N/A
O crime do padre Amoro	E. de Queirós	1875	620	141700	2362	N/A
Uma familia ingleza	J. Diniz	1875	360	122008	994	N/A

Table 12 – *small*-ELTeC-por, 4 ouvrages, 1795 pages

La Très grande Bibliothèque (TGB)

Ouvrage	Auteur	Année	Pages	Tokens	spaCy_lg	stanza
La princesse Pallianci	C. L. Bazancourt	1852	340	36 423	304	263
Meryem, scènes de la vie algérienne. Marcel.	C. Perrier [Bentégeat]	1863	360	85 077	662	512
Wilmina, ou L'enfant des Apennins	L. G. de Caudemberg	1820	242	36218	353	188
Les fourmis du parc de Versailles raisonnant ensemble dans leurs fourmilières	C. Lambert	1803	72	10 173	57	47
Œuvres complètes de Pierre Loti	P. Loti	1893-1911	588	133 129	2040	2136
La confession d'un enfant du siècle / Alfred de Musset; avec un portrait... par Eugène Lami...	A. de Musset	1879	494	92 140	578	269
Le Parnasse envahi, petit poème allégorique au sujet du sacre de S. M. Charles X.	E. Rullier	1825	71	10 261	165	38
La Comtesse de Rudolstadt	G. Sand	1861	340	102 423	618	505
Diégarias, drame en 5 actes et en vers	V. Séjour	1844	38	18 603	970	293
Le département de l'Oise : Compiègne et Marat, fragment historique	A. Sorel	1865	19	6 277	108	105

Table 13 – *small*-TGB-RevueCorpus, 10 ouvrages, 2564 pages.

La Très grande Bibliothèque (TGB)

Ouvrage	Auteur	Année	Pages	Tokens	spaCy_lg	stanza
L'Alsace et la Lorraine	L. Longret	1873	2	357	13	12
La Grèce libre	A. Bignan	1821	20	1 027	35	19
Poësies diverses	Inconnu	1745	10	1 502	32	11
Les dernières Étrivières [...]	B. Bonafoux	1877	22	2 320	29	20
M. de L'Espinasse [...]	D. L. Baric	1851	20	3 058	102	91
Adélaïde de Mariendal, drame en cinq actes	Inconnu	1783	100	15 344	276	217
Œuvres du seigneur de Brantome. Tome 14	P. de Bourdeille Sgr de Brantôme	1779	255	49 084	844	507
Souvenirs d'un vieux mélomane	A. Pontmartin	1879	350	61 872	659	598
La lyre des petits enfants	A. Cordier	1857	357	62 639	646	447

Table 14 – *small*-TGB-RevueTAL, 9 ouvrages, 1136 pages.

Les hapax : indice de la contamination de l'OCR

➔ Loi de Zipf, transcription qualité ↗, CER : Kraken = 0.0886, avec Tess. fr = 0.0496

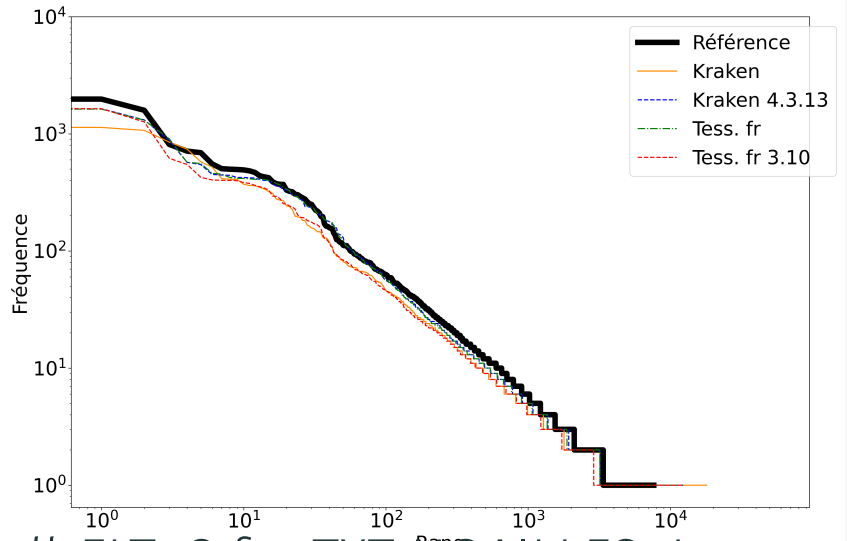


Figure 17 – *small*-ELTeC-fra TXT NOAILLES, *La nouvelle esperance*

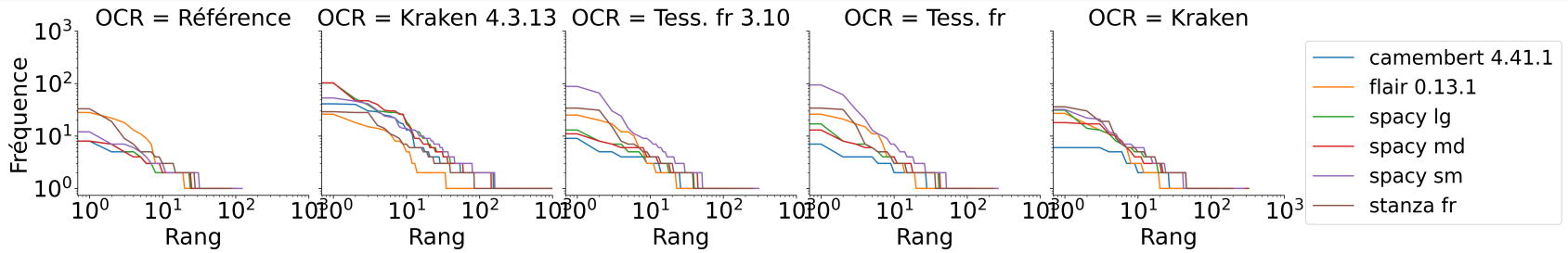


Figure 18 – *small*-ELTeC-fra REN NOAILLES, *La nouvelle esperance*

Cluster : des pistes pour le liage des entités contaminées ?

➜ Clusters intéressants, le centroïde et un VP.

Version	Centroid	Cluster members
Réf.[a]	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken	Montparnasse	Montparnasse, boulevard Montparnasse, theatre Montparnasse, Gymnase, Debarrassez, WWt3, rs5, ytP
Kraken[b]	Goderville	Gdoderville, Gloderville, Goderville, Barville, Fourville, OD0

➜ Mais parfois le centroïde peut être un FP.

Version	Centroid	Cluster members
Réf.[a]	PION	Lyon, Odéon, PION, Rio
Kraken [a]	Fougeroux	Broum, Fougeroux, Luxembourg, MY, Perou, Vaudoux, lesFougeroux

- erreur Clustering,
- erreur Clustering + interférence OCR,
- erreur Clustering + bruit REN,
- EN LOC.

[a] modèle REN `spaCy_lg`, *"Le petit chose"*, Daudet, 1868.

[b] modèle REN `spaCy_lg`, *"Une vie"*, G. de Maupassant, 1883