CORPUS

Corpus

26 | 2025
Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques

Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ?

Ljudmila Petkovic, Caroline Koudoro-Parfait, Marie-Sophie Desmarest and Gaël Lejeune

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/9059
DOI: 10.4000/1364s
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Ljudmila Petkovic, Caroline Koudoro-Parfait, Marie-Sophie Desmarest and Gaël Lejeune, "Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ?", *Corpus* [Online], 26 | 2025, Online since 14 January 2025, connection on 31 January 2025. URL: http://journals.openedition.org/corpus/9059 ; DOI: https://doi.org/10.4000/1364s

Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ?

Ljudmila Petkovic, Caroline Koudoro-Parfait, Marie-Sophie Desmarest and Gaël Lejeune

1. Contexte

1 Face au volume croissant des données issues de la numérisation et de la reconnaissance optique de caractères (OCR) émergent des problématiques relatives à la qualité de ces données et à leur exploitabilité scientifique étant donné le bruit ou les « contaminations » (terme repris de Hamdi *et al.* 2022) que l'on trouve dans les transcriptions OCR. Le bruit désigne alors toutes les erreurs produites par le système OCR (insertion, suppression et substitution de caractères). En aval, les modèles entraînés sur des données textuelles correctement orthographiées (Eshel *et al.* 2017), peinent à maintenir leurs performances sur ces données moins standardisées. Une des manières de contourner ce problème est de corriger les erreurs pour améliorer la qualité intrinsèque des corpus. Néanmoins, si certaines interférences des dispositifs d'OCR sont systématiques (Stanislawek *et al.* 2019), lorsqu'elles sont singulières, la correction devient difficile. En outre, il est montré que la correction produisait de nouvelles erreurs (phénomène de sur-correction) provenant soit des algorithmes de basse complexité (Huynh *et al.* 2020)[1], soit des grands modèles de langue (Boros *et al.* 2024). Ceci motive la recherche sur le traitement du bruit issu de la phase d'OCR et son influence sur la Reconnaissance d'Entités Nommées (REN), dans le cadre de laquelle nous proposons une nouvelle approche pour exploiter les sorties d'OCR bruitées.

2 Nos analyses s'appuient sur deux corpus pour lesquels nous comparons les entités nommées reconnues par spaCy (Montani *et al.* 2023)[2] sur les versions de référence d'une part et des versions OCR, corrigées automatiquement ou non.

3 Dans le cadre applicatif qui nous intéresse, nous nous positionnons du point de vue d'un·e chercheur·se ne disposant pas des données standards, corrigées et annotées manuellement (*Gold Standard*), mais des données réelles (éventuellement imparfaites) que l'on considère comme référence, et qui sont annotées automatiquement (*Silver Standard*)[3]. En l'absence de données *Gold*, les données *Silver* permettent d'évaluer efficacement s'il est nécessaire de corriger en détails un corpus en amont de son annotation automatique. Nous proposons des analyses manuelles et automatiques, qui reprendront les expériences effectuées par Koudoro-Parfait *et al.* (2022) avec l'outil NERVAL[4], ainsi que les expérimentations menées par Petkovic *et al.* (2022) avec l'outil JamSpell[5]. Notre objectif est d'établir (i) une typologie des contaminations d'OCR et (ii) une typologie des erreurs de la correction automatique. Nous montrerons qu'une partie imporatnte du bruit n'est en vérité pas imputable à l'OCR mais bien à la REN.

4 Nous dressons dans la section suivante un état de l'art sur influence de l'OCR et de la correction automatique d'OCR sur les traitements réalisés en aval. La section 3 décrit les corpus utilisés dans cette étude. Nous analysons la relation entre la qualité de l'OCR, de sa correction et la tâche de REN dans les sections 4 et 5. Nous proposons enfin dans la section 6 d'exploiter de multiples versions bruitées pour aider à éliminer les entités nommées fautives (faux positifs) sans supervision (sans *Gold Standard* donc). Nous présentons des résultats prometteurs de filtrage automatique qui nous permettent de tracer de nouvelles perspectives de recherche sur l'utilisation de données textuelles bruitées[6].

2. État de l'art

5 En vue de dépasser les limites de l'exploitation de données OCR, la communauté TAL s'est penchée sur la correction des erreurs d'OCR (Sagot & Gábor 2014). Selon Kukich (1992), l'origine des questions concernant la correction automatique de texte remonte à Damerau (1964) qui a analysé ce phénomène sur les textes artificiellement bruités. Ce dernier a constaté que 80 % d'erreurs de tous les mots mal orthographiés contenaient une seule instance d'une des quatre erreurs possibles (insertion, suppression, substitution et transposition) tout en implémentant le premier algorithme de correction orthographique (distance minimale d'édition). Ce constat, bien qu'il ne s'applique pas directement aux erreurs issues des textes océrisés, est à entendre comme une première tentative de cerner la problématique de la robustesse de la correction automatique face à des variations textuelles : bruit artificiel, contamination OCR ou erreurs d'orthographe d'internautes.

6 Malgré de nombreuses avancées dans ce domaine jusqu'à présent, cette pratique reste toujours questionnée (Nguyen *et al.* 2021). Même s'il n'existe pas une classification unanimement établie (Bassil & Alwani 2012, Edwards 2016, Nguyen *et al.* 2020, Nguyen *et al.* 2021), nous avons identifié trois méthodes courantes : fondées sur des lexiques, sur des modèles de langue, et sur un apprentissage automatique spécifique (Hládek *et al.* 2020, Petkovic *et al.* 2022).

7 Oger *et al.* (2012) proposent de classer le bruit OCR en deux catégories : (i) les erreurs de non-mots (angl. *non-word errors*) qui représentent les termes absents d'un lexique, p. ex. « Devonshire »[7] → « Dconshire » et (ii) les erreurs de mots réels (angl. *real-word errors*) ou erreurs sémantiques (angl. *semantic/context-sensitive errors*) (Azmi *et al.* 2019), p. ex. « Gélons »[8] → « Gelons », terme existant dans la langue (ou *a minima* dans des lexiques de référence) mais incorrect dans le contexte donné. Cette seconde catégorie est plus rare, et souvent liée à la sur-correction automatique, p. ex. « M. Eyssette » → « M. Assiette », où un mot connu remplace un mot inconnu.

8 Un certain nombre de membres de la communauté TAL s'intéresse à la manière d'évaluer l'impact des erreurs d'OCR sur la REN et les tâches en aval (Chiron *et al.* 2017, Hamdi *et al.* 2020, Tual 2023, Van Strien *et al.* 2020) pour répondre à deux questions : (i) s'il est bien nécessaire de corriger les erreurs de l'OCR et (ii) si l'absence de la correction détériore véritablement les performances des outils de TAL en aval (segmenteurs, outil de REN etc.).

9 La REN est un moyen efficace d'améliorer l'accès aux informations contenues dans de vastes corpus (Van Strien *et al.* 2020). D'ailleurs, Chiron *et al.* (2017) ont montré qu'un nombre important de requêtes sur Gallica étaient affectées par des termes mal océrisés et absents des usuels. Koudoro-Parfait *et al.* (2021) ont montré que les systèmes de REN présentaient une certaine robustesse face à la variabilité, au bruitage des données. Certaines entités dont la forme est donc « contaminée » sont en effet reconnues par des outils tels que spaCy ou Stanza, p. ex. quand la forme « Iuda » apparaît dans l'OCR au lieu de « India ». Des tâches comme le *topic modeling* sont également perturbées par le bruitage (Van Strien *et al.* 2020). Enfin, Evershed et Fitch (2014) soulignent l'importance de la correction automatique (sur un corpus de journaux), qui a réduit de plus de 50 % le nombre de faux positifs (FP) en Recherche d'Information grâce à une réduction du *Word Error Rate*. Alex *et al.* (2012) montrent par ailleurs que des corrections ciblées[9] contribuent à une amélioration significative des résultats de la REN. Notre contribution s'attachera à déterminer si la correction de l'OCR permet d'améliorer significativement les résultats de la tâche de REN.

3. Les corpus utilisés

3.1. La Très Grande Bibliothèque (TGB) : le corpus d'entraînement

10 La TGB[10] est une ressource documentaire qui comprend des œuvres en français, issues des collections Gallica transcrites par OCR. Elle comprend 128 441 textes en XML-TEI[11] sur différentes thématiques (littérature, histoire, droit, philosophie, etc.). Nous avons extrait 10 œuvres de la catégorie « Littérature française », pour constituer notre corpus dont la composition et la densité[12] en EN sont présentées dans le tableau 1. Les textes sont produits par OCR et non corrigés. La qualité de l'OCR est hétérogène, le corpus provenant de plusieurs campagnes avec différents outils.

Tableau 1. Statistiques sur le corpus TGB, la densité en EN est le pourcentage de tokens couverts par celles-ci

Ouvrage	Auteur	Année	Pages	Tokens	spaCy-lg	Densité EN (%)

La princesse Pallianci	Bazancourt, C. L.	1852	340	36 423	**1 438**	3.95
Meryem, scènes de la vie algérienne. Marcel	Bentégeat, C. P.	1863	360	85 077	**3 768**	4.43
Wilmina, ou L'enfant des Apennins	Girard de Caudemberg, L.	1820	242	36 218	**1 533**	4.23
Les fourmis du parc de Versailles raisonnant ensemble dans leurs fourmilières	Lambert, C.	1803	72	10 173	**244**	2.40
Œuvres complètes de Pierre Loti	Loti, P.	1893-1911	588	133 129	**4 153**	3.12
La confession d'un enfant du siècle / Alfred de Musset ; avec un portrait... par Eugène Lami...	Musset, A. de	1879	494	92 140	**2 478**	2.69
Le Parnasse envahi, petit poème allégorique au sujet du sacre de S. M. Charles X.	Rullier, E.	1825	71	10 261	**830**	8.09
La Comtesse de Rudolstadt	Sand, G.	1861	340	102 423	**3 653**	3.57
Diégarias, drame en 5 actes et en vers	Séjour, V.	1844	38	18 603	**3 869**	20.8
Le département de l'Oise : Compiègne et Marat, fragment historique	Sorel, A.	1865	19	6 277	**315**	5.02

3.2. European Literary Text Collection (ELTeC) : le corpus de test

11 Notre second corpus est extrait de ELTeC, créé dans le cadre de l'action COST *Distant Reading for European Literary History*[13]. La majorité des textes sont de très bonne qualité si l'on se réfère aux règles suivies pour la composition du corpus[14].

Tableau 2. Statistiques sur le corpus ELTeC français

Ouvrage	**Auteur**	**Année**	**Pages**	**Tokens**	**spaCy-lg**	**Densité EN (%)**
Mon village	Adam, J.	1860	200	20 938	**906**	4.33
Les trappeurs de l'Arkansas	Aimard, G.	1858	450	91 119	**2 149**	2.36
La Belle rivière	Aimard, G.	1894	339	137 392	**3 686**	2.68
Marie-Claire	Audoux, M.	1925	120	35 780	**835**	2.33

Albert Savarus. Une fille d'Ève	Balzac, H. de	1853	60	79 924	**2 121**	2.65
La petite Jeanne	Carraud, Z.	1884	220	53 212	**2 294**	4.31
Le château de Pinon, vol. I	Dash, G. A.	1844	332	44 246	**2 162**	4.89
Le petit chose	Daudet, A.	1868	292	86 482	**3 158**	3.65
L'Éducation sentimentale	Flaubert, G.	1880	520	150 494	**6 313**	4.19
Une vie	Maupassant, G. de	1883	337	75 745	**1 725**	2.28
La nouvelle espérance	Noailles, A. de	1903	325	54 272	**1 544**	2.84

12 Pour chacun des deux corpus, TGB et ELTeC, nous disposons donc d'une version de référence (ci-après Réf.), et de deux versions OCR (ci-après v-OCR) produites avec Kraken et Tesseract pour chaque texte[15]. Dans le cas de la TGB, en l'absence du corpus *Gold Standard*, le terme « référence » est à entendre comme une approximation de la vérité terrain, car elle n'est pas exempte de bruit d'OCR à l'origine des campagnes de numérisation. D'où la contrainte de considérer la TGB presqu'au même niveau qu'ELTeC du point de vue de la qualité de l'OCR.

4. Observations sur les entités nommées bruitées

4.1. Les outils d'OCR et de REN utilisés

13 Pour obtenir les versions OCR, nous avons utilisé deux modèles : Tesseract (Smith 2007) avec le modèle français entraîné sur des données Google (LSTM tessdata_best) et Kraken (Kiessling 2019) avec le modèle de base. Pour la REN nous nous sommes appuyés sur spaCy qui utilise les plongements *Bloom* et un réseau de neurones convolutif[16]. Trois modèles de langue *small*, *medium* et *large* sont proposés (cf. tableau 3). Nous avons privilégié l'usage du modèle français *large* (spaCy-lg), plus efficace.

Tableau 3. Description des modèles de langue de spaCy

	Plongements	**Taille jeu de données**	**Données d'entraînement**
small		15 MB	UD French Sequoia v2.8, WikiNER, spaCy lookups data
medium	✔	43 MB	UD French Sequoia v2.8, WikiNER, spaCy lookups data, Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia)
large	✔	545 MB	UD French Sequoia v2.8, WikiNER, spaCy lookups data, Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia)

4.2. Évaluation de la REN sur données bruitées : une typologie adaptée

14 En l'absence de *Gold Standard* de REN sur des versions bruitées multiples, nous avons constitué un *Silver Standard* à partir de l'annotation automatique avec spaCy des données de référence (cf. tableaux 1 et 2). Notre *Silver Standard* comporte potentiellement du bruit, étant donné que les entités extraites n'ont pas été manuellement vérifiées, mais ceci permet de montrer à quel point le bruit dans les entités des v-OCR est véritablement dû à la qualité de celles-ci. Les entités détectées dans la Réf. permettent de juger si les entités détectées dans la v-OCR sont massivement bruitées ou si ce bruit n'était pas déjà présent dans la version de Réf., et donc imputable à la REN. Pour justifier l'utilisation du *Silver Standard*, nous avons calculé le F-score entre les annotations manuelles (*Gold*)[17] et automatiques (*Silver*) avec l'outil NERVAL, en obtenant le score de 0.67 entre l'accord inter-annotateur (annotation obtenue par vote majoritaire)[18] et la référence[19]. Afin de bien définir l'évaluation de la REN sur données bruitées, il faut tenir compte des phénomènes suivants dans v-OCR :

- du bruit au niveau des blocs : du texte absent de Réf. (notes de bas de pages)
- du bruit au niveau des mots : les contaminations d'OCR
- du silence également, certaines pages de texte ayant été peu ou pas du tout transcrites.

15 Koudoro-Parfait *et al.* (2021) identifient deux types principaux d'erreurs : (i) des FP liés au bruit de l'OCR (*non-word errors*), soit les EN récupérées dans l'OCR mais pas dans la Réf. et (ii) des FP qui sont en fait des entités dont la forme est contaminée (plus proche des *context-sensitive errors* évoquées plus haut), mais dont le contenu renvoie à une vraie EN. Conséquemment, une partie des FP sont en fait des faux FP, p. ex. « Mlorlincourtl » est une forme contaminée de « Morlincourt », il faut le considérer donc comme un vrai positif, car son contenu se réfère à une vraie EN. Nous proposons une traduction et une réorganisation de la typologie classique positif/négatif, selon que les cas présentés participent du bruit ou du silence dans la REN. Cette catégorisation vise à ce que l'évaluation dans le cas d'usage d'un *Silver Standard* soit plus fine que les critères binaires classiques de l'évaluation propres à un *Gold.* Celle-ci engendre trois sous-catégories en plus des FP, les vrais positifs (VP, entités correctes), faux négatifs (FN, entités manquantes ou silence) et enfin vrais négatifs (VN, non-entités). Pour bien expliciter la limite de ces catégories dans un contexte bruité il faut garder en mémoire qu'une entité contaminée comme « Rh6ne » pour « Rhône » est un FP strictement car « Rh6ne » n'existe dans le texte de Réf. mais c'est un VP, car il est bien une variante contaminée de « Rhône ». La typologie propose donc d'ajouter des catégories adaptées au contexte bruité[20] pour mieux évaluer bruit et silence que nous présentons ci-après.

- **CAS DE SOUS-ÉVALUATION DU BRUIT ET DU SILENCE DE LA REN**
  Faux VP : EN détectées à tort dans les deux versions
  Faux VN : EN manquantes dans les deux versions
- **CAS DE SUR-ÉVALUATION DU BRUIT ET DU SILENCE DE LA REN**
  Faux FP : EN dans v-OCR mais pas dans Réf.[21]
  Faux FN : EN détectées à tort dans Réf.

16 Nous exemplifions cette catégorisation immédiatement après sous la forme d'ensembles GOLD (EN obtenues manuellement), SILVER (EN obtenues sur la référence) et OCR. Pour la lisibilité de l'exemple à cette étape, on montre l'entité hors contexte, en s'appuyant sur une liste d'occurrences. Néanmoins, pour chaque EN ayant subi une

contamination OCR nous avons bien observé le contexte d'apparition et nous l'avons comparé avec la Réf. pour nous assurer qu'il s'agit bien de la même EN.

- **GOLD** : {"département du Cher", "Languedoc", "Marseille", "Rhône", "Lyon", "La grande route"}
- **SILVER** : {"département du Cher", "Languedoc", "Eysette", "Marseille", "Rhône", Annou}
- **OCR** : {"tCher", "Languedoc", "Marseille", "Eysette", "Rh6ne", "Rhone", "Lyon"}

Tableau 4. Comparaison des résultats GOLD/SILVER et OCR

	GOLD *vs.* SILVER	**GOLD *vs.* OCR**	**SILVER *vs.* OCR**
VP	• *département du Cher, Languedoc, Marseille, Rhône*	• *Languedoc, Marseille, Lyon*	• *Languedoc, Marseille* • ***Eysette (Faux VP)***
FP	• *Eysette, Annou*	• *Eysette* • ***tCher, Rh6ne, Rhone (Faux FP)***	• ***tCher, Rh6ne, Rhone (Faux FP)***
FN	• Lyon, *La grande route*	• département du Cher, *La grande route*	• ***Annou (Faux FN)*** • département du Cher
VN	• –	• –	• ***La grande route (Faux VN)***[22]

17 Le tableau 4 montre les limites d'une évaluation classique sur des données bruitées. Nous proposons donc, en suivant la typologie évoquée plus haut, une évaluation plus fine qui tient compte des particularités de ce cas d'étude. En effet, dans une évaluation GOLD *vs.* SILVER, « bleu » et « aller » sont des FP. Alors que dans l'évaluation SILVER *vs.* OCR « bleu » est un faux VP car c'est une entité erronée. Le verbe « aller » quant à lui est un faux FN car il a été récupéré de manière fautive par l'outil de REN sur la Réf. mais pas sur l'OCR. Dans une évaluation portant sur des données bruitées par rapport à un GOLD ou un SILVER, « Parisl », « Rh6ne », « Rhone », et « Lile » sont des FP. Néanmoins, ce sont des formes contaminées des VP « Paris », « Rhône », et « Lille », que le regard humain peut identifier comme tel. « Mont-parnasse » serait un faux VN car en fait il faudrait le rapprocher de la forme correcte Montparnasse.

5. Correction automatique des textes bruités et REN

5.1. JamSpell : un outil pour la correction automatique

18 Pour corriger les transcriptions OCR, nous avons utilisé la version 0.0.12 de l'outil JamSpell basé sur un modèle de langue trigramme de mots et utilisant l'alphabet de la langue[23]. Ce choix d'utilisation dans cette étude est dû à la prise en compte du contexte lors de la correction, sa rapidité de traitement de textes, ainsi qu'à son utilisation relativement facile, bien qu'il existe des outils de correction automatique d'OCR basés sur des architectures de l'état de l'art ou nécessitant l'utilisation du GPU. Une partie des fonctionnalités ainsi que les modèles de langue pour le français (Jspll-fr) sont accessibles gratuitement sur le web[24]. Nous avons entraîné 40 % de la partie du corpus

ELTeC non utilisée pour cette étude pour entraîner un nouveau modèle JamSpell (Jspll-ELTeC).

5.2. Évaluer les sorties de REN sur données bruitées

5.2.1. Typologie pour l'évaluation manuelle

19 Les v-OCR et les Réf. ont été annotées en EN spaCy-lg, les EN de la Réf. constituent le *Silver Standard* (pseudo-vérité de terrain, soit une annotation automatique qui nous sert de référence en l'absence de *Gold*). Le tableau 5 montre des exemples de variations produites par le correcteur automatique sur des contextes contenant l'EN « Meunet-sur-Vatan », où nous pouvons constater que les différentes contaminations de cette EN ne sont pas toutes détectées par le modèle de REN. Nous voyons également que la correction automatique n'améliore systématiquement ni (i) la qualité de l'entrée avec par exemple « Meunet » incorrectement corrigé en « Meuret » par JamSpell ni (ii) la qualité de la sortie puisque la sur-correction de « Yatan » en « Satan/Avant » sur les versions Kraken aboutit à ce que l'entité « Neunet-sur-Yatan » n'est même pas partiellement reconnue.

Tableau 5. Influence de la correction de l'OCR sur la REN avec spaCy-lg. (*La petite Jeanne*, Carraud)

Version	**Contexte**	**spaCy-lg**
Référence	*à l'assemblée de Meunet-sur-Vatan ;*	*Meunet-sur-Vatan*
Kraken "brut"	***a** l'assem**hl6e**' de **N**eunet-sur- **Y**atan' '*	***Y**atan*
Kraken Jspll-fr	*a l'assembl**6e**' de **N**euner-sur- **S**atan' ;*	⌀
Kraken ELTeC-fr	*a l'assemhl6e' de **N**eunet-sur-**Avant'** ;*	⌀
Tess. fr. "brut"	*à l'assemblée' de Meunet-sur-| Vatan* ;*	*Meunet-sur-|*
Tess. fr. Jspll-fr	*à l'assemblée' de Meu**r**et-sur- | Vatan* ;*	⌀
Tess. fr. ELTeC-fr	*à l'assemblée' de Meunet-sur-| Vatan* ;*	*Meunet-sur-_*

20 Afin d'aller plus loin, nous proposons dans le tableau 6 une typologie des erreurs et réussites de la correction automatique avec des exemples concrets de leur influence sur la REN. Cette typologie est principalement motivée par la volonté de distinguer les cas où il y a une correction justifiée et correcte (MOBC), une correction manquante (MONC), des corrections erronées et sur-corrections (MOMC, BOIC) qui sont particulièrement problématiques pour le cas qui nous intéresse, par exemple « Arbres » pour « Brûlés » et « Martincourt » pour « Morlincourt ».

Tableau 6. Typologie de l'influence de la correction de l'OCR sur la REN pour les configurations avec spaCy-lg. (*Belle rivière*, Aimard et *Mon village*, Adam). Les erreurs sont mises en gras

		Version	**Contexte**	**spaCy-lg**

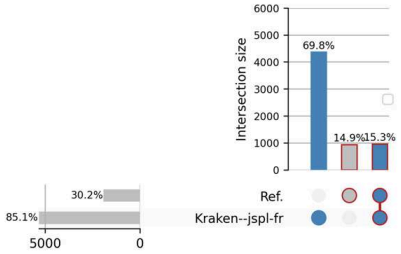
Type de correction (acronyme)	**Description du type de correction**			
MOBC	mal océrisé bien corrigé	Réf. Kraken Jspll-fr Jspll-ELTeC	*ennemis de l'Angleterre* *ennemis de l'Angle**l**terre* *ennemis de l'Angleterre* *ennemis de l'Angle**l**terre*	*Angleterre* *Angle**l**terre* *Angleterre* *Angle**l**terre*
MOMC	mal océrisé mal corrigé	Réf. Kraken Jspll-fr Jspll-ELTeC	*le surnom de Bois-Brûlés* *le surnom de Bois-Br**t**l**e**s* *le surnom de Bois-Br**a**l**e**s* *le surnom de Bois-**Arbres***	*Bois-Brûlés* *Bois-Br**t**l**e**s* *Bois-Br**a**l**e**s* *Bois-**Arbres***
MONC	mal océrisé non corrigé	Réf. Kraken Jspll-fr Jspll-ELTeC	*cure de Morlincourt* *cure de M**l**orlincourt**l*** *cure de M**l**orlincourt**l*** *cure de M**l**orlincourt**l***	*Morlincourt* *M**l**orlincourt**l*** *M**l**orlincourt**l*** *M**l**orlincourt**l***
BOIC	bien océrisé indûment corrigé	Réf. Kraken Jspll-fr Jspll-ELTeC	*en retournant de Morlincourt* *en retournant de Morlincourt* *en retournant de **Martincourt*** *en retournant de Morlincourt*	*Morlincourt* *Morlincourt* ***Martincourt*** *Morlincourt*

21 Nous étudions les intersections entre les EN détectées dans Réf. et les v-OCR pour généraliser notre évaluation. Les intersections sont calculées œuvre par œuvre, avant d'être additionnées. Ainsi, l'EN « Paris » repérée dans une v-OCR ne compte que pour la Réf. correspondante (on raisonne sur les types, document par document). Les intersections dans la figure 1 représentent les EN de v-OCR qui ont une orthographe identique à une EN de la Réf. correspondante, après correction automatique.

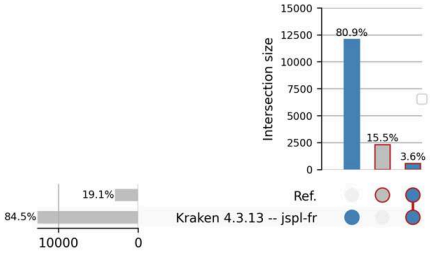
22 Les v-OCR Kraken et Tesseract sont corrigées avec le modèle pré-entraîné de Jspll-fr. Il est notable que les intersections sont meilleures pour les figures 1c et 1d, pour lesquelles la qualité de l'OCR Tesseract est meilleure que Kraken.

Figure 1. Intersections entre les EN détectées dans Réf. (*Silver Standard*) et les v-OCR pour les configurations Kraken (4a-4b) et Tess. fr. (4c-4d) corrigées par JamSpell pré-entraîné, spaCy-lg sur le corpus ELTeC français et la TGB

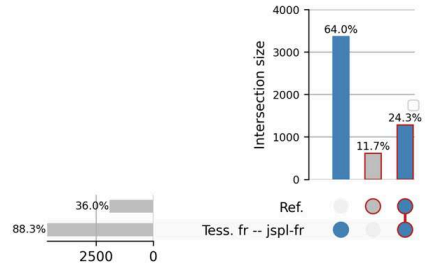
(a) ELTeC : Kraken–Jspll-fr – spaCy-lg.



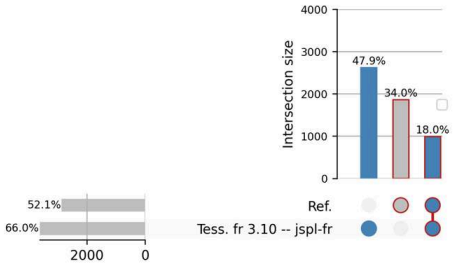
(b) TGB : Kraken–Jspll-fr – spaCy-lg.



(c) ELTeC : Tess. fr.–Jspll-fr spaCy-lg.



(d) TGB : Tess. fr.–Jspll-fr spaCy-lg.

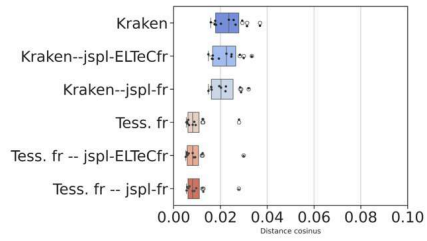


5.2.2. Évaluation automatique sur les occurrences

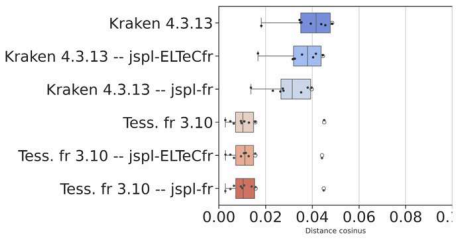
23 Les diagrammes de Venn présentés ci-dessus permettent un travail sur les types (la présence est traitée, mais pas la fréquence). Pour travailler sur les occurrences nous utilisons des mesures de distance textuelle. Nous avons privilégié la métrique cosinus qui est une des mesures de référence quand il est question de distance textuelle (Buscaldi *et al.* 2020). Elle est calculée sur des 2 et 3-grammes de caractères constituant les vecteurs d'effectifs pour être robuste à la contamination des entités. Dans les figures 2, plus la boîte est proche de zéro, plus les sorties comparées sont similaires. Les figures 2a et 2b illustrent les résultats obtenus pour les Réf. et les v-OCR pour tous les corpus avec une distance cosinus. Les figures 2c et 2d montrent les résultats en comparant les sorties de REN (en construisant les vecteurs d'effectif uniquement sur les EN identifiées de part et d'autre) obtenues sur les Réf. et v-OCR. Les résultats de la configuration Tess. fr—Jspll-fr peuvent être un signe du bruit ou des sur-corrections, étant donné les valeurs réparties dans un espace important au sein de la boîte (figure 2c). Généralement, nous observons ce qui se passe entre les expériences sur les corpus hétérogène et homogène (dont la qualité de l'OCR est très variable ou très bonne, respectivement), et nous constatons un saut qualitatif entre les différentes versions OCR de la TGB.

Figure 2. Distances cosinus entre les Réf. et les v-OCR (4a-4b) et entre les EN extraites par spaCy-lg sur les Réf. et les v-OCR (4c-4d)

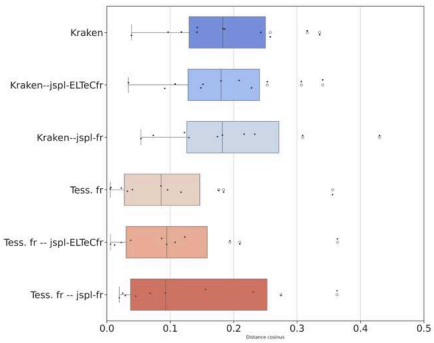
(a) ELTeC-fra texte Réf. – OCR, cosinus.



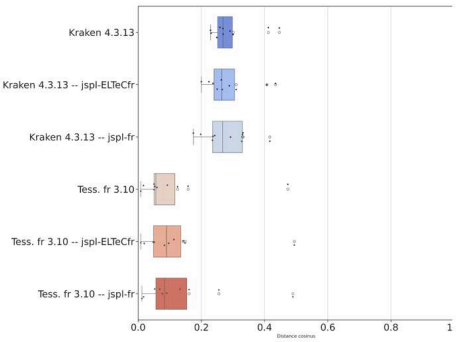
(b) TGB texte Réf. – OCR, cosinus.



(c) ELTeC-fra spaCy-lg Réf. – OCR, cosinus.



(d) TGB spaCy-lg Réf. – OCR, cosinus.



6. Exploiter les versions bruitées pour filtrer les entités

24 La figure 3 montre la variation de la précision et du rappel en fonction de la longueur (ci-après LEN pour *length*) et de la fréquence documentaire (ci-après DF pour *Document Frequency*) des entités. DF a une influence plus marquée sur les résultats que la fréquence absolue. Cela se produit car les formes contaminées ne se répètent pas d'un texte à l'autre et les formes contaminées sont tout de même beaucoup moins fréquentes que les formes stables. La partie gauche montre que les entités hapax sont les plus bruitées. Corollaire, la probabilité qu'une entité soit un VP augmente avec sa fréquence documentaire (elle est trouvée dans plusieurs hypothèses). D'autre part, les entités de longueur 1 et 2 sont très susceptibles d'être bruitées. Au-delà d'une longueur de 20 caractères, il y a aussi une proportion importante de bruit, mais ces entités sont aussi moins nombreuses. La figure 4 est une carte thermique montrant l'évolution de la précision sur notre *Silver Standard* selon différentes combinaisons (DF, longueur). Nous pouvons voir de nouveau que les entités hapax sont très souvent des faux positifs et d'autant plus si elles sont courtes. On observe que la zone constituée par les entités avec DF < 6 et LN < 5 correspond à des entités rarement présentes dans la référence et constitue sans doute majoritairement du bruit. Ceci nous invite à construire un filtre qui exploite ces caractéristiques pour écarter les entités contaminées.

Figure 3. Évaluation de la qualité de la REN (Précision et Rappel) dans les versions non corrigées selon la fréquence documentaire (DF) et la longueur (LEN) des formes

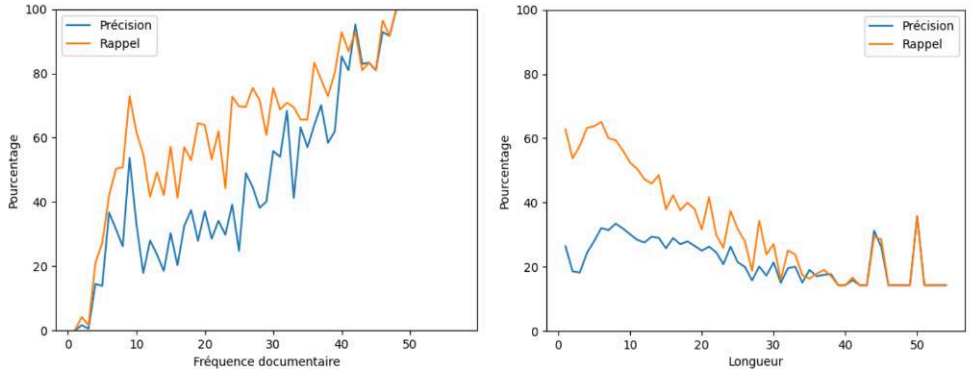
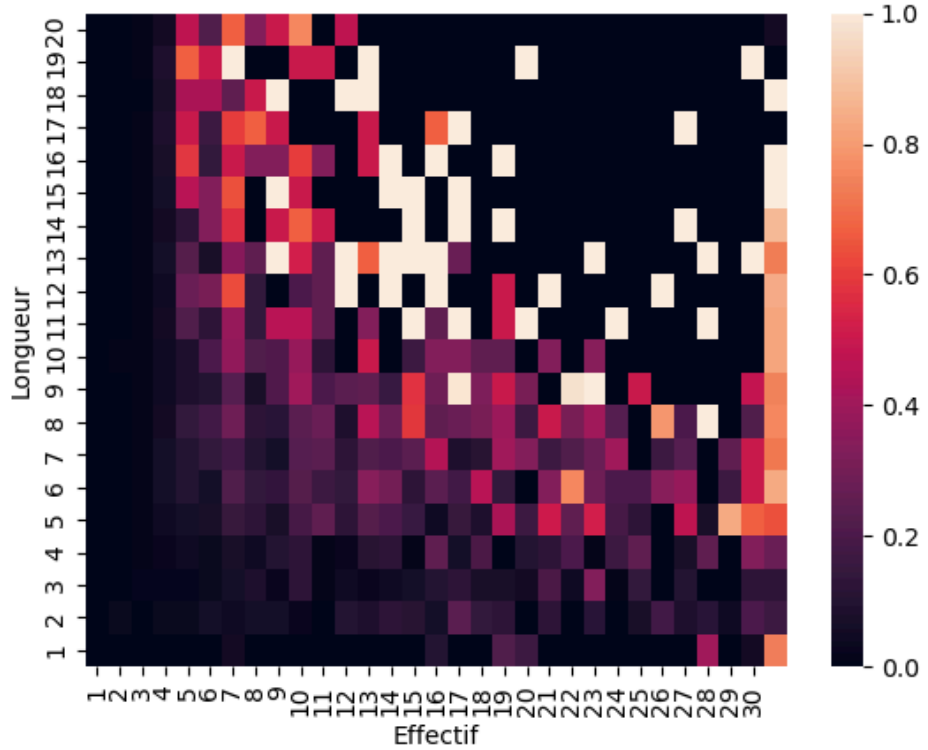


Figure 4. Carte thermique de précision pour tous les couples (DF, LEN)



25 Nous avons utilisé notre *Silver Standard* pour construire un modèle de filtrage des EN contaminées. Nous utilisons simplement la longueur et la DF comme caractéristiques et nous choisissons d'utiliser un arbre de décision pour l'interprétabilité. Le tableau 7 présente les résultats de ce classifieur. S'il n'est pas particulièrement performant pour détecter les vrais positifs, nous pouvons voir qu'il a une précision supérieure à 99 % dans la détection des faux positifs avec un Rappel de 65 %. Ce faisant, nous ne perdons que 10 % des vrais positifs (rappel de 90,74 %). Autrement dit, un tel filtre élimine 65 % du bruit tout en occasionnant moins de 10 % de silence.

Tableau 7. Résultat d'un arbre de décision entraîné sur la TGB et appliqué sur les données ELTeC

	Précision	*Rappel*	*F-mesure*	*Effectif*

Faux positifs	0.9918	0.6509	0.7860	72 895
Vrais positifs	0.1314	0.9074	0.2296	4 242

26 Nous présentons ci-dessous des exemples de faux positifs filtrés par notre classifieur pour différents couples [LEN,DF] :

[LEN= 2, DF=1] : {'KT', 'dp', 'Iw', 'av', 'H '}
[LEN =6, DF=1] : {'A AN S', 'DAMENE', 'AECAAE', 'Jactes', 'DNIITE'}
[LEN=10, DF=1] : {'Alala Bale', 'RTEE E EME', 'Du Sal als', 'lieis Srew', 'DEE ELBCEW'}

27 Nous présentons cette fois des exemples de vrais positifs qui ont bien été conservés par le classifieur (en italiques les erreurs issues de la référence, des faux VP) :

[LEN=10, DF= 1] : {'*Huit à Dix*', *'Qui chante', 'Regardez !'*, 'Barbette !', 'Madagascar'}
[LEN= 6, DF=10] : {'Girard', *'Restez'*, 'Tribou', 'Peyrol', 'christ'}
[LEN=10, DF=10] : {'Villenauxe', 'Larsillois', 'la Blanche', 'Neufchâtel', 'Christophe'}

28 Ces séries d'exemples montrent que la fréquence documentaire est un indice très utile pour filtrer les FP, ce qui permet à la fois de valoriser les versions bruitées en tant que « ressource lexicale » et d'aider aussi à exploiter ces versions pour elles-mêmes en rendant les séries d'EN extraites moins bruitées.

7. Conclusion

29 Dans cet article, nous avons exploité des données *Silver Standard* afin d'examiner à grande échelle l'influence du bruit des données OCR sur la Reconnaissance d'Entités Nommées, en nous plaçant au plus près des conditions réelles rencontrées dans les humanités, notamment en ne choisissant que des outils librement disponibles et simples à utiliser (*off-the-shelf* en résumé). Nos premiers résultats confirment que la correction automatique introduit des biais dans les données textuelles. Le F-score sur la REN entre la version OCR et la version OCR corrigée diminue de 12 points de pourcentage avec Tesseract dans le cas d'un OCR de bonne qualité (*Character Error Rate* ou CER de 0.03, Daudet, *Le petit chose*) et de 4 points pour un OCR de moindre qualité (CER de 0.13 Maupassant, *Une vie*). En revanche, les résultats avec Kraken, OCR moins performant sur ces corpus, baissent de 8 points dans le premier cas et augmentent de 23 points dans le second. Ce qui illustre le fait que le gain de la correction automatique n'est présent et significatif que lorsque l'OCR est de très mauvaise qualité. Ces résultats nous ont amenés à présenter une nouvelle manière d'exploiter des sorties d'OCR bruitées. Nous avons établi (i) une typologie des contaminations d'OCR et (ii) une typologie des erreurs de la correction automatique. Nous avons montré que la correction automatique de données en entrée n'a offert que peu d'améliorations de la qualité de la REN, très probablement du fait de la sur-correction. Cette affirmation découle surtout des résultats présentés sur la figure 2c au niveau de la configuration Tess. fr—Jspll-fr, mais également des travaux de Koudoro-Parfait *et al.* (2024). Bien que les corrections ne dégradant pas les résultats soient repérées (p. ex. la forme *leanne* bien corrigée en le prénom *jeanne*, si l'on ignore la sensibilité à la casse), elles semblent être plus rares. En revanche, le filtrage des entités en sortie semble être une piste plus prometteuse. Nous avons montré que les entités incorrectes sont massivement des entités courtes et hapax en corpus et que la fréquence documentaire et la longueur des entités étaient de bons critères pour exclure des entités candidates avec une précision

de plus de 98 %. Nous avons montré ainsi qu'exploiter plusieurs versions bruitées, issues de différents outils d'OCR et ayant subi un impact de l'outil de REN, offre une réelle plus-value et que, en l'absence d'outils de correction automatique véritablement efficaces, il existe une alternative à la coûteuse correction manuelle des textes océrisés. Ceci montre une nouvelle manière d'exploiter des données bruitées existantes pour valoriser d'autres corpus bruités, notamment dans l'optique d'utiliser la DF et la longueur comme les critères discriminatoires pour filtrer les EN. Nous pensons explorer cette idée en comparant des données PDF de différentes qualités afin de voir si les versions OCR qui en seront tirées pourraient être complémentaires. Combiner des versions bruitées, comme le fait par exemple Riddell (2022) pour enrichir les possibilités de filtrage des entités sera une autre piste que nous comptons suivre.

BIBLIOGRAPHY

Azmi A. M., Almutery M. N. & Aboalsamh H. A. (2019). « Real-Word Errors in Arabic Texts: A Better Algorithm for Detection and Correction », *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(8) : 1308-1320. https://doi.org/10.1109/TASLP.2019.2918404.

Alex B., Grover C., Klein E. & Tobin R. (2012). « Digitised Historical Text : Does it have to be mediOCRe ? », *in KONVENS*, 401-409. http://www.oegai.at/konvens2012/proceedings/59_alex12w/.

Bassil Y. & Alwani M. (2012). « OCR Post-Processing Error Correction Algorithm using Google Online Spelling Suggestion », *arXiv preprint arXiv:1204.0191*. https://doi.org/10.48550/arXiv.1204.0191.

Boros E., Ehrmann M., Romanello M., Najem-Meyer S. & Kaplan F. (2024). « Post-correction of Historical Text Transcripts with Large Language Models : An Exploratory Study », *LaTeCH-CLfL 2024*, 133-159. https://infoscience.epfl.ch/record/307961?v=pdf.

Boros E., Nguyen N. K., Lejeune G. & Doucet A. (2022). « Assessing the impact of OCR noise on multilingual event detection over digitised documents », *International Journal on Digital Libraries* 23(3) : 241-266. https://doi.org/10.1007/s00799-022-00325-2.

Buscaldi D., Felhi G., Ghoul D., Le Roux J., Lejeune G. & Zhang X. (2020). « Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? », *in Actes de Traitement Automatique des Langues Naturelles (TALN, 27e édition). Atelier DÉfi Fouille de Textes*, 14-25. https://aclanthology.org/2020.jeptalnrecital-deft.2.pdf.

Chiron G., Doucet A., Coustaty M., Visani M. & Moreux J.-P. (2017). « Impact of OCR Errors on the Use of Digital Libraries : Towards a Better Access to Information », *in 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, Canada, 1-4. IEEE. https://doi.org/10.1109/JCDL.2017.7991582.

Cohen J. (1960). « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement* 20(1) : 37-46. https://doi.org/10.1177/001316446002000104.

Damerau F. J. (1964). « A technique for computer detection and correction of spelling errors », *Communications of the ACM* 7(3) : 171-176. https://doi.org/10.1145/363958.363994.

Edwards L. D. M. (2016). *Conception de formes de relecture dans les chaînes éditoriales numériques*, Thèse de doctorat, Université de Technologie de Compiègne. https://tel.archives-ouvertes.fr/tel-01562039.

Eshel Y., Cohen N., Radinsky K., Markovitch S., Yamada I. & Levy O. (2017). « Named Entity Disambiguation for Noisy Text », https://arxiv.org/abs/1706.09147.

Evershed J. & Fitch K. (2014). « Correcting noisy OCR : Context beats confusion », *in Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 45-51. https://doi.org/10.1145/2595188.2595200.

Hamdi A., Jean-Caurant A., Sidère N., Coustaty M. & Doucet A. (2020). « Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition », *in Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, Proceedings*, 87-101. https://doi.org/10.1007/978-3-030-54956-5_7.

Hamdi A., Linhares Pontes E., Sidère N., Coustaty M. & Doucet A. (2022). « In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking », *Natural Language Engineering* 29(2) : 425-448. https://doi.org/10.1017/S1351324922000110.

Hládek D., Staš J. & Pleva M. (2020). « Survey of Automatic Spelling Correction », *Electronics* 9(10). https://doi.org/10.3390/electronics9101670.

Honnibal M. (*s.d.*). « spaCy's NER model », https://spacy.io/universe/project/video-spacys-ner-model.

Montani I., Honnibal M., Van Landeghem S., Boyd B., Peters H., O'Leary McCann P. & Geovedi J. (2023). « explosion/spaCy : v3. 5.1 : spancat for multi-class labeling, fixes for textcat+ transformers and more », *Zenodo*. https://zenodo.org/records/10009823.

Huynh V.-N., Hamdi A. & Doucet A. (2020). « When to Use OCR Post-correction for Named Entity Recognition ? », *in 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, 33-42. https://doi.org/10.1007/978-3-030-64452-9_3.

Kádár Á., Miranda Lj., Slocum V. & Van Landeghem S. (2023). « The Tale of Bloom Embeddings and Unseen Entities », https://explosion.ai/blog/technical-report.

Kiessling B. (2019). « Kraken – a Universal Text Recognizer for the Humanities », *in ADHO, Éd., Actes de Digital Humanities Conference*. https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX.

Koudoro-Parfait C., Lejeune G. & Roe G. (2021). « Spatial Named Entity Recognition in Literary Texts : What is the Influence of OCR Noise ? », *in Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 13-21. https://dl.acm.org/doi/10.1145/3486187.3490206.

Koudoro-Parfait C., Lejeune G. & Buth R. (2022). « Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique », *in Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, 45-55. https://aclanthology.org/2022.jeptalnrecital-humanum.6.

Koudoro-Parfait C., Petkovic L. & 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 : Robustesse et limites des modèles de traitement automatique des langues* 64(2) [*à paraître*].

Kukich K. (1992). « Techniques for Automatically Correcting Words in Text », *ACM Computing Surveys (CSUR)* 24(4) : 377-439. https://doi.org/10.1145/146370.146380.

Nguyen T. T. H., Jatowt A., Coustaty M. & Doucet A. (2021). « Survey of Post-OCR Processing Approaches », *ACM Computing Surveys (CSUR)* 54(6) : 1-37. https://doi.org/10.1145/3453476.

Nguyen T. T. H., Jatowt A., Nguyen N. V., Coustaty M. & Doucet A. (2020). « Neural Machine Translation with BERT for Post-OCR Error Detection and Correction », *in JCDL '20 : Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 333-336. https://doi.org/10.1145/3383583.3398605.

Oger S., Rouvier M., Camelin N., Kessler R., Lefèvre F. & Torres-Moreno J. M. (2012). « Système du LIA pour la campagne DEFT2010 », *Expérimentations et évaluations en fouille de textes : Un panorama des campagnes DEFT*, Coll. Systèmes d'information et organisations documentaires, Hermès, Chapitre 9. https://hal.archives-ouvertes.fr/hal-01433469/.

Petkovic L., Alrahabi M. & Roe G. (2022). « Impact de la correction automatique de l'OCR/HTR sur la reconnaissance d'entités nommées dans un corpus bruité », *Journal of Information Sciences (JIS)* 21(2) : 42-57. https://doi.org/10.34874/IMIST.PRSM/jis-v21i2.36599.

Qi P., Zhang Y., Zhang Y., Bolton J. & Manning C. D. (2020). « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages », https://doi.org/10.48550/arXiv.2003.07082.

Rebholz-Schuhmann D., Yepes A. J., Li C., Kafkas S., Lewin I., Kang N. & Hahn U. (2011). « Assessment of NER solutions against the first and second CALBC Silver Standard Corpus », *Journal of Biomedical Semantics* 2 : 1-12. https://doi.org/10.1186/2041-1480-2-S5-S11.

Riddell A. (2022). « Reliable editions from unreliable components : estimating ebooks from print editions using profile hidden markov models », *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1-5. https://doi.org/10.1145/3529372.3533292.

Sagot B. & Gábor K. (2014). « Named Entity Recognition and Correction in OCRized Corpora (Détection et correction automatique d'entités nommées dans des corpus OCRisés) », *Traitement Automatique des Langues Naturelles, TALN 2014, Marseille, France, 1-4 juillet 2014, articles courts, The Association for Computer Linguistics*, 437-442. https://aclanthology.org/F14-2009/.

Smith R. (2007). « An Overview of the Tesseract OCR Engine », *in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2. IEEE, 629-633. https://doi.org/10.1109/ICDAR.2007.4376991.

Stanislawek T., Wróblewska A., Wójcicka A., Ziembicki D. & Biecek P. (2019). « Named Entity Recognition – Is There a Glass Ceiling? », *in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 624-633. https://aclanthology.org/K19-1058.

Tual S., Abadie N., Chazalon J., Duménieu B. & Carlinet E. (2023). « A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents », *in International Conference on Document Analysis and Recognition*. Cham : Springer Nature Switzerland, 115-131. https://arxiv.org/abs/2302.10204.

Van Strien D., Beelen K., Ardanuy M., Hosseini K., McGillivray B. & Colavizza G. (2020). « Assessing the impact of OCR quality on downstream NLP tasks », *SCITEPRESS – Science and Technology Publications*. https://doi.org/10.17863/CAM.52068.

NOTES

1. Cf. l'outil SymSpell https://github.com/mammothb/symspellpy.

2. Nous avons aussi évalué Stanza (Qi *et al.* 2020), dont les résultats sont disponibles sur le dépôt GitHub : https://github.com/These-SCAI2023/NER_GEO_COMPAR. Par ailleurs, ce dépôt fait partie intégrante du projet doctoral de Caroline Koudoro-Parfait en cours.

3. Nous reprenons le terme *Silver Standard* en s'appuyant sur la définition de Rebholz-Schuhmann *et al.* (2011) : il s'agit de données de référence obtenues automatiquement ce qui facilite les comparaisons à grande échelle.

4. NERVAL est un outil d'évaluation de la REN sur des textes bruités en termes du calcul de la précision, du rappel et du F-score (cf. https://gitlab.com/teklia/nerval). Malgré certains biais d'évaluation, cet outil apparaît néanmoins comme un moyen assez fiable pour dépasser les problèmes d'alignements entre les résultats des différentes configurations à comparer pour calculer le F-score.

5. JamSpell est un outil de correction automatique des sorties OCR, basé sur les modèles de langue trigrammes statistiques (grain mot), en s'appuyant sur l'alphabet de la langue. Cf. https://github.com/bakwc/JamSpell.

6. Les données utilisées dans le cadre de cet article sont disponibles sur le dépôt GitHub https://github.com/These-SCAI2023/ARTICLE_Revue-Corpus_16092024.

7. Toponyme extrait de *Home influence*, Aguillar 1847.

8. Peuples de Sarmatie, dans le contexte « Tel sur les monts glacés des farouches Gelons », *Œuvres de Boileau*, 1836.

9. Suppression des césures et remplacement des « s longs » ⟨ſ⟩ par des « s ».

10. https://api.bnf.fr/documents-de-gallica-produits-au-format-tei-par-obvil

11. XML-TEI désigne un format de balisage s'appuyant sur les recommandations du consortium *Text Encoding Initiative* (*TEI*) pour l'encodage de ressources numériques, notamment de documents textuels. Cf. https://tei-c.org/.

12. Calculée comme le ratio entre le nombre de candidats des EN et le nombre de tokens, exprimé en pourcentages. Étant donné la variation de l'OCR dans nos corpus, nous considérons que cette densité estimée peut corréler avec le bruit d'OCR, dans le cas des faux positifs, par exemple.

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

14. https://github.com/distantreading/WG1/wiki/E5C-discussion-paper

15. Contrairement aux cas des données synthétiques, où le bruit artificiel (dégradations des caractères, floutage, etc.) est injecté dans les textes (Boros *et al.* 2022), nos corpus sont composés de vraies transcriptions OCR.

16. Ces plongements désignent un réseau neuronal convolutionnel profond avec des connexions résiduelles, capable de représenter un grand nombre de tokens de manière efficace en termes de mémoire, et ainsi d'offrir un bon équilibre entre efficacité, précision et adaptabilité (Honnibal s.d., Kádár *et al.* 2023).

17. L'annotation manuelle étant une tâche coûteuse, nous avons annotés 6 000 tokens pour 3 versions (Réf. et 2 v-OCR) de 2 textes : Carraud et Daudet https://github.com/These-SCAI2023/ELteC_GOLD.

18. Il s'agit de la mesure de la concordance entre chaque annotation produite par différents annotateurs, basée sur le coefficient Kappa (κ) de Cohen (Cohen 1960).

19. Cf. les données disponibles sur le dépôt GitHub https://github.com/ljpetkovic/ELTeC_GOLD_REVUE_TAL/tree/main/NERVAL/Revue_Corpus.

20. Une partie du bruit est déjà présente dans la version de Réf. ; ce n'est pas dû aux données mais à la méthode de REN.

21. Notamment EN contaminées dans v-OCR.

22. Ce cas de figure est totalement silencieux : il n'apparaît que si l'on dispose d'un *Gold Standard*.

23. Le code est librement disponible sur notre dépôt GitHub https://github.com/These-SCAI2023/EXPE24_CORRECTION_06022024.

24. https://github.com/bakwc/JamSpell?tab=readme-ov-file#download-models

ABSTRACTS

This paper presents the results of a research work that aims to determine whether the upstream OCR correction can significantly improve the results of the Named Entity Recognition (NER) task. The experiments were applied to the ELTeC and Very Big Library (TGB) corpora. Our objective is to establish (i) a typology of OCR contaminations from the Kraken and Tesseract tools and (ii) a typology of errors in the automatic correction produced by the JamSpell tool. As part of our evaluation, we study the intersections between the NEs detected by the spaCy tool in the reference texts and the two OCR versions on the one hand, and on the other hand the cosine similarity used to measure the textual distance between these versions. Our results show that automatic correction introduces biases in the textual data and that filtering the output NEs seems to be a more promising approach. Finally, we find that incorrect NEs are overwhelmingly short and hapax entities in corpus, and that the entity length and document frequency are the discriminatory criteria to exclude candidate NEs with a precision of over 98%.

Cet article présente les résultats d'un travail de recherche qui vise à déterminer si la correction de l'OCR, en amont, permet d'améliorer significativement les résultats de la tâche de la Reconnaissance d'Entités Nommées (REN). Les expériences ont été appliquées aux corpus ELTeC et Très Grande Bibliothèque (TGB). Notre objectif est d'établir (i) une typologie des contaminations d'OCR issues des outils Kraken et Tesseract et (ii) une typologie des erreurs de la correction automatique produite par l'outil JamSpell. Dans le cadre de notre évaluation, nous étudions les intersections entre les EN détectées par l'outil spaCy dans les textes de référence et les deux versions d'OCR d'une part, et d'autre part la mesure cosinus utilisée pour calculer la distance textuelle entre ces versions. Nos résultats montrent que la correction automatique introduit des biais dans les données textuelles et que le filtrage des entités en sortie semble être une piste plus prometteuse. Enfin, nous constatons que les EN incorrectes sont massivement des entités courtes et hapax en corpus, ainsi que la longueur des entités et la fréquence documentaire sont les critères discriminatoires pour exclure des EN candidates avec une précision de plus de 98 %.

INDEX

Mots-clés: reconnaissance d'entités nommées, correction automatique d'OCR, filtrage des sorties d'OCR, spaCy, JamSpell
Keywords: named entity recognition, automatic OCR correction, filtering OCR results, spaCy, JamSpell

AUTHORS

LJUDMILA PETKOVIC

Faculté des Lettres, Sorbonne Université
Équipe-projet Observatoire des Textes, des Idées et des Corpus (ObTIC)
Sorbonne Center for Artificial Intelligence (SCAI)
Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599

CAROLINE KOUDORO-PARFAIT

Faculté des Lettres, Sorbonne Université
Équipe-projet Observatoire des Textes, des Idées et des Corpus (ObTIC)
Sorbonne Center for Artificial Intelligence (SCAI)
Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599
Sens Texte Informatique Histoire (STIH)

MARIE-SOPHIE DESMAREST

Faculté des Lettres, Sorbonne Université
Sens Texte Informatique Histoire (STIH)
Centre d'expérimentation en méthodes numériques pour les recherches en sciences humaines et sociales (CERES)

GAËL LEJEUNE

Faculté des Lettres, Sorbonne Université
Sens Texte Informatique Histoire (STIH)
Centre d'expérimentation en méthodes numériques pour les recherches en sciences humaines et sociales (CERES)