Défis et enjeux de la fouille automatique de textes anciens : de la constitution de corpus à la REN sur textes bruités

Caroline Koudoro-Parfait
parfait@uni-trier.de

19 mars 2026

Universität Trier Fachbereich III - Neuere Geschichte

Les projets présentés aujourd'hui

- Les pistes et le projet mené dans le cadre de la bourse Postdoctorale avec Damien Tricoire (Université de Trèves) ;
- OCR sur des images de basses qualité avec Toufik Boubehziz (Arts et Métiers) et Gaël Lejeune (Sorbonne Université) ;
- Les Insinuations du Châtelet de Paris avec Simon Dagenais (Université de Trèves) → accepté Conf. DH 2026 ;
- MEGV avec Marion Philipe *et al.* (Université de Genève) ;

Bourse Postdoctorale Fondation Humboldt

Relier les sociétés métropolitaines et coloniales grâce aux archives : les défis numériques de l'extraction d'informations dans la recherche socio-historique

mots clés : Archives, HTR, Données bruitées, REN, Analyse de réseaux

- Université de Trèves, Centre des humanités numériques ;
- Damien Tricoire, Historien moderniste ;

✎ site web du projet : `https://carolinekparfait.github.io/`

Premières pistes

☞ Les mobilités féminines entre les Antilles Française – Martinique, Guadeloupe et Guyane – et la métropole aux 17ème et 18ème siècle.

☞ Méthodologie :

- Inventorier des travaux d'historiens existants → de nombreux échanges de mails
- Premiers pas dans les archives

Premiers pas dans les archives

Constats :

- **Nombreux projets** disponibles sur les lieux des archives → **non numérisés**, **non interrogeables** ;
- les associations de généalogistes : **bases de données** énormes → **non publiques**
- Interrogation du catalogue (*par ex.* AN) :
  - nécessite connaissance **Regex**
  - chronophage
- Salles virtuelles → documents **1 par 1**
- demandes de lots possibles mais :
  - nécessitent des **conventions** parfois lourdes à mettre en place (ANOM)
  - les images numérisées **n'appartiennent pas toujours** aux services d'archives qui les mets en ligne (ANOM)
- Les agents des archives **toujours très heureux** de nous renseigner

Premiers pas dans les archives

Les projets :

➢ AD Loire Atlantique :

- les *Glanes Antillaises dans le notariat Nantais* de 1659 à 1830, dépouillement [2] entre 1988 et 2010
  ➜ dépôt numérisation :
  `https://github.com/carolineKParfait/AD_Loire-Atlantique_18122025`

➢ AD Seine Maritime :

- *Les engagés pour les antilles 1634 1715*, G. Debien
- *Les engagés pour les antilles la Rochelle*, G. Debien
- *Les femmes des premiers colons aux antilles 1635-1680*, G. Debien
- *Les engagés de Nantes 1636-1660*, J. TANGUY
  ➜ dépôt numérisation :
  `https://github.com/carolineKParfait/AD_Seine-Maritime_17032026`

* Numérisation : CZUR Aura pro

Système pour l'interrogation croisée de bases de données

➢ Inventorier les plateformes existantes :

- FranceArchives [1]
  - Renvoie vers les pages des services d'archives
  - Pas de récupération de manifests IIIF

➢ Enquête auprès d'historien·ne·s pour comprendre leurs méthodologies :

- recherches dans les archives,
- requêtages dans les bases de données.

1. https://francearchives.gouv.fr/

OCR sur des images de basse qualité

Corpus ELTeC [2] : français, allemand et espagnol

Corpus	Ouvrages	nb Mots	DPI moyen	Résumé des remarques
*small*_ELTeC-de	7	1 688 613	30.61	Majoritairement binarisé, souvent Fraktur
*small*_ELTeC-fr	11	791 952	71.71	Binarisation fréquente, DPI variable, artefacts (verso visible)
*small*_ELTeC-sp	6	465 403	36.34	Mélange binarisé / numérisé, qualité hétérogène

Table 1 – Corpus de travail

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

OCR sur des images de basse qualité

➢ Moteurs OCR pour les 3 langues :

- PaddleOCR : fr, sp, german
- Tesseract : fra, spa, deu
- Kraken : catmus-print-fondue-large.mlmode, german print.mlmodel

➢ Expé. compression images avec JPEG, PNG, RAW et BMP → JPEG meilleur rapport temps/espace stockage

Seuil minimal de qualité pour l'efficacité de l'OCR

3 expériences :

1. Phase 1. Référence = La résolution image à 100%
2. Phase 2. Vérité terrain = textes issus de ELTeC
3. Phase 3. Évaluation du bruit par l'analyse de la richesse lexicale

Seuil minimal pour l'efficacité de l'OCR ➢ Phase 1.

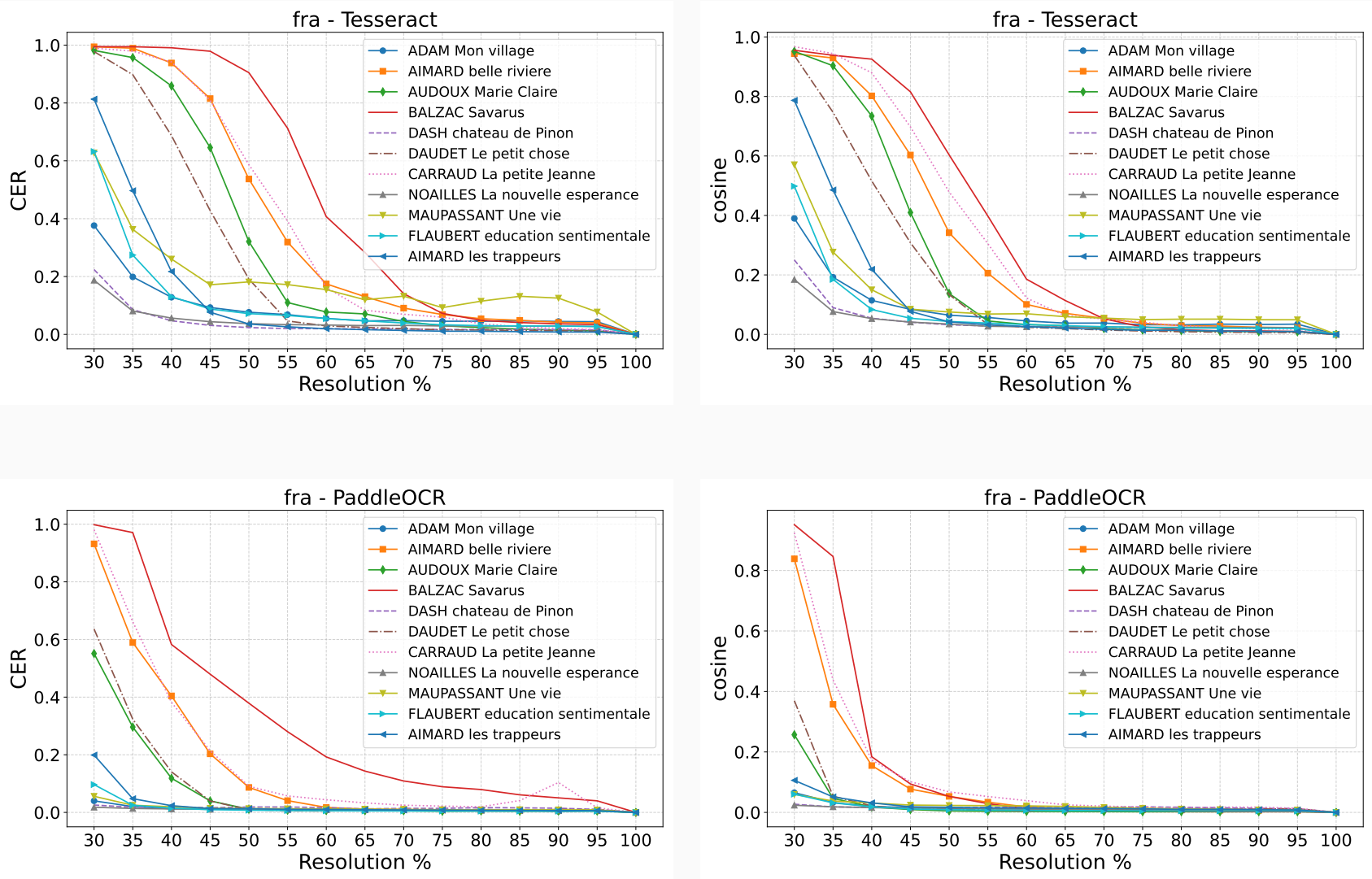
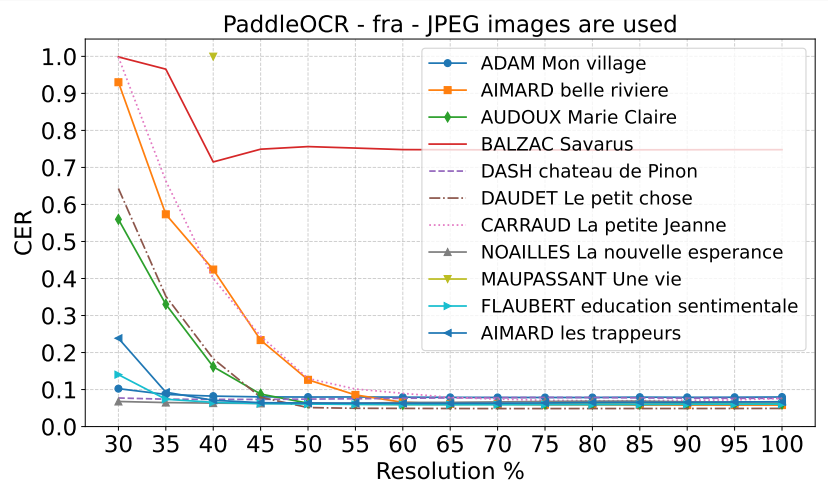
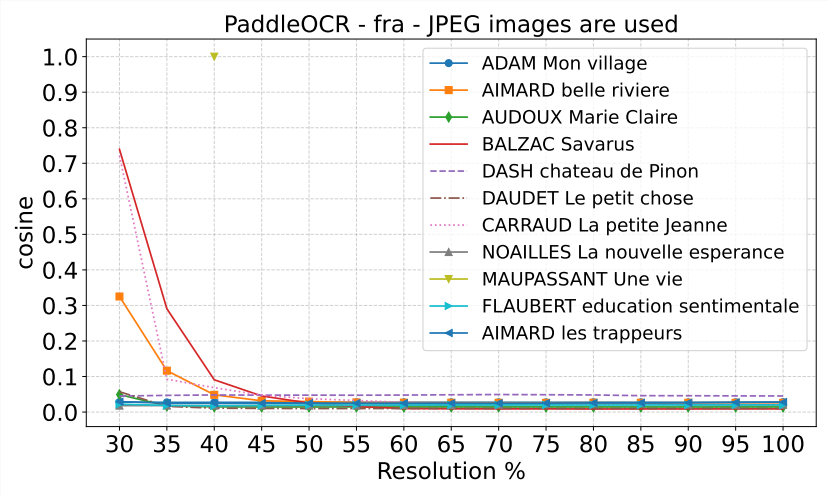


Figure 1 – Précision moyenne pondérée pour différentes résolutions `JPEG` (en %), évaluée à l'aide de la distance cosinus et du CER sur *small* ELTeC-fra.

Seuil minimal pour l'efficacité de l'OCR ➢ Phase 2.



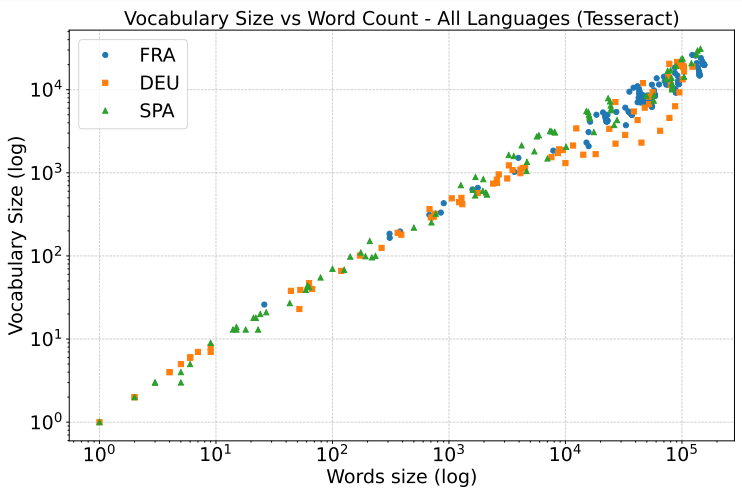
(a) CER



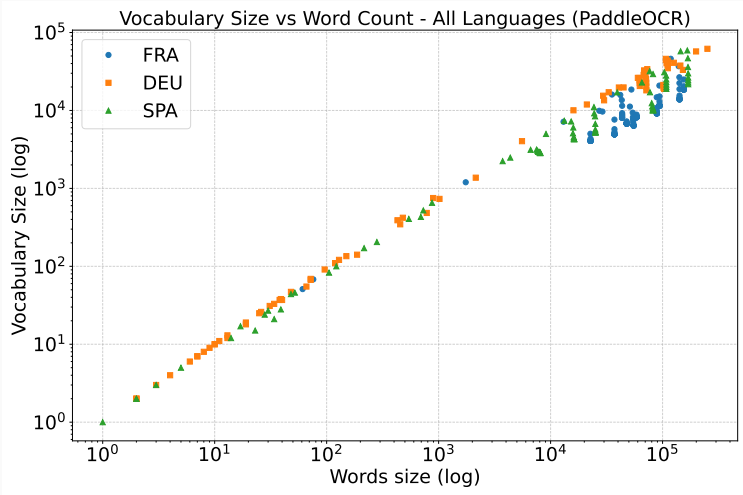
(b) cosine distance

Figure 2 – Précision de PaddleOCR sur le corpus français, comparée à la vérité terrain

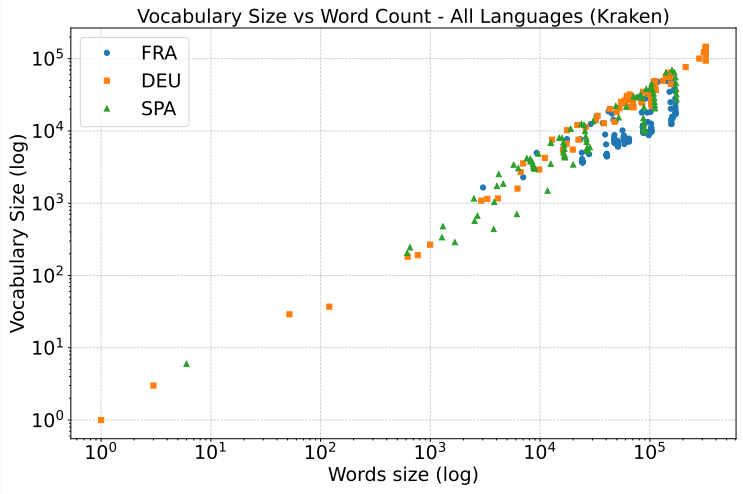
Observer la richesse lexicale ➢ Phase 3.



(a) Tesseract



(b) PaddleOCR



(c) Kraken

A retenir

- Le seuil minimal pour le corpus français qui a le DPI moyen le plus haut est entre 40%-50%
- Le seuil minimal pour le corpus allemand qui a le DPI moyen le plus bas est autour 60%
- PaddleOCR semble le moteur le plus robuste
- La compression JPEG à le meilleur rapport temps/espace de stockage

Les Insinuations du Châtelet de Paris : 1539 et 1791

Origine édit promulgué par François Ier

Nature Inventaire détaillé d'enregistrements d'actes notariés, provenant du Châtelet de Paris :

- contrats de mariage,
- testaments,
- donations

Les Insinuations du Châtelet de Paris : 1539 et 1791

Origine édit promulgué par François Ier

Nature Inventaire détaillé d'enregistrements d'actes notariés, provenant du Châtelet de Paris :

- contrats de mariage,
- testaments,
- donations

ça permet Riches sources d'informations prosopographiques :

- structures sociales,
- les réseaux,
- les professions
- les relations avec des individus de haut statut

Les Insinuations du Châtelet de Paris : 1539 et 1791

Origine édit promulgué par François Ier

Nature Inventaire détaillé d'enregistrements d'actes notariés, provenant du Châtelet de Paris :

- contrats de mariage,
- testaments,
- donations

ça permet Riches sources d'informations prosopographiques :

- structures sociales,
- les réseaux,
- les professions
- les relations avec des individus de haut statut

En chiffres L'inventaire manuscrit du XIX^{e} siècle a été numérisé par les AN :

- 89 987 enregistrements
- plus de 180 000 individus

Le corpus de travail

➢ Chantier de dématérialisation des instruments de recherche des Archives Nationales en 2010.

➢ Transcription manuelle de :

- 32 documents
- nombre de tokens : 4 120 712
- nombre de caractères : 26 217 581

➢ Notre annotation en entités nommées :

- 3 documents
- 36 000 tokens
- 1 129 entités nommées

Préparation d'une vérité terrain pour la REN

Catégories de spaCy LOC, PER, ORG, MISC

Choix pour notre cas d'usage :

- Professions = MISC
- Enseignes de magasin = LOC

  → elles servent principalement d'adresse.

Token	Annot
logé	O
à	O
...	...
l'	B-LOC
enseigne	I-LOC
du	I-LOC
Croissant	I-LOC
...	...

Table 2 – Exemple d'annotation pour une enseigne

➥ Accord inter-annotateur, **Kappa de Cohen** : 0.83

Archéologie de spaCy

Label	GT	spaCy 2.3.0			spaCy 3.8.0		
		sm	md	lg	sm	md	lg
B-PER	416	313	301	294	**119**	**114**	276
B-LOC	433	316	348	364	322	288	335
B-MISC	265	**9**	**7**	**7**	133	175	87
B-ORG	15	22	17	11	31	15	14

Table 3 – Comparison of entity-type annotations across different versions of spaCy. GT = Ground Truth.

Archéologie de spaCy

spaCy model		Cosinus		CER
Version	Size	Characters (2–3-gram)	word	
spaCy 2.3	sm	0.117	0.155	0.633
	md	0.**112**	**0.149**	**0.619**
	lg	0.118	0.164	0.626
spaCy 3.8	sm	0.312	0.55	0.784
	md	0.**319**	**0.693**	**0.795**
	lg	0.170	0.260	0.634

Table 4 – Results of spaCy models according to the cosine metric and the CER.

Un corpus sur les violences dans les territoires coloniaux

➢ Projet « Masculinités esclavagistes : Genre et Violence dans la Caraïbe française (XVIII^{e}siècle) » (MEGV) [3], dirigé par Marie Houllemare (Université de Genève).

➢ Transcription/Correction manuelle de 11 sous-corpus :

Corpus	Documents	Tokens	Caractères
anom_col-e-soldats-1b	51	8 178	49 777
anom_col-e-soldats-1c	46	5 754	34 513
anom_col-e-soldats-2	61	7 113	42 785
anom_col-e-soldats-3	67	7 955	48 247
anom_col-e-soldats_1a	50	6 279	38 067
anom_col-e-vrac	105	2 1216	125 760
anom_col-e-vrac2	68	1 1915	71 207
an_corresp-fougeu-conflans	49	8 151	45 947
an_corresp-fougeu-conflans-2	58	10 008	57 607
an_corresp-prov-roch	66	18 826	112 434
an_corresp-prov-roch-2	16	2 798	16 736
Total	**637**	**108 193**	**630 080**

Table 5 – Statistiques du corpus MEGV

3. https://www.unige.ch/masculinitesesclavagistes/

Archéologie de spaCy et stanza

➢ Entraînement du modèle FoNDUE-GD-v2

➢ Annotation des entités nommées *silver standard* :

versions moteurs de REN	spaCy 2.3.5	spaCy 3.8.11	Stanza 1.2.1	Stanza 1.11.0
versions modèles de langue	fr_core_news_lg-2.3.0	fr_core_news_lg-3.8.0	fr Wikiner	fr wikinergold_charlm
date maj	juin 2020	sept. 2024	oct. 2017	déc. 2025
Model size	545 MB	545 MB	NC	62.9 MB
Sources	• UD French Sequoia v2.5 [1] • WikiNER • OSCAR (Common Crawl)[3] • Wikipedia	• UD French Sequoia v2.8 • WikiNER • Explosion fastText Vectors (cbow, OSCAR + Wikipedia) • spaCy lookups data (Explosion)	WikiNER (silver standard)	WikiNER (gold standard)
F_1-score REN	85.63	84.18	92.9	NC
Catégories	LOC, PER, ORG, MISC	→ *idem*	→ *idem*	→ *idem*

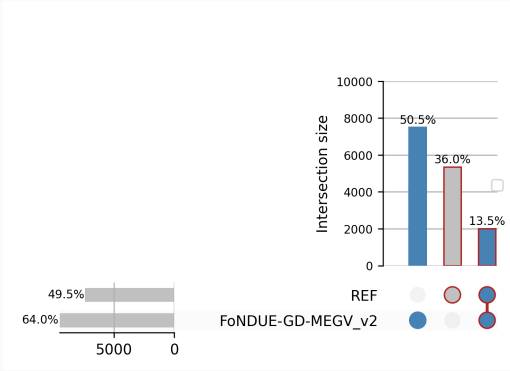
Table 6 – Description des modèles de REN évalués

Analyses qualitatives

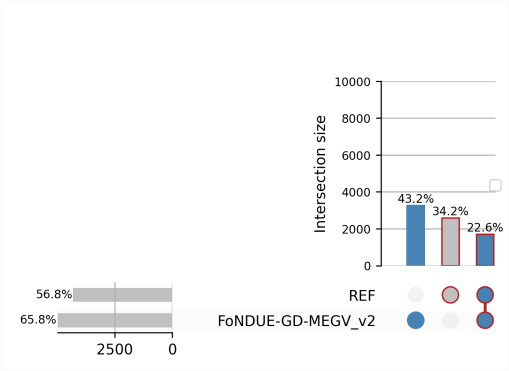
Version	*Contexte* + ***EN***	`lg-2.3.0`	`lg-3.8.0`	`WikiNER S-std`	`WikiNER S-gld`
Réf.	***Croiriéz*** *vous, [...]*	[e]PER	[e]PER	-	-
FoNDUE-GD-v2	***Croiriez*** *vous [...]*	[e]PER	[e]PER	-	-
Réf.	***Madame la comtesse de Conflans*** *[...]*	[b]PER	[a]()	[c]PER / **Conflans** LOC	[b]PER
FoNDUE-GD-v2	***la Comtesse de Conflai****[...]*	[b]PER	[a]()	MISC	[b]PER
Réf.	*[...] Adjoint de l'Administration Coloniale* ***Ed. Poncet***	[f]PER	PER	[b]PER	[d]ORG
FoNDUE-GD-v2	*[...] Adjoint de l' Aidministration Coloniale.* ***Ed Poniet***	[f]LOC	()	[b]PER	[d]PER

Table 7 – Exemples des cas les plus fréquents d'erreurs dans la REN sur données bruitées.

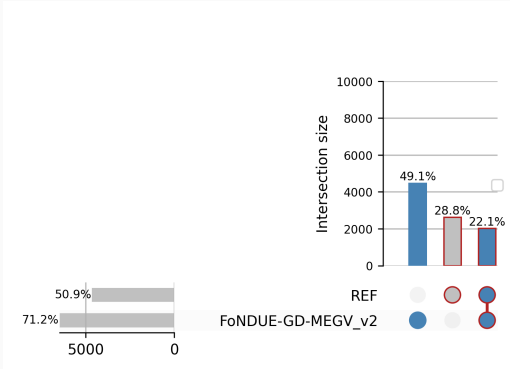
Analyses quantitatives : intersections



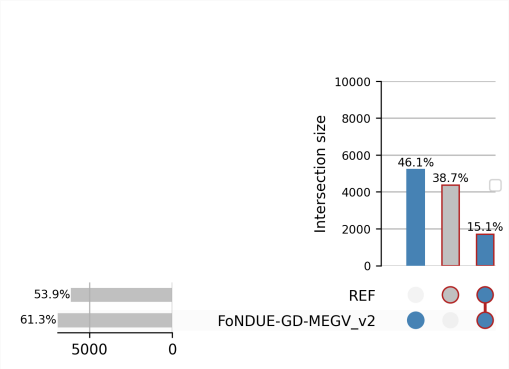
(a) spaCy lg-2.3.0



(b) Stanza WikiNER



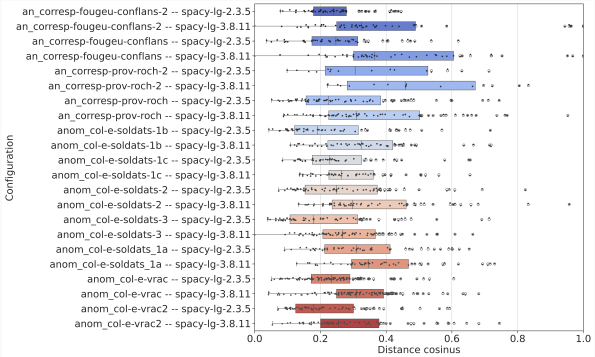
(c) spaCy lg-3.8.0



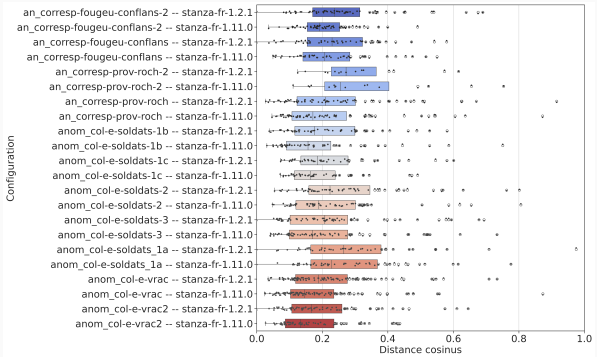
(d) Stanza `wikinergold_charlm`

Figure 4 – Représentation des intersections pour tout le corpus pour toutes les étiquettes avec les systèmes spaCy et Stanza dans

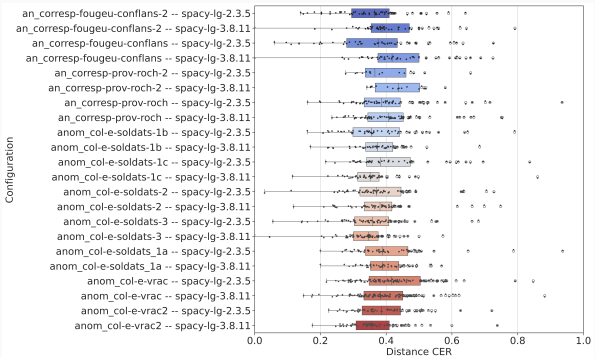
Analyses automatiques : CER & Cosinus



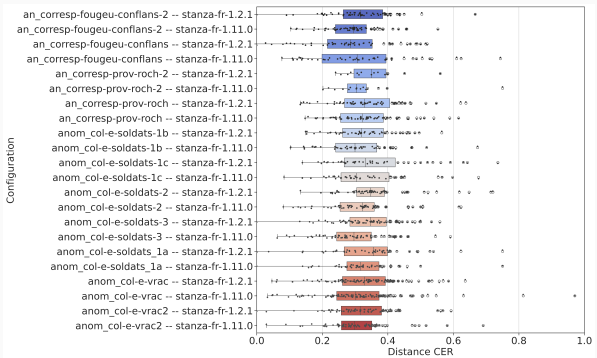
(a) 2 versions de SPACY distance cosinus



(b) 2 versions de STANZA distance cosinus



(c) 2 versions de SPACY CER



(d) 2 versions de STANZA CER

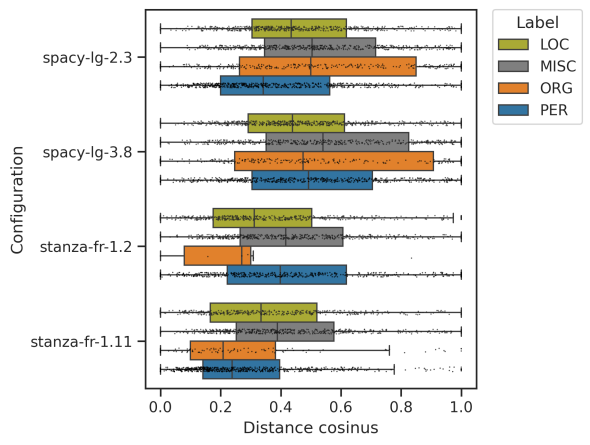
Figure 5 – Distance cosinus et CER calculés sur chaque sous-corpus

Analyses automatiques : CER & Cosinus

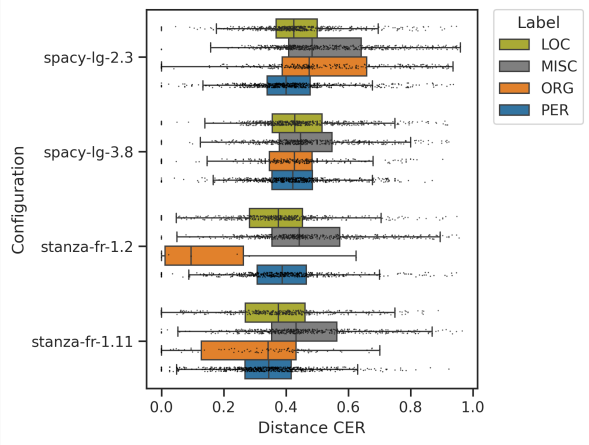
→ spaCy 2.3.0 plus performant que 3.8.0

→ Stanza `wikinergold_charlm` globalement plus adapté à notre cas d'usage

Analyses automatiques : CER & Cosinus sur les catégories



(a) Distance cosinus



(b) CER

Figure 6 – Distance cosinus calculée pour les catégories PER, LOC, ORG et MISC sur l'entièreté du corpus pour les modèles SPACY 2.3.0 et 3.8.0 et STANZA 1.2.0 et 1.11.0.

→ Les catégories MISC et ORG sont plus susceptibles d'erreurs

Contributions

- Numérisation de corpus produits par des historien·ne·s
- Images **basse qualité** peuvent produire des **OCR correctes**
- **Dernière version** d'un système de REN n'est **pas forcément la plus adaptée**

Perspectives

- Numérisation de corpus produits par des historien·ne·s → **emploi numérique** : bases de données
- Images **basse qualité** → Évaluer des tâches de TAL en aval
- Archéologie des outils de REN → Trouver les meilleures **combinaisons** pour les cas d'usage

Références

[1] Marie Candito et Djamé Seddah. **“Le corpus Sequoia : annotation syntaxique et exploitation pour l’adaptation d’analyseur par pont lexical (The Sequoia Corpus : Syntactic Annotation and Use for a Parser Lexical Domain Adaptation Method) [in French]”.** In : *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*. Sous la dir. de Georges Antoniadis, Hervé Blanchon et Gilles Sérasset. Grenoble, France : ATALA/AFCP, juin 2012, p. 321-334. URL : https://aclanthology.org/F12-2024/.

Références ii

[2] Françoise Loré et Jean-Marie Loré. ***Les Premières glanes antillaises dans le notariat nantais de 1673 à 1747***. 15 p. : ill. ; 30 cm. 1988.

[3] Pedro Javier Ortiz Suárez et al. **"Establishing a New State-of-the-Art for French Named Entity Recognition"**. eng. In : *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Sous la dir. de Nicoletta Calzolari et al. Marseille, France : European Language Resources Association, mai 2020, p. 4631-4638. ISBN : 979-10-95546-34-4. URL : https://aclanthology.org/2020.lrec-1.569/.