Analyse multilingue de l'impact de la correction automatique de la ROC sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires

Auteur(s)

A définir par la commande `\author{...}`. *Dans le cas de plusieurs auteurs, séparer chaque auteur par* `\andauthor`. *Dans le cas d'adresses différentes des auteurs, indexer chaque auteur avec des astérisques comme* * *ou* ** *obtenues par* `\fup{*}` *ou* `\fup{**}`.

RÉSUMÉ. *L'extraction d'informations de textes issus de reconnaissance optique de caractères (ROC) interroge sur la possibilité d'exploiter des données bruitées. Notre contribution est double, nous nous attacherons : d'une part, à déterminer si la correction de la ROC permet d'améliorer significativement les résultats de la tâche de reconnaissance d'entités nommées (REN) sur des corpus de langue française, anglaise et portugaise, d'autre part, à montrer les limites des évaluations strictes (F-score ou intersections), tout en proposant des stratégies d'évaluation plus souple. Nous présentons plusieurs typologies et protocoles d'évaluation pour la REN sur des données bruitées et sur des données bruitées corrigées automatiquement.*

MOTS-CLÉS : *ROC_1, REN_2, Correction automatique de ROC_3.*

TITLE. *Multilingual analysis of the impact of automatic OCR correction on spatial recognition of spatial named entities in literary corpora*

ABSTRACT. *The extraction of information from texts produced by optical character recognition (OCR) raises questions about the possibility of exploiting noisy data. Our contribution is twofold: firstly, to determine whether OCR correction can significantly improve the results of the Named Entity Recognition (NER) task on French, English and Portuguese language corpora, and secondly, to show the limitations of strict evaluations (F-score or intersections), while proposing more flexible evaluation strategies. We present several typologies and evaluation protocols for NER on noisy data and on automatically corrected noisy data.*

KEYWORDS: *OCR_1, NER_2, Automatic OCR $Correction_3$.*

R^{e} soumission à *TAL VV-NN*, le 21/05/2023

1. Introduction

Les techniques de traitement automatique des langues (TAL), combinées aux méthodes des humanités numériques (HN), rendent possibles l'exploration et l'exploitation de corpus numérisés à grande échelle. Ces deux domaines trouvent souvent leurs applications dans l'extraction d'informations (reconnaissance d'entités nommées, abr. REN) d'une part, et d'autre part dans la valorisation des corpus patrimoniaux (reconnaissance optique de caractères, abr. ROC). Les institutions publiques européennes, internationales et des membres indépendants de la communauté scientifique ont mené des campagnes de numérisation et de publication des transcriptions (ROC) d'œuvres littéraires sur le web. Ils ont mis à disposition de la communauté de vastes corpus dont la qualité est hétérogène. En effet, si ces initiatives rendent l'accès aux textes plus aisé, force est de constater que la ROC génère du bruit. Le bruit désigne toutes les erreurs produites par le système de ROC : insertion, suppression, mais aussi substitution d'un ou plusieurs caractères. Le bruit dans les sorties de la ROC peut être provoqué par des taches, du texte disposé sur deux colonnes, l'emploi de certaines polices typographiques, etc.

Par ailleurs, une grande majorité des outils de TAL utilisés en aval de la ROC sont entraînés sur des données préparées (non bruitées). Ainsi, les scientifiques des sciences humaines et sociales (SHS) et HN qui utilisent ces outils sur leurs données en conditions réelles rencontrent des difficultés liées à l'inadaptation des outils aux données bruitées. De fait, les erreurs commises par les systèmes de REN sont souvent imputées au caractère bruité des transcriptions de la ROC, ce qui induit l'idée que la correction des données en entrée est la seule manière pertinente d'améliorer les résultats de la REN. S'il est possible de corriger automatiquement des erreurs régulières produites par la ROC, l'apparition d'erreurs singulières rend difficile la correction. De plus, comme le soulignent (Huynh et al., 2020a) et (Petkovic et al., 2022), s'il est possible d'améliorer les résultats de la REN en corrigeant automatiquement les sorties de la ROC, celle-ci produit ses propres erreurs. Enfin, à la complexité de la REN à partir de la ROC s'ajoute la variation de la langue employée (diachronique, diatopique) et la variation du genre (littéraire, critique). L'état de l'art en REN révèle un faible intérêt pour les langues autres que l'anglais ((Lejeune et al., 2015), (Rahimi et al., 2019)), notamment pour des langues moins bien dotées comme le portugais.

Nous souhaitons déterminer si la correction de la ROC, en amont, permet d'améliorer significativement les résultats de la REN en aval. Nous proposons en section 2 un état de l'art portant sur la correction de la ROC et sur la REN à partir des transcriptions bruitées. Puis, en section 3, nous présentons les corpus littéraires (TGB[1] et ELTeC[2]) sur lesquels nos analyses s'appuient. La section 4 présente différentes méthodes d'évaluation manuelles et automatiques de l'impact des contaminations[3] de la

1. http://obvil.lip6.fr/tgb/
2. https://www.distant-reading.net/eltec/
3. Nous adoptons le terme « contamination » proposé par (Hamdi et al., 2022) pour qualifier les entités dont l'orthographe a été modifiée à cause de la transcription fautive de la ROC.

ROC sur la REN effectuée avec l'outil `spaCy`[4] (Montani et al., 2023), ainsi qu'une typologie des contaminations. La section 5 comprend les évaluations manuelles et automatiques de la REN sur des corrections de la ROC produites avec JamSpell[5] (outil de correction automatique), et une typologie des contaminations de la correction de la ROC sur la REN. Enfin, nous exposons nos conclusions et les pistes de recherches dans la section 6.

2. Correction de la ROC dans la perspective d'appliquer la REN en aval

Face au volume croissant des données issues de la numérisation et de la ROC, des problématiques relatives à la qualité de ces données et à leur exploitabilité scientifique émergent, étant donné les erreurs dans les transcriptions de la ROC. Les scientifiques rencontrent ainsi des difficultés pour appliquer des outils informatiques, généralement entraînés sur des données textuelles correctement orthographiées (Eshel et al., 2017), à des données textuelles bruitées. Un des remèdes consiste à corriger les données délivrées par la ROC (Sagot et Gábor, 2014), idéalement de manière automatique, lesquelles seront ensuite exploitées dans les différentes tâches du TAL. Or, si certaines interférences des dispositifs de ROC sont systématiques (Stanislawek et al., 2019), lorsqu'elles sont singulières, cet exercice devient difficile à réaliser. En outre, ainsi que le soulignent (Huynh et al., 2020b), la correction peut, elle aussi, produire des erreurs. Les erreurs de la ROC peuvent être regroupées en deux catégories (Oger et al., 2012) : celles des erreurs lexicales (angl. *non-word errors*) qui ne représentent pas des mots valides de la langue, p. ex. si le mot "Morlincourt[6]" est écrit "Mlolincourtl", et celles, beaucoup plus rares, des erreurs grammaticales (angl. *real-word errors*) (Wisniewski et al., 2010) auxquelles pourraient s'ajouter les erreurs sémantiques (angl. *semantic/context-sensitive errors*), quand p. ex. "Gélons[7]" devient "Gelons", grammaticalement correct mais incorrect dans le contexte donné. Les erreurs liées à la correction automatique sont principalement des erreurs sémantiques (Azmi et al., 2019), p. ex., "M. Eyssette[8]" devient "M. Cassette".

Si le domaine de la correction automatique de texte est très actif et remonte à plusieurs décennies, depuis les travaux de (Damerau, 1964) jusqu'à nos jours (Nguyen et al., 2021), il n'existe pas de classification unanime pour une approche standard de correction des textes bruités ((Bassil et Alwani, 2012), (Dumas Milne Edwards, 2016), (Nguyen et al., 2020), Néanmoins, trois grandes méthodes se démarquent : méthodes exploitant des lexiques, méthodes sur des modèles de langue statistiques, et méthodes à base d'apprentissage automatique (Petkovic, 2022).

4. https://spacy.io/
5. https://github.com/bakwc/JamSpell
6. Toponyme français extrait de *Mon village*, J. Adam, 1860.
7. Peuples de Sarmatie, voisins du Borysthène, dans le contexte « Tel sur les monts glacés des farouches Gelons », *Œuvres de Boileau*, T. 2, Boileau, 1836.
8. Nom d'un personnage du roman *Le petit chose*, de A. Daudet, 1868, corrigé automatiquement avec l'outil JamSpell.

Une des questions qui préoccupe actuellement la communauté TAL concerne l'évaluation de l'incidence des erreurs de la ROC sur la REN ((Chiron et al., 2017), (Hamdi et al., 2020), (Tual et al., 2023)). La REN, et particulièrement l'identification des entités nommées (EN) de lieux (van Strien et al., 2020), est un moyen efficace pour améliorer l'accès aux informations contenues dans de vastes corpus. D'ailleurs, (Chiron et al., 2017), ont montré qu'un nombre important de requêtes d'utilisateurs de la plateforme Gallica[9] était affecté par des termes mal transcrits et non répertoriés dans les dictionnaires habituels. Les erreurs de la ROC impactent également d'autres tâches (segmentation de phrases, analyse de dépendances, modélisation de sujets et réglage fin du modèle de langage neuronal) ; par exemple, la tâche de modélisation des sujets (angl. *topic models*) est impactée par la mauvaise qualité de la ROC, car les modèles produits divergent de ceux corrigés à la main (van Strien et al., 2020). Par ailleurs, (Evershed et Fitch, 2014) soulignent l'importance de la correction automatique des erreurs de la ROC dans un corpus de journaux avec le logiciel `overProof`[10], le taux d'erreur de mots réduit de plus de 60% a permis de réduire de plus de 50% le nombre d'articles manqués lors d'une recherche par mot-clé.

Lors de leurs expériences (Koudoro-Parfait et al., 2021)[11] ont noté que les systèmes de REN ont une certaine robustesse face à la variabilité dans les données. Certaines EN dont la forme est dite « contaminée » (Hamdi et al., 2022) ont malgré tout été reconnues par des outils tels que `spaCy` ou `stanza`, p. ex. "Mlorlincourtl" (forme contaminée de "Morlincourt") est repéré et correctement labellisé. On peut supposer qu'il n'est pas nécessaire de corriger la totalité des EN et que la correction de seulement certaines EN[12] (Alex et al., 2012) améliore les résultats de la REN.

3. Corpus d'évaluation de l'impact de la correction automatique sur la REN

Pour chaque corpus ELTeC français, anglais et portugais (cf. tableaux 2a, 2b et 2c), ainsi que pour les textes de la Très Grande Bibliothèque (TGB) (tableau 1) nous disposons des textes que nous nommons « référence » (version orthographiquement correcte des textes). Les textes extraits des collections ELTeC sont généralement de très bonne qualité. Concernant, les textes de la TGB la qualité est plus hétérogène. Pour chaque texte nous possédons deux transcriptions différentes : (i) Kraken[13] (Kiessling et al., 2019), (ii) la version transcrite de Tesseract[14] pour chacune des langues des corpus (anglais (Tess. en), français (Tess. fr) et portugais (Tess. pt)).

9. https://gallica.bnf.fr
10. http://overproof.projectcomputing.com/
11. https://github.com/These-SCAI2023/NER_GEO_COMPAR
12. Suppression des césures et remplacement des "s longs" par des "s".
13. https://kraken.re/3.0/index.html
14. https://doc.ubuntu-fr.org/tesseract-ocr

3.1. *La Très Grande Bibliothèque (TGB)*

La TGB est une bibliothèque de documents français qui met a disposition des œuvres, issues des collections Gallica, transcrites par la ROC. Le corpus compte 128 441 textes au format XML-TEI et 58 287 auteurs, et couvre différentes thématiques (littérature, histoire, philosophie, etc.). La TGB est constituée de 95 479 œuvres, datées du XIX^{e} s., 7 294 du XVIII^{e} s., 54 du XX^{e} s. et 24 du XVII^{e} s.[15] D'après une gestionnaire du service SINDBAD[16] de la BnF, plusieurs moteurs ROC ont été utilisés, Abby étant le principal depuis 2019 et celui utilisé en interne. Néanmoins, certains prestataires utilisaient des solutions internes, ou un mix de ROC, et certains marchés incluaient une phase de correction humaine post-ROC. Si certains textes ont un taux de confiance de ROC indiqué de 100%, d'autres n'en disposent pas et aucune autre information sur la performance n'est fournie ; or, nous observons que la qualité de la ROC est assez hétérogène pour ce corpus. Nous avons extrait une dizaine d'œuvres des catégories *Littérature (Belles-lettres)* et *Langues romanes, Français*, pour constituer notre corpus de PDF comportant des traits permettant d'illustrer les difficultés de l'application de la ROC à des textes anciens (transcription de décorations ou de caractères en capitales stylisées). Les informations générales et le nombre d'EN de lieux reconnues par l'outil `spaCy` dans les textes que nous avons sélectionnés comme textes de référence sont présentées dans le tableau 1.

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

TABLEAU 1. *Statistiques sur le corpus TGB. (`spaCy_lg` : modèle large de `spaCy`. La dernière colonne indique le nombre total des EN).*

3.2. *European Literary Text Collection (ELTeC)*

Entre 2017 et 2022 l'action COST *Distant Reading for European Literary History* (CA16204) a constitué une collection de corpus de textes littéraires dans plusieurs

15. Un nombre considérable de documents représentant des rééditions de textes plus anciens.
16. https://www.bnf.fr/fr/une-question-pensez-sindbad

langues européennes dont certains ont pour source le site web du projet Gutenberg[17], Gallica mais aussi la Bibliothèque électronique du Québec[18]. Les textes disponibles sont tous de bonne qualité. L'action COST a mis en place une liste de critères[19] permettant la sélection des œuvres entrant dans le périmètre d'une collection ELTeC. La question de la qualité intrinsèque du texte n'étant pas clairement mise en question, on peut en conclure qu'implicitement il est attendu que les textes romanesques soient le plus possible exempts de fautes de ROC. Le but de l'action est de rendre disponible des œuvres romanesques pour la conception, l'évaluation et l'utilisation d'outils et de méthodes d'analyse multilingues des textes littéraires. ELTeC compile des corpus de romans pour plus d'une vingtaine de langues européennes. Les collections française, anglaise et portugaise comprennent chacune une centaine de romans transcrits par la ROC publiés entre le milieu du XIXe siècle et le début du XXe siècle. Les romans sont disponibles en plusieurs formats : le format texte brut (.txt), un encodage TEI et un encodage TEI enrichi par une annotation morphosyntaxique. Pour cette étude, nous avons travaillé avec des textes collectés dans les collections française (11), anglaise (9) et portugaise (4). Le corpus portugais est d'une taille restreinte du fait de difficultés à rassembler des PDFs d'une qualité équivalente à ceux des corpus français et anglais. Les tableaux 2a, 2b et 2c présentent les informations générales sur les corpus conçus à partir des collections ELTeC. Ils comprennent le nombre d'EN de lieux reconnues dans chacun d'eux par l'outil `spaCy`.

4. Problématiques d'évaluation de l'impact de la ROC sur la REN

4.1. *Outils de ROC et REN utilisés dans le cadre de cette étude*

4.1.1. *Les outils de ROC*

Les transcriptions issues de la ROC ont été produites avec deux systèmes disponibles gratuitement : Kraken (Kiessling, 2019) et Tesseract (Smith, 2007). Bien qu'il existe un modèle pour le français du XVIIe siècle (Gabay et al., 2020)[20], ainsi que le modèle Gallicorpora (Sagot et al., 2022), nous ne les avons pas jugés adaptés à notre corpus, et donc nous utilisons le modèle de base de Kraken, version 3.0. Ce modèle permet d'opérer la segmentation[21] et la transcription[22]. Concernant Tesseract, nous avons utilisé le modèle LSTM `tessdata_best`, sur la version 4.1.2 du système, entraîné sur des données Google. Tesseract propose une analyse de la mise en page

17. https://www.gutenberg.org/
18. http://beq.ebooksgratuits.com/, Jean-Yves Dupuis 1998-2018.
19. https://github.com/distantreading/WG1/wiki/E5C-discussion-paper
20. https://github.com/Heresta/OCR17plus/tree/main/Model
21. Le modèle de segmentation est constitué d'un réseau d'étiquetage par classification des phrases (angl. *seed-labeling network*).
22. Le modèle de transcription fonctionne comme un classifieur de séquences sans segmentation qui utilise un réseau neuronal artificiel pour mapper une image d'une ligne de texte (séquence d'entrée), en une séquence de caractères (séquence de sortie).

Ouvrage	Auteur	Année	Pages	Tokens	spaCy_lg
Mon village	J. Adam	1860	200	20 938	213
Marie-Claire	M. Audoux	1925	120	35 780	101
Le château de Pinon, vol. I	G. A. Dash	1844	332	44 246	271
La petite Jeanne	Z. Carraud	1884	220	53 212	316
La nouvelle espérance	A. de Noailles	1903	325	54 272	182
Une vie	G. de Maupassant	1883	337	75 745	302
Albert Savarus. Une fille d'Ève	H. de Balzac	1853	60	79 924	682
Le petit chose	A. Daudet	1868	292	86 482	744
Les trappeurs de l'Arkansas	G. Aimard	1858	450	91 119	646
La Belle rivière	G. Aimard	1894	339	137 392	1 004
L'Éducation sentimentale	G. Flaubert	1880	520	150 494	1 304

(a) *corpus français.*

Ouvrage	Auteur	Année	Pages	Tokens	spaCy_lg
Auriol	W. H. Ainsworth	1844	246	46 388	82
Wuthering Heights	E. Brontë	1847	764	94 986	140
Coningsby	B. Disraeli	1844	983	101 778	634
Mary Barton	E. Gaskell	1848	423	161 568	290
Home influence	G. Aguillar	1847	628	171 342	205
Modern Flirtations vol. 1	C. Sinclair	1841	386	189 057	502
The Life and Adventures of M. Armstrong	F. Trolloppe	1840	387	189 392	187
Vanity Fair	W. M. Thackeray	1848	624	298 568	1 492
The Mysteries of London	G. Reynolds	1844	840	810 167	2 019

(b) *corpus anglais.*

Ouvrage	Auteur	Année	Pages	Tokens	spaCy_lg
Quattro Novelas	A. Castro Osorio	1908	272	50 766	353
Casa de Ramires	E. de Queiroz	1900	543	107 441	3 881
Uma familia ingleza	J. Diniz	1875	360	122 008	994
O crime do padre Amoro	E. de Queiroz	1875	620	141 700	2362

(c) *corpus portugais.*

TABLEAU 2. *Statistiques sur les corpus ELTeC. La dernière colonne indique le nombre total d'EN.*

intégrée à travers la segmentation des cadres (angl. *box segmentation*), ce qui rend le traitement des mises en page complexes plus difficile (Reul et al., 2017). Le modèle de base de Tesseract est un modèle conçu pour l'anglais, et il existe un modèle français et un modèle portugais. Quatre modèles de langue neuronaux pour la ROC ont été utilisés dans le cadre de ces expériences : Kraken de base, Tess. en, fr et pt contemporain.

4.1.2. *L'outil de REN*

Pour effectuer la tâche de REN[23] nous avons utilisé la chaîne de traitements de la boîte à outils pour le TAL `spaCy` dans sa version 3.5.1. Le système `spaCy` contient une stratégie d'intégration de mots utilisant des fonctionnalités de sous-mots et les plongements "Bloom" (angl. *Bloom embeddings*)[24], ainsi qu'un réseau neuronal convolutif avec des connexions résiduelles, ce qui peut expliquer sa robustesse lors de l'extraction des EN contaminées. `spaCy` propose des modèles de langue du type *large* pour le français[25], le portugais[26] et l'anglais[27].

Les modèles français et portugais sont chacun entraînés sur UD adaptés à leur propre langue (UD French Sequoia v2.8 et UD Portuguese Bosque v2.8 respectivement), ainsi que sur WikiNER et Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia) ; le modèle français s'appuie aussi sur `spaCy` lookups data. Le modèle anglais, quant à lui, est entraîné sur Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl), WordNet 3.0, OntoNotes 5 et ClearNLP Constituent-to-Dependency Conversion[28]. Nous avons favorisé l'usage du modèle *large* (`spaCy_lg`) plutôt que du modèle *small* (`spaCy_sm`) car la différence principale entre les deux modèles pour les trois langues tient à l'ajout de la vectorisation et des plongements de mots (angl. *embeddings*) dans l'entraînement du modèle *large*.

4.2. *Moins d'hapax : indice de la performance de la REN sur données bruitées ?*

Dans un premier temps, nous proposons une évaluation de l'outil de REN `spaCy` sur données bruitées. Nous comparons les résultats obtenus automatiquement sur les corpus ELTeC français, anglais et portugais et la TGB (annotations automatiques de référence), et ceux obtenus sur leurs transcriptions de ROC[29], sans nous appuyer sur un *gold standard*. Nous observons que les fautes d'orthographes provoquées par la ROC ne sont pas systématiquement un frein à la bonne extraction des noms de lieux, comme en témoigne le tableau 3.

En revanche, la concaténation des tokens d'une EN semble être une contamination plus préjudiciable à sa bonne détection. Par ailleurs, l'étude de (Koudoro-Parfait et al., 2021) laisse entendre que (i) le contexte contaminé autour d'une EN pourrait être un facteur de non détection et (ii) un contexte parfaitement propre ne serait pas

23. Nous avons également effectué les évaluations pour `stanza` (Qi et al., 2020), v. 1.5.0, dont les résultats sont disponibles sur le dépôt GitHub : anonymous
24. Il s'agit de la structure de données probabiliste qui permet de réduire la dimension des vecteurs (*cf.* https://explosion.ai/blog/bloom-embeddings).
25. `fr_core_news_lg`, https://spacy.io/models/fr
26. `pt_core_news_lg`, https://spacy.io/models/pt
27. `en_core_web_lg`, https://spacy.io/models/en
28. Tailles modèles : fr – 545 Mo, en – 560 Mo, pt – 541 Mo.
29. Dépôt GitHub avec nos données : https://github.com/anonymous

Type d'impact	**Contexte**	`spaCy_lg`
Contamination orthographique interne à l'entité	*il en est tombe au sort cinq de Sainl-Bruncl le* *duranta todo o tompe em qne ostivesso em Portngal*	Sainl-Bruncl le. Portngal
Ajout d'un caractère minuscule au début de l'entité	*Aux kEtats-Unis*	()
Ponctuation substituée par un caractère collé à l'entité	*about Manchesterl A pretty state*	Manchesterl
Entité tronquée	*dans l'intérieur de l'Améri-* *et le golfe de Cali-foruie._n*	Améri— golfe de Cali-
Mots concaténés	*[...] larue Saint-Honoré ;* *afriver aMorlincourt' tot*	_ Saint-Honoré ()

TABLEAU 3. *Proposition de typologie pour l'évaluation de la REN sur des données issues de la ROC.*

la garantie que l'EN soit reconnue par le système. Il semble que ces faits soient vérifiables pour les trois langues sur lesquelles nos expériences ont porté. Par ailleurs, certaines entités même très contaminées sont identifiées, comme p. ex. *"ancehester"* pour « Manchester ».

	Anglais			Français			Portugais		
	Reynolds	Troll.	Brontë	Daudet	Adam	Maup.	Diniz	Queiroz	Osorio
CER Tess+ lang.	0.10	0.11	0.25	0.03	0.10	0.13	0.06	0.09	0.10
WER Tess+ lang.	0.18	0.13	0.28	0.05	0.22	0.21	0.14	0.24	0.21
Réf.	495	83	40	209	71	172	625	919	231
Tess+ lang.	1 373	165	107	295	298	347	951	873	428
Variation	+177%	+98%	+168%	+41%	+319%	+102%	+52%	-5%	+85%

TABLEAU 4. *Nombre de types d'EN identifiées par `spaCy_lg` dans les corpus ELTeC en fonction de différentes qualités de la ROC déterminées par le CER calculé sur le modèle Tess. adapté à la langue du corpus.*

Le tableau 4 qui répertorie le nombre de types d'EN reconnues par `spaCy` selon la qualité des versions de ROC, illustre le fait que sur les versions de ROC les systèmes de REN récupèrent plus de types d'EN différents, donc que la qualité du texte en entrée influe sur la quantité des types d'EN récupérés en sortie. La qualité des versions de ROC a été évaluée en appliquant les métriques *Character Error Rate* (CER) et *Word Error Rate* (WER) sur les textes de référence et les versions de ROC. On note que plus le WER est élevé, plus la qualité de la transcription baisse, plus le nombre de type d'EN en sortie est élevé. On peut en conclure que (i) le système de REN ramène plus de bruit ou Faux Positif (FP) en sortie quand la ROC est moins bonne et (ii) le nombre des hapax augmente selon que la qualité de la transcription diminue. Dans la quantité d'EN surnuméraire détectée sur les versions de ROC par rapport à la référence, figurent des FP mais aussi des formes contaminées des entités, qui sont des hapax, et qui comptent chacune pour un type différent d'EN en plus du type initial de l'EN. Ces phénomènes sont illustrés dans le tableau 5 qui recense une annotation

manuelle[30] des Vrais Positifs (VP) et des FP par type d'EN récupérées par `spaCy_lg` sur l'ensemble des EN de référence et celui des versions pour Kraken et Tess. fr.

	REF	Kraken	Tess. fr
	`spaCy_lg`	`spaCy_lg`	`spaCy_lg`
Nb. types	203	467	237
VP	96	**107**	96
FP	107	360	141
VP hapax	60	70	57

TABLEAU 5. *Annotation manuelle des VP et FP sur les EN types reconnues par `spaCy` pour Daudet.*

On retrouve bien (i) plus de FP et (ii) plus de VP qui sont des hapax sur les transcriptions OCR que sur la référence. Il y a donc plus de types différents d'entités sur la sortie de la ROC que sur la sortie de la réf., car, comme l'illustre le tableau 6, les variantes d'une entités peuvent être nombreuses.

Version	**Modèle REN**	**Entité**	**# Manque**	**Entité**	**# Manque**
Réf.	`spaCy_lg`	Ormeaux : 5	N/A	Nouvelle-France" : 17	N/A
Kraken	`spaCy_lg`	Ormaeuux : 1, Ormenux : 1, Ormeuux : 2	1	Nouvelle-Fance : 4, Nouvelle-France : 8, Nouvelle-Frnce : 1, Nouvelle-lFrance : 2	2
Tess fr	`spaCy_lg`	Ormeaux : 3, Ormenux : 1	1	Nouvelle-France : 5, Nouvelle—France : 8	3

TABLEAU 6. *REN sur des formes contaminées de l'EN "Ferme des Ormeaux",* La petite Jeanne*, Carraud.*

L'analyse du tableau 4 montre qu'il existe des problèmes relatifs à la déperdition de données lors de la transcription de ROC. Des EN n'ont pas été transcrites par l'outil de ROC et donc apparaissent comme du silence. Néanmoins, il ne s'agit pas d'un FN de l'outil de REN, mais d'un FN de l'outil de ROC. Le cas Reynolds, pour lequel seuls 111 types d'EN ont été récupérés dans la configuration Kraken-`spaCy_lg`, en est l'exemple, plus de 90% des pages n'ont pas été transcrites à cause du flou sur les pages concernées. D'autres textes ont connu le même sort dans de très moindre proportions. Nous n'avons pas mesuré l'impact de cette non transcription car elle était en faible proportion sur tout le corpus.

Enfin, il peut arriver, plus rarement, que des entités ne soient pas détectées sur la version de référence (réf.). mais le soient sur la version de ROC. Il peut s'agir d'une part du fait que la version de ROC contient du texte en plus de la version de réf., par exemple les notes de bas de pages transcrites par la ROC mais non prises en compte par les auteurs de la référence, ou, d'autre part d'une erreur du système même en contexte non bruité. Dans ces deux cas, il ne s'agit pas véritablement de FP, et ces cas particuliers viennent poser les limites de l'évaluation de la tâche de REN sur données bruitées.

30. Nous n'avons pas annoté tous les corpus en raison du temps limité.

4.3. *Usage des intersections : une évaluation trop stricte ?*

Pour automatiser nos analyses et pouvoir les conduire à plus grande échelle nous avons décidé de calculer et représenter les intersections entre les ensembles des EN reconnues sur la version de réf. et celles obtenues sur les versions de ROC. Les versions de ROC et les textes de référence ont été annotés automatiquement avec `spaCy_lg`. Nous nous servons de ces derniers comme vérité terrain [31]. Nous avons calculé les intersections pour chacun des corpus (ELTeC français, anglais et portugais, et TGB) de manière globale. Pour ce faire nous avons fait correspondre les entités de chacun des textes de référence avec celles de leurs versions de ROC. Ainsi, dans le cas du corpus ELTeC français l'EN "Paris" repérée dans le texte de référence pour Daudet, n'est pas la même que l'EN "Paris" récupérée dans le texte de référence de Noailles. Il en va de même pour les différentes autres configurations (où, sous le terme, « configuration » on désigne la combinaison d'un modèle de ROC et d'un modèle de REN, p. ex. les résultats de la configuration Kraken-`spaCy_lg`). *cf.* le tableau 9).

La figure 1 rend compte de cette évaluation stricte dans laquelle chaque token de l'ensemble de réf. est comparé avec chaque token de l'ensemble de ROC. Pour être considéré comme un VP une EN de l'ensemble de ROC doit être orthographiée de manière identique à une EN de l'ensemble de réf. Dans ce graphique, le cercle de gauche - "EN Réf" - qui comprend les EN de la réf. et l'intersection entre les deux cercles sont considérés comme représentant les VP. Le cercle de droite - "EN_Kraken/Tess." - qui représente les EN récupérées uniquement sur la version de ROC est considéré comme comportant les FP de la sortie de REN.

(a) *Kraken –`spaCy-lg`* **(b)** *Tess. fr. – `spaCy_lg`*

FIGURE 1. *Intersections pour les configurations Kraken-`spaCy_lg` et Tess. fr.-`spaCy_lg`, pour le corpus ELTeC français.*

Cependant, lorsque nous avons observé les résultats il s'est avéré que (i) les EN issues de la ROC portant des contaminations et n'étant pas strictement orthographiées comme leur pendant de référence et (ii) les EN qui sont bien récupérées sur la ROC et pas sur la réf. même rares sont considérées par la machine comme des FP. Entre autres

31. Jeu de données de REN constitué pour l'évaluation.

exemples l'EN "*Ormeaux*" n'est pas strictement identique à sa version contaminée "*Ormaeuux*" ou "*Dconshire*" pour "*Devonshire*". Ces EN contaminées sont comptées comme des FP et ajoutées à la liste des hapax, ce qui vient gonfler artificiellement le nombre des FP dans l'ensemble des EN de la ROC. Il s'agit en fait de Faux Faux Positif (FFP), autrement dit de VP masqués par la rectitude de l'alignement inhérent au mode d'évaluation adopté. Ces différents cas sont décrits en détail dans la sous-section 4.4. Cet état de fait crée donc un biais dans nos évaluations, et cela pour les trois langues évaluées. Enfin, dans l'ensemble des EN reconnues sur les versions de ROC nous remarquons que la majorité des EN présentes dans les résultats obtenus sur la ROC sont effectivement présentes dans les résultats obtenus sur les versions de référence, comme le présente le tableau 5. Il n'y a donc pas de véritable déperdition des VP.

4.4. *Typologie des contaminations de la ROC pour une évaluation fine*

En regard des différentes observations que nous venons d'apporter et parce que nous souhaitons rendre compte de la complexité de ces cas réels, tels que nous les exposons dans les parties 4.2 et 4.3 ainsi que dans le tableau 7, nous proposons d'établir une typologie pour l'évaluation des contaminations de la ROC sur la REN, élargissant la classification standard des vrais/faux positifs/négatifs. Si les FP sont qualifiés de bruit et les FN de silence, nous avons repéré que l'évaluation par calcul des intersections cache des phénomènes de surestimation ou sous-estimation du bruit et du silence dans les données. Cette typologie permettrait d'établir quels sont les vrais bruits, autrement dit les vrais FP et les vrais silences les vrais VN.

CAS ATTENDUS

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

Vrais négatifs (VN) : Aucune EN à reconnaître dans les deux versions.

Faux positif (FP) : EN détectées à tort dans la version de ROC (bruit de la REN).

Faux négatif (FN) : EN manquantes dans la version de ROC (silence de la REN).

SOUS ÉVALUATION DU BRUIT ET DU SILENCE DE LA REN

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

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

SUR ÉVALUATION DU BRUIT ET DU SILENCE DE LA REN

Faux faux positifs (FFP) : EN détectées dans les versions de ROC mais pas dans le texte de référence (EN manquantes dans la référence$^{(i)}$ ou EN contaminées détectées dans la version de ROC$^{(ii)}$).

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

Type	Version	Contexte	spaCy-lg
FVP	Réf.	*[...] better than the milk-and-water lagrime*	lagrime
	Kraken	*[...] better than the _ilk- and-water lagrime*	lagrime
FVN	Réf.	*[...] l'été dans leur **propriété des Peuples***	()
	Kraken	*[...] l'ete dans leur **pro- priete des Peuples***	()
FFP (i)	Réf.	*[...] a sua entrada para o **colegio militar***	()
	Kraken	*[...] a s_a entrada para o colegio milita*	colegio milita
FFP (ii)	Réf.	*[...] e na vespera delle ir para Coimbra*	Coimbra
	Kraken	*[...] e na vespera delle ir para Coimhra*	Coimhra
FFN	Réf.	*[...] fleurs emblématiques que les Bachagas*	Bachagas
	Kraken	*[...] fleurs emble- matiques que les Bach'agas*	()

TABLEAU 7. *Exemples de cas réels d'EN justifiant de la typologie d'évaluation de l'impact des erreurs de la ROC sur la REN.*

4.5. *Évaluation supervisée des contaminations de la ROC sur un corpus annoté*

Afin d'évaluer de manière supervisée l'influence du bruit de la ROC sur la REN, nous avons annoté un échantillon du corpus ELTeC français. Nous avons choisi de nous limiter aux quatre catégories présentes dans `spaCy` (Lieux, Personnes, Organisations et Divers). Nous avons tout d'abord annoté un échantillon de 3 000 tokens d'une œuvre puis réalisé une adjudication pour régler les désaccords. Nous avons ensuite annoté 5 000 tokens de 3 versions (Réf., Tess. et Kraken) de deux œuvres (Daudet et Maupassant). L'accord inter-annotateur (Kappa de Fleiss (Fleiss et al., 2013)) était de 0.905 sur la version de réf., ce qui est, significativement plus élevé que le score obtenu sur les versions de ROC : 0.877. Nous avons pu observer que les désaccords étaient plus nombreux sur l'annotation des versions de ROC du fait des problèmes de tokenisation.

Grâce à un système de vote majoritaire, nous avons fusionné les annotations pour obtenir un *gold standard* sur chaque version de chaque œuvre. Nous avons évalué `spaCy_lg` sur cet échantillon, dont les résultats sont présentés dans le tableau 8.

	GLOBALE			LIEUX		
Souple	Rappel	Précision	F-sc.	Rappel	Précision	F-sc.
Kraken	49.57	73.72	59.28	48.84	52.50	50.60
Tesseract	51.53	77.63	61.94	56.41	57.89	57.14
Référence	49.78	77.55	60.64	53.49	53.49	53.49
Stricte	Rappel	Précision	F-sc.	Rappel	Précision	F-sc.
Kraken	18.26	58.82	27.87	43.24	45.71	44.44
Tesseract	21.00	68.66	32.16	43.33	44.83	44.07
Référence	21.62	69.57	32.98	41.18	45.16	43.08

TABLEAU 8. *Evaluation de `spaCy_lg` sur un échantillon annoté de 10 000 tokens dans trois versions textuelles différentes en configuration souple et en configuration stricte (GLOBALE : tous les types d'entités, LIEUX : seulement les lieux).*

À cet effet, nous proposons deux types d'évaluation que nous appelons ici « stricte » et « souple ». La première configuration ne prend en compte que les correspondances exactes des EN (p. ex. le mot contaminé *Acques* sera considéré comme

un FP, même s'il renvoie au VP *Jacques*). En revanche, la configuration « souple » considère une EN comme correcte, quelle que soit sa taille (une partie ou l'intégralité de l'EN, auquel cas la forme contaminée *Acques* sera considéré comme un VP. Par rapport à l'évaluation dite « globale » (qui concerne tous les types d'EN), nous pouvons remarquer que les résultats obtenus sur les versions Tess. fr sont meilleurs que ceux obtenus sur les versions Kraken dans les configurations stricte et souple. Plus étonnamment, les résultats de Tess. fr sont meilleurs par rapport à ceux de la référence dans le cadre de l'évaluation souple. Nous remarquons que là aussi la faiblesse apparente des résultats de la REN obtenus sur des versions de ROC est principalement due à des problèmes d'alignement entre les tokens contaminés et les tokens de référence. Enfin, le fait que le F-score soit meilleur sur les EN de lieux que sur le globale dans la configuration stricte peut s'expliquer par le fait que les EN de lieux comptent le plus souvent un token dans nos corpus (p. ex. *Paris*), au contraire des EN de personnes. Par exemple, *Daniel Eyssette*, qui est écrit *Daniel Ey-sset-te*[32]. Ainsi, la deuxième partie de l'EN est divisée en trois éléments distincts, ce qui a vraisemblablement posé problème à l'évaluation globale et diminué les F-scores. Dès lors, le rappel sur les noms de personnes descend beaucoup plus ; nous obtenons moins de VP, et les FP augmentent moins sur les lieux.

Nous concluons de ces travaux préliminaires d'évaluation de l'impact du bruit de la ROC sur la tâche de REN que les erreurs de la ROC ne sont pas toujours un frein à la bonne conduite de la tâche de REN, et que la présence de nombreux hapax dans une sortie de REN peut être le signe qu'il existe des formes contaminées d'EN. Néanmoins, nous constatons qu'il est difficile d'évaluer de manière stricte le silence et le bruit réel dans les sorties de REN sur données bruitées par la ROC, puisque l'alignement entre les versions de référence et de ROC du fait des formes contaminées des EN est une tâche ardue. Jusque là nous n'avons pas pu trouver de manière convaincante de calculer un F-score et nous commençons à entrevoir les limites de l'usage des intersections.

5. Analyse de l'impact des corrections de la ROC sur la REN

5.1. *Outils de la correction des sorties ROC utilisées dans le cadre de cette étude*

Nous avons utilisé la version 0.0.12 de JamSpell[33] (Jspll) pour la correction automatique des transcriptions de la ROC. Jspll est un outil développé en Python qui exploite un modèle de langue trigramme statistique (grain mot), en s'appuyant sur l'alphabet de la langue. Une partie des fonctionnalités, ainsi que les modèles de langue pour le français et l'anglais sont accessibles gratuitement sur le web, le modèle portugais est disponible uniquement dans la version payante, de fait cette option n'a pas été testée. Nous avons entraîné un modèle de langue pour Jspll pour chacune des trois

32. Les tirets représentent des retours à la ligne.
33. *Cf.* doc. : https://habr.com/en/articles/346618/

langues. Pour ce faire, nous avons sélectionné 40% de chacun des corpus mis à disposition par ELTeC et nous en avons exclu les textes utilisés pour notre étude. Nous avons procédé aux évaluations des différentes configurations présentées dans le tableau 9.

ROC	Kraken	Tess. en	Tess. fr	Tess. pt
REN	*sp*	*sp*	*sp*	*sp*
non corr.	✓	✓	✓	✓
Jspll-pretr.	✓	✓	✓	×
Jspll-ELTeC	✓	✓	✓	✓

TABLEAU 9. *Ensemble des configurations que nous évaluons dans cette étude. `spaCy_lg` : sp.*

5.2. *Typologie des contaminations de corrections de la ROC*

En observant les exemples de correction de la ROC présentés dans les tableaux 10 et 11, nous constatons des fluctuations au niveau de la performance du correcteur automatique. Notons le cas particulier de l'EN "Meunet-sur-Vatan" dont on constate différentes déclinaisons en fonction du type de correcteur automatique (tableau 10). Nous nous apercevons que les différentes versions de cette EN, contaminée par les différentes OCRisations et sur-corrections, n'ont pas du tout été extraites par `spaCy_lg`.

Version	**Contexte**	`spaCy_lg`
Réf.	*[…] à l'assemblée de Meunet-sur-Vatan ;*	Meunet-sur-Vatan
Kraken	*[…] a l'assemhl6e' de Neunet-sur- Yatan'*	Yatan
Kraken Jspll-fr	*[…] a l'assembl6e' de Neuner-sur- Satan' ;*	()
Kraken ELTeC-fr	*[…] a l'assemhl6e' de Neunet-sur-Avant' ;*	()
Tess fr	*[…] à l'assemblée' de Meunet-sur-| Vatan* ;*	Meunet-sur-|
Tess fr Jspll-fr	*[…] à l'assemblée' de Meuret-sur- | Vatan* ;*	()
Tess fr ELTeC-fr	*[…] à l'assemblée' de Meunet-sur-| Vatan* ;*	Meunet-sur_

TABLEAU 10. *Exemples illustrant l'impact de la correction de la ROC sur la REN avec `spaCy_lg`.* La petite Jeanne*, Carraud.*

Version	**Contexte**	`spaCy_lg`
Réf.	*[...] before you went to India.*	India
Kraken	*[...] before you went to Iudia.*	Iudia
Kraken Jspll	*[...] before you went to India*	India
Kraken ELTeC	*[...] before you went to Iudia.*	Iudia

TABLEAU 11. *Exemples illustrant l'impact de la correction de la ROC sur la REN avec `spaCy_lg`.* Vanity Fair*, Thackeray.*

Similairement, nous observons dans le tableau 11 que le modèle de la correction automatique de la ROC par Jspll, entraîné sur le corpus ELTeC, n'a pas eu d'impact sur l'extraction de l'EN "Iudia", puisqu'elle n'avait pas été corrigée en l'EN de référence

"India". Par contre, le modèle Jspll pré-entraîné a bien corrigé la même EN, ce qui a permis son extraction sous forme correcte. À partir des exemples présentés dans les tableaux 10 et 11, nous déduisons une typologie des corrections automatiques de la ROC, résumée dans le tableau 12. Cela permet de distinguer les différents cas de figure où les corrections en question ont soit amélioré les sorties de ROC (MOBC), soit les ont incorrectement modifiées (MOMC, BOIC) ou même ignorées (MOI).

Type Acronyme	Définition
MOBC	mal océrisées bien corrigées
MOMC	mal océrisées mal corrigées
MOI	mal océrisées ignorées
BOIC	bien océrisées indûment corrigées

TABLEAU 12. *Typologie de l'impact de la correction de la ROC sur la REN.*

Pour illustrer ce propos, quelques exemples sont indiqués dans le tableau 13, parmi lesquels se distinguent les sur-corrections "Conspire" (au lieu de "Devonshire" dans ELTeC anglais), ainsi que "Martincourt" (au lieu de "Morlincourt" dans ELTeC français).

Type	Version	Contexte	`spaCy_lg`
MOBC	Kraken	*[…] when they were in Lonlon*	()
	Jspll	*[…] when they were in London*	London
	ELTeC	*[…] when they were in London*	London
MOMC	Kraken	*[…] flowery lanes peeuliar to Dconshire ;*	Dconshire
	Jspll	*[…] flowery lanes peculiar to Conspire ;*	()
	ELTeC	*[…] flowers lanes peculiar to Dconshire ;*	Dconshire
MOI	Kraken	*cure de Mlorlincourtl*	Mlorlincourtl
	Jspll-fr	*cure de Mlorlincourtl*	Mlorlincourtl
	ELTeC-fr	*cure de Mlorlincourtl*	Mlorlincourtl
BOIC	Kraken	*en retournant de Morlincourt*	Morlincourt
	Jspll-fr	*en retournant de* ***Martincourt***	**Martincourt**
	ELTeC-fr	*en retournant de Morlincourt*	Morlincourt

TABLEAU 13. *Exemples illustrant la typologie de l'impact de la correction de la ROC sur la REN pour les configurations avec* `spaCy_lg`*. Formes de références des entités : London, Devonshire, Morlincourt.* Home influence*, Aguillar et* Mon village*, Adam.*

5.3. *Analyses quantitatives des contaminations de la ROC et de leurs corrections*

5.3.1. *Le CER médian : indice de la performance de la correction automatique ?*

Les colonnes *Brut* du tableau 14 montrent qu'il y a globalement plus de types d'EN récupérées par `spaCy` sur Kraken (CER médian pour en : 0.36, fr : 0.10, pt : 0.16) que sur Tess. (CER médian en : 0.20, fr : 0.05, 0.09), ce qui vient étayer l'hypothèse que

plus la qualité de la transcription est mauvaise plus les variations orthographiques des EN peuvent être nombreuses et plus il y a d'hapax dans les résultats de la REN pour les trois langues. Les CER médians sont moins bons pour l'anglais que pour les deux autres langues à cause du texte de Reynolds.

	Kraken						Tesseract					
	Brut	Jspll pré-entraîné		Jspll-ELTeC			Brut	Jspll pré-entraîné		Jspll ELTeC		
Anglais	3 121	1 691	54%	1 446	46%		2 789	2 160	77%	1 882	65%	
CER médian	0.364	0.362		0.364			0.205	0.261		0.243		
Français	7 059	5 317	75%	**6 755**	**95%**		4 884	4 634	95%	**4 917**	**100,7%**	
CER médian	0.098	0.102		0.104			0.050	0.053		0.055		
Portugais	10 108	N/A	N/A	6 147	61%		4 006	N/A	N/A	3 594	90%	
CER médian	0.159	N/A		0.155			0.093	N/A		0.096		

TABLEAU 14. *Nombre d'EN types avec les pourcentages d'EN issues des EN brutes et repérées par `spaCy_lg` pour les corpus ELTeC anglais, français et portugais. N/A – modèle Jspll pré-entraîné pour le portugais non disponible.*

Nous venons à penser que si la correction automatique fonctionne bien, le nombre des hapax sera réduit dans les sorties de REN, puisque la variabilité du vocabulaire (tokens types et EN types) sera réduite. Pour affiner notre analyse, nous avons décider d'utiliser le CER médian plutôt que le CER moyen, car la moyenne est assujettie aux aberrations plus que la médiane. En partant de ce postulat, nous observons que :

– l'observation des CER médians des versions Tess. corrigées par rapport à la version Tess. brute montre que la qualité baisse, alors que celle des CER médians pour les versions Kraken corrigées par rapport à la version Kraken brute montre une stagnation. Les résultats sont plus explicites pour l'anglais et le portugais. Il y aurait un effet de seuil concernant la qualité des versions de ROC au-delà de laquelle la correction automatique serait moins efficace. Autrement dit, plus une transcription serait de bonne qualité, moins la correction serait pertinente.

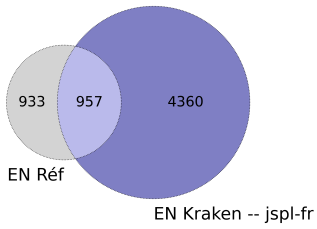
– même si la correction automatique permet de diminuer le vocabulaire et le nombre des hapax (-48% du vocabulaire pour l'anglais sur Kraken Jspll-pré-entraîné et -54% pour l'anglais Jspll-ELTeC), il semble, au vu des CER médians, que la qualité des transcriptions ne soit pas grandement améliorée par les modèles de correction automatique (0.362% et 0.364%).

– en effet, les CER médians ne montrent pas de performance significative des modèles de correction automatique, et même il semble qu'ils soient dégradés dans certains cas. La correction automatique avec Jspll ELTeC sur Tesseract pour le français voit le nombre des hapax augmenter, 100,7% des EN récupérées, ce qui semble indiquer qu'il y a de nouvelles EN (hapax) récupérées, ce qui peut être dû au phénomène de sur-correction (BOIC, tableau 13).

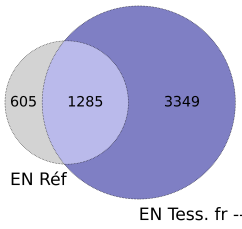
Enfin, concernant le corpus ELTeC portugais, comme nous l'avons souligné précédemment, nous avons extrait uniquement les EN corrigées avec notre modèle Jspll-ELTeC pour le portugais, et la quantité d'EN reconnues est moindre que celle trouvée sur la sortie brute de ROC, donc la correction semble avoir été pertinente.

5.3.2. *Calculs des intersections, toujours plus de problèmes d'alignements.*

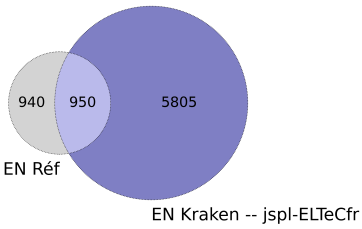
Nous reprenons ici la stratégie d'évaluation stricte par calcul des intersections comme décrite dans la section 4.3. Les graphiques de la figure 2 représentent les intersections entre les EN issues des textes de référence et celles provenant des versions de ROC (Kraken ou Tesseract) corrigées avec le modèle pré-entraîné de Jspll (2a-2b), et le modèle entraîné avec le corpus ELTeC français (2c-2d). La figure 2d montre que la configuration Tesseract Jspll-ELTeC-fr `spaCy_lg` a permis de récupérer le plus grand nombre d'EN en commun. Il est notable que la correction automatique avec le modèle pré-entraîné de Jspll ou le modèle entraîné sur une partie du corpus ELTeC adapté à la langue du corpus testé ne sont pas un gain pour l'intersection.



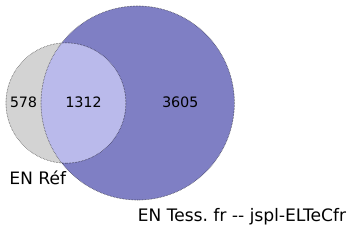
(a) *Kraken–Jspll pretrain-fr –* `spaCy_lg`



(b) *Tess. fr.–Jspll pretrain-fr –* `spaCy_lg`



(c) *Kraken–Jspll ELTeC-fr –* `spaCy_lg`



(d) *Tess. fr.–Jspll ELTeC-fr –* `spaCy_lg`

FIGURE 2. *Intersections pour les configurations Kraken et Tess. fr. corrigées par JamSpell pré-entraîné et modèle ELTeC,* `spaCy_lg` *sur le corpus ELTeC français.*

Ce fait peut être lu à l'aune des observations présentées dans le tableau 13 rapportant la typologie des erreurs de corrections. Autrement dit, la correction automatique ne transforme pas toutes les EN contaminées par la ROC en EN corrigées strictement associables avec les EN du groupe de référence. Ainsi les BOIC ("Morlincourt" qui devient "Martincourt"), se cumulant aux EN contaminées non corrigées MOI ("Morlincourtl" qui reste "Morlincourtl"), n'améliorent pas les résultats obtenus par calcul des intersections. Il semble que pour les corpus ELTeC français, anglais et portugais et celui de la TGB la correction automatique avec le modèle entraîné sur une partie de chaque corpus ELTeC fasse perdre 5% des EN dans l'intersection avec Kraken et

10% avec Tesseract, alors que concernant les modèles pré-entraînés on perd 3% avec Kraken et 9% avec Tesseract. Cette expérience est l'occasion de démontrer les limites d'une évaluation stricte de la REN sur des textes bruités et leurs versions corrigées. Nos observations manuelles montrent que les contaminations de la ROC d'une part et de la correction automatique d'autre part ne sont pas véritablement un frein à la REN, mais l'évaluation automatique des résultats n'est pas triviale.

5.4. *Comment dépasser les problèmes d'alignements ?*

5.4.1. *Mesures de distance textuelle*

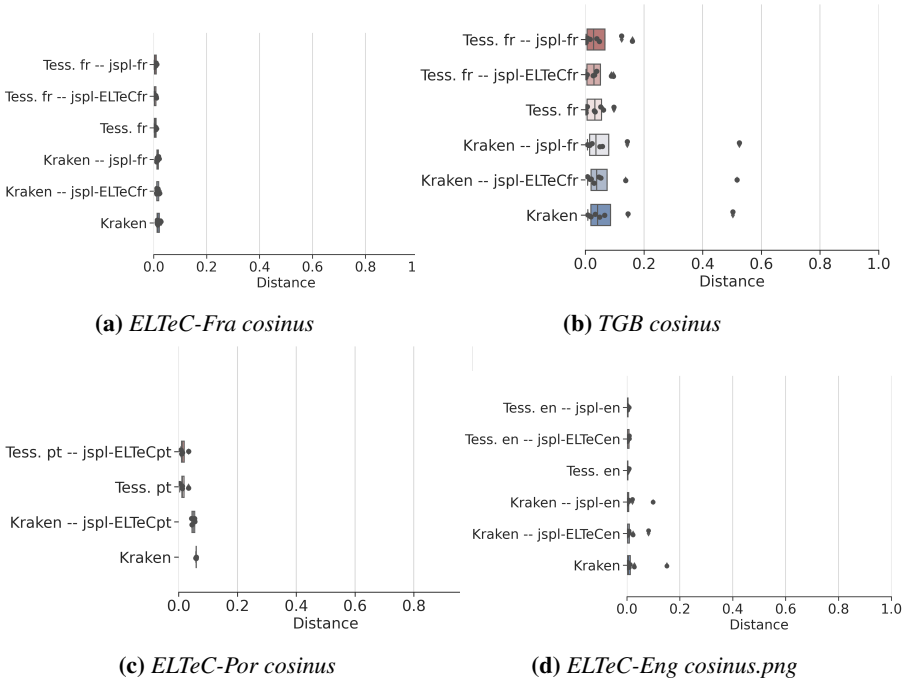
Dans le but d'approfondir nos évaluations et de dépasser les verrous de l'évaluation stricte, nous employons des mesures de distance textuelle afin de rendre plus souple nos critères d'évaluation des résultats de la REN sur les sorties de ROC bruitées et leurs corrections automatiques. Nous avons privilégié les métriques de Jaccard[34] et cosinus, calculées sur les bigrammes et trigrammes de caractères[35], car elles sont considérées comme des mesures de référence quand il est question de (dis)similarité textuelle (Buscaldi et al., 2020).

Pour lire les figures 3 et 4, il faut noter que plus la boîte est proche de zéro, plus les sorties comparées sont similaires. La figure 3 illustre les résultats obtenus pour les textes de réf. et les différentes versions de ROC pour tous les corpus avec une distance cosinus. La figure 4 montre les résultats en comparant les sorties de REN obtenues sur les textes de réf. et celles des différentes configurations évaluées (tableau 9). Après une observation des différentes mesures de distance présentées pour chaque corpus évalué, on note l'écart constant et considérable entre les résultats de Jaccard et cosinus. Les résutats pour Jaccard sont souvent proches de 1, tandis que ceux de cosinus sont proches de 0 pour les mêmes configurations. Il semble que les métriques cosinus et Jaccard ne mesurent pas la même chose.

La consultation manuelle des résultats de la REN montre que la différence entre les résultats de Jaccard et cosinus peut s'expliquer par le fait que la première mesure prend en compte le vocabulaire, alors que la seconde s'intéresse au nombre d'occurrences d'une EN. Concrètement, cela signifie pour la distance de Jaccard que si le vocabulaire entre les sorties de deux ensembles comparés est différent même en moindre proportion, le résultat est proche de 1. Pour Jaccard, s'il manque un mot du vocabulaire, cela impactera beaucoup le résultat, alors que pour cosinus ce n'est pas le cas. En effet, les résultats pour la mesure cosinus dépendent du nombre d'occurrences de chaque EN dans les groupes comparés. Autrement dit, s'il y a beaucoup d'occurrences d'une EN dans la configuration de réf. mais qu'elle n'apparaît pas en même quantité dans la configuration ROC, les résultats pour cosinus grimpent en flèche. L'observation des résultats roman par roman pour la mesure cosinus nous permet d'étayer cette

34. Nous présentons uniquement des résultats pour cosinus ; pour les résultats Jaccard, *cf.* le dépôt GitHub : https://github.com/anonymous

35. Nous avons vectorisé le texte avec la librairie `CountVectorizer` pour les deux distances.



(a) *ELTeC-Fra cosinus* **(b)** *TGB cosinus*

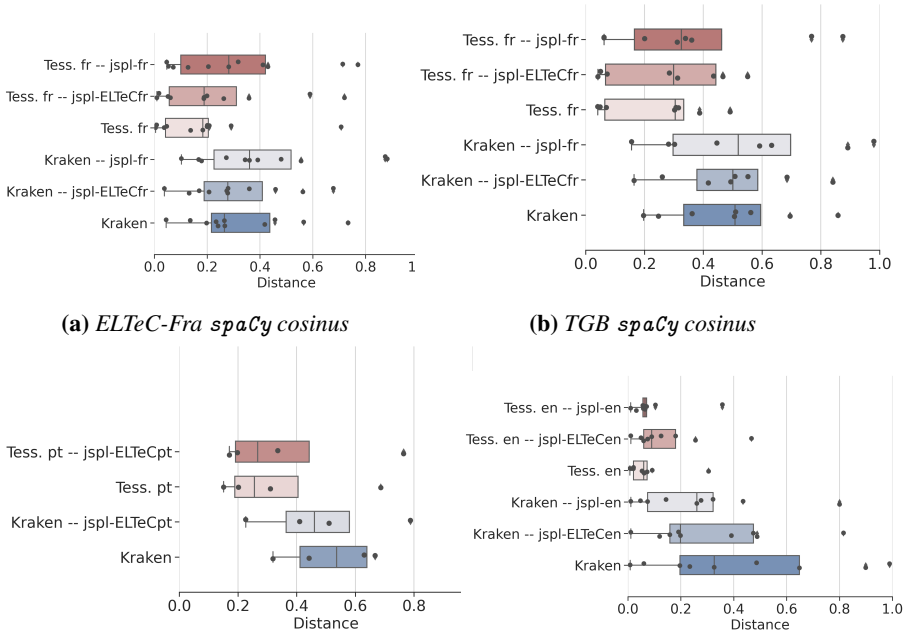
(c) *ELTeC-Por cosinus* **(d)** *ELTeC-Eng cosinus.png*

FIGURE 3. *Distances calculées entre les textes de référence et les versions de ROC.*

hypothèse, en effet on dénombre, p. ex. 290 occurrences du terme “INGLEZA” pour la configuration Tesseract-pt–`spaCy_lg`, alors qu’il apparaît 3 fois seulement dans les résultats de la configuration de réf.[36] – dans ce dernier cas la valeur de cosinus est très élevée et dépasse celle de Jaccard (cos. : 0.69, Jaccard : 0.67). Nous observons un comportement analogue pour chacun des corpus analysés, nous en concluons que la distance de Jaccard a tendance à sur-estimer la distance entre deux ensembles et être plus susceptible au bruit.

La figure 4 laisse apercevoir que les résultats de la REN sur les versions Kraken des textes sont moins bons de manière générale que pour les versions produites avec Tesseract. Toutefois, il semblerait que la correction automatique soit un peu plus efficace sur les versions de Kraken avec le modèle Jspll-ELTeC que sur les versions Tesseract, car l’écart entre les boîtes est plus grand. Cependant, les résultats des distances obtenus pour ces versions corrigées de Kraken restent inférieurs à ceux observés pour les versions Tesseract avec et sans corrections. Ces différents constats laissent à penser que plus une version de ROC est bruitée, plus le correcteur automatique intervient et produit de bonnes corrections (figure 4c). À l’inverse, si une version de ROC est peu bruitée, alors le correcteur automatique aura tendance à moins bien corriger, voire à

36. *Uma familia ingleza*, Diniz.



(a) *ELTeC-Fra* `spaCy` *cosinus*

(b) *TGB* `spaCy` *cosinus*

(c) *ELTeC-Por* `spaCy` *cosinus*

(d) *ELTeC-Eng* `spaCy` *cosinus*

FIGURE 4. *Distance Cosinus pour* `spaCy_lg` *sur chaque corpus globalement.*

sur-corriger. On peut observer ce phénomène concernant les résultats de la REN sur Tesseract qui sont moins bons sur les versions Tesseract corrigées, en le mettant en regard avec nos observations sur le tableau 14 – c'est une deuxième manière d'analyser le phénomène de sur-correction.

5.4.2. NERVAL *: Précision, rappel, f-score et effet de seuil*

Dans le but de calculer la précision, le rappel et d'obtenir un f-score nous avons utilisé l'outil NERVAL[37], évalué par (Koudoro-Parfait et al., 2022). Si cette évaluation présente quelques biais de l'outil, NERVAL apparaît tout de même comme un très bon moyen de dépasser les problèmes d'alignements entre les résultats des différentes configurations à comparer pour calculer le f-score. NERVAL est développé en Python, et est conçu pour l'évaluation de sorties de REN sur du texte bruité avec la distance de Levenshtein. Les fichiers des textes de référence et des versions de ROC et de ROC corrigées sont annotés au format IOB avec `spaCy_lg`. Les fichiers des textes de références ainsi annotés font office de vérité de terrain. Les premières observations des résultats semblent confirmer que la correction automatique n'est pas forcément un gain pour la REN, en effet le f-score pour les configurations de Tesseract dans les tableaux

37. (Miret et Kermorvant, 2021), https://gitlab.com/teklia/nerval

15 et 16 perd en moyenne 0.06 points. À l'inverse, le f-score sur les configurations de Kraken semble légèrement augmenter, ce constat venant illustrer le phénomène de creux que nous évoquions dans la partie 5.3.1.

	#Entités		Évaluation par NERVAL			
Version	ROC	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken	1 122	944	566	0.50	0.76	*0.61*
Tess fr	860	944	646	0.75	0.87	**0.81**
Kraken + Jspll-fr	1 027	944	471	0.46	0.63	**0.53** ⇓
Tess fr + Jspll-fr	794	944	532	0.67	0.72	***0.69*** ⇓
Kraken + ELTeC-fr	1 055	944	548	0.52	0.74	*0.61* ⇑
Tess fr + ELTeC-fr	838	944	621	0.74	0.84	*0.79* ⇓

TABLEAU 15. *Résultat de* NERVAL *sur* Le petit chose*, Daudet.*

	#Entités			Évaluation par NERVAL			
Version	Label	ROC	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken	LOC	180	168	89	0.50	0.53	0.51
Tess		161	168	130	0.81	0.77	**0.79**
Kraken	GPE	1 925	1 324	824	0.43	0.62	0.51
Tess		1 464	1 324	1 080	0.74	0.82	**0.78**
Kraken + Jspll-en	LOC	158	168	105	0.67	0.63	0.64⇑
Tess+ Jspll-en		152	168	119	0.79	0.71	0.75 ⇓
Kraken + Jspll-en	GPE	1 542	1 324	910	0.59	0.69	0.64⇑
Tess + Jspll-en		1 411	1 324	1 030	0.73	0.78	0.75⇓
Kraken + ELTeC-en	LOC	176	168	99	0.56	0.59	0.58⇑
Tess + ELTeC-en		158	168	120	0.76	0.71	0.74 ⇓
Kraken + ELTeC-en	GPE	1 149	1 324	743	0.65	0.56	0.60⇑
Tess + ELTeC-en		1 131	1 324	868	0.77	0.66	0.71⇓

TABLEAU 16. *Résultat de* NERVAL *sur* Vanity Fair*, Thackeray.*

6. Conclusion

Dans ce travail, nous avons mené des expériences sur la correction automatique des contaminations de la ROC, avec l'objectif de mesurer l'impact de ces corrections sur la reconnaissance d'EN spatiales. Notre étude s'appuie sur les corpus littéraires ELTeC (ouvrages en anglais, français et portugais), ainsi que sur celui de la TGB (ouvrages en français), dont les versions de ROC ont été corrigées à l'aide de deux modèles de l'outil JamSpell (l'un fournit par défaut pour l'anglais et le français et l'autre entraîné sur le corpus ELTeC selon la langue adéquate). Les résultats ont montré que, contre-intuitivement, la correction automatique introduit des biais (notamment les sur-corrections) dans les données textuelles et que le gain apporté par les corrections justes n'était pas considérable. En revanche, pour les ouvrages en anglais et en français transcrits avec Tesseract, nous avons constaté un gain de performance léger mais quasi systématique du correcteur JamSpell–ELTeC. A contrario le modèle pré-entraîné semble plus souvent apporter des sur-corrections. Dans la suite de notre travail, nous nous appuierons sur l'utilisation d'un autre outil qui utilise des réseaux de neurones, qui serait susceptible de corriger automatiquement les contaminations de la ROC de manière plus probante.

7. Bibliographie

Alex B., Grover C., Klein E., Tobin R., « Digitised historical text : Does it have to be mediOCRe ? », in J. Jancsary (ed.), 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012, vol. 5 of Scientific series of the ÖGAI, ÖGAI, Wien, Österreich, p. 401-409, 2012.

Azmi A., Almutery M., Aboalsamh H., « Real-Word Errors in Arabic Texts : A Better Algorithm for Detection and Correction », IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, p. 1-1, 05, 2019.

Bassil Y., Alwani M., « Post-Editing Error Correction Algorithm for Speech Recognition using Bing Spelling Suggestion », CoRR, 2012.

Buscaldi D., Felhi G., Ghoul D., Le Roux J., Lejeune G., Zhang X., « Calcul de similarité entre phrases : quelles mesures et quels descripteurs ?(Sentence Similarity : a study on similarity metrics with words and character strings) », Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, p. 14-25, 2020.

Chiron G., Doucet A., Coustaty M., Visani M., Moreux J.-P., « Impact of OCR errors on the use of digital libraries Towards a better access to information », 2017 ACM IEEE Joint Conference on Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, Toronto, Canada, June, 2017.

Damerau F. J., « A Technique for Computer Detection and Correction of Spelling Errors », Commun. ACM, vol. 7, n° 3, p. 171–176, mar, 1964.

Dumas Milne Edwards L., Conception de formes de relecture dans les chaînes éditoriales numériques, Theses, Université de Technologie de Compiègne, January, 2016.

Eshel Y., Cohen N., Radinsky K., Markovitch S., Yamada I., Levy O., « Named Entity Disambiguation for Noisy Text », CoRR, 2017.

Evershed J., Fitch K., « Correcting noisy OCR : Context beats confusion », Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, p. 45-51, 2014.

Fleiss J. L., Levin B., Paik M. C., Statistical Methods for Rates and Proportions, John Wiley & Sons, 2013.

Gabay S., Clérice T., Reul C., « OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more) », May, 2020, working paper or preprint.

Hamdi A., Jean-Caurant A., Sidère N., Coustaty M., Doucet A., « Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition », 24th International Conference on Theory and Practice of Digital Libraries 2020, Lyon, France, p. 87-101, August, 2020.

Hamdi A., Linhares Pontes E., Sidère N., Coustaty M., Doucet A., « In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking », Natural Language Engineering, March, 2022.

Huynh V.-N., Hamdi A., Doucet A., « When to Use OCR Post-correction for Named Entity Recognition ? »,

22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, p. 33-42, November, 2020a.

Huynh V.-N., Hamdi A., Doucet A., « When to Use OCR Post-correction for Named Entity Recognition ? », in E. Ishita, N. L. S. Pang, L. Zhou (eds), Digital Libraries at Times of Massive Societal Transition, Springer International Publishing, Cham, p. 33-42, 2020b.

Kiessling B., « Kraken-an universal text recognizer for the humanities », ADHO, Éd., Actes de Digital Humanities Conference, 2019.

Kiessling B., Tissot R., Stokes P., Ezra D. S. B., « eScriptorium : An open source platform for historical document analysis », 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 2, IEEE, p. 19-19, 2019.

Koudoro-Parfait C., Lejeune G., Buth R., « Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique (Resolution of entity linking issues on noisy OCR output : automatic disambiguation tracks) », Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN), p. 45-55, 2022.

Koudoro-Parfait C., Lejeune G., Roe G., « Spatial Named Entity Recognition in Literary Texts : What is the Influence of OCR Noise ? », in L. Moncla, C. Brando, K. McDonough (eds), GeoHumanities@SIGSPATIAL 2021 : Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021, ACM, p. 13-21, 2021.

Lejeune G., Brixtel R., Doucet A., Lucas N., « Multilingual Event Extraction for Epidemic Detection », Artificial Intelligence in Medicine, vol. 65, nº 2, p. 131-143, October, 2015.

Miret B., Kermorvant C., « Nerval : a python library for named-entity recognition evaluation on noisy texts », , Teklia, 2021.

Montani I., O'Leary McCann P., Geovedi J., O'Regan J., Samsonov M., Orosz G., de Kok D., Blättermann M., Altinok D., Kristiansen S. L., Kannan M., Mitsch R., Bournhonesque R., Edward, Miranda L., Baumgartner P., Hudson R., Bot E., Roman, Fiedler L., Daniels R., Phatthiyaphaibun W., Howard G., Tamura Y., « explosion/spaCy : v3.5.1 : spancat for multi-class labeling, fixes for textcat+transformers and more », March, 2023.

Nguyen N. K., Boros E., Lejeune G., Doucet A., « Impact Analysis of Document Digitization on Event Extraction », Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI), 19th International Conference of the Italian Association for Artificial Intelligence, -, Roma, Italy, p. to appear, 2020.

Nguyen T. T. H., Jatowt A., Coustaty M., Doucet A., « Survey of Post-OCR Processing Approaches », ACM Comput. Surv., jul, 2021.

Oger S., Rouvier M., Camelin N., Kesler R., Lefevre F., Torres-Moreno J.-M., Système du LIA pour la campagne DEFT2010, Lavoisier, 01, 2012.

Petkovic L., « Impact de la correction automatique de l'OCR/HTR sur la tâche de reconnaissance d'entités nommées dans un corpus bruité », Actes de la journée d'étude sur la robustesse des systemes de TAL, vol. 1, p. 22, 2022.

Petkovic L., Alrahabi M., Roe G., « Impact de la correction automatique de l'OCR/HTR sur la reconnaissance d'entités nommées dans un corpus bruité », JIS - Journal of Information Sciences, vol. 21, nº 2, p. 42-57, December, 2022.

Qi P., Zhang Y., Zhang Y., Bolton J., Manning C. D., « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages », Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations, 2020.

Rahimi A., Li Y., Cohn T., « Massively Multilingual Transfer for NER », Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, p. 151-164, July, 2019.

Reul C., Dittrich M., Gruner M., « Case Study of a highly automated Layout Analysis and OCR of an incunabulum : 'Der Heiligen Leben'(1488) », Proceedings of the 2nd international conference on digital access to textual cultural heritage, p. 155-160, 2017.

Sagot B., Gábor K., « Named Entity Recognition and Correction in OCRized Corpora (Détection et correction automatique d'entités nommées dans des corpus OCRisés) [in French] », Traitement Automatique des Langues Naturelles, TALN 2014, Marseille, France, 1-4 Juillet 2014, articles courts, The Association for Computer Linguistics, p. 437-442, 2014.

Sagot B., Romary L., Bawden R., Suárez P. J. O., Christensen K., Gabay S., Pinche A., Camps J.-B., « Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue », DataLab de la BnF : Restitution des travaux 2022, 2022.

Smith R., « An overview of the Tesseract OCR engine », Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, IEEE, p. 629-633, 2007.

Stanislawek T., Wróblewska A., Wójcicka A., Ziembicki D., Biecek P., « Named Entity Recognition - Is There a Glass Ceiling ? », Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), p. 624-633, November, 2019.

Tual S., Abadie N., Chazalon J., Duménieu B., Carlinet E., « A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents », CoRR, 2023.

van Strien D., Beelen K., Ardanuy M., Hosseini K., McGillivray B., Colavizza G., « Assessing the Impact of OCR Quality on Downstream NLP Tasks. », In Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1 : ARTIDIGH, p. 484 - 496, 2020.

Wisniewski G., Max A., Yvon F., « Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia », Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs, ATALA, Montréal, Canada, p. 121-130, July, 2010.