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

Caroline Koudoro-Parfait* — Ljudmila Petkovic* — Glenn Roe*

* Sorbonne Université, Observatoire des textes, des idées et des corpus (OBTIC), Paris, France

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 souples. 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, REN, correction automatique de ROC.

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, NER, Automatic OCR Correction.

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, REN) d’une part, et d’autre part dans la valorisation des corpus patrimoniaux (reconnaissance optique de caractères, 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 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 des 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 (TGB¹ et ELTeC²) sur lesquels nos analyses s’appuient. La section 4 présente différentes méthodes d’évaluation manuelles et automatiques de l’impact des contaminations³ 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 spaCy⁴ (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 JamSpell⁵ (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) : celle des erreurs lexicales (*non-word errors*) qui ne représentent pas des mots valides de la langue, par exemple si le mot « Morlincourt »⁶ est orthographié comme « Mlolincourtl », et celle, beaucoup plus restreinte, des erreurs grammaticales (*real-word errors*) (Wisniewski *et al.*, 2010) auxquelles pourraient s’ajouter les erreurs sémantiques (*semantic/context-sensitive errors*), quand p. ex. « Gélongs »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 : les méthodes exploitant des lexiques, les méthodes sur des modèles de langue statistiques, et les 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 Gallica⁹ étaient affectées 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 (*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 mots-clés.

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. « MlorlincourtI » (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 EN¹² (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 (fr), anglais (en) et portugais (pt), ainsi que pour les textes de la Très Grande Bibliothèque (TGB), nous disposons des textes que nous nommons « référence » (version des textes sans corrections manuelles, ci-après « réf. »). Nous présentons ces corpus dans les tableaux 1, 2a, 2b et 2c. 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, deux transcriptions différentes ont été produites à partir des documents PDF des œuvres, notamment en utilisant : (i) *Kraken*¹³ (Kiessling *et al.*, 2019) et (ii) *Tesseract*¹⁴, et cela pour chacune des langues des corpus : anglais (*Tess.en*), français (*Tess.fr*) et portugais (*Tess.pt*). Nous nommons ces corpus d'évaluation *small* suivi du nom du corpus et du code ISO 639-1 de la langue soit : *small-ELTeC-{fr, en, pt}* et *small-TGB-fr*.

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 à disposition des œuvres, issues des collections Gallica, transcrites par la ROC. Le corpus est constitué de 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 siècle, 7 294 du XVIII^e siècle, 54 du XX^e siècle et 24 du XVII^e siècle.¹⁵ D’après une gestionnaire du service SINDBAD¹⁶ 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 spaCy¹⁷ dans les textes que nous avons sélectionnés comme textes de réf. sont présentés dans le tableau 1.

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

TABLEAU 1. Statistiques sur le corpus small-TGB-fr (*spaCy_lg* est le modèle large de *spaCy*. Cette 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 langues européennes dont certains ont pour source le site web du projet Gutenberg¹⁸,

15. Un nombre considérable de documents représentant des rééditions de textes plus anciens.

16. <https://www.bnf.fr/fr/une-question-pensez-sindbad>

17. Cet outil et ses configurations utilisées seront détaillés dans la partie 4.1.2.

18. <https://www.gutenberg.org/>

Gallica mais aussi la Bibliothèque électronique du Québec¹⁹. Les textes disponibles sont tous de bonne qualité. L'action COST a mis en place une liste de critères²⁰ 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 XIX^e siècle et le début du XX^e 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 PDF 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 XVII^e siècle (Gabay et al., 2020)²¹, 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 segmentation²² et la transcription²³. Concernant Tesseract, nous avons utilisé le modèle neuronal 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 intégrée à travers le réseau neuronal pour la segmentation des cadres (*box segmentation*), ce qui rend le traitement des mises en page complexes plus difficile (Reul

19. <http://beq.ebooksgratuits.com/>, Jean-Yves Dupuis 1998-2018.

20. <https://github.com/distantreading/WG1/wiki/E5C-discussion-paper>

21. <https://github.com/Heresta/OCR17plus/tree/main/Model>

22. Le modèle de segmentation est constitué d'un réseau neuronal d'étiquetage par classification des phrases (*seed-labeling network*).

23. Le modèle de transcription fonctionne comme un classifieur de séquences sans segmentation qui utilise un réseau neuronal 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
<i>Mon village</i>	J. Adam	1860	200	20 938	213
<i>Marie-Claire</i>	M. Audoux	1925	120	35 780	101
<i>Le Château de Pinon, vol. 1</i>	G. A. Dash	1844	332	44 246	271
<i>La Petite Jeanne</i>	Z. Carraud	1884	220	53 212	316
<i>La Nouvelle Espérance</i>	A. de Noailles	1903	325	54 272	182
<i>Une vie</i>	G. de Maupassant	1883	337	75 745	302
<i>Albert Savarus. Une fille d'Ève</i>	H. de Balzac	1853	60	79 924	682
<i>Le Petit Chose</i>	A. Daudet	1868	292	86 482	744
<i>Les Trappeurs de l'Arkansas</i>	G. Aimard	1858	450	91 119	646
<i>La Belle Rivière</i>	G. Aimard	1894	339	137 392	1 004
<i>L'Éducation sentimentale</i>	G. Flaubert	1880	520	150 494	1 304

(a) corpus small-ELTeC-fr

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

(b) corpus small-ELTeC-en

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

(c) corpus small-ELTeC-pt

TABLEAU 2. Statistiques sur les corpus small-ELTeC français, anglais et portugais. La dernière colonne indique le nombre total d'EN.

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 neuro-naux pour la ROC ont été utilisés dans le cadre de ces expériences : Kraken de base, Tess.en, Tess.fr et Tess.pt contemporain.

4.1.2. L'outil de REN

Pour effectuer la tâche de REN, 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 » (*Bloom embeddings*)²⁴, ainsi qu'un réseau neuronal convolutif avec des connexions résiduelles, ce qui peut expliquer sa robustesse lors de l'extraction

24. Il s'agit de la structure de données probabiliste qui permet de réduire la dimension des vecteurs (<https://explosion.ai/blog/bloom-embeddings>).

des EN contaminées. spaCy propose des modèles de langue du type *large* pour le français²⁵, le portugais²⁶ et l’anglais²⁷.

Les modèles français et portugais sont chacun entraînés sur les dépendances universelles (*Universal Dependencies*, UD) adaptées à leur propre langue (UD French Sequoia v2.8 et UD Portuguese Bosque v2.8 respectivement), ainsi que sur WikiNER et sur Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia); le modèle français s’appuie aussi sur spaCy *lookups data*, soit les ressources supplémentaires mises à disposition pour chaque langue comme point d’entrée²⁸. Le modèle anglais, quant à lui, est entraîné sur Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl), sur WordNet 3.0, sur OntoNotes 5 et sur ClearNLP Constituent-to-Dependency Conversion²⁹. Nous avons favorisé l’usage du modèle *large* (spaCy_lg) plutôt que celui 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 (*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 bruitées³⁰. Nous comparons les résultats obtenus automatiquement sur les corpus small-ELTeC-{fr, en, pt} et small-TGB-fr (annotations automatiques de réf.), et ceux obtenus sur leurs transcriptions de ROC, sans nous appuyer sur un *gold standard*. Nous observons que les fautes d’orthographe 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 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, p. ex. « *ancehester* » pour « Manchester ».

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. Cela évite le téléchargement des ressources volumineuses pour toutes les langues par défaut, <https://github.com/explosion/spacy-lookups-data>

29. Les tailles des modèles sont les suivantes : 545 Mo pour le français, 560 Mo pour l’anglais et 541 Mo pour le portugais.

30. Tous les résultats des expériences sont disponibles sur le dépôt GitHub : https://github.com/These-SCAI2023/EXPE42_EVALUATION_ELTeC-fra_09072024

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

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

	small-ELTeC-en			small-ELTeC-fr			small-ELTeC-pt		
	Reynolds	Troll.	Brontë	Daudet	Adam	Maup.	Diniz	Queiroz	Osorio
CER Tess. + lang.	0,10	0,12	0,25	0,03	0,09	0,12	0,06	0,08	0,10
WER Tess. + lang.	0,18	0,14	0,28	0,05	0,24	0,19	0,14	0,24	0,20
Réf.	515	83	40	209	71	172	626	919	231
Tess. + lang.	1 236	165	107	295	298	347	952	873	428
Variation	+140 %	+98 %	+168 %	+41 %	+319 %	+101 %	+52 %	-5 %	+85 %

TABLEAU 4. Nombre de types d'EN identifiées par spaCy_lg dans les corpus small-ELTeC-{fr, en, pt} en fonction de différentes qualités de la ROC déterminées par le CER calculé sur le modèle Tesseract (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 transcriptions de la ROC produites par Tesseract, illustre le fait que les systèmes de REN récupèrent plus de types d'EN différents sur les transcriptions de la ROC, et donc que la qualité du texte en entrée influe sur la quantité des types d'EN récupérés en sortie. La qualité des transcriptions de la ROC a été évaluée en appliquant les métriques *Character Error Rate* (CER) et *Word Error Rate* (WER) sur les textes de réf. et sur les transcriptions. On note que plus le WER est élevé, plus la qualité de la transcription baisse, et plus le nombre de types d'EN en sortie est élevé. On peut en conclure que (i) le système de REN ramène plus de bruit ou faux positifs (FP) en sortie quand la ROC est moins bonne et (ii) le nombre des hapax augmente si la qualité de la transcription diminue. Dans la quantité d'EN surnuméraire détectée sur les transcriptions de la ROC par rapport à la réf., 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 d'un échantillon du corpus des vrais positifs (VP) et des FP par type d'EN récupérées par spaCy_lg sur l'ensemble des EN de réf. et des versions pour Kraken et Tess.fr.

spaCy_lg			
	Réf.	Kraken	Tess.fr
Nb. types	209	455	295
VP	106	71	100
FP	103	283	105
FFP → VP	×	50	17

TABLEAU 5. Annotation manuelle des VP et des FP sur les EN types reconnues par spaCy pour Daudet, small-ELTec-fr

On retrouve bien (i) plus de FP sur Kraken que sur Tess.fr et (ii) pour Tess.fr, presque le même nombre de VP que sur la réf. et 71 VP pour Kraken. Il est aussi vrai que pour Kraken nous répertorions 50 entités qui sont des formes contaminées des entités de la réf. Dans une évaluation automatique, elles sont annotées comme des FP. Néanmoins, lors de l’annotation manuelle nous les avons annotées comme des faux faux positifs (FFP) (nous explicitons ce nommage dans la partie 4.4), autrement dit des VP, par exemple « Batignolleslo » pour « Les Batignolles ». Enfin, on relève qu’il y a plus de types différents d’entités sur la sortie de la ROC que sur la sortie de la réf., car les variantes d’une entité peuvent être nombreuses (tableau 6).

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

TABLEAU 6. REN sur des formes contaminées de l’EN « Ferme des Ormeaux », La Petite Jeanne, Carraud, small-ELTec-fr

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 faux négatif (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 moindres 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 réf., mais le soient sur la transcription de la 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 page transcrites par la ROC mais non prises en compte par les auteurs de la réf., 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.

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 de représenter les intersections entre les ensembles des EN reconnues sur la version de réf. et celles obtenues sur les versions de ROC en s'appuyant sur les diagrammes *UpSetplot*³¹. Les versions de ROC et les textes de réf. ont été annotés automatiquement avec *spaCy_1g*. Nous nous servons de ces derniers comme vérité terrain³². 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 réf. avec celles de leurs versions de ROC. Ainsi, dans le cas du corpus *small-ELTeC-fr*, l'EN « Paris » repérée dans le texte de réf. pour Daudet, n'est pas la même que l'EN « Paris » récupérée dans le texte de réf. 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_1g*, 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. Chacune des sous-figures 1a et 1b comprend deux sous-graphiques. Dans le sous-graphique situé à droite, la première série de cercles représente les EN de la réf. et la seconde comprend les EN de la version de ROC. La troisième colonne représente l'intersection entre la réf. et la ROC, et donc les deux cercles liés avec une ligne représentent les VP. Le cercle noir associé à un outil de ROC marque les EN récupérées uniquement sur la version de ROC, qui sont considérées comme les FP de la sortie de REN, alors que le cercle noir associé à la réf. désigne les EN récupérées uniquement sur la réf., soit les VP. Le sous-graphique situé à gauche permet de calculer le nombre de VP d'une autre manière. Par exemple, pour la figure 1a, en additionnant 11,4 % et 12,3 % de VP, nous en obtenons 23,7 %. En revanche, nous constatons plus de VP et moins de FP pour la configuration *Tess.fr - spaCy_1g* en comparaison avec son pendant de *Kraken*.

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

31. <https://upsetplot.readthedocs.io/en/stable/>

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

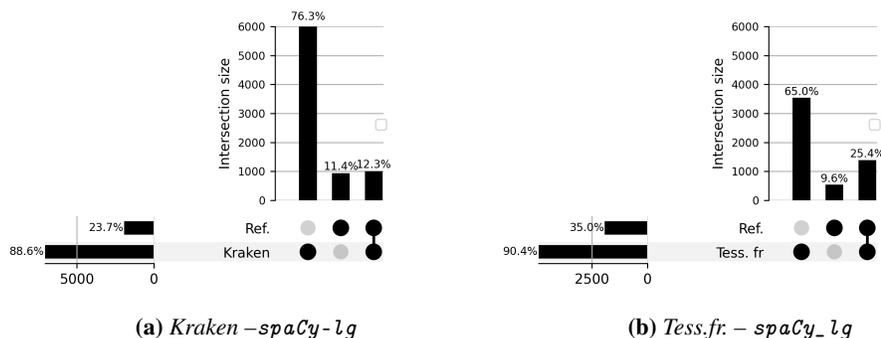


FIGURE 1. *Intersections pour les configurations Kraken-spaCy_lg et Tess.fr. - spaCy_lg, pour le corpus small-ELTec-fr*

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 positifs (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 réf., 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 de 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, les vrais silences et 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 positifs (FP) : EN détectées à tort dans la version de ROC (bruit de la REN).

Faux négatifs (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.

SURÉVALUATION 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 réf. (EN manquantes dans la réf. ou EN contaminées détectées dans la version de ROC).

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

Type	Version	Contexte	spaCy_1g
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	Réf.	[...] a sua entrada para o colegio militar	()
	Kraken	[...] a s_a entrada para o colgio milita	colgio milita
FFP	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

L'analyse des annotations des sorties de REN avec spaCy_1g pour le texte de Daudet, révèle 50 syntagmes pour Kraken et 70 pour Tess.fr qui sont des FVP, c'est-à-dire reconnus comme des EN sur la réf. et sur les transcriptions de la ROC alors qu'ils n'en sont pas. Il s'agit donc de bruit dans la sortie de REN pour les transcriptions de la ROC mais aussi pour la réf. Ces FVP sont par exemple des verbes (« Allons » ou « Parlons »). Nous avons aussi relevé 45 FFN sur la sortie de réf. comparée à Kraken. Ces FFN sont des syntagmes reconnus à tort comme des EN par spaCy_1g sur la réf. mais pas sur les transcriptions de la ROC, par exemple le syntagme « *sanglotaient là-bas* ». Ces évaluations apportent la confirmation que les outils de REN sont susceptibles de faire des erreurs même sur des textes dits « propres ». Ainsi les contaminations de la ROC sur la REN ne sont pas les seules causes des erreurs de 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 *small-ELTeC-fi*³³. 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

33. https://github.com/ljpetkovic/ELTeC_GOLD_REVUE_TAL

Maupassant). L'accord inter-annotateur, calculé à partir du coefficient 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_1g sur cet échantillon, dont les résultats sont présentés dans le tableau 8.

Soutle	GLOBALE : tous types d'EN			LIEUX		
	Rappel	Précision	F-score	Rappel	Précision	F-score
Kraken	49,57	73,72	59,28	48,84	52,50	50,60
Tess.	51,53	77,63	61,94	56,41	57,89	57,14
Réf.	49,78	77,55	60,64	53,49	53,49	53,49
Stricte	Rappel	Précision	F-score	Rappel	Précision	F-score
Kraken	18,26	58,82	27,87	43,24	45,71	44,44
Tess.	21,00	68,66	32,16	43,33	44,83	44,07
Réf.	21,62	69,57	32,98	41,18	45,16	43,08

TABLEAU 8. Évaluation de *spaCy_1g* sur un échantillon annoté de *small-ELTec-fr* de 10 000 tokens dans trois versions textuelles différentes, en configurations souple et stricte

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 réf. 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 réf. Enfin, le fait que le F-score soit meilleur sur les EN de lieux que sur tous les types d'entités 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 »³⁴. 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.

34. Les tirets représentent des retours à la ligne.

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 réf. et de ROC est une tâche ardue, du fait des formes contaminées des EN. Les résultats présentés dans cette partie montrent les limites des évaluations strictes qui s'appuient sur le F-score ou sur les intersections, ce qui nous incite à proposer des stratégies d'évaluation plus souples.

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 JamSpell³⁵ (Jspl) pour la correction automatique des transcriptions de la ROC. Jspl est un outil développé en C++ 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 Jspl pour chacune des trois 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	spaCy_lg			
non corr.	✓	✓	✓	✓
Jspl-pretr.	✓	✓	✓	×
Jspl-ELTeC	✓	✓	✓	✓

TABLEAU 9. Ensemble des configurations que nous évaluons dans cette étude

5.2. Typologie des contaminations de corrections de la ROC

En observant les exemples de correction de la ROC (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 surcorrections, n'ont pas du tout été extraites par spaCy_lg.

35. <https://habr.com/en/articles/346618/>

Version	Contexte	spaCy_lg
Réf.	[...] à l'assemblée de Meunet-sur-Vatan ;	Meunet-sur-Vatan
Kraken	[...] a l'assembl6e' de Neunet-sur- Yatan'	Yatan
Kraken Jsp11-fr	[...] a l'assembl6e' de Neuner-sur- Satan' ;	()
Kraken ELTeC-fr	[...] a l'assembl6e' de Neunet-sur-Avant' ;	()
Tess.fr	[...] à l'assemblée de Meunet-sur- Vatan* ;	Meunet-sur-
Tess.fr Jsp11-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, small-ELTeC-fr

Version	Contexte	spaCy_lg
Réf.	[...] before you went to India.	India
Kraken	[...] before you went to Iudia.	Iudia
Kraken Jsp11-en	[...] before you went to India	India
Kraken ELTeC-en	[...] 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, small-ELTeC-en

Similairement, nous observons dans le tableau 11 que le modèle de la correction automatique de la ROC par Jsp11, 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 dans l'EN de réf. « India ». Par contre, le modèle Jsp11 préentraîné 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 surcorrections « Conspire » (au lieu de « Devonshire » dans ELTeC anglais), ainsi que « Martincourt » (au lieu de « Morlincourt » dans ELTeC français).

Type	Version	Contexte	spaCy_lg	EN réf.
MOBC	Kraken-en	[...] when they were in <i>Lonlon</i>	()	London
	JspI-en	[...] when they were in London	London	
	Kraken-ELTeC-en	[...] when they were in London	London	
MOMC	Kraken-en	[...] flowery lanes peuliar to <i>Dconshire</i> ;	<i>Dconshire</i>	Devonshire
	JspI-en	[...] flowery lanes peculiar to <i>Conspire</i> ;	()	
	Kraken-ELTeC-en	[...] flowers lanes peculiar to <i>Dconshire</i> ;	<i>Dconshire</i>	
MOI	Kraken-fr	cure de <i>Mlorlincourt</i>	<i>Mlorlincourt</i>	Morlincourt
	JspI-fr	cure de <i>Mlorlincourt</i>	<i>Mlorlincourt</i>	
	Kraken-ELTeC-fr	cure de <i>Mlorlincourt</i>	<i>Mlorlincourt</i>	
BOIC	Kraken-fr	en retournant de <i>Morlincourt</i>	Morlincourt	Morlincourt
	JspI-fr	en retournant de <i>Martincourt</i>	Martincourt	
	Kraken-ELTeC-fr	en retournant de <i>Morlincourt</i>	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*. Home influence, Aguillar et Mon village, Adam, *small-ELTeC*-{en, fr}

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 : 0,26, fr : 0,05, pt : 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				Tess.					
	Brut	JspI préentraîné	JspI- ELTeC		Brut	JspI préentraîné	JspI- ELTeC			
<i>small-ELTeC-en</i>	3 121	1 691	-46 %	1 446	-54 %	2 653	2 072	-22 %	1 776	-34 %
CER médian	0,364	0,362		0,364		0,256	0,261		0,265	
<i>small-ELTeC-fr</i>	7 062	5 320	-25 %	6 758	-5 %	4 888	4 637	-5 %	4 921	+0,7 %
CER médian	0,098	0,102		0,104		0,050	0,054		0,055	
<i>small-ELTeC-pt</i>	10 110	N/A	N/A	6 147	-39 %	4 007	N/A	N/A	3 594	-10 %
CER médian	0,157	N/A		0,155		0,093	N/A		0,096	

TABLEAU 14. Comparaison du nombre d'EN types repérées par *spaCy_lg* sur les transcriptions ROC non-corrigées (*Brut*) et corrigées (*JspI préentraîné/ELTeC*) pour les corpus *small-ELTeC*-{en, fr, pt}

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 décidé 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 :

– même si la correction automatique permet de diminuer le vocabulaire et le nombre des hapax (-46 % du vocabulaire pour l'anglais sur Kraken *JspI préentraîné*)

et -54 % pour l’anglais JspI-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 %);

– l’observation des CER médians des versions Tess. corrigées par rapport à la version Tess. brute montre une très légère baisse de qualité, alors que celle des CER médians pour les versions Kraken corrigées par rapport à la version Kraken brute montre une stagnation. En effet, les CER médians ne montrent pas de performance significative des modèles de correction automatique, et même il semble qu’ils soient dégradés. La correction automatique avec JspI-ELTeC sur Tess. pour le français voit le nombre des hapax augmenter, +0,7 % des EN récupérées, ce qui semble indiquer qu’il y a de nouvelles EN (hapax) récupérées, ce qui peut être dû au phénomène de surcorrection (BOIC, tableau 13);

– on remarque que la baisse du nombre d’EN types, et donc des hapax est plus significative sur les textes de version Kraken que Tess., ce qui laisse penser qu’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.

Enfin, concernant le corpus *small-ELTeC-pt*, comme nous l’avons souligné précédemment, nous avons extrait uniquement les EN corrigées avec notre modèle JspI-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’alignement*

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 réf. et celles provenant des versions de ROC (Kraken ou Tesseract) corrigées avec le modèle préentraîné de JspI (2a-2b), et le modèle entraîné avec le corpus ELTeC français (2c-2d). La figure 2b montre que la configuration Tess.fr–JspI *pretrain-fr* – spaCy_1g a permis de récupérer le plus grand nombre d’EN en commun. Il est notable que la correction automatique avec le modèle Kraken préentraîné de JspI ou le modèle entraîné sur une partie du corpus ELTeC adapté à la langue du corpus testé n’apporte pas de gain considérable pour l’intersection. Toutefois, il faut prendre en considération que le nombre de FP est diminué pour le modèle Kraken préentraîné de JspI, contrairement à celui de Kraken sans corrections. En comparant les figures 1 et 2, nous constatons également que la configuration Tess. sans corrections récupère le plus d’EN en commun.

Ce fait peut être lu à l’aune des observations présentées dans le tableau 13 rapportant la typologie des erreurs de correction. 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 réf. Ainsi, les BOIC (« Morlincourt » qui devient « Martincourt »), se cumulant aux EN contaminées non corrigées MOI (p. ex. « Morlincourt » qui reste « Morlincourt »), n’améliorent pas les résultats obtenus par calcul des intersections. Pour les corpus *small-ELTeC-{fr, en, pt}* et *small-TGB-fr*, il

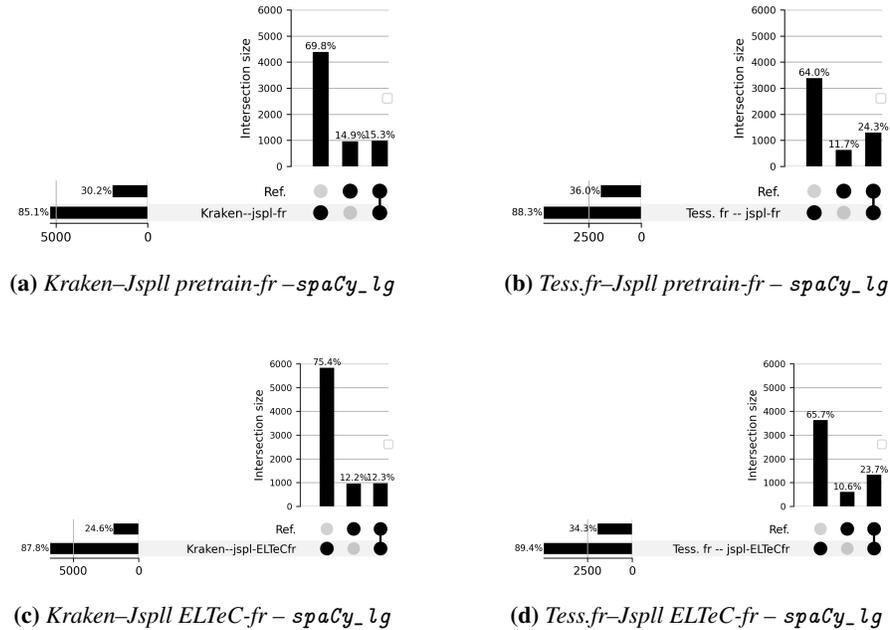


FIGURE 2. Intersections pour les configurations *Kraken* et *Tess.fr* corrigées par *JamSpell* préentraîné et modèle *ELTeCfr*, *spaCy_lg* sur le corpus *small-ELTeC-fr*

semble que 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 préentraînés, 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 sur 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'alignement ?

5.4.1. Mesures de distance textuelle

Afin de dépasser les verrous de l'évaluation stricte tels que décrits précédemment, nous employons des mesures de distance textuelle pour rendre plus souples nos critères d'évaluation des résultats de la REN sur les sorties de ROC bruitées et sur leurs corrections automatiques. Nous avons privilégié la distance cosinus, calculée entre la REN sur tous les textes de réf. et leurs versions ROC, car elle est considérée comme

une mesure de référence quand il est question de (dis)similarité textuelle (Buscaldi et al., 2020). La distance cosinus est calculée sur les bigrammes et les trigrammes de caractères³⁶.

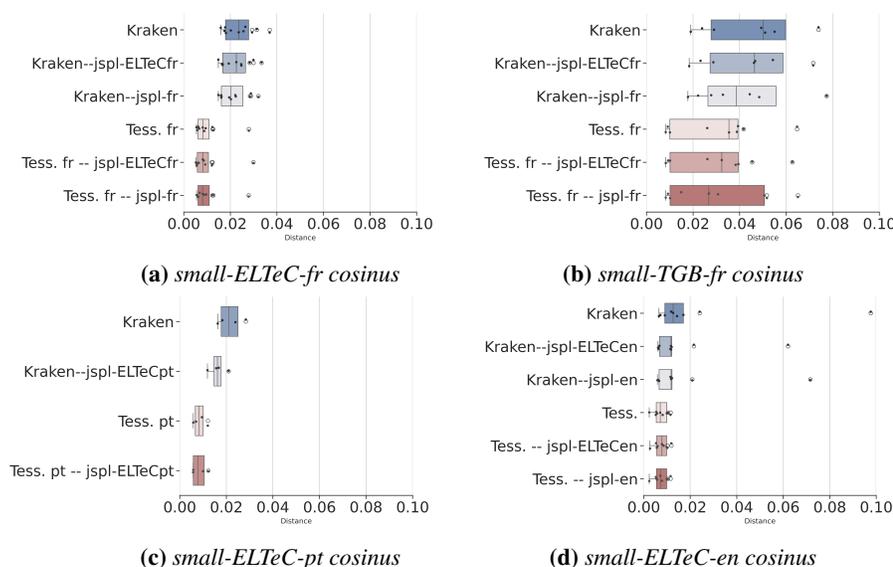


FIGURE 3. Distances cosinus calculées entre les textes de réf. complets et les versions de ROC des mêmes textes

Pour lire les figures 3 et 4, il faut noter que plus la boîte à moustache 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. Nous constatons que les versions de ROC pour Kraken corrigées avec tous les modèles comportent une légère amélioration ; à l'inverse, les versions corrigées de Tess. semblent ne pas connaître d'améliorations, phénomène illustré sur la figure 3b. 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). Cette figure laisse apercevoir que les résultats de la REN sur les versions Kraken sont moins bons que pour celles produites avec Tess. Il apparaît que les corrections automatiques sur les versions Kraken apportent un gain tandis que celles effectuées sur les versions de Tess. ne semblent pas faire évoluer significativement les résultats. Il semblerait que la correction automatique avec le modèle Jspl-ELTeC soit légèrement plus efficace sur les versions de Kraken que sur les versions Tess., car l'écart entre les bords droits et gauches des boîtes est plus petit. Ce constat laisse à penser que le correcteur automatique produit plus de bonnes corrections dans le cas de versions de ROC bruitées (figure 4d, Kraken). À l'inverse, si une version de ROC est peu bruitée, le correcteur automatique aura tendance à moins

³⁶. Nous avons vectorisé le texte avec la librairie `CountVectorizer`.

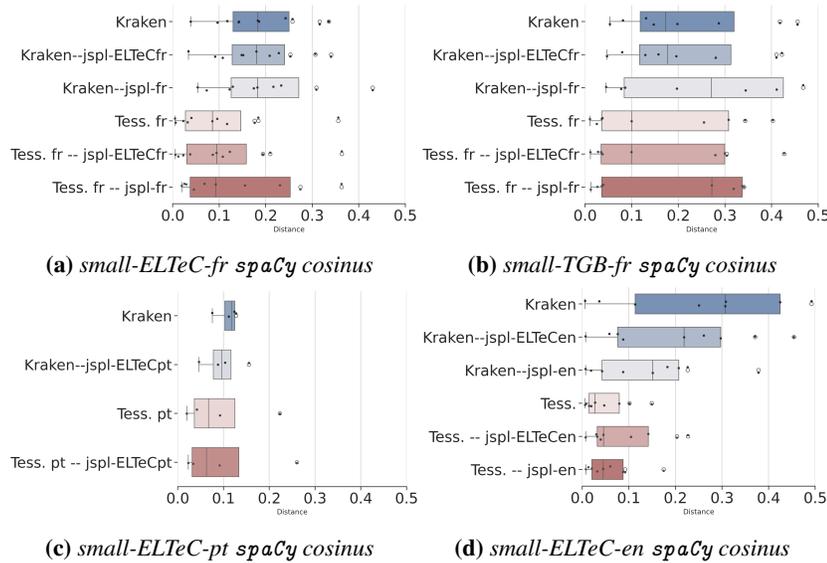


FIGURE 4. Distance cosinus pour *spaCy_lg* sur chaque corpus globalement

bien corriger, voire à surcorriger. On peut observer ce phénomène concernant les résultats de la REN sur Tess. qui sont moins bons sur les versions Tess. corrigées (figure 4d, Tess.), en les mettant en regard avec nos observations sur le tableau 14 – c’est une deuxième manière d’analyser le phénomène de surcorrection. Enfin, en comparant les modèles préentraînés et ceux entraînés par notre équipe, les graphiques 4a et 4b montrent que la correction automatique avec le modèle Jspl-fr provoque beaucoup de variations, des surcorrections, autant sur Kraken que sur Tess. car les boîtes sont plus larges, et le modèle Jspl-ELTeC-fr semble produire moins de variations, contrairement à celui de l’anglais. Finalement, il apparaît que pour le français il est préférable d’entraîner un modèle sur nos données, alors que le modèle préentraîné pour l’anglais donne des résultats relativement meilleurs que notre modèle ELTeC-en.

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 NERVAL³⁷, 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’alignement 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 réf. et des versions de ROC et de ROC corrigées sont annotés au format IOB avec *spaCy_lg*. Les fichiers des textes de réf.

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

ainsi annotés font office de vérité 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 Tess. dans les tableaux 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 évoquons dans la partie 5.3.1.

Version	# Entités		Évaluation par NERVAL			
	ROC	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken	1 122	744	566	0,50	0,76	0,61
Tess.fr	860	744	646	0,75	0,87	0,81
Kraken + Jspil-fr	1 027	744	471	0,46	0,63	0,53 ↓
Tess.fr + Jspil-fr	794	744	532	0,67	0,72	0,69 ↓
Kraken + ELTeC-fr	1 055	744	548	0,52	0,74	0,61 ↑
Tess.fr + ELTeC-fr	838	744	621	0,74	0,84	0,79 ↓

TABLEAU 15. Résultat de NERVAL sur Le Petit Chose, Daudet, small-ELTeC-fr

Version	Label	# Entités		Évaluation par NERVAL			
		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 + Jspil-en	LOC	158	168	105	0,67	0,63	0,64 ↑
Tess. + Jspil-en		152	168	119	0,79	0,71	0,75 ↓
Kraken + Jspil-en	GPE	1 542	1 324	910	0,59	0,69	0,64 ↑
Tess. + Jspil-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, small-ELTeC-en

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 REN spatiales. Nous avons établi (i) une typologie pour l'évaluation plus fine de l'impact des contaminations de la ROC sur les sorties de REN pour pallier le problème de l'évaluation automatique stricte, selon laquelle des cas particuliers d'EN contaminées étaient considérés comme des FP alors qu'il s'agit de VP, et, (ii) une typologie des contaminations de correction automatique de la ROC afin de rendre compte des fluctuations au niveau de la performance du correcteur. 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 que nous avons générées ont été corrigées à l'aide de deux modèles de l'outil JamSpell : l'un fourni 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 des surcorrections, dans les données textuelles et que le gain apporté par les corrections justes n'était pas considérable. Par ailleurs, les résultats du correcteur automatique sont plus significatifs dans le cas des textes plus bruités. Pour preuve, nous observons une réduction plus importante du nombre d'hapax dans les sorties de REN sur l'outil de ROC Kraken, qui est moins performant que Tesseract. Enfin, nous concluons que l'évaluation automatique de l'impact de la ROC sur la REN n'est pas une tâche triviale, et que sa complexité s'étend sur l'évaluation de la REN sur des textes de ROC corrigés ; il apparaît nécessaire de croiser les résultats de multiples méthodes d'évaluation pour affiner notre propos, c'est pourquoi nous employons différentes métriques ainsi que l'outil NERVAL. Dans la suite de notre travail, nous nous appuyons sur l'utilisation d'un autre outil de correction 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.
- Damerou F. J., « A Technique for Computer Detection and Correction of Spelling Errors », Commun. ACM, vol. 7, n^o 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, n° 2, p. 131-143, October, 2015.
- Miret B., Kermorvant C., « Nerval : a python library for named-entity recognition evaluation on noisy texts », 2021. <http://gitlab.com/tekliia/ner/nerval>.
- 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. DOI : 10.5281/zenodo.1212303.
- 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 », Expérimentations et évaluations en fouille de textes, 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., Atrahabi 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, n^o 2, p. 42-57, December, 2022.
- 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.