Exploration de la collection Française d'ELTeC avec l'outil de cartographie textuelle Epiméthée

Caroline Koudoro-Parfait
caroline.parfait@sorbonne-universite.fr

20 Mars 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

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**

→ Reconnaissance optique de caractère (OCR) ;

Variabilité dans le contexte d'usage

→ **qualité de la transcription OCR** ;

Angles Expérimentaux

- Comment adapter les outils de TAL aux pratiques des chercheurs en Littérature ?
- Quelle chaîne de traitement pour assister les chercheurs en Lettres ?
- Quelles stratégies pour dépasser les contaminations OCR ?

Les données réelles

Données : Textes Littéraires

- ✱ European Literary Text Collection (ELTeC) [1]
  [Schöch et al., 2021] :
  - 22 langues, $\approx$ 100 romans par langue ;
  - période : 1840 à 1920 ;

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

Des corpus pour une évaluation multilingue

✂ Corpus constitués

Corpus	Books	Pages	Words	# Named Entities : LOC		# Évaluations		
				spaCy_lg	flair	qti.	qli	Cluster
small-ELTeC-fra	11	3 195	829 604	5 765	4 814	✓	✓	✓
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

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]
- **Correction automatique**, problèmes de sur-correction ✗ [Koudoro-Parfait et al., 2024]
- **Linking**, Lier des entités contaminées à leur forme standard ✗
- **Cartes** Une vue d'ensemble des entités ✓ [Koudoro-Parfait and Lejeune, 2024]

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 ;

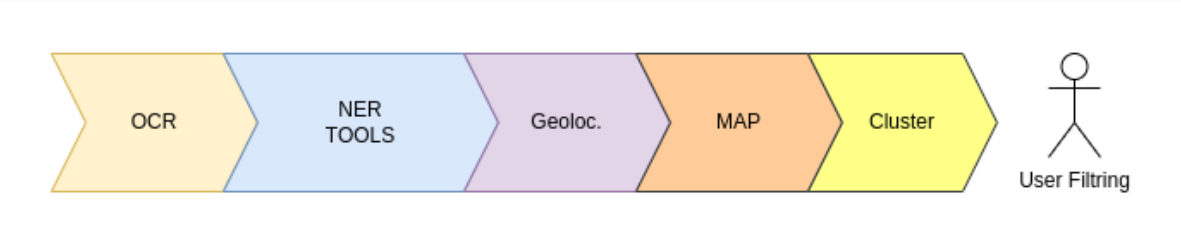


Figure 2 – Chaîne de traitement Épiméthée, `https://github.com/These-SCAI2023/EPIMETHEE`

☞ Interface : filtre manuel par l'utilisateur·ice ;

☞ Récupération (CSV), réemploi et analyse littéraire.

Impact des contaminations OCR sur la REN

Difficultés rencontrées par les modèles d'OCR

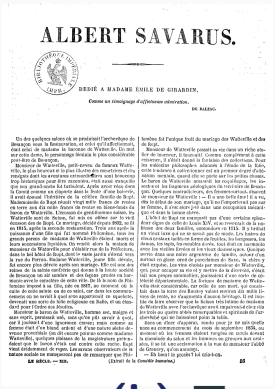
ENFANCE DE JEANNE. 3

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



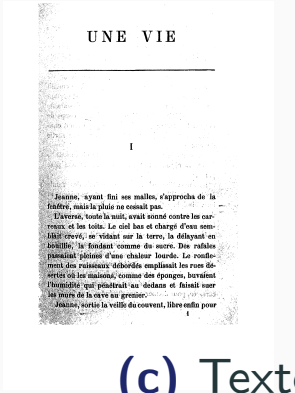
ALBERT SAVARUS.

DÉDIÉ A MADAME ÉMILE DE GIRARDIN,

Comme un témoignage d'affectueuse admiration.

DE BALZAC.

(b) Texte en colonnes



UNE VIE

I

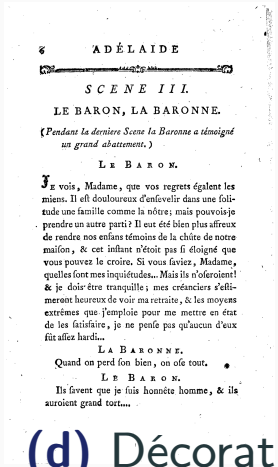
Jeanne, ayant fini ses malles, s'approcha de la fenêtre, mais la pluie ne cessait pas.

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 ruissellement 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

(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.

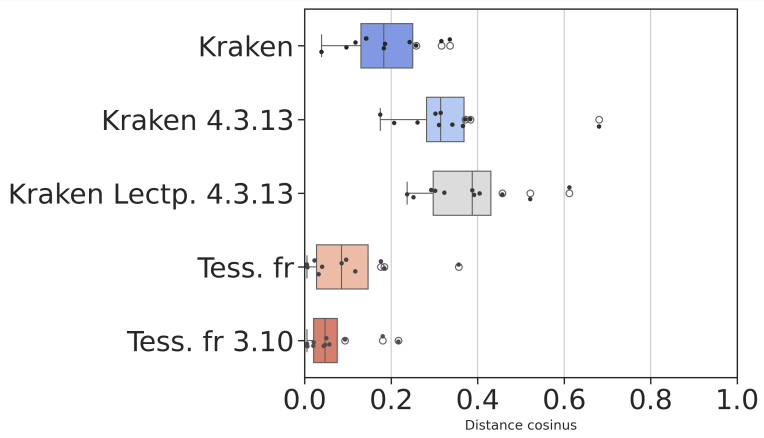
* Z. Carraud, *La petite Jeanne*, 1884.

Difficultés d'entity matching et silence.

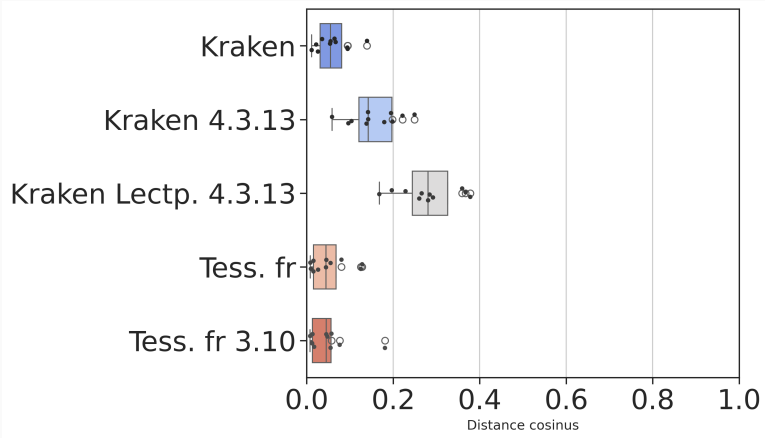
Entités contaminées → Faux Positif ou **Vrai Positif contaminé**

	Context	spaCy_lg	Flair
Ref	*There are some fields near Manchester*	Manchester	Manchester
Krak	*_ are some fields near ancehester [...]*	ancehester	()
Tess. en	*N/A*	N/A	N/A
Ref	*toute la noblesse de Picardie*	Picardie	Picardie
Krak	***toutela** noblesse de **V**icardie*	**V**icardie	**V**icardie
Tess. fr	*toute la noblesse de Picardie*	Picardie	Picardie

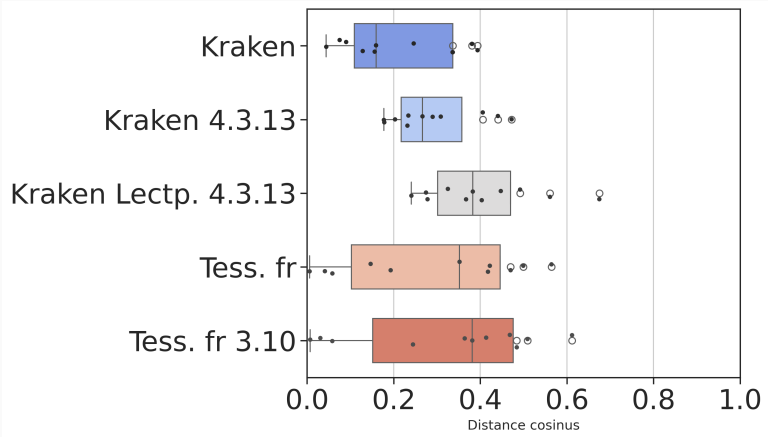
Meilleure configuration selon cosinus



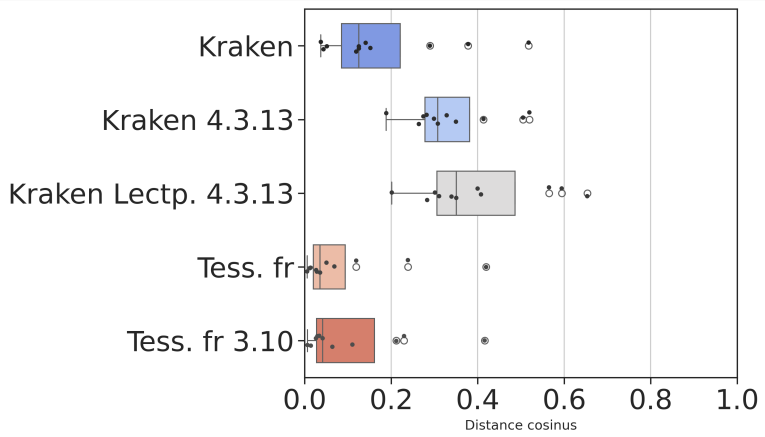
(a) spacy_lg, cosinus



(b) Flair, cosinus.



(c) stanza, cosinus



(d) CamemBert, cosinus

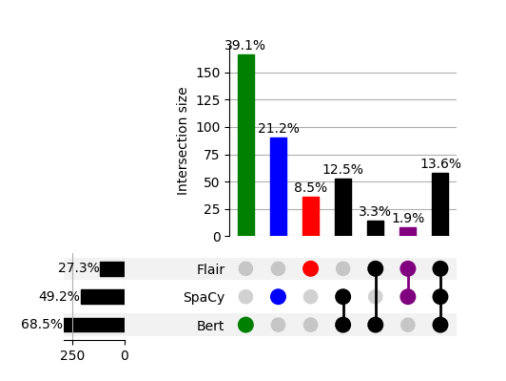
Figure 4 – Distances cosinus pour la REN sur le corpus *small*-ELTeC-fra (global).

Points clés

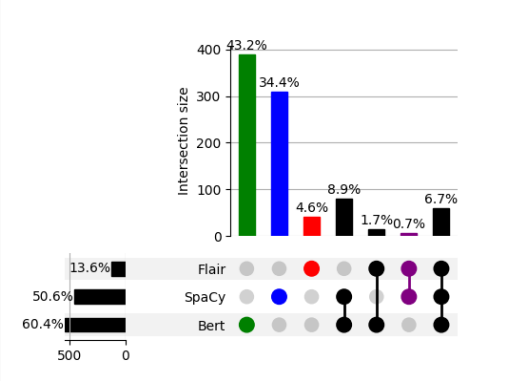
- Toutes les erreurs d'OCR ne se valent pas ;
- Meilleures configurations : `Tesseract` – `spaCy_lg` ou `Tesseract` – `flair` ;

Stratégies pour aider les utilisateur·ice·s

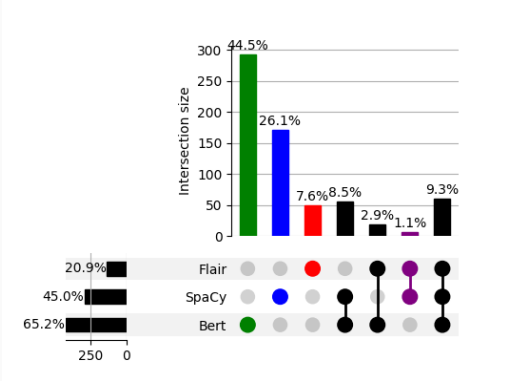
Combinaison de $\neq$ systèmes de REN



(a) Référence



(b) Kraken



(c) Tess_fr

Figure 5 – Combination of three NER tools on high quality OCR (Tess.fr : CER = 0.03, WER = 0.05 ; Kraken : CER = 0.05, WER = 0.18). Each column represents the percentage of NEs found by each combination of tool(s) (exclusive). Whilst each row gives the individual, and inclusive, coverage of each subset. Daudet, 1868.

Cluster : des pistes pour le liage des entités contaminées ?

Figure 6 – 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 peut être un FP.

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

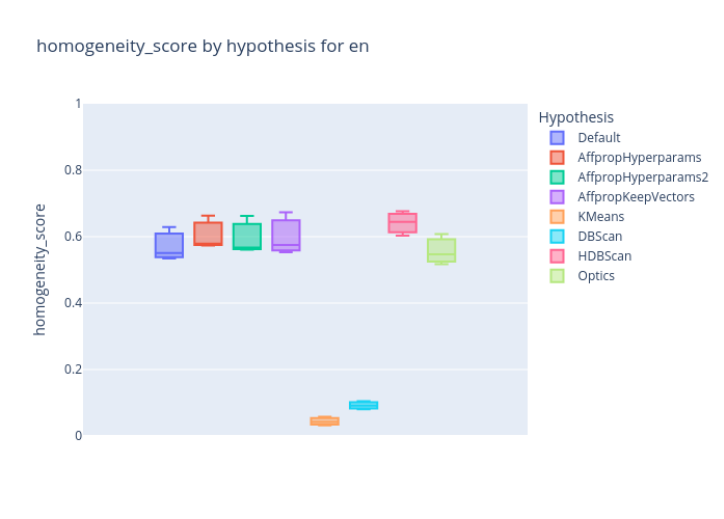
Table 3 – Statistics on the annotated subset

Évaluation quantitative de $\neq$ algo. de clustering

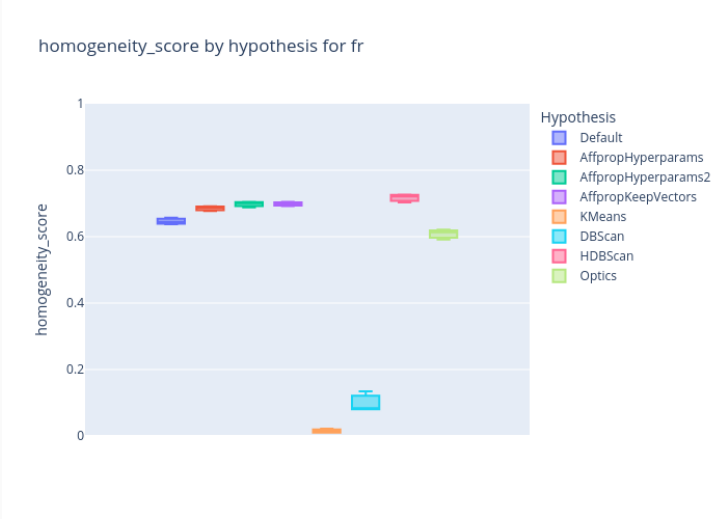
Les algo. évalués + CountVectorizer, n-gram(min, max)

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

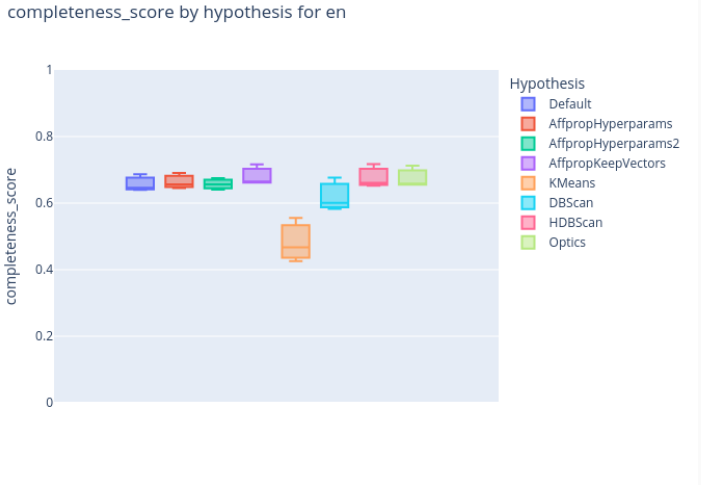
Évaluation quantitative de ≠ algo. de clustering



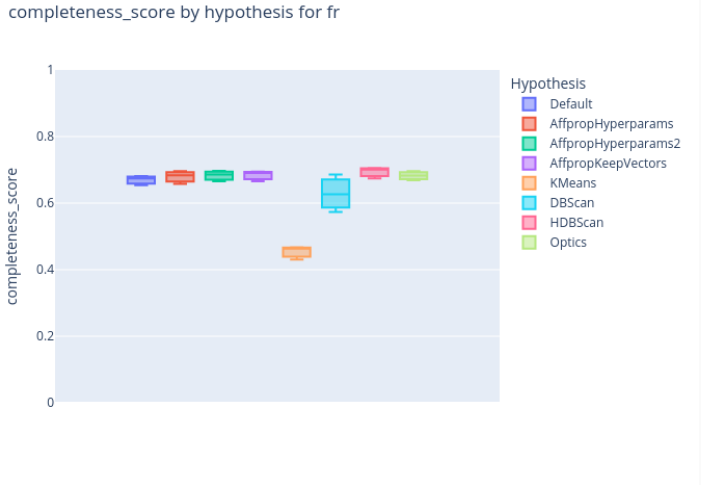
(a) homogeneity score, en.



(b) homogeneity score, fr.



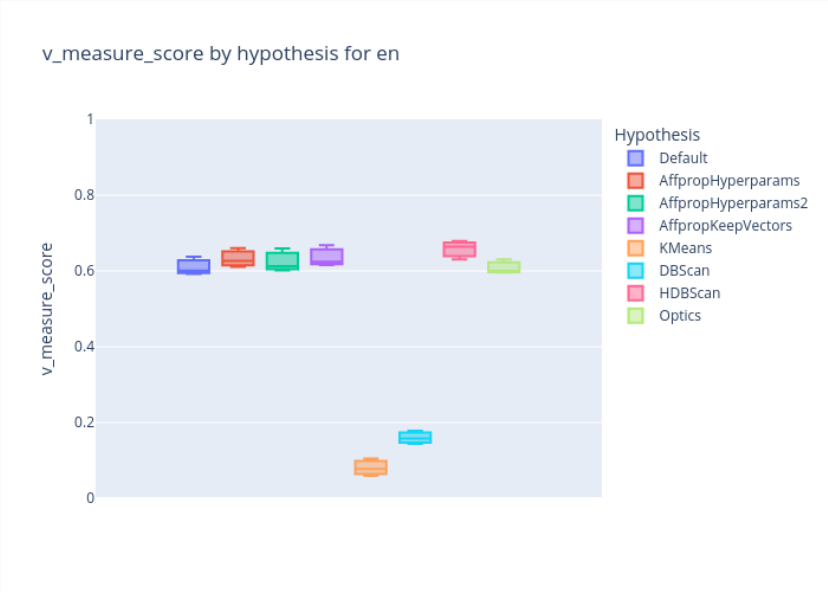
(c) completeness score, en.



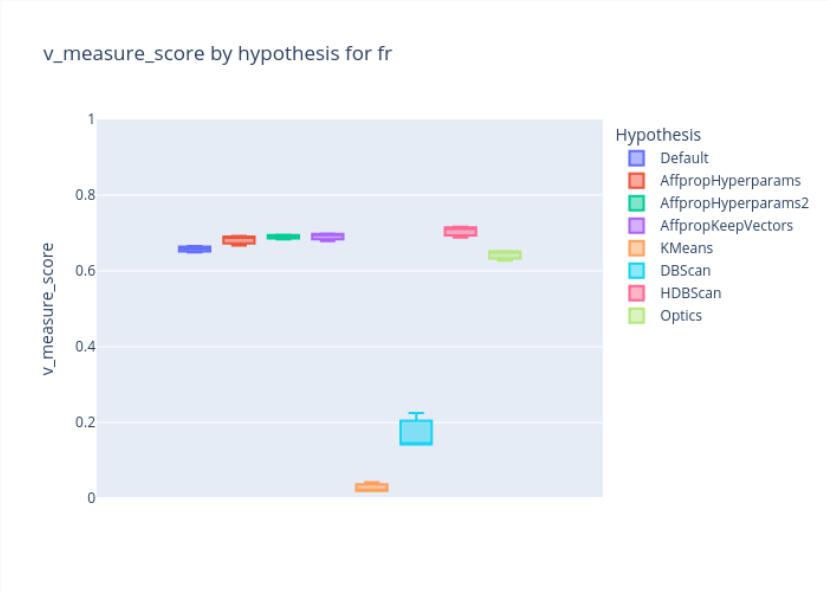
(d) completeness score, fr.

Figure 7 – Evaluation of clustering algorithms for *small*-ELTeC-eng and

Évaluation quantitative de ≠ algo. de clustering



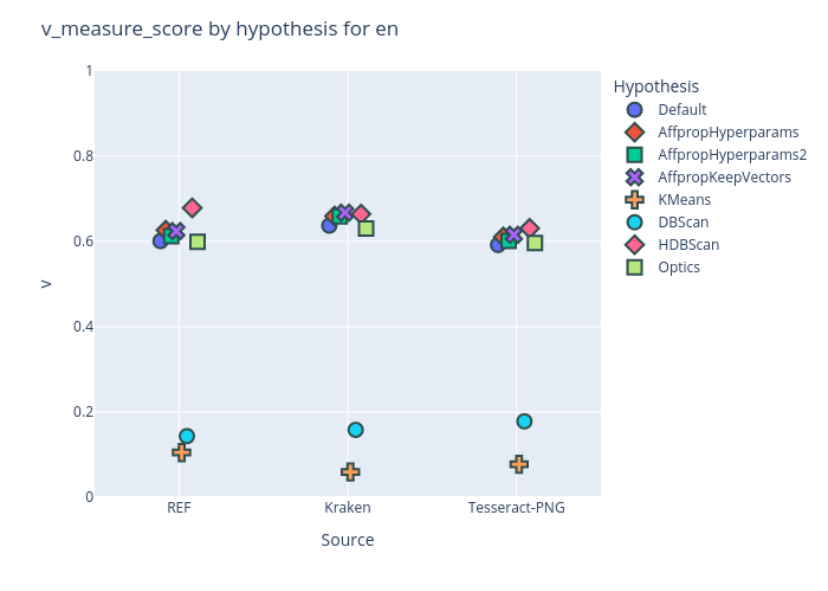
(a) v-measure, en.



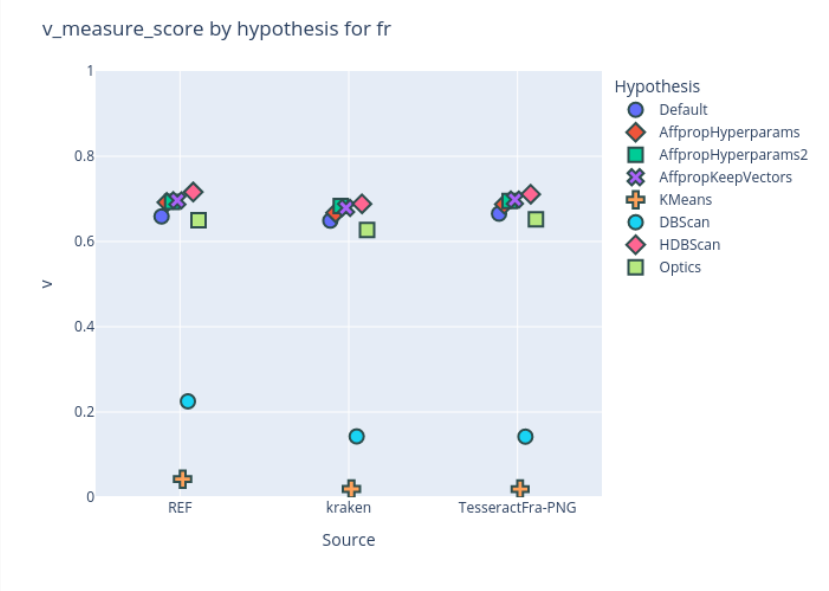
(b) v-measure, fr.

Figure 8 – Evaluation of clustering algorithms for *small*-ELTeC-eng and *small*-ELTeC-fra.

Évaluation quantitative de ≠ algo. de clustering



(a) v-measure, en.



(b) v-measure, fr

Figure 9 – Evaluation of various clustering algorithms on different versions of text (reference and OCR) using v-measure, for the *small*-ELTeC-eng and *small*-ELTeC-fra.

Explorer l'espace littéraire européen

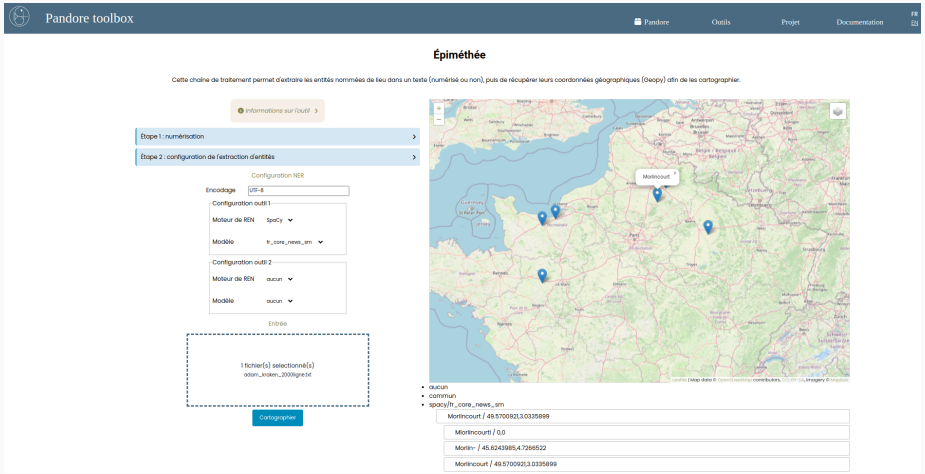
Systèmes d'OCR et de REN intégrés dans Épiméthée

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

	fr	en	pt
spaCy	✓	✓	✓
Flair	✓	✓	✗

Épiméthée : un système de bout-en-bout

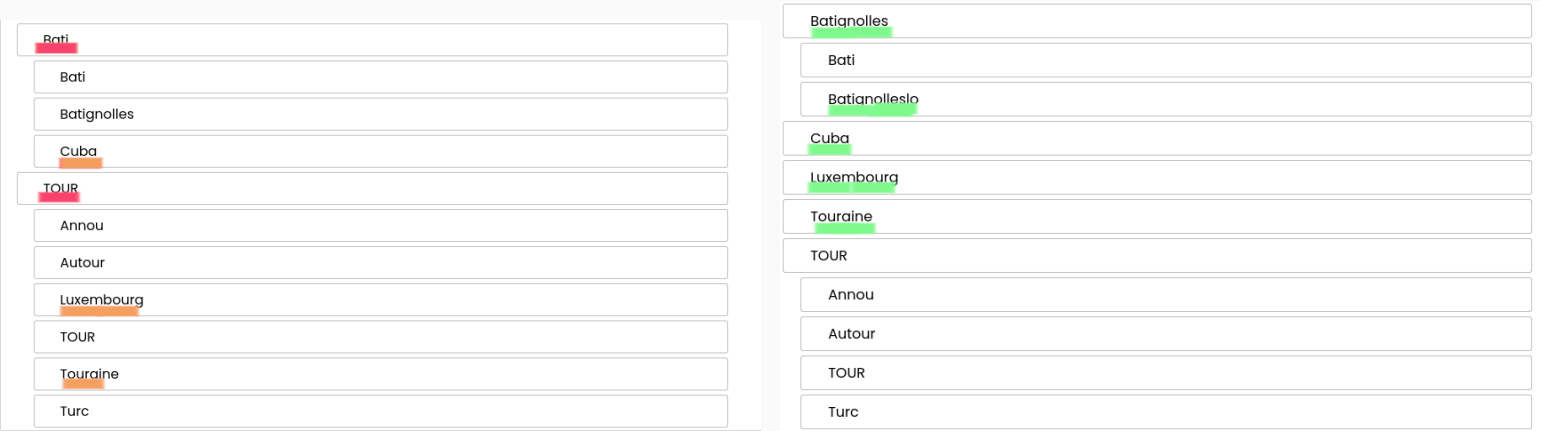
- ☛ Conception du projet ;
- ☛ Conduite du projet ;
- ☛ Gestion équipe ;



✍ Interface Épiméthée [2]

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

Épiméthée : un système de bout-en-bout



(a) Output from kraken **(b)** Output from kraken man. corr.

Figure 10 – Results of the Épiméthée pipeline with the spaCy_lg and flair models, on kraken version. Before and after user filtering *man. corr.* = *manually corrected by the user* ; ● centroid error compared to candidates, ● candidate error, ● manual filtering.

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 ?

Quel orient chez les Orientalistes ?

Cartes obtenues à partir de la chaîne de traitement Épiméthée et filtrage sur les romans de :

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

`https://drive.google.com/drive/folders/1h3sUycwQKexjMuC01bLMJdTEsvJRiIDD?usp=drive_link`

Et après ?

Contributions

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

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

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

➡ Contribuer à une géographie littéraire du XIX^{e} siècle : Quel orient chez les Orientalistes ? ;

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. ;

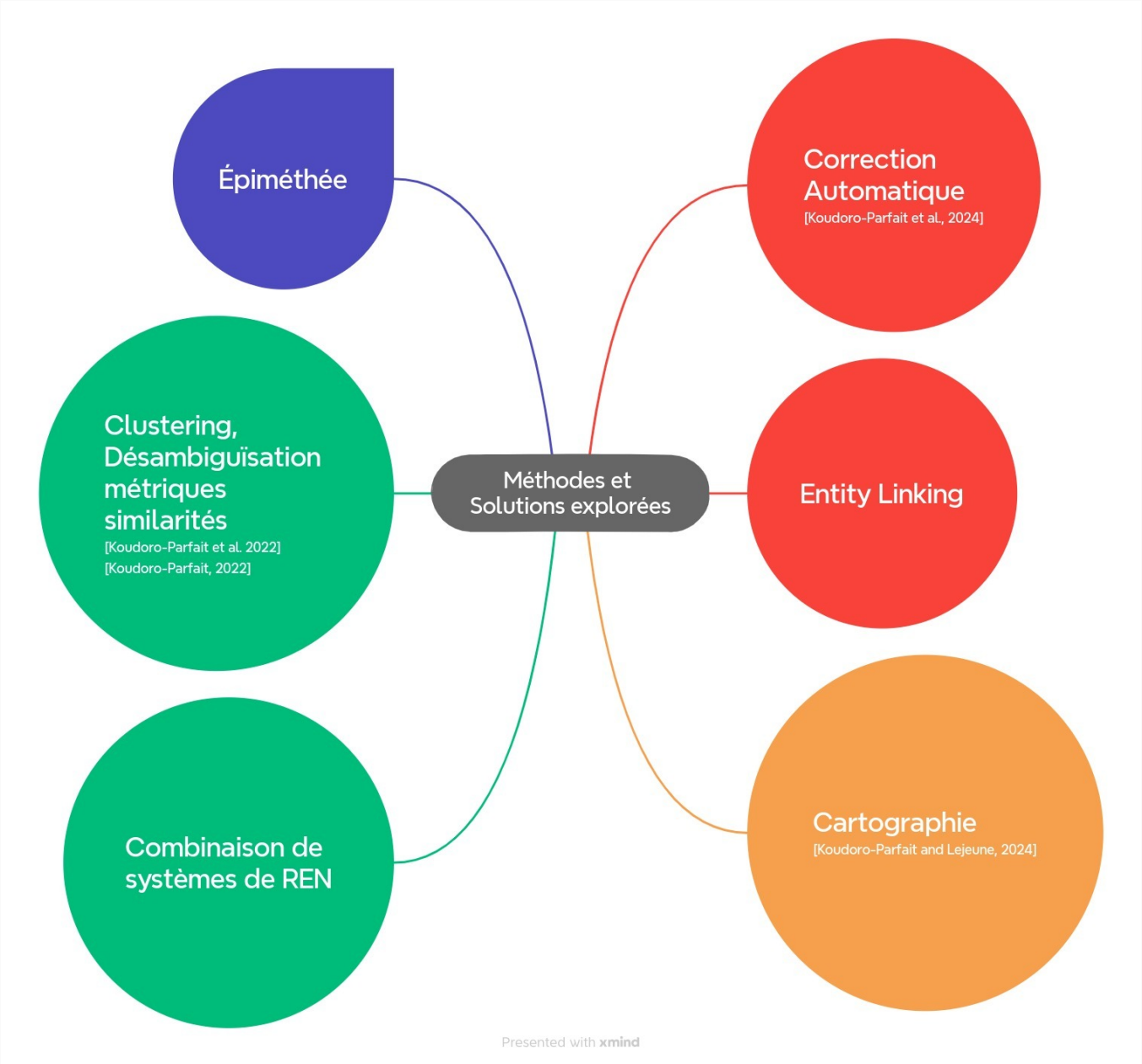
➡ Assistance utilisateur·ice·s :

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

Perspectives

- Intégrer : observations de la **fréquence** documentaire et **longueur** ;
- Liage avec des bases de données de toponymes anciens ;
- Améliorer la Géolocalisation auto. ;

Merci de votre attention !



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 5 – *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 6 – *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 7 – *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 8 – *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 9 – *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

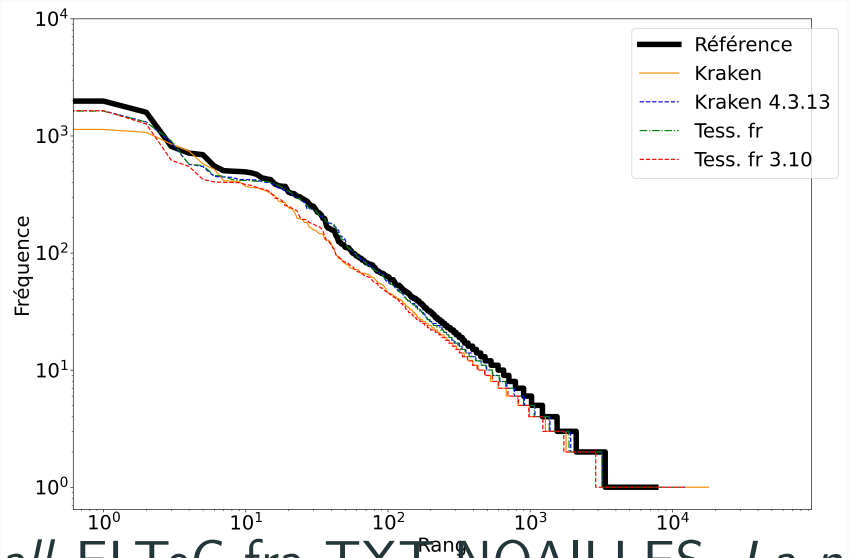


Figure 11 – *small*-ELTeC-fra TXT NOAILLES, *La nouvelle esperance*

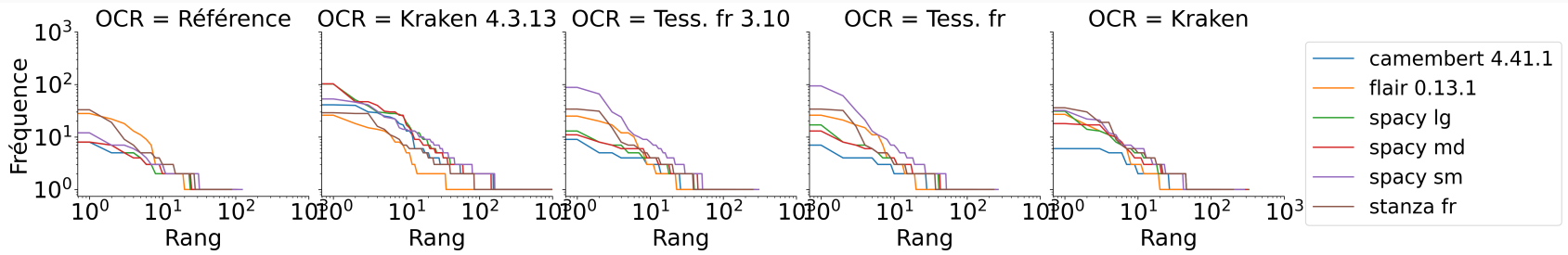


Figure 12 – *small*-ELTeC-fra REN NOAILLES, *La nouvelle esperance*