Traitement automatique des langues

Robustesse et limites des systèmes de TAL

sous la direction de
Caio Corro
Gaël Lejeune
Vlad Niculae

Vol. 64 - n°2 / 2023

ATALA

Robustesse et limites des systèmes de TAL

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
64

n°2
2023

Robustesse et limites des systèmes de TAL

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 64 – n°2 / 2023

ROBUSTESSE ET LIMITES DES SYSTÈMES DE TAL

Table des matières

Introduction au numéro spécial — Robustesse et limites des modèles de traitement automatique des langues

Caio Corro* — **Gaël Lejeune**** — **Vlad Niculae*****

* *INSA Rennes, IRISA, Inria, CNRS, Université de Rennes*
** *STIH/CERES, Sorbonne Université Paris, France*
*** *Language Technology Lab, IVI, FNWI, University of Amsterdam*

RÉSUMÉ. *Les chercheurs en traitement automatique des langues (TAL) sont amenés à traiter des tâches et des données de plus en plus variées. Ce numéro de la revue TAL s'intéresse à la capacité des systèmes de TAL à s'adapter à la variation des données, à leur robustesse. Les articles présentés ici s'intéressent à deux types de données qui questionnent la robustesse : les données générées par des utilisateurs et les données issues de la reconnaissance optique de caractères.*

MOTS-CLÉS : *robustesse, normalisation lexicale, contenus produits par les utilisateurs, correction automatique d'OCR, reconnaissance optique de caractères, reconnaissance d'entités nommées.*

TITLE. *Introduction to the special issue on robustness and limits of NLP systems*

ABSTRACT. *Researchers in natural language processing (NLP) are required to address a variety of tasks and data. This issue of the TAL journal focuses on the ability of NLP systems to adapt to data variability, to their robustness. The articles presented here explore two types of data that challenges robustness: user-generated content and data derived from optical character recognition.*

KEYWORDS: *robustness, lexical normalisation, user generated content, automatic OCR correction, optical character recognition, named entity recognition.*

1. Introduction

Les méthodes modernes d'apprentissage automatique ont permis d'atteindre des résultats spectaculaires en traitement automatique des langues, notamment à partir de la construction de modèles préentraînés comme les réseaux de plongements contextuels, c'est-à-dire BERT (Devlin *et al.*, 2019) et ses dérivés, et les (grands/giga) modèles de langue (Radford *et al.*, 2018). D'un point de vue scientifique, les avancées se sont faites sans changement concret de paradigme depuis une trentaine d'année. En effet, la production scientifique en traitement automatique des langues et en apprentissage automatique peut être caractérisée par son caractère routinier, par exemple :

– le web ou *world wide web* met à disposition un très grand nombre de données qui peuvent être exploitées par les modèles de TAL (Resnik, 1999 ; Grefenstette, 1999 ; Keller *et al.*, 2002 ; Ortiz Suárez *et al.*, 2019). Ces données sont souvent complétées par des données de meilleure qualité mais en quantité moindre, comme des productions journalistiques (Charniak, Eugene *et al.*, 2000) ;

– les réseaux de neurones sont utilisés depuis longtemps comme fondement pour les modèles de langue (Bengio *et al.*, 2000 ; Schwenk et Gauvain, 2002) ;

– l'apprentissage des paramètres d'un réseau de neurones nécessite de développer des méthodes d'optimisation pouvant s'appliquer sur des fonctions non convexes avec un très grand nombre de paramètres, par exemple en tirant bénéfice de la structure géométrique des données, ce qui est le cas de la descente de gradient naturelle (Amari, 1998), mais également de Adam (Kingma et Ba, 2015) qui sert de fondement à la plupart des approches d'optimisation actuelles ;

– l'utilisation de très grands jeux de données pour l'apprentissage de modèles génératifs nécessite de mettre en place des architectures neuronales qui peuvent pleinement profiter de la parallélisation des calculs, c'est par exemple le cas des architectures attentionnelles (Vaswani *et al.*, 2017) qui sont aujourd'hui hégémoniques, mais la masse des données était déjà un enjeu pour les *self-organizing maps* (Seiffert et Michaelis, 2001), pour lesquels la question de la parallélisation matérielle des calculs pour l'entraînement était importante : « *In general there are two main reasons to implement artificial neural networks on parallel hardware. [...] the second motive becomes evident when dealing with demanding real-world applications, when training times are increasing up to and above the pain threshold* » (Seiffert, 2004).

Il est d'ailleurs parfois difficile de savoir à partir de deux titres d'article qui ont presque 20 ans de différence lequel est le plus ancien :

– dans les actes d'HLT/EMNLP 2005 : « *Training Neural Network Language Models On Very Large Corpora* » (Schwenk et Gauvain, 2005) ;

– dans les actes d'un atelier d'EMNLP 2024 : « *Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations* » (Csordás *et al.*, 2024).

Ceci étant, ces dernières années ont quand même été marquées par un tournant sur plusieurs plans : passage à l'échelle des méthodes de TAL (autant en termes de

données utilisées qu'en taille des modèles), amélioration des performances pour de nombreuses tâches cibles, en particulier la tâche de génération de textes qui est aujourd'hui utilisée pour réaliser toutes sortes d'autres tâches, et enfin démocratisation de l'accès à ces modèles (où les interfaces conversationnelles sont devenues incontournables pour les utilisateurs).

La communauté a développé de nombreux jeux de données (ou *benchmarks*) sur lesquels les résultats évoluent très rapidement, donnant l'impression que de nombreux problèmes liés au traitement automatique des langues sont « résolus » ou en passe de l'être. La production scientifique du domaine s'apparente beaucoup à la création d'une collection de timbres[1], où chacun y va de son nouveau modèle, de son nouveau jeu de données, de sa nouvelle tâche, etc.

Pourtant, la question de la capacité des méthodes de TAL à être efficaces, voire simplement utilisables, sur différents types de données et cas d'usage reste ouverte. Par exemple, des applications comme la pharmacovigilance ou encore l'épidémiologie digitale nécessitent d'analyser en continu de larges volumes de données produits sur les réseaux sociaux, souvent écrits dans une langue non standard. Cela requiert d'une part d'avoir des méthodes d'analyse robustes, rapides et de préférence à faible consommation énergétique, mais d'autre part de penser finement les métriques d'évaluation qui peuvent être difficiles à agréger en un seul nombre : quelques faux positifs peuvent avoir un impact négligeable, alors qu'un faux négatif peut être catastrophique ; détecter un nouveau cluster de cas de rhumes semble moins primordial qu'un nouveau cluster de cas de virus Ebola. En traduction automatique, des métriques standards comme le score BLEU (Papineni *et al.*, 2002), mais également les métriques automatisées plus modernes comme COMET (Rei *et al.*, 2020), peuvent passer à côté d'éléments importants, comme une terminologie spécifique mais cruciale (même si elle ne concerne que peu de mots) ou la consistance de traduction des termes au sein d'un document (Semenov *et al.*, 2023). Il nous semble donc essentiel que la communauté s'intéresse à la robustesse concrète des méthodes de TAL au-delà de la collection de modèles et la surenchère de *benchmarks*.

2. Contexte de l'appel et relecture

Une journée d'études organisée par l'ATALA en 2022[2] avait été l'occasion pour la communauté française de présenter un éventail de travaux sur la robustesse et sur les limites des modèles actuels de TAL (Corro et Lejeune, 2022), c'est-à-dire sur la capacité de ces modèles d'offrir des performances comparables sur des données et des cas d'usage variés (Yu *et al.*, 2022). Cette journée d'études était focalisée sur les données dites « non standards », qui avaient été définies de manière large comme des données présentant des variations vis-à-vis d'un état de langue attendu : variation de la langue

1. Pour reprendre les termes d'une citation attribuée de façon incertaine à Ernest Rutherford : « *all science is either physics or stamp collecting* ».

2. `https://www.atala.org/content/robustesse-des-systemes-de-tal`

en diachronie, variations régionales, variation dans l'ordre des mots, *code-switching*, *user generated content*, orthographe inconsistante, données accidentellement bruitées suite à un prétraitement, données incomplètes ou encore présence d'un vocabulaire de domaine spécialisé. Plusieurs problématiques liées à la robustesse avaient été abordées telles que la reproductibilité des résultats, la portabilité des algorithmes, ou encore l'influence de la qualité des données. La variété des tâches abordées (reconnaissance d'entités nommées, traduction automatique, reconnaissance automatique de la parole, *web scraping* ou encore similarité sémantique) a naturellement amené à proposer de faire de cette thématique un numéro spécial de la revue TAL.

À la suite de cette journée, nous avons donc lancé un appel pour un numéro thématique de la revue TAL, visant à questionner la robustesse et les limites des modèles de TAL, en particulier en ce qui concerne les trois points suivants :

– données « non standards » : utilisation de modèles sur des données présentant des variations vis-à-vis d'un certain attendu en termes d'état de langue ;

– données hors domaine : utilisation de modèles sur des données d'un domaine différent par rapport aux données d'entraînement ;

– généralisation à des structures linguistiques non observées à l'entraînement : généralisation compositionelle (Kim et Linzen, 2020), généralisation structurelle (Yao et Koller, 2022) ou encore généralisation du genre (Stanovsky *et al.*, 2019), entres autres.

En terme de thématique, nous avons ouvert l'appel à un large éventail de contributions :

– identification et évaluation des phénomènes linguistiques problématiques pour les modèles neuronaux et les autres systèmes de TAL ;

– analyse et correction de la propagation des erreurs dans les systèmes fondés sur une analyse en cascade ;

– retours d'expérience sur l'utilisation de systèmes de TAL qui se sont révélés non fonctionnels sur des types de données particuliers ;

– critique de jeux de données utilisés pour l'apprentissage ou pour l'évaluation ;

– construction de jeux de données permettant d'évaluer la robustesse aux variations linguistiques ;

– augmentation artificielle de données pour améliorer la robustesse des modèles ;

– adaptation hors domaine ou apprentissage avec des domaines peu représentés dans les données ;

– architectures neuronales et méthodes d'entraînement améliorant la robustesse des modèles.

De plus, toutes les tâches standards du traitement automatique des langues pouvaient être considérées et les travaux portant sur d'autres langues que le français étaient les bienvenus, car nous avions pour objectif d'identifier les cas d'usage intéressants pour la recherche sur la robustesse.

3. Articles acceptés

Ce numéro comporte deux articles en français sélectionnés à l'issue du processus de relecture, portant tous les deux sur le prétraitement des entrées dans une chaîne d'analyse textuelle afin d'en normaliser le contenu : la normalisation lexicale pour le premier article et la correction des erreurs d'OCR (*optical character recognition*) pour le second.

Le premier d'entre eux, « *Étude sur la normalisation lexicale de contenus produits par les utilisateurs* » écrit par Lydia Nishimwe, Benoît Sagot et Rachel Bawden, propose un état de l'art sur la normalisation lexicale des contenus produits par des utilisateurs (ou UGC pour *User Generated Content*). La normalisation est ici définie comme la transformation des formes non standards par leur variantes standards, comme « jvien » en « je viens ». L'article propose de classifier les travaux de la littérature en deux catégories : les méthodes de correction de mot et les méthodes de traduction de phrase. De plus, l'article détaille les métriques et les corpus utilisés pour évaluer la qualité de la normalisation. Il montre que la normalisation n'améliore pas systématiquement les résultats, puisqu'elle introduit elle-même du bruit et que son efficacité dépend de la tâche et de la langue. Les autrices et l'auteur expliquent que les approches fondées sur des modèles comme BERT ont tendance à être moins sensibles aux données non standards mais que la normalisation se justifie sans doute encore sur des contextes avec peu de ressources (langues ou domaines peu dotés).

Le second article, « *Analyse multilingue de l'impact de la correction automatique de l'OCR sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires* » écrit par Caroline Koudoro-Parfait, Ljudmila Petkovic et Glenn Roe s'intéresse quant à lui à la normalisation des données produites par OCR. Il propose une analyse multilingue (français, anglais et portugais) de la robustesse de systèmes de reconnaissance d'entités nommées (REN) utilisés sur des données bruitées obtenues par OCR, en utilisant des outils existants pour les différentes étapes de traitement et de normalisation. Les autrices et l'auteur partent du constat que l'OCR génère des erreurs qui posent des problèmes à des systèmes de REN entraînés sur des données non bruitées. L'objectif de l'article est double : (1) évaluer si la correction automatique des transcriptions (ou post-correction OCR), censée améliorer la qualité de l'entrée, améliore la qualité de la REN et (2) proposer une critique des métriques d'évaluation strictes qui pourraient contribuer à sous-évaluer la robustesse des systèmes de REN. L'article montre que les outils ont une forte tendance à la surcorrection, avec des modifications erronées qui affectent la qualité (et le nombre) d'entités correctes extraites. Une typologie de l'impact de la correction sur la REN est également proposée.

4. Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs et relectrices, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro.

5. Bibliographie

Amari S.-i., « Natural Gradient Works Efficiently in Learning », *Neural Computation*, vol. 10, nº 2, p. 251-276, 1998.

Bengio Y., Ducharme R., Vincent P., « A Neural Probabilistic Language Model », *in* T. Leen, T. Dietterich, V. Tresp (eds), *Advances in Neural Information Processing Systems*, vol. 13, MIT Press, 2000.

Charniak, Eugene, Blaheta, Don, Ge, Niyu, Hall, Keith, Hale, John, Johnson, Mark, « BLLIP 1987-89 WSJ Corpus Release 1 », 2000.

Corro C., Lejeune G. (eds), *Actes de la journée d'étude sur la robustesse des systemes de TAL*, 2022.

Csordás R., Potts C., Manning C. D., Geiger A., « Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations », *in* Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, H. Chen (eds), *Proceedings of the 7th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Miami, Florida, US, p. 248-262, November, 2024.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *in* J. Burstein, C. Doran, T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Grefenstette G., « The World Wide Web as a Resource for Example-Based Machine Translation Tasks », *Proceedings of Translating and the Computer 21*, Aslib, London, UK, November 10-11, 1999.

Keller F., Lapata M., Ourioupina O., « Using the Web to Overcome Data Sparseness », *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, p. 230-237, July, 2002.

Kim N., Linzen T., « COGS : A Compositional Generalization Challenge Based on Semantic Interpretation », *in* B. Webber, T. Cohn, Y. He, Y. Liu (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 9087-9105, November, 2020.

Kingma D. P., Ba J., « Adam : A Method for Stochastic Optimization », *Proceedings of the International Conference on Learning Representations*, 2015.

Ortiz Suárez P. J., Sagot B., Romary L., « Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures », Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, p. 9 - 16, 2019.

Papineni K., Roukos S., Ward T., Zhu W.-J., « Bleu : a Method for Automatic Evaluation of Machine Translation », *in* P. Isabelle, E. Charniak, D. Lin (eds), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311-318, July, 2002.

Radford A., Narasimhan K., Salimans T., Sutskever I. *et al.*, « Improving language understanding by generative pre-training », 2018.

Rei R., Stewart C., Farinha A. C., Lavie A., « COMET : A Neural Framework for MT Evaluation », *in* B. Webber, T. Cohn, Y. He, Y. Liu (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 2685-2702, November, 2020.

Resnik P., « Mining the Web for Bilingual Text », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, College Park, Maryland, USA, p. 527-534, June, 1999.

Schwenk H., Gauvain J.-L., « Connectionist language modeling for large vocabulary continuous speech recognition », *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. I-765-I-768, 2002.

Schwenk H., Gauvain J.-L., « Training Neural Network Language Models on Very Large Corpora », *in* R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (eds), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, British Columbia, Canada, p. 201-208, October, 2005.

Seiffert U., « Artificial neural networks on massively parallel computer hardware », *Neurocomputing*, vol. 57, p. 135-150, 2004. New Aspects in Neurocomputing : 10th European Symposium on Artificial Neural Networks 2002.

Seiffert U., Michaelis B., « Multi-Dimensional Self-Organizing Maps on Massively Parallel Hardware », *Advances in Self-Organising Maps*, Springer London, London, p. 160-166, 2001.

Semenov K., Zouhar V., Kocmi T., Zhang D., Zhou W., Jiang Y. E., « Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies », *in* P. Koehn, B. Haddow, T. Kocmi, C. Monz (eds), *Proceedings of the Eighth Conference on Machine Translation*, Association for Computational Linguistics, Singapore, p. 663-671, December, 2023.

Stanovsky G., Smith N. A., Zettlemoyer L., « Evaluating Gender Bias in Machine Translation », *in* A. Korhonen, D. Traum, L. Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 1679-1684, July, 2019.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

Yao Y., Koller A., « Structural generalization is hard for sequence-to-sequence models », *in* Y. Goldberg, Z. Kozareva, Y. Zhang (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, p. 5048-5062, December, 2022.

Yu Y., Khan A. R., Xu J., « Measuring robustness for NLP », *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3908-3916, 2022.

Étude sur la normalisation lexicale de contenus produits par les utilisateurs

Lydia Nishimwe — Benoît Sagot — Rachel Bawden

Inria, Paris, France

RÉSUMÉ. *L'essor du traitement automatique des langues (TAL) se vit dans un monde où l'on produit de plus en plus de contenus en ligne. En particulier sur les réseaux sociaux, les textes publiés par les internautes sont remplis de phénomènes « non standard » tels que les fautes d'orthographe, l'argot, les marques d'expressivité, etc. Ainsi, les modèles de TAL, en grande partie entraînés sur des données « standard », voient leur performance diminuer lorsqu'ils sont appliqués aux contenus produits par les utilisateurs* (User-Generated Content, UGC). *L'une des approches pour atténuer cette dégradation est la normalisation lexicale : les mots non standard sont remplacés par leurs formes standard. Dans cet article, nous réalisons un état de l'art de la normalisation lexicale des UGC. Nous discutons de ses avantages, limites et perspectives de travaux de recherche, ainsi que de sa pertinence dans l'avenir du TAL : les modèles actuels étant déjà très robustes aux UGC, la normalisation lexicale reste utile dans des contextes de ressources limitées, ou pour des études sociolinguistiques.*

MOTS-CLÉS : *normalisation lexicale, contenus produits par les utilisateurs, réseaux sociaux.*

TITLE. *A Study on the Lexical Normalisation of User-Generated Content*

ABSTRACT. *The boom of natural language processing (NLP) is taking place in a world where more and more content is produced online. On social networks especially, the textual content published by users is full of "non-standard" phenomena such as spelling mistakes, jargon, marks of expressiveness, etc. Therefore, NLP models, which are largely trained on "standard" data, suffer a decline in performance when applied to user-generated content (UGC). One approach to mitigate this degradation is through lexical normalisation, where non-standard words are replaced by their standard forms. In this paper, we review the state of the art of lexical normalisation of UGC. We discuss its advantages, limitations and research perspectives, and its relevance in the future of NLP: while current models are already very robust to UGC, lexical normalisation remains useful in resource-limited contexts or for sociolinguistic studies.*

KEYWORDS: *lexical normalisation, user-generated content (UGC), social media.*

1. Introduction

Pour développer des systèmes de traitement automatique des langues (TAL) capables de traiter les « contenus produits par les utilisateurs » (*User-Generated Content, UGC*)[1], il est nécessaire de se pencher soit sur les moyens de rendre les modèles robustes aux variations linguistiques associés aux UGC, soit sur la normalisation de ces contenus afin qu'ils ressemblent le plus possible à la langue standard sur laquelle ces modèles sont généralement entraînés. Dans cet article, nous étudions la seconde de ces deux approches. Nous nous consacrons ainsi à la tâche de normalisation lexicale des UGC, qui consiste à remplacer les formes non standard par leurs variantes standard (« normalisées »). À titre d'exemple, le tableau 1 illustre quelques phrases non standard issues de corpus d'UGC (voir section 3.3.1) et leurs normalisations.

Corpus	Phrase non standard	Phrase normalisée
MTNT (Michel et Neubig, 2018)	*C crtn ke* si l'*fccté is prmrdial pr* toi, les *sltns snt rars*.	*C'est certain que* si l'*efficacité est primordiale pour* toi, les *solutions sont rares*.
PFSMB (Rosales Núñez *et al.*, 2021a)	Trop content *jvien* de battre mon record sur Flappy Bird ! *Jai fai* 19 ! *Mdr*	Trop content *je viens* de battre mon record sur Flappy Bird ! *J'ai fait* 19 ! *Mort de rire*
MultiLexNorm (van der Goot *et al.*, 2021)	*nvr* met a girl like her *bfor*	*never* met a girl like her *before* (Traduction : *Je n'ai jamais rencontré de fille comme elle.*)
RoCS-MT (Bawden et Sagot, 2023)	*someone pls lmk*	*Someone please let me know.* (Traduction : *Quelqu'un pourrait-il me conseiller ?*)

TABLEAU 1. *Exemples de phrases non standard issues de corpus d'UGC en français (partie supérieure) et en anglais (partie inférieure du tableau)*

Nous commençons par un état de l'art du domaine : nous décrivons d'abord les spécificités des UGC (section 2.1) et les problèmes qu'ils posent pour les systèmes de TAL (section 2.2). Nous détaillons ensuite les méthodes proposées dans la littérature pour la normalisation des UGC (section 3.1), mais également pour des tâches connexes telles que la correction orthographique, la normalisation phonétique, la correction de transcriptions automatiques, et la normalisation des variantes dialectales (section 3.2). Nous poursuivons avec un bref panorama des jeux de test et des métriques pour la

1. D'autres appellations rencontrées dans la littérature sont : *langage texto* (Choudhury *et al.*, 2007), *textes bruités* (Formiga et Fonollosa, 2012), *communication médiée par les réseaux* (Chanier *et al.*, 2014), *textes bruités générés par les utilisateurs* (Baldwin *et al.*, 2015) et, par calque de l'anglais, *contenus générés par les utilisateurs* (Nishimwe, 2023).

tâche en question (section 3.3). Enfin, nous concluons par une discussion des limites (section 4.1) et des perspectives (section 4.2) de la normalisation.

2. Le TAL et les UGC : une relation amour-haine

2.1. *Les UGC sur les réseaux sociaux*

Sproat *et al.* (2001) ont utilisé le terme de « mots non standard » pour décrire des mots et symboles (chiffres, abréviations, dates, devises monétaires, acronymes) qui ne se trouvent pas dans un dictionnaire, ou dont la prononciation ne peut se déduire des règles usuelles[2]. Avec l'expansion des messages textuels envoyés par téléphone (*Short Message Service, SMS*) au tournant du XXIe siècle, d'autres phénomènes non standard sont apparus dans les textes écrits : la simplification de l'orthographe (p. ex. la suppression d'accents[3]), de la grammaire (p. ex. l'omission de pronoms) et de la syntaxe (p. ex. l'omission de signes de ponctuation), la substitution phonétique (p. ex. *a 2m1* pour *à demain*), l'utilisation d'émoticônes, etc.

Après les SMS, les textes non standard ont connu un essor sur les réseaux sociaux, les forums de discussion, les tchats et d'autres plateformes où les internautes interagissent. Cela a marqué l'émergence des UGC, qui ont été largement qualifiés de « bruités »[4] dans le domaine du TAL. Pour quantifier cette affirmation, Baldwin *et al.* (2013) ont mené une étude linguistique et statistique sur un corpus d'UGC provenant de sources différentes et ont démontré qu'il était effectivement moins standard qu'un corpus composé de textes édités. Par ailleurs, Eisenstein (2013) a expliqué des raisons fréquentes pour lesquelles les utilisateurs écrivent « si mal », à savoir : l'illettrisme, le nombre de caractères limité (p. ex. sur Twitter), le système de saisie du texte (clavier externe p. opp. clavier tactile avec autocomplétion), des phénomènes pragmatiques, et certaines variables sociales.

Certains mots non standard présents dans les UGC sont propres aux réseaux sociaux utilisés, comme les hashtags (*#JeuxOlympiques*), les mentions (*@gouvernementFR*) et leur métalangage (*RT* pour *Retweet*). De plus, le langage des UGC évolue constamment : il y a des néologismes qui sont créés en permanence (*burka* + *bikini* → *burkini*) ; et la façon dont les contenus diffèrent de la norme évolue avec le temps (p. ex. , le français SMS des années 2000 diffère de celui des années 2020). D'autres phénomènes souvent observés sont l'emploi de mots empruntés d'autres langues ou

2. D'autres termes similaires employés dans la littérature sont : *mots bruités* (Contractor *et al.*, 2010), *mots mal formés* (Han et Baldwin, 2011), *tokens non standard* (Liu *et al.*, 2012).

3. En réalité, la suppression des diacritiques date du tout début de l'informatique avec le code ASCII. Cependant, les claviers actuels permettent l'usage aisé des caractères accentués, faisant de leur omission un choix de simplification.

4. Nous éviterons d'utiliser ce terme car il est ambigu et peut être confondu avec d'autres notions de bruit de corpus (p. ex. dans la phase de collecte de données). De plus, il sous-entend un jugement négatif sur la façon d'écrire des internautes.

même le mélange de plusieurs langues (l'alternance codique), ou encore l'utilisation du *leet speak*[5] pour censurer des jurons ou des propos offensants (*!d10t* pour *idiot*).

Dresser une liste exhaustive de tous les phénomènes non standard spécifiques aux UGC n'est pas une tâche aisée, cependant quelques tentatives ont été faites. Par exemple, Seddah *et al.* (2012) ont proposé une classification des phénomènes UGC rencontrés dans des forums de discussion et réseaux sociaux français. Ils les ont définis selon trois axes : (1) les phénomènes ergographiques qui visent à simplifier l'écriture comme l'omission d'accents, la phonétisation, et certaines fautes d'orthographe (*son* pour *sont*) ; (2) les phénomènes transversaux comme la contraction (*nimp* pour *n'importe quoi*) et la segmentation typographique (*c a dire* pour *c'est-à-dire*, *N.U.L.* pour *nul*) ; (3) les marques d'expressivité comme l'étirement des graphèmes ou de ponctuation (*superrr!!!*) et les émoticônes. Sanguinetti *et al.* (2020) se sont appuyés sur cette classification et y ont rajouté les phénomènes d'autocensure, ainsi qu'un quatrième axe des phénomènes d'influence de langues étrangères comme la translitération, la formation de nouveaux verbes et l'autocorrection. Baldwin et Li (2015) ont élaboré une taxonomie basée sur les substitutions, insertions et suppressions des mots et des signes de ponctuation. Par ailleurs, van der Goot *et al.* (2018) ont élaboré une taxonomie des spécificités UGC en anglais. Ils ont considéré trois types d'« anomalies » : (1) les anomalies non intentionnelles comme les fautes typographiques, orthographiques ou de segmentation ; (2) les anomalies intentionnelles telles que les abréviations d'expressions (*mdr* pour *mort de rire*), les répétitions, les contractions, les transformations phonétiques et l'argot ; (3) les anomalies de catégorie inconnue.

2.2. *L'impact des UGC sur le TAL*

Les modèles de TAL étant traditionnellement entraînés sur des données standard, ils s'attendent à traiter des données du même type pendant l'inférence. En présence de phénomènes UGC, la performance de plusieurs tâches de TAL a longtemps été négativement affectée, à savoir : l'analyse syntaxique (Foster, 2010 ; Seddah *et al.*, 2012), la synthèse vocale (Pennell et Liu, 2010), l'étiquetage morphosyntaxique (Ritter *et al.*, 2011), la détection de thèmes (Muñoz-García *et al.*, 2012), la tokénisation (Aminian *et al.*, 2012), la reconnaissance d'entités nommées (Moon *et al.*, 2018), l'analyse des dépendances (Zhang *et al.*, 2013 ; van der Goot, 2019a), la traduction automatique (Belinkov et Bisk, 2017 ; Michel et Neubig, 2018 ; Rosales Núñez *et al.*, 2021a ; Bawden et Sagot, 2023 ; Popović *et al.*, 2024), l'analyse de sentiments (van Hee *et al.*, 2017 ; Kumar *et al.*, 2020), etc.

Pour pallier la dégradation de performance des modèles de TAL causée par la présence de phénomènes UGC, Eisenstein (2013) a recensé deux approches principales : (1) la normalisation, qui vise à adapter les données à ce que les modèles attendent, et (2) l'adaptation de domaine, qui consiste à adapter les modèles

5. `https://fr.wikipedia.org/wiki/Leet_speak`

aux données, par exemple en entraînant sur des données UGC réelles (Nguyen *et al.*, 2020) ou synthétiques (Karpukhin *et al.*, 2019). Une autre approche consiste à utiliser une architecture de modèle (de TAL ou de normalisation) qui encourage des représentations plus robustes, p. ex. en passant à l'échelle des caractères (Riabi *et al.*, 2021 ; Rosales Núñez *et al.*, 2021b) ou des segments de phrases (Rosales Núñez *et al.*, 2019a), ou à une architecture variationnelle (Rosales Núñez *et al.*, 2023).

Le choix d'approche dépend fortement de contraintes de quantité des données d'entraînement (annotées) et de ressources matérielles disponibles. D'une part, la normalisation permet d'utiliser directement les modèles de TAL sans avoir à les entraîner de nouveau ou à les affiner sur les UGC. Elle est plus économique et plus flexible. Par exemple, à défaut d'un modèle de traduction robuste, il est plus simple d'entraîner un modèle de normalisation, tâche moins complexe que la traduction (Wang et Ng, 2013). D'autre part, l'adaptation de domaine est plus coûteuse mais obtient les meilleures performances : les (très) grands modèles de TAL actuels sont très performants et, en particulier, plus robustes aux UGC (Bawden et Sagot, 2023 ; Peters et Martins, 2024) car ils sont plus complexes et entraînés sur beaucoup plus de données, issues en partie voire en totalité d'Internet, et donc ont été exposés à plus de phénomènes UGC.

L'approche privilégiée a longtemps été la normalisation lexicale, qui consiste à remplacer les mots non standard par leurs formes standard. Cette définition de la tâche est globalement acceptée dans la littérature (Sproat *et al.*, 2001 ; Han et Baldwin, 2011 ; Ling *et al.*, 2013). En revanche, la définition de ce qui est « standard » ou non dépend du domaine d'application (Costa Bertaglia et Volpe Nunes, 2016). De même, la portée de la tâche peut varier selon les cas d'usage et, plus on l'élargit, plus la tâche devient complexe. Ainsi, elle se limite généralement à faire des remplacements 1-à-1 (*ke* $\rightarrow$ *que*), 1-à-n (*jvien* $\rightarrow$ *je viens*), n-à-1 (*N.U.L.* $\rightarrow$ *nul*) et, plus rarement, n-à-m (*c t* $\rightarrow$ *c'était*) (Chanier *et al.*, 2014). Dans certains cas, elle peut inclure d'autres transformations afin d'obtenir une phrase entièrement grammaticale (Zhang *et al.*, 2013 ; Bawden et Sagot, 2023). Par conséquent, les guides d'annotations des corpus de normalisation d'UGC dépendent aussi de la tâche considérée (voir section 4.1.2).

Appliquée en amont sur les données UGC, la normalisation a permis d'améliorer la performance de modèles dans plusieurs tâches de TAL telles que la traduction automatique (Hassan et Menezes, 2013), la reconnaissance d'entités nommées simples (Nguyen *et al.*, 2016) ou imbriquées (Plank *et al.*, 2020), l'étiquetage morphosyntaxique (van der Goot *et al.*, 2017 ; van der Goot et Çetinoğlu, 2021), l'analyse de dépendances (van der Goot *et al.*, 2020), ou encore la compréhension d'UGC par des locuteurs non natifs (Ehara, 2021). Cependant, la normalisation n'est pas une solution toujours bénéfique. Par exemple, van der Goot *et al.* (2017) ont argumenté que, certes, la normalisation améliorait la performance sur l'étiquetage morphosyntaxique, mais pas plus qu'une bonne méthode d'initialisation de plongements de mots. Vielsted *et al.* (2022) ont montré qu'elle n'augmentait ni la robustesse ni la performance de leur modèle de classification d'actes de dialogue. Bien

qu'elle ait ses limites (section 4.1), la normalisation a fait l'objet de nombreux travaux dans la littérature (section 3) et a suivi une évolution similaire à celle de plusieurs tâches de TAL, jusqu'au point de voir sa pertinence remise en cause dans l'écosystème actuel du TAL (section 4.2).

3. La normalisation lexicale : un chevalier blanc ?

3.1. *Méthodes*

3.1.1. *Deux approches principales*

Les approches de normalisation lexicale sont catégorisées selon deux perspectives : (1) la correction de mots et (2) la traduction de la phrase[6]. D'une part, la correction consiste à remplacer les mots erronés (ici, non standard) dans la phrase par leur version correcte (standard). Elle peut se baser directement sur leur forme explicite (orthographe) ou sur une représentation implicite (p. ex. phonétique). D'autre part, la traduction consiste à récrire une phrase dans une autre langue. Dans le cas de la normalisation, il s'agit d'une variante non standard de la même langue.

3.1.2. *Un point de départ commun : le modèle du canal bruité*

De même que pour les tâches de correction orthographique et de traduction automatique statistique, l'approche qui a traditionnellement été utilisée pour la normalisation lexicale est celle du modèle du canal bruité (Shannon, 1948). Soient $\mathcal{V}$ un vocabulaire de mots standard, $S = s_1 \dots s_N$ une phrase standard de longueur N, et $T = t_1 \dots t_M$ une phrase non standard de longueur M résultant de la « corruption » de S par le modèle du canal bruité. La normalisation consiste alors à trouver la phrase $\hat{S}$ qui maximise la probabilité $P(S|T)$ d'obtenir une phrase standard $S \in \mathcal{V}^N$ à partir de la phrase non standard T. En appliquant le théorème de Bayes, nous obtenons :

$$\hat{S} = \text{argmax}_{S \in \mathcal{V}^N} P(S|T) = \text{argmax}_{S \in \mathcal{V}^N} P(T|S)P(S) \quad [1]$$

où $P(S)$ est souvent qualifié de modèle de langue et $P(T|S)$ de modèle d'erreur. Pour la traduction, le modèle d'erreur est souvent appelé modèle de traduction.

En pratique, dans l'approche de normalisation par traduction, cette formule peut être décomposée pour normaliser des sous-groupes de mots, c.-à-d. des segments, et non toute la phrase (Aw *et al.*, 2006). Dans l'approche par correction, la probabilité $P(T|S)$ peut être factorisée en $\prod_i P(t_i|s_i)$ pour traiter un mot à la fois. Cette

6. Kobus *et al.* (2008) ont ajouté une troisième perspective ou « métaphore », la transcription de la parole, considérant que l'orthographe des SMS se rapproche plus « d'approximations alphabétiques et syllabiques de formes phonétiques ». En effet, il est de même pour beaucoup de phénomènes non standard observés dans les UGC (voir section 2.1). En pratique cependant, les méthodes de normalisation fusionnent cette métaphore avec l'une des deux autres par le biais d'un module de correction phonétique ou d'un passage intermédiaire de l'échelle de graphèmes à l'échelle de phonèmes.

factorisation implique un alignement 1-à-1 entre les deux phrases, une hypothèse sous-optimale car elle ne tient pas compte des insertions et des suppressions. Cette limite est résolue en ajoutant le mot vide ϵ au vocabulaire $\mathcal{V}$ (Choudhury *et al.*, 2007).

Avec un ensemble bien choisi de normalisations candidates $\mathcal{C}^N \subset \mathcal{V}^N$, l'équation 1 devient :

$$\hat{S} = \text{argmax}_{S \in \mathcal{C}^N} P(T|S)P(S) \quad [2]$$

Cette hypothèse a donné lieu à plusieurs méthodes en deux étapes : (1) la génération de candidats de correction ou de traduction, et (2) la sélection du meilleur candidat.

Cependant, avec la montée en popularité de l'apprentissage profond, l'approche du modèle du canal bruité a été remplacée par des méthodes neuronales plus directes et plus performantes. Ce sont des modèles qui assurent implicitement la génération et la sélection de candidats (parmi les mots de leurs vocabulaires). Il s'agit surtout de modèles de langue par masquage pour la correction, et de modèles de type encodeur-décodeur pour la traduction.

3.1.3. *La normalisation vue comme une tâche de correction de mots*

L'une des approches classiques de génération de candidats est de concevoir un système qui combine plusieurs modules de correction, soit pour normaliser différents types de mots non standard (Sproat *et al.*, 2001 ; Zhang *et al.*, 2013 ; Stewart *et al.*, 2018), soit pour aborder la normalisation sous plusieurs angles (orthographique, phonétique, sémantique), par exemple avec des modules de correction à partir de règles graphémiques, phonémiques ou morphologiques (Han et Baldwin, 2011 ; Cerón-Guzmán et León-Guzmán, 2016 ; Jiang *et al.*, 2022), de distance d'édition et de similarité phonétique par rapport aux mots d'un lexique (Ruiz *et al.*, 2014 ; Cotelo *et al.*, 2015 ; Ahmed, 2015 ; Supranovich et Patsepnia, 2015), de correcteurs orthographiques (Liu *et al.*, 2012) ou de plongements de mots (van der Goot et van Noord, 2017 ; Stewart *et al.*, 2018). En particulier, les méthodes statistiques se sont montrées efficaces pour la correction d'erreurs : Choudhury *et al.* (2007) ont développé un modèle bigramme à base d'un modèle de Markov caché pour corriger les erreurs dans le langage texto. Ensuite, Xu *et al.* (2015) se sont appuyés sur cette approche et l'ont adaptée au chinois en proposant un modèle à base de champs aléatoires conditionnels pour segmenter les mots non standard en syllabes.

Une autre approche consiste à détecter d'abord les mots non standard, et ensuite à générer leurs candidats de normalisation. Cela a pour avantage de ne pas traiter tous les mots de la phrase source, mais uniquement ceux qui ont besoin de correction. Une stratégie simple pour détecter les mots non standard est de repérer les mots hors vocabulaire ou d'utiliser un correcteur orthographique (Melero *et al.*, 2016). Cependant, elle ne suffit pas car les orthographes non standard peuvent être des mots standard (p. ex. les homonymes), et il faut prendre en compte le contexte pour lever l'ambiguïté (Aw *et al.*, 2006). Ainsi, des méthodes plus complexes de détection ont été explorées, notamment : (1) des modèles statistiques, dont un classifieur de type machine à vecteur de support (Han et Baldwin, 2011) et un modèle à base de champs aléatoires conditionnels (Supranovich et Patsepnia, 2015) pour détecter si

un mot est « mal formé » dans son contexte, et (2) des modèles neuronaux, dont un modèle de réseaux de neurones à propagation avant (Leeman-Munk *et al.*, 2015), un modèle de réseaux de neurones convolutifs (Tian *et al.*, 2017) et le modèle de langue préentraîné BERT (Devlin *et al.*, 2019) affiné pour la classification de mots (Scherrer et Ljubešić, 2021 ; Nishimwe, 2023).

Une fois l'ensemble de candidats généré (sous forme de liste ou de treillis), le meilleur candidat est sélectionné. Cette étape est le plus souvent assurée par le biais de modèles de langue n-grammes (Sproat *et al.*, 2001 ; Ahmed, 2015 ; Ruiz *et al.*, 2014), ou d'une combinaison de ceux-ci. Par exemple, Melero *et al.* (2016) ont proposé un module de sélection constitué d'une interpolation linéaire de quatre modèles de langue encodant des informations linguistiques différentes ; Han et Baldwin (2011) ont combiné un modèle de langue et un modèle de dépendances à base de caractéristiques lexicales et morphophonémiques. Une autre méthode consiste à remplacer les mots non standard par leurs correspondances dans une base de données selon des règles définies (Clark et Araki, 2011) ou des règles apprises sur un corpus de fautes lexicales (Baranes et Sagot, 2014 ; Stewart *et al.*, 2019). D'autres approches utilisent des algorithmes de recherche pour la sélection du meilleur candidat, à savoir : un algorithme de recherche en faisceau intégrant les différentes normalisations (Wang et Ng, 2013), un graphe de normalisation où les nœuds correspondent aux candidats produits par des modules de génération de remplacements (Zhang *et al.*, 2013), et un algorithme de Viterbi (1967) à bigrammes (Beckley, 2015). Par ailleurs, des approches statistiques ont aussi été proposées : le modèle MoNoise (van der Goot et van Noord, 2017) utilise une forêt aléatoire pour sélectionner la meilleure normalisation parmi les candidats générés par ses différents modules. Sa version ultérieure a obtenu à l'époque la meilleure performance sur plusieurs langues (van der Goot, 2019b) et a été utilisée comme modèle de référence dans MultiLexNorm (van der Goot *et al.*, 2021), une campagne d'évaluation de normalisation lexicale multilingue.

Par la suite, des modèles neuronaux ont été proposés. Par exemple, Sproat et Jaitly (2016) et Stewart *et al.* (2019) ont utilisé des modèles de réseaux de neurones récurrents bidirectionnels qui prédisent pour chaque mot de la source, soit sa forme corrigée, soit un token spécial pour signifier qu'il doit rester inchangé. Cependant, grâce au succès des modèles de langue préentraînés, l'approche neuronale privilégiée consiste à masquer les mots à normaliser et à les prédire par des modèles de langue par masquage. Notamment, Muller *et al.* (2019) ont apporté des modifications à l'architecture de BERT et ont affiné ce dernier pour la normalisation en tant que tâche de prédiction de tokens. Kubal et Nagvenkar (2021) ont quant à eux affiné un modèle BERT multilingue (Devlin *et al.*, 2019) pour la normalisation comme une tâche d'étiquetage de séquences et l'ont combiné avec une technique d'alignement de mots. Ainsi, ils ont pu utiliser le même modèle pour effectuer la normalisation sur plusieurs langues. Nishimwe (2023) a utilisé une combinaison linéaire d'un modèle de langue par masquage préentraîné et de la distance de Damerau-Levenshtein (Damerau, 1964) pour choisir les candidats de normalisation.

En général, les approches récentes sont plus performantes (pour les langues bien dotées) car elles sont basées sur de grands modèles de langue qui leur permettent de prendre en compte le contexte environnant pour sélectionner le meilleur candidat. Cependant, une liste de candidats prédéfinie reste contraignante. Pour surmonter cela, une solution est de générer la correction plutôt que de la sélectionner. Par exemple, Samuel et Straka (2021) ont utilisé un modèle de langue génératif à base d'octets, ByT5 (Xue *et al.*, 2022), dont ils ont poursuivi l'entraînement sur des données UGC artificielles et naturelles. Scherrer et Ljubešić (2021) ont combiné un modèle BERT pour la détection avec un modèle de traduction statistique à base de caractères pour générer la correction. Ces deux méthodes ont respectivement eu la première et la deuxième place lors de la campagne d'évaluation MultiLexNorm, et étaient les seules à surpasser le modèle de référence MoNoise (van der Goot, 2019b). Par ailleurs, pour les langues moins dotées où il manque assez de données pour entraîner ces modèles de langue neuronaux, les approches statistiques comme MoNoise sont à privilégier, voire des approches heuristiques à base de lexiques pour les langues très peu dotées.

3.1.4. *La normalisation vue comme une tâche de traduction de la phrase*

Alors que l'approche par correction normalise la phrase par des changements locaux sur les mots, l'approche par traduction effectue un traitement global de toute la phrase. Plusieurs méthodes de traduction automatique classique ont été explorées pour la normalisation de SMS et d'UGC sur les réseaux sociaux, notamment : des modèles de traduction statistique à base de segments (Aw *et al.*, 2006), à base de caractères (Pennell et Liu, 2011 ; Formiga et Fonollosa, 2012) ou à base d'une combinaison des deux (Ling *et al.*, 2013). Des modèles hybrides ont aussi été proposés. Par exemple, Kobus *et al.* (2008) ont combiné un modèle à base de segments avec un module de transduction phonémique pour proposer des hypothèses pour les mots hors vocabulaire, et un modèle de langue pour sélectionner le meilleur candidat. Par ailleurs, Li et Liu (2012) ont combiné un correcteur orthographique avec un modèle de traduction à base de blocs de caractères générés selon des règles phonétiques chinoises. Kogkitsidou et Antoniadis (2016) ont proposé un modèle qui, d'une part, produit une représentation intermédiaire de SMS par l'application de grammaires locales et, d'autre part, utilise un modèle de traduction automatique à base de règles pour convertir cette représentation vers une forme standard.

Par la suite, des modèles de traduction automatique neuronale ont été utilisés. Par exemple, Tiwari et Naskar (2017) ont proposé un modèle encodeur-décodeur de réseaux de neurones récurrents à mémoire court et long terme avec un mécanisme d'attention. Lourentzou *et al.* (2019) ont introduit un modèle hybride encodeur-décodeur à base de mots et de caractères, la composante à base de caractères étant entraînée sur des exemples antagonistes synthétiques. Plus récemment, Bucur *et al.* (2021) se sont servis du modèle de traduction multilingue préentraîné mBART (Liu *et al.*, 2020) pour proposer un modèle de normalisation au niveau de la phrase.

L'avantage des méthodes par traduction est qu'elles sont flexibles quant aux types de normalisation à réaliser (remplacements de plusieurs mots, réorganisation

des mots). Cependant, cette approche est limitée par le manque de ressources parallèles UGC et doit souvent reposer sur des techniques d'augmentation des données d'entraînement (Ling *et al.*, 2013 ; Tiwari et Naskar, 2017). À ce sujet, Matos Veliz *et al.* (2019b) ont comparé deux modèles de traduction automatique, statistique et neuronale, pour la normalisation de divers UGC en anglais et en néerlandais. Ils ont conclu que, pour la traduction statistique, il est mieux d'entraîner le modèle de langue sous-jacent sur un corpus issu d'un domaine similaire à celui des UGC et que, pour la traduction neuronale, il est préférable d'ajouter plus de données d'entraînement que de les augmenter artificiellement. Ils ont aussi proposé d'envisager une approche modulaire pour le modèle statistique, et une technique d'augmentation de données basée sur des règles pour le modèle neuronal. De plus, l'approche par traduction est parfois considérée comme « excessive » car la normalisation n'effectue pas beaucoup de transformations de la phrase source, contrairement à la traduction (Choudhury *et al.*, 2007). Ainsi, cette approche introduirait beaucoup plus de complexité que nécessaire (Kobus *et al.*, 2008). Elle peut aussi introduire beaucoup plus d'erreurs car la traduction n'est pas contrainte. En pratique, l'approche par correction des mots reste à privilégier car plus ciblée et performante : les derniers bons modèles de normalisation par l'approche de traduction (Lourentzou *et al.*, 2019 ; Bucur *et al.*, 2021) ne surpassent pas MoNoise.

3.2. *Sous-tâches et tâches connexes*

D'autres tâches étudiées dans la littérature ont un lien plus ou moins proche avec la normalisation lexicale et les travaux correspondants sont donc pertinents ici. Nous distinguons : (1) les sous-tâches, qui permettent de corriger une partie des phénomènes non standard, et (2) les tâches connexes, qui sont théoriquement semblables à la normalisation lexicale.

3.2.1. *Sous-tâches*

La correction orthographique consiste à remplacer des mots mal orthographiés dans un texte. Le plus souvent, il s'agit de fautes d'orthographe (cognitives) ou de typographie qui produisent des mots hors vocabulaire (Kukich, 1992). Dans le cas des UGC, ces erreurs ne sont pas toujours des fautes, mais peuvent être des choix intentionnels de l'auteur. Bien que la correction orthographique puisse normaliser certains mots non standard, elle ne suffit pas pour corriger certains phénomènes UGC comme les acronymes, les agglutinations et les abréviations qui couvrent plusieurs mots (Aw *et al.*, 2006 ; Han et Baldwin, 2011).

La normalisation phonétique consiste à corriger les erreurs d'ordre phonétique (qui constituent l'un des phénomènes non standard les plus observés dans les UGC). Elle est souvent couplée avec d'autres types de correction. En effet, certaines méthodes décrites dans la section 3.1 intègrent un module de calcul de similarité phonétique. Cette tâche est particulièrement utile pour normaliser les UGC dans les langues riches en homophonies comme le français (Rosales Núñez *et al.*, 2019b) ou le chinois (Qin

et al., 2021). Elle a aussi été appliquée à la correction orthographique dans les moteurs de recherche pour le commerce en ligne (Yang *et al.*, 2022).

La correction grammaticale vise à corriger les erreurs d'ordre grammatical, faisant le pendant de la normalisation lexicale qui vise à corriger les erreurs d'ordre lexical. Elle est aussi souvent découpée en deux sous-tâches : détection et correction. En pratique, la frontière entre erreur lexicale et erreur grammaticale n'est pas bien définie dans les UGC car certains phénomènes peuvent appartenir aux deux classes. Le choix revient aux annotateurs des données : certains essaient de se limiter à corriger les mots non standard d'un point de vue lexical, même si la phrase résultante reste agrammaticale (van der Goot *et al.*, 2021), alors que d'autres préfèrent garder un minimum de correction grammaticale comme l'insertion de mots manquants (p. ex. pronoms personnels sujets et verbes auxiliaires) et de signes de ponctuation (Wang et Ng, 2013 ; Zhang *et al.*, 2013 ; Bawden et Sagot, 2023).

3.2.2. *Tâches connexes*

Les textes résultant de la reconnaissance optique de caractères (*Optical Character Recognition, OCR*) doivent souvent être corrigés en post-traitement car ils contiennent des caractères mal reconnus et donc des mots non standard. Par ailleurs, les transcriptions résultant de la reconnaissance automatique de la parole (*Automatic Speech Recognition, ASR*) contiennent des mots non standard provenant des phonèmes mal compris. Ainsi, les tâches de correction post-OCR et post-ASR sont respectivement comparables à celles de correction orthographique et phonétique, et sont souvent abordées par les mêmes approches.

La normalisation de variantes dialectales et historiques est comparable à la normalisation lexicale, en assimilant grossièrement le langage non standard des UGC à un « dialecte » du langage standard. En particulier, certains travaux sur la normalisation de dialectes (Partanen *et al.*, 2019) et de créoles (Liu *et al.*, 2022), de textes produits par des locuteurs non natifs (Sarkar *et al.*, 2020 ; Alam et Anastasopoulos, 2020), et de langue non contemporaine (Ljubešić *et al.*, 2016 ; Bawden *et al.*, 2022) peuvent s'avérer intéressants.

3.3. *Évaluation*

Bien que la normalisation soit une solution potentielle pour le problème des mots non standard dans les UGC, elle reste une tâche qui est difficile à évaluer en raison du manque de ressources annotées d'une part, et du manque d'homogénéité dans le choix des conventions d'annotation et des métriques utilisées d'autre part.

3.3.1. *Données*

Malgré l'abondance d'UGC sur Internet, peu de données parallèles annotées pour la normalisation lexicale sont disponibles. Néanmoins, la campagne d'évaluation MultiLexNorm (van der Goot *et al.*, 2021) comprend des données annotées en

douze langues issues d'autres campagnes d'évaluation. D'autres données parallèles annotées sont disponibles en anglais et en néerlandais (De Clercq *et al.*, 2014), et en japonais (Higashiyama *et al.*, 2021). Stewart *et al.* (2019) ont publié un corpus annoté de rapports d'accidents industriels. Il est aussi important de noter le manque d'homogénéité dans les corpus annotés par rapport à la quantité de mots non standard et aux choix de normalisation. Tous ces corpus sont alignés au niveau des mots (c.-à-d. qu'ils comportent les correspondances explicites entre les mots non standard et leurs normalisations). Notamment, cela leur permet d'être directement évalués par les métriques de classification (voir section 3.3.2).

Il existe aussi des données UGC pour l'évaluation des tâches en aval. Celles-ci ne sont en général pas alignées au niveau des mots. Par exemple, le projet CoMeRe (Chanier *et al.*, 2014) rassemble des corpus français issus de communication médiée par les réseaux (SMS, forums, Twitter, etc.), annotés en parties du discours, en passant par une étape de normalisation. Plusieurs corpus parallèles sont aussi disponibles pour évaluer la traduction d'UGC (Michel et Neubig, 2018 ; Rosales Núñez *et al.*, 2019a ; Berard *et al.*, 2019 ; McNamee et Duh, 2022). Sluyter-Gäthje *et al.* (2018) ont annoté un corpus à la fois pour la traduction et pour l'analyse de sentiments. Plus récemment, Bawden et Sagot (2023) ont fourni un jeu de test pour évaluer la traduction depuis l'anglais non standard, et y ont inclus la normalisation de ces données.

Pour pallier le problème de manque de données, des techniques d'augmentation ont été utilisées pour générer des textes non standard artificiels. À partir de données parallèles entre anglais non standard et chinois standard, Ling *et al.* (2013) ont utilisé les sorties de systèmes de traduction du chinois vers l'anglais pour obtenir des textes en anglais standard alignés avec la source. Dekker et van der Goot (2020) et Samuel et Straka (2021) ont inséré des phénomènes UGC dans des textes non standard à partir de règles et de dictionnaires de mots et expressions UGC usuels. Dhole *et al.* (2023) ont mis en place le projet NL-Augmenter qui permet d'effectuer des transformations sur des textes pour générer des données artificielles pour les tâches de TAL.

3.3.2. *Métriques*

Plusieurs types de métriques ont été utilisés pour évaluer la normalisation lexicale :

– les métriques basées sur le comptage d'opérations d'édition (substitution, insertion, suppression) : le taux d'erreur de caractères (Ljubešić *et al.*, 2016 ; Matos Veliz *et al.*, 2019a) et le taux d'erreur de mots (Sproat *et al.*, 2001 ; Kobus *et al.*, 2008 ; Matos Veliz *et al.*, 2019b) ;

– les métriques de classification : l'exactitude, la précision, le rappel et la F-mesure (Baldwin *et al.*, 2015), la précision sur les mots hors vocabulaire (Alegria *et al.*, 2013), le taux de couverture c.-à-d. la capacité du modèle à toujours prédire la forme correcte dans ses n premiers candidats (Liu *et al.*, 2012) ;

– BLEU (Papineni *et al.*, 2002), qui est une métrique de traduction (Aw *et al.*, 2006 ; Kobus *et al.*, 2008 ; Han et Baldwin, 2011 ; Nishimwe, 2023).

Ces métriques ne sont pas sans détracteurs. D'une part, elles sont des métriques de surface qui pénalisent toutes les fausses normalisations de la même façon. Elles ne donnent donc pas une idée de la qualité de la phrase normalisée. Par exemple, étant donnée une phrase non standard *hello ppl* et sa normalisation attendue *hello people* (en français : *salut les gens*), le même score sera accordé à un modèle qui remplace l'abbréviation *ppl* par un autre candidat d'expansion comme *perplexity* (*perplexité*, dans le domaine du TAL), par un synonyme comme *everyone* (*tout le monde*) ou par un signe de ponctuation ! Par ailleurs, BLEU, qui est basé sur le calcul de chevauchement de n-grammes, est parfois considéré trop complexe pour une tâche où l'ordre des mots ne change pas (Kobus *et al.*, 2008 ; van der Goot, 2019c), étant donné qu'il corrèle presque parfaitement avec les métriques à base d'opérations d'édition pour la normalisation lexicale (Ljubešić *et al.*, 2016). Nishimwe (2023) a proposé d'envisager une combinaison de BLEU avec COMET (Rei *et al.*, 2020), une métrique neuronale de traduction qui compare le sens de deux textes et qui est plus robuste aux variations de surface. Avec COMET, une phrase non standard peut obtenir un score élevé sans être normalisée tant qu'elle conserve le sens de la phrase standard. En revanche, COMET pénalise par des scores plus faibles les normalisations erronées qui dégradent le sens de la phrase source. Il complémente donc BLEU et les autres métriques de surface : *hello everyone* serait donc moins pénalisée que *hello perplexity*.

D'autre part, les scores sont difficiles à comparer sur plusieurs corpus. En effet, ce qui constitue un « mot à normaliser » dépend des guides d'annotations. De plus, le taux de ces mots varie d'un corpus à l'autre : une exactitude élevée sur un corpus peut donc être insuffisante sur un autre. Un autre problème identifié par Reynaert (2008) est l'erreur de traitement des mots normalisés incorrectement par le système : ils étaient pénalisés à la fois dans la précision et dans le rappel[7], p. ex. dans les évaluations faites par Baldwin *et al.* (2015) et van der Goot et van Noord (2017). Par conséquent, van der Goot (2019c) a défini les éléments de la matrice de confusion pour la normalisation comme :

– *vrais positifs (VP)* : les mots normalisés par les annotateurs et correctement normalisés par le système ;

– *faux positifs (FP)* : les mots inchangés par les annotateurs, mais normalisés par le système ;

– *vrais négatifs (VN)* : les mots inchangés par les annotateurs et le système ;

– *faux négatifs (FN)* : les mots normalisés par les annotateurs, mais inchangés ou incorrectement normalisés par le système ;

et a défini le « taux de réduction de l'erreur » (*Error Reduction Rate, ERR*), qui peut être décrit comme l'exactitude normalisée par le nombre de mots à remplacer. L'ERR

7. Une métrique qui pourrait éviter ce problème est le *Slot Error Rate*, qui permet d'associer des coûts aux erreurs de détection et de correction des mots à normaliser (Makhoul *et al.*, 1999).

d'un modèle peut donc être calculé à partir de l'exactitude du modèle *Identité* (qui ne change rien dans la phrase source) :

$$\text{ERR} = \frac{\%\text{exactitude} - \%\text{exactitude}_{\text{Identité}}}{100 - \%\text{exactitude}_{\text{Identité}}} = \frac{\text{VP} - \text{FP}}{\text{VP} + \text{FN}} \quad [3]$$

L'ERR[8] a été utilisé dans la campagne d'évaluation MultiLexNorm (van der Goot *et al.*, 2021). Il permet de comparer la performance d'un modèle sur plusieurs jeux de données différents, voire plusieurs langues. Cependant, comme l'exactitude, il ne distingue pas entre les faux positifs et les faux négatifs ; l'utilisation de la précision et du rappel, en plus de l'ERR, est donc préférable. En outre, ces métriques de classification nécessitent une correspondance entre les mots à normaliser et leurs remplacements, à la fois dans les données d'évaluation et dans les sorties des modèles. Si l'on n'en dispose pas, il faut prévoir une étape supplémentaire (automatique ou manuelle) pour effectuer cet alignement (Bucur *et al.*, 2021). Par ailleurs, les autres métriques peuvent être appliquées directement sur les phrases entières sans correspondances explicites entre les mots et leurs normalisations, mais elles sont moins descriptives.

Enfin, toutes ces métriques nécessitent une normalisation de référence. Ainsi, une normalisation n'est correcte que si elle a été prévue dans les guides d'annotations des corpus. Par exemple, un modèle qui normalise *mdr* en *mort de rire*, pourtant correct, va être pénalisé si la normalisation de référence ne l'a pas fait[9]. Une autre limite est qu'elles ne tiennent pas compte de la performance de la tâche que l'on souhaite réaliser en aval. Par exemple, Zhang *et al.* (2013) ont préconisé l'utilisation d'une métrique conjointe entre la normalisation et l'analyse de dépendances. Ce type de métrique permet de mettre en évidence les transformations qui ont un impact sur la tâche considérée (p. ex. la restauration de la ponctuation et des majuscules ou la réorganisation des mots).

4. Discussion

4.1. *Limites*

4.1.1. *La normalisation peut introduire du bruit*

La normalisation d'UGC présente encore quelques difficultés. Certains phénomènes non standard restent difficiles à normaliser, particulièrement les abréviations, agglutinations et acronymes, en raison de leur ambiguïté et de la grande différence de nombre de caractères avec leurs normalisations. De plus, une fois un modèle de normalisation entraîné, il reste figé dans le temps et peut peiner à se généraliser aux

8. Voir (van der Goot, 2019c) pour la démonstration des égalités dans l'équation 3.

9. Dans (van der Goot *et al.*, 2021), l'expression équivalente en anglais *lol* n'a pas été remplacée par *laughing out loud* dans la référence.

nouvelles expressions émergeant sur les réseaux sociaux. Celles-ci varient beaucoup d'une personne à l'autre et d'une plateforme à l'autre (Dekker et van der Goot, 2020).

Bien qu'une bonne normalisation puisse améliorer la performance de modèles de TAL sur les UGC, une mauvaise normalisation ou une surnormalisation peuvent être une source de bruit et de propagation de l'erreur, et entraîner une dégradation de la performance en aval (Matos Veliz *et al.*, 2019a). Par exemple, le tableau 2 illustre la traduction anglais-français par le modèle NLLB (NLLB Team *et al.*, 2022) à 600 millions de paramètres[10] d'une phrase issue du corpus RoCS-MT, de sa version standard de référence, et de sa normalisation effectuée par le modèle ÚFAL[11] (Samuel et Straka, 2021), vainqueur de la campagne d'évaluation MultiLexNorm. Nous observons que la qualité de la traduction se dégrade après la normalisation car ÚFAL remplace l'abréviation *uni* par *united* et non *university*.

	Phrase source	Traduction
n. s.	wld rly appreciate if yall can help me out, esp those currently in *uni* or left alr.	J'apprécierai si vous pouvez m'aider, surtout ceux qui sont actuellement à l'université ou à l'extérieur.
réf.	I would really appreciate if you all could help me out, especially those who are currently at *university* or have already left.	Je serais vraiment reconnaissant si vous pouviez tous m'aider, surtout ceux qui sont actuellement à l'université ou qui ont déjà quitté.
norm.	would really appreciate if y'all can help me out, especially those currently in *united* or left alr.	J'apprécierais vraiment si vous pouviez m'aider, surtout ceux qui sont actuellement *en Alger*.

TABLEAU 2. *Phrase de RoCS-MT (Bawden et Sagot, 2023) en anglais non standard* (n. s.)*, sa version standard de référence* (réf.)*, sa normalisation par le modèle ÚFAL* (norm.)*, et leurs traductions en français par le modèle NLLB*

4.1.2. *La normalisation est une tâche difficile à définir*

Il n'y a pas de définition unique de la portée de la normalisation lexicale. Pourtant, cette dernière permet de définir les guides d'annotations des données d'entraînement et d'évaluation. Zhang *et al.* (2013) ont suggéré que le niveau de normalisation adéquat dépend de la tâche de TAL effectuée en aval, et que celle-ci ne peut être dissociée ni de la création des jeux de données, ni de la conception et de l'évaluation du modèle de normalisation. Baldwin et Li (2015) ont aussi montré que les transformations effectuées pendant la normalisation n'avaient pas la même importance selon la tâche considérée. Par exemple, les corpus MultiLexNorm et RoCS-MT n'ont pas le même niveau de normalisation car ils ont été conçus pour des tâches différentes : l'étiquetage morphosyntaxique et l'analyse de dépendances pour MultiLexNorm, et la

10. `https://huggingface.co/facebook/nllb-200-distilled-600M`
11. `https://huggingface.co/ufal/byt5-small-multilexnorm2021-en`

traduction automatique pour RoCS-MT. Alors que MultiLexNorm se limite à faire des remplacements 1-à-1, 1-à-n et n-à-1 même si la phrase reste agrammaticale, RoCS-MT effectue plus de transformations pour obtenir une phrase grammaticale : dans le tableau 1, la normalisation de la phrase de MultiLexNorm se limite à la correction de mots alors que celle de RoCS-MT corrige aussi la casse et la ponctuation.

4.1.3. *La normalisation est une tâche dépendante de la langue*

Bien que toutes les langues présentent des phénomènes non standard dans les UGC, ces phénomènes se manifestent différemment d'une langue à l'autre et d'un système d'écriture à l'autre, et certains sont spécifiques à une langue et n'ont pas d'équivalence dans une autre (Ling *et al.*, 2013). Il est donc difficile d'implémenter un modèle de normalisation universel multilingue, surtout avec le manque de données annotées. Par exemple, Bucur *et al.* (2021) ont entraîné une version de leur modèle de normalisation pour chacune des langues de MultiLexNorm, ainsi qu'un modèle unique multilingue, et ont eu de meilleures performances avec les modèles séparés.

4.1.4. *La normalisation fait perdre des informations stylistiques*

Normaliser une phrase non standard peut entraîner une perte d'informations stylistiques telles que les nuances pragmatiques, le registre, les émotions et la mise en évidence de certains mots. Ces informations peuvent être utiles pour des tâches comme la synthèse vocale (Baldwin et Chai, 2011), l'analyse de sentiments (van Hee *et al.*, 2017) et la traduction (Bawden et Sagot, 2023). Par exemple, le tableau 3 illustre la normalisation de phrases non standard issues du corpus PFSMB (Rosales Núñez *et al.*, 2021a) par l'agent conversationnel ChatGPT-3.5[12]. Ce dernier ne se limite pas à remplacer les mots non standard mais peut aussi rendre les phrases moins familières ou moins grossières (p. ex. les deux dernières phrases).

Phrase non standard	Phrase normalisée
c un *truk* que j'ai *compri* aussi *o* fil des années, *c* que si *c* pas pour nous on y arrivera pas.	*C'est* un *truc* que j'ai *compris* aussi *au* fil des années, *c'est* que si *ce n'est* pas pour nous, on n'y arrivera pas.
kontinut kom sa tu *er tro bele* tu *va avoare onkore plin* de *sukser*	*Continue comme ça*, tu *es vraiment belle*. Tu *vas avoir encore plein* de *succès*.
Waw *c koi* ce plat *de ouf* , *c* pour 150 *Personne omoin*	Waw, *c'est quoi* ce plat *incroyable*, *c'est* pour *au moins* 150 *personnes*.
56 *a* ce *fdp* de *flappy bird* ! *Jvais* saigner *loiseau*	56 *à* ce *fichu Flappy Bird* ! *Je vais* saigner *l'oiseau*.

TABLEAU 3. *Phrases issues de PFSMB (Rosales Núñez* et al.*, 2019a) et leurs normalisations par ChatGPT-3.5*

12. `https://chat.openai.com`

4.2. *Perspectives*

Une question importante se pose alors : *la normalisation lexicale reste-elle une tâche pertinente (en soi ou en amont d'autres tâches de TAL) ?* Cette question est d'autant plus d'actualité aujourd'hui puisque les modèles de TAL sont très performants et, en particulier, plus robustes aux UGC. Par exemple, Bawden et Sagot (2023) ont montré que le modèle GPT-4 (OpenAI, 2023) traduit déjà bien les phrases non standard de RoCS-MT. De plus, ChatGPT-3.5 se montre aussi capable de réaliser la normalisation (voir tableau 3), et Pan *et al.* (2024) ont montré que les grands modèles de langue *(Large Language Models, LLM)* pouvaient apprendre à traduire des textes UGC à partir de quelques exemples de démonstration. Alors, *est-ce que ces LLM génératifs ont rendu la normalisation obsolète ?* Oui, mais seulement pour les quelques langues très dotées pour lesquelles ils sont très performants. Pour la majorité des langues (qui sont moins ou peu dotées), la performance sur les données standard est loin d'être satisfaisante (Ignat *et al.*, 2024). Il en découle que ce problème est exacerbé sur les données UGC. Par ailleurs, il est judicieux de noter qu'une contamination est possible pour ces modèles, c.-à-d. qu'ils aient pu voir les données d'évaluation dans leurs données d'entraînement (parfois non rendues publiques), ce qui rend leur évaluation difficile. En outre, la question de la capacité à généraliser de ces modèles reste ouverte. Sont-ils vraiment plus robustes aux variations lexicales, ou ont-ils vu assez d'instances non standard des mots pour les considérer comme standard ? Seront-ils robustes aux néologismes et aux nouveaux phénomènes UGC qui émergeront d'ici quelques années ?

Qu'en est-il donc de l'avenir de la recherche sur la normalisation lexicale ? Nous identifions encore deux intérêts de cette tâche : en soi, elle permet d'étudier les aspects sociolinguistiques et phonologiques des textes issus des réseaux sociaux (Eisenstein, 2013 ; Chanier *et al.*, 2014) et, en amont d'autres tâches, elle permet de continuer à utiliser des modèles de TAL économiques (plus petits et donc moins robustes aux UGC) dans des situations de ressources limitées. Dans ces contextes, nous jugeons bénéfique de continuer de faire de la recherche sur la normalisation lexicale, et nous proposons des axes de travaux de recherche qui répondent aux trois premières limites identifiées dans la section 4.1 :

1) créer plus de corpus parallèles d'entraînement et d'évaluation : par le biais d'annotations manuelles, de la fouille automatique de textes alignés, ou des techniques plus sophistiquées d'augmentation artificielle de données (y compris des LLM) ;

2) améliorer les protocoles et les métriques d'évaluation ;

3) étendre et améliorer les modèles de normalisation sur d'autres langues, surtout les moins dotées.

Cependant, le problème de la perte d'informations stylistiques reste inévitable avec la normalisation qui, par définition, vise à supprimer les variations lexicales dans les textes UGC. Il peut être contourné en adoptant une approche d'adaptation de domaine.

5. Conclusion

Cet article a pour vocation de faire une étude de la tâche de normalisation lexicale des contenus produits par les utilisateurs (UGC). Dans un premier temps, nous avons présenté les UGC sur les réseaux sociaux et nous avons montré qu'ils sont un fléau pour les modèles de TAL entraînés sur des données standard, en raison de leur multitude de phénomènes de langage non standard. Dans un second temps, nous avons présenté la normalisation lexicale et montré qu'elle est l'une des approches pratiques pour pallier ce problème. Nous avons effectué un état de l'art de ses méthodes principales et évoqué ses avantages mais aussi ses limites, en particulier la difficulté d'évaluation et le manque de ressources. Enfin, nous avons conclu par une discussion sur la pertinence de la tâche dans l'avenir du TAL : les modèles actuels étant déjà très robustes aux UGC, la normalisation lexicale reste utile sous certaines contraintes (ressources matérielles limitées, langues peu dotées), ou pour des études sociolinguistiques. Dans ces contextes, nous avons ouvert la porte à des perspectives de travaux de recherche.

Remerciements

Un grand merci aux relecteurs de la Revue TAL pour leurs commentaires précieux. Ce travail a été financé par les chaires de Rachel Bawden et de Benoît Sagot dans l'institut PRAIRIE, lui-même financé par l'Agence Nationale de la Recherche dans le cadre du programme « Investissements d'avenir » sous la référence ANR-19-P3IA-0001.

6. Bibliographie

Ahmed B., « Lexical normalisation of Twitter Data », *Proceedings of the 2015 Science and Information Conference*, IEEE, London, UK, p. 326-328, 2015.

Alam M. M. I., Anastasopoulos A., « Fine-Tuning MT systems for Robustness to Second-Language Speaker Variations », *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Association for Computational Linguistics, Online, p. 149-158, 2020.

Alegria I., Aranberri N., Fresno-Fernández V., Gamallo P., Padró L., Vicente I. S., Turmo J., Zubiaga A., « Introducción a la Tarea Compartida Tweet-Norm 2013 : Normalización Léxica de Tuits en Español », *Proceedings of the XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural*, Madrid, Spain, p. 38-46, 2013.

Aminian M., Avontuur T., Balemans I., Elshof L., Newell R., Noord N. V., Ntavelos A., van Zaanen M., Azar E. Z., « Assigning Part-of-Speech to Dutch Tweets », *Proceedings of the LREC 2012 Workshop @NLP can u tag #user_generated_content ?!*, Istanbul, Turkey, p. 9-14, 2012.

Aw A., Zhang M., Xiao J., Su J., « A Phrase-Based Statistical Model for SMS Text Normalization », *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Association for Computational Linguistics, Sydney, Australia, p. 33-40, 2006.

Baldwin T., Chai J., « Beyond Normalization : Pragmatics of Word Form in Text Messages », *Proceedings of 5th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, p. 1437-1441, 2011.

Baldwin T., Cook P., Lui M., MacKinlay A., Wang L., « How Noisy Social Media Text, How Diffrnt Social Media Sources ? », *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, p. 356-364, 2013.

Baldwin T., de Marneffe M. C., Han B., Kim Y.-B., Ritter A., Xu W., « Shared Tasks of the 2015 Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition », *Proceedings of the Workshop on Noisy User-generated Text*, Association for Computational Linguistics, Beijing, China, p. 126-135, 2015.

Baldwin T., Li Y., « An In-depth Analysis of the Effect of Text Normalization in Social Media », *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, p. 420-429, 2015.

Baranes M., Sagot B., « Analogy-based Text Normalization : the case of unknowns words (Normalisation de textes par analogie : le cas des mots inconnus) [in French] », *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, Association pour le Traitement Automatique des Langues, Marseille, France, p. 137-148, 2014.

Bawden R., Poinhos J., Kogkitsidou E., Gambette P., Sagot B., Gabay S., « Automatic Normalisation of Early Modern French », *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 3354-3366, 2022.

Bawden R., Sagot B., « RoCS-MT : Robustness Challenge Set for Machine Translation », *Proceedings of the Eighth Conference on Machine Translation*, Association for Computational Linguistics, Singapore, p. 198-216, 2023.

Beckley R., « Bekli :A Simple Approach to Twitter Text Normalization. », *Proceedings of the Workshop on Noisy User-generated Text*, Association for Computational Linguistics, Beijing, China, p. 82-86, 2015.

Belinkov Y., Bisk Y., « Synthetic and Natural Noise Both Break Neural Machine Translation », *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada, 2017.

Berard A., Calapodescu I., Dymetman M., Roux C., Meunier J.-L., Nikoulina V., « Machine Translation of Restaurant Reviews : New Corpus for Domain Adaptation and Robustness », *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong, p. 168-176, 2019.

Bucur A.-M., Cosma A., Dinu L. P., « Sequence-to-Sequence Lexical Normalization with Multilingual Transformers », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Association for Computational Linguistics, Online, p. 473-482, 2021.

Cerón-Guzmán J. A., León-Guzmán E., « Lexical Normalization of Spanish Tweets », *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16

Companion, International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, p. 605-610, 2016.

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C., Hriba L., Longhi J., Seddah D., « The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres », *Journal for Language Technology and Computational Linguistics*, vol. 29, n° 2, p. 1–30, 2014.

Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar S., Basu A., « Investigation and modeling of the structure of texting language », *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, n° 3-4, p. 157-174, 2007.

Clark E., Araki K., « Text Normalization in Social Media : Progress, Problems and Applications for a Pre-Processing System of Casual English », *Procedia - Social and Behavioral Sciences*, vol. 27, p. 2-11, 2011.

Contractor D., Faruquie T. A., Subramaniam L. V., « Unsupervised cleansing of noisy text », *Proceedings of Coling 2010 : Posters*, Beijing, China, p. 189-196, 2010.

Costa Bertaglia T. F., Volpe Nunes M. d. G., « Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization », *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, Osaka, Japan, p. 112-120, 2016.

Cotelo J., Cruz F., Troyano J., Ortega F., « A modular approach for lexical normalization applied to Spanish tweets », *Expert Systems with Applications*, vol. 42, n° 10, p. 4743-4754, 2015.

Damerau F. J., « A technique for computer detection and correction of spelling errors », *Communications of the ACM*, vol. 7, n° 3, p. 171-176, 1964.

De Clercq O., Schulz S., Desmet B., Hoste V., « Towards Shared Datasets for Normalization Research », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, p. 1218-1223, 2014.

Dekker K., van der Goot R., « Synthetic Data for English Lexical Normalization : How Close Can We Get to Manually Annotated Data ? », *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, p. 6300-6309, 2020.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, p. 4171-4186, 2019.

Dhole K. D., Gangal V., Gehrmann S., et al., « NL-Augmenter : A Framework for Task-Sensitive Natural Language Augmentation », *The Northern European Journal of Language Technology (NEJLT)*, vol. 9, n° 1, p. 60-100, 2023.

Ehara Y., « To What Extent Does Lexical Normalization Help English-as-a-Second Language Learners to Read Noisy English Texts ? », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 451-456, 2021.

Eisenstein J., « What to do about bad language on the internet », *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Atlanta, Georgia, p. 359-369, 2013.

Formiga L., Fonollosa J. A. R., « Dealing with Input Noise in Statistical Machine Translation », *Proceedings of COLING 2012 : Posters*, Mumbai, India, p. 319-328, 2012.

Foster J., « "cba to check the spelling" : Investigating Parser Performance on Discussion Forum Posts », *Human Language Technologies : The 2010 Annual Conference of the*

North American Chapter of the Association for Computational Linguistics, Los Angeles, California, p. 381-384, 2010.

Han B., Baldwin T., « Lexical Normalisation of Short Text Messages : Makn Sens a #twitter », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, Portland, Oregon, USA, p. 368-378, 2011.

Hassan H., Menezes A., « Social Text Normalization using Contextual Graph Random Walks », *in* H. Schuetze, P. Fung, M. Poesio (eds), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Sofia, Bulgaria, p. 1577-1586, 2013.

Higashiyama S., Utiyama M., Watanabe T., Sumita E., « User-Generated Text Corpus for Evaluating Japanese Morphological Analysis and Lexical Normalization », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Online, p. 5532-5541, 2021.

Ignat O., Jin Z., Abzaliev A., Biester L., Castro S., Deng N., Gao X., Gunal A. E., He J., Kazemi A., Khalifa M., Koh N., Lee A., Liu S., Min D. J., Mori S., Nwatu J. C., Perez-Rosas V., Shen S., Wang Z., Wu W., Mihalcea R., « Has It All Been Solved ? Open NLP Research Questions Not Solved by Large Language Models », *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, p. 8050-8094, 2024.

Jiang N., Luo C., Lakshman V., Dattatreya Y., Xue Y., « Massive Text Normalization via an Efficient Randomized Algorithm », *Proceedings of the ACM Web Conference 2022*, Virtual Event, Lyon France, p. 2946-2956, 2022.

Karpukhin V., Levy O., Eisenstein J., Ghazvininejad M., « Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, p. 42-47, 2019.

Kobus C., Yvon F., Damnati G., « Normalizing SMS : are Two Metaphors Better than One ? », *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, p. 441-448, 2008.

Kogkitsidou E., Antoniadis G., « L'architecture d'un modèle hybride pour la normalisation de SMS (A hybrid model architecture for SMS normalization) », *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, Paris, France, p. 355-363, 2016.

Kubal D., Nagvenkar A., « Multilingual Sequence Labeling Approach to solve Lexical Normalization », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 457-464, 2021.

Kukich K., « Techniques for automatically correcting words in text », *ACM Computing Surveys*, vol. 24, n° 4, p. 377–439, 1992.

Kumar A., Makhija P., Gupta A., « Noisy Text Data : Achilles' Heel of BERT », *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Online, p. 16-21, 2020.

Leeman-Munk S., Lester J., Cox J., « NCSU_SAS_SAM : Deep Encoding and Reconstruction for Normalization of Noisy Text », *Proceedings of the Workshop on Noisy User-generated Text*, Beijing, China, p. 154-161, 2015.

Li C., Liu Y., « Improving Text Normalization using Character-Blocks Based Models and System Combination », *Proceedings of COLING 2012*, Mumbai, India, p. 1587-1602, 2012.

Ling W., Dyer C., Black A. W., Trancoso I., « Paraphrasing 4 Microblog Normalization », *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, p. 73-84, 2013.

Liu F., Weng F., Jiang X., « A Broad-Coverage Normalization System for Social Media Language », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Jeju Island, Korea, p. 1035-1044, 2012.

Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L., « Multilingual Denoising Pre-training for Neural Machine Translation », *Transactions of the Association for Computational Linguistics*, vol. 8, p. 726-742, 2020.

Liu Z., Ni S., Aw A. T., Chen N. F., « Singlish Message Paraphrasing : A Joint Task of Creole Translation and Text Normalization », *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, p. 3924-3936, 2022.

Ljubešić N., Zupan K., Fišer D., Erjavec T., « Normalising Slovene data : historical texts vs. user-generated content », *Proceedings of the 13th Conference on Natural Language Processing*, Bochum, Germany, p. 146-155, 2016.

Lourentzou I., Manghnani K., Zhai C., « Adapting Sequence to Sequence models for Text Normalization in Social Media », *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*, Münich, Germany, p. 335-345, 2019.

Makhoul J., Kubala F., Schwartz R., Weischedel R., « Performance measures for information extraction », *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia, p. 249-252, 1999.

Matos Veliz C., De Clercq O., Hoste V., « Benefits of Data Augmentation for NMT-based Text Normalization of User-Generated Content », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, p. 275-285, 2019a.

Matos Veliz C., De Clercq O., Hoste V., « Comparing MT Approaches for Text Normalization », *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, INCOMA Ltd., Varna, Bulgaria, p. 740-749, 2019b.

McNamee P., Duh K., « The Multilingual Microblog Translation Corpus : Improving and Evaluating Translation of User-Generated Text », *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, p. 910-918, 2022.

Melero M., Costa-Jussà M. R., Lambert P., Quixal M., « Selection of correction candidates for the normalization of Spanish user-generated content », *Natural Language Engineering*, vol. 22, nº 1, p. 135-161, 2016.

Michel P., Neubig G., « MTNT : A Testbed for Machine Translation of Noisy Text », *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, p. 543-553, 2018.

Moon S., Neves L., Carvalho V., « Multimodal Named Entity Recognition for Short Social Media Posts », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, p. 852-860, 2018.

Muñoz-García Ó., Navarro C., Avontuur T., Azar Z., Balemans I., Elshof L., Newell R., Noord N. V., Ntavelos A., Maynard D., Bontcheva K., Rout D., Strassel S., Ismael S., Song Z., Lee H., « Comparing User Generated Content Published in Different Social Media Sources », *Proceedings of the LREC 2012 Workshop @NLP can u tag #user_generated_content ?!*, Istanbul, Turkey, p. 1-8, 2012.

Muller B., Sagot B., Seddah D., « Enhancing BERT for Lexical Normalization », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, p. 297-306, 2019.

Nguyen D. Q., Vu T., Tuan Nguyen A., « BERTweet : A pre-trained language model for English Tweets », *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, Online, p. 9-14, 2020.

Nguyen V. H., Nguyen H. T., Snasel V., « Text normalization for named entity recognition in Vietnamese tweets », *Computational Social Networks*, vol. 3, n° 1, p. 10, 2016.

Nishimwe L., « Normalisation lexicale de contenus générés par les utilisateurs sur les réseaux sociaux », *Actes des* 16^e *Rencontres Jeunes Chercheurs en RI (RJCRI) et* 25^e *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, Paris, France, p. 160-183, 2023.

NLLB Team, Costa-jussà M. R., Cross J., Çelebi O., et al., « No Language Left Behind : Scaling Human-Centered Machine Translation », *CoRR*, 2022.

OpenAI, « GPT-4 Technical Report », *CoRR*, 2023.

Pan L., Leng Y., Xiong D., « Can Large Language Models Learn Translation Robustness from Noisy-Source In-context Demonstrations ? », *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, p. 2798-2808, 2024.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a Method for Automatic Evaluation of Machine Translation », *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, p. 311-318, 2002.

Partanen N., Hämäläinen M., Alnajjar K., « Dialect Text Normalization to Normative Standard Finnish », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, p. 141-146, 2019.

Pennell D. L., Liu Y., « Normalization of text messages for text-to-speech », *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 4842-4845, 2010.

Pennell D., Liu Y., « A Character-Level Machine Translation Approach for Normalization of SMS Abbreviations », *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, p. 974-982, 2011.

Peters B., Martins A. F. T., « Did Translation Models Get More Robust Without Anyone Even Noticing ? », *CoRR*, 2024.

Plank B., Jensen K. N., van der Goot R., « DaN+ : Danish Nested Named Entities and Lexical Normalization », *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), p. 6649-6662, 2020.

Popović M., Lapshinova-Koltunski E., Koponen M., « Effects of different types of noise in user-generated reviews on human and machine translations including ChatGPT », *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, San Ġiljan, Malta, p. 17-30, 2024.

Qin W., Li X., Sun Y., Xiong D., Cui J., Wang B., « Modeling Homophone Noise for Robust Neural Machine Translation », *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, ON, Canada, p. 7533-7537, 2021.

Rei R., Stewart C., Farinha A. C., Lavie A., « COMET : A Neural Framework for MT Evaluation », *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, p. 2685-2702, 2020.

Reynaert M., « All, and only, the Errors : more Complete and Consistent Spelling and OCR-Error Correction Evaluation », *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Riabi A., Sagot B., Seddah D., « Can Character-based Language Models Improve Downstream Task Performances In Low-Resource And Noisy Language Scenarios ? », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 423-436, 2021.

Ritter A., Clark S., Mausam, Etzioni O., « Named Entity Recognition in Tweets : An Experimental Study », *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., p. 1524-1534, 2011.

Rosales Núñez J. C., Seddah D., Wisniewski G., « Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content », *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland, p. 2-14, 2019a.

Rosales Núñez J. C., Seddah D., Wisniewski G., « Phonetic Normalization for Machine Translation of User Generated Content », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Association for Computational Linguistics, Hong Kong, China, p. 407-416, 2019b.

Rosales Núñez J. C., Seddah D., Wisniewski G., « Understanding the Impact of UGC Specificities on Translation Quality », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 189-198, 2021a.

Rosales Núñez J. C., Seddah D., Wisniewski G., « Multi-way Variational NMT for UGC : Improving Robustness in Zero-shot Scenarios via Mixture Density Networks », *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands, p. 447-459, 2023.

Rosales Núñez J. C., Wisniewski G., Seddah D., « Noisy UGC Translation at the Character Level : Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 199-211, 2021b.

Ruiz P., Cuadros M., Etchegoyhen T., « Lexical Normalization of Spanish Tweets with Rule-Based Components and Language Models », *Procesamiento del Lenguaje Natural*, vol. 52, p. 45-52, 2014.

Samuel D., Straka M., « ÚFAL at MultiLexNorm 2021 : Improving Multilingual Lexical Normalization by Fine-tuning ByT5 », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 483-492, 2021.

Sanguinetti M., Bosco C., Cassidy L., Çetinoğlu Ö., Cignarella A. T., Lynn T., Rehbein I., Ruppenhofer J., Seddah D., Zeldes A., « Treebanking User-Generated Content : A Proposal for a Unified Representation in Universal Dependencies », *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, p. 5240-5250, 2020.

Sarkar R., Mahinder S., KhudaBukhsh A., « The Non-native Speaker Aspect : Indian English in Social Media », *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Online, p. 61-70, 2020.

Scherrer Y., Ljubešić N., « Sesame Street to Mount Sinai : BERT-constrained character-level Moses models for multilingual lexical normalization », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 465-472, 2021.

Seddah D., Sagot B., Candito M., Mouilleron V., Combet V., « The French Social Media Bank : a Treebank of Noisy User Generated Content », *Proceedings of COLING 2012*, Mumbai, India, p. 2441-2458, 2012.

Shannon C. E., « A mathematical theory of communication », *The Bell System Technical Journal*, vol. 27, n° 3, p. 379-423, 1948.

Sluyter-Gäthje H., Lohar P., Afli H., Way A., « FooTweets : A Bilingual Parallel Corpus of World Cup Tweets », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

Sproat R., Black A. W., Chen S., Kumar S., Ostendorf M., Richards C. D., « Normalization of non-standard words », *Computer Speech & Language*, vol. 15, n° 3, p. 287-333, 2001.

Sproat R., Jaitly N., « RNN Approaches to Text Normalization : A Challenge », *CoRR*, 2016.

Stewart M., Liu W., Cardell-Oliver R., « Word-level Lexical Normalisation using Context-Dependent Embeddings », *CoRR*, 2019.

Stewart M., Liu W., Cardell-Oliver R., Wang R., « Short-Text Lexical Normalisation on Industrial Log Data », *2018 IEEE International Conference on Big Knowledge (ICBK),* , vol. , p. 113-122, 2018.

Supranovich D., Patsepnia V., « IHS_RD : Lexical Normalization for English Tweets », *Proceedings of the Workshop on Noisy User-generated Text*, Beijing, China, p. 78-81, 2015.

Tian T., Tellier I., Dinarelli M., Cardoso P., « Détection des mots non-standards dans les tweets avec des réseaux de neurones (Detecting non-standard words in tweets with neural networks) », *Actes des* 24^e *Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts*, Orléans, France, p. 174-182, 2017.

Tiwari A. S., Naskar S. K., « Normalization of Social Media Text using Deep Neural Networks », *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, Kolkata, India, p. 312-321, 2017.

van der Goot R., « An In-depth Analysis of the Effect of Lexical Normalization on the Dependency Parsing of Social Media », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China, p. 115-120, 2019a.

van der Goot R., « MoNoise : A Multi-lingual and Easy-to-use Lexical Normalization Tool », *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Association for Computational Linguistics, Florence, Italy, p. 201-206, 2019b.

van der Goot R., Normalization and parsing algorithms for uncertain input, PhD thesis, University of Groningen, 2019c.

van der Goot R., Plank B., Nissim M., « To normalize, or not to normalize : The impact of normalization on Part-of-Speech tagging », *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Association for Computational Linguistics, Copenhagen, Denmark, p. 31-39, 2017.

van der Goot R., Ramponi A., Caselli T., Cafagna M., De Mattei L., « Norm It ! Lexical Normalization for Italian and Its Downstream Effects for Dependency Parsing », *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, p. 6272-6278, 2020.

van der Goot R., Ramponi A., Zubiaga A., Plank B., Muller B., San Vicente Roncal I., Ljubešić N., Çetinoğlu Ö., Mahendra R., Çolakoğlu T., Baldwin T., Caselli T., Sidorenko W., « MultiLexNorm : A Shared Task on Multilingual Lexical Normalization », *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, p. 493-509, 2021.

van der Goot R., van Noord G., « MoNoise : Modeling Noise Using a Modular Normalization System », *Computational Linguistics in the Netherlands Journal*, vol. 7, p. 129-144, 2017.

van der Goot R., van Noord R., van Noord G., « A Taxonomy for In-depth Evaluation of Normalization for User Generated Content », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, p. 684-688, 2018.

van der Goot R., Çetinoğlu Ö., « Lexical Normalization for Code-switched Data and its Effect on POS Tagging », *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, Online, p. 2352-2365, 2021.

van Hee C., van de Kauter M., de Clercq O., Lefever E., Desmet B., Hoste V., « Noise or music ? Investigating the usefulness of normalisation for robust sentiment analysis on social media data », *Traitement Automatique des Langues*, vol. 58, n° 1, p. 63-87, 2017.

Vielsted M., Wallenius N., van der Goot R., « Increasing Robustness for Cross-domain Dialogue Act Classification on Social Media Data », *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, Gyeongju, Republic of Korea, p. 180-193, 2022.

Viterbi A., « Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm », *IEEE Transactions on Information Theory*, vol. 13, n° 2, p. 260-269, 1967.

Wang P., Ng H. T., « A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation », *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Atlanta, Georgia, p. 471-481, 2013.

Xu K., Xia Y., Lee C.-H., « Tweet Normalization with Syllables », *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, Beijing, China, p. 920-928, 2015.

Xue L., Barua A., Constant N., Al-Rfou R., Narang S., Kale M., Roberts A., Raffel C., « ByT5 : Towards a Token-Free Future with Pre-trained Byte-to-Byte Models », *Transactions of the Association for Computational Linguistics*, vol. 10, p. 291-306, 2022.

Yang F., Bagheri Garakani A., Teng Y., Gao Y., Liu J., Deng J., Sun Y., « Spelling Correction using Phonetics in E-commerce Search », *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, Dublin, Ireland, p. 63-67, 2022.

Zhang C., Baldwin T., Ho H., Kimelfeld B., Li Y., « Adaptive Parser-Centric Text Normalization », *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Sofia, Bulgaria, p. 1159-1168, 2013.

Annexe : Méthodologie de recherche bibliographique

Nous avons utilisé diverses méthodes pour établir la bibliographie de notre état de l'art, notamment :

1) la recherche par mots-clés en anglais (*lexical normalization, user-generated content*, *corpus*, *dataset*, *social media*) sur les moteurs de recherche bibliographique *Google Scholar*[13] et *Semantic Scholar*[14], et sur le site *ACL Anthology*[15] qui indexe les articles des principales conférences de TAL ;

2) le parcours manuel des tous les articles de toutes les éditions de l'atelier *W-NUT (Workshop on Noisy User-generated Text)* jusqu'à 2024 inclus ;

3) l'exploration des réseaux d'articles construits sur le site *Connected Papers*[16] à partir de quelques articles clés (Seddah *et al.*, 2012 ; Eisenstein, 2013 ; Rosales Núñez *et al.*, 2021a ; van der Goot *et al.*, 2021).

13. https://scholar.google.com
14. https://www.semanticscholar.org
15. https://aclanthology.org
16. https://www.connectedpapers.com

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 (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) : 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. « 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 : 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 Gallica[9] 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. « 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 (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 a 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 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`[17] dans les textes que nous avons sélectionnés comme textes de réf. sont présentés 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 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 Gutenberg[18],

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 Québec[19]. Les textes disponibles sont tous de bonne qualité. L'action COST a mis en place une liste de critères[20] 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 segmentation[22] et la transcription[23]. 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
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 small-ELTeC-fr*

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 small-ELTeC-en*

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 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 neuronaux 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 français[25], le portugais[26] et l'anglais[27].

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'entrée[28]. 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 Conversion[29]. 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 bruitées[30]. 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	*il en est tombe au sort cinq de Sainl-Bruncl*le *durant*a *todo o t*ompe *em q*ne *ostivesso em Portn*gal	Sainl-Bruncl*l*e. Portngal
Ajout d'un caractère minuscule au début de l'entité	*Aux k*Etats-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 a*Morlincourt' *tot*	_ 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	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, 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.

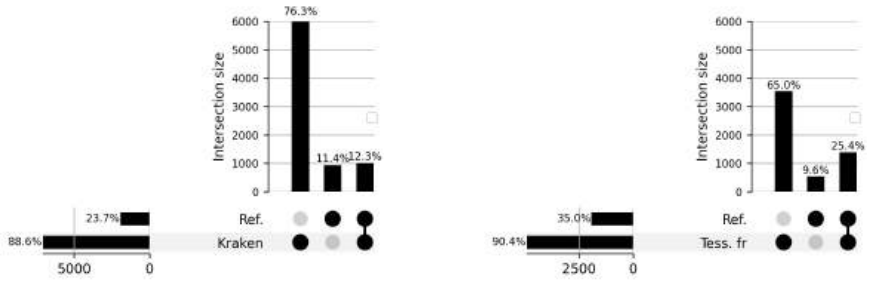
4.3. *Usage des intersections : une évaluation trop stricte ?*

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_lg`. Nous nous servons de ces derniers comme vérité terrain[32]. 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_lg`, 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_lg` 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.



(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 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_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	Réf.	*[...] a sua entrada para o **colegio militar***	()
	Kraken	*[...] a s_a entrada para o colcgio milita*	colcgio 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_lg` 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_lg` 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-fr[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_lg` sur cet échantillon, dont les résultats sont présentés dans le tableau 8.

	GLOBALE : tous types d'EN			LIEUX		
Souple	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_lg` 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'appuyent 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 JamSpell[35] (Jspll) pour la correction automatique des transcriptions de la ROC. Jspll 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 Jspll 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.	✓	✓	✓	✓
Jspll-pretr.	✓	✓	✓	×
Jspll-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'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, small-ELTec-fr*

Version	Contexte	spaCy_lg
Réf.	*[...] before you went to India.*	India
Kraken	*[...] before you went to Iudia.*	Iudia
Kraken Jspll-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 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 dans l'EN de réf. « India ». Par contre, le modèle Jspll 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 Lonlon*	()	London
	Jspll-en	*[…] when they were in London*	London	
	Kraken-ELTeC-en	*[…] when they were in London*	London	
MOMC	Kraken-en	*[…] flowery lanes peeuliar to Dconshire ;*	Dconshire	Devonshire
	Jspll-en	*[…] flowery lanes peculiar to Conspire ;*	()	
	Kraken-ELTeC-en	*[…] flowers lanes peculiar to Dconshire ;*	Dconshire	
MOI	Kraken-fr	*cure de Mlorlincourtl*	Mlorlincourtl	Morlincourt
	Jspll-fr	*cure de Mlorlincourtl*	Mlorlincourtl	
	Kraken-ELTeC-fr	*cure de Mlorlincourtl*	Mlorlincourtl	
BOIC	Kraken-fr	*en retournant de Morlincourt*	Morlincourt	Morlincourt
	Jspll-fr	*en retournant de **Martincourt***	**Martincourt**	
	Kraken-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`. 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	Jspll préentraîné		Jspll-ELTeC		Brut	Jspll préentraîné		Jspll-ELTeC	
small-ELTec-en	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	
small-ELTec-fr	7 062	5 320	-25 %	**6 758**	**-5 %**	4 888	4637	-5 %	**4 921**	**+0,7 %**
CER médian	0,098	0,102		0,104		0,050	0,054		0,055	
small-ELTec-pt	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 (Jspll préentrainé/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 Jspll préentraîné

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 %) ;

– 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 Jspll-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) ;

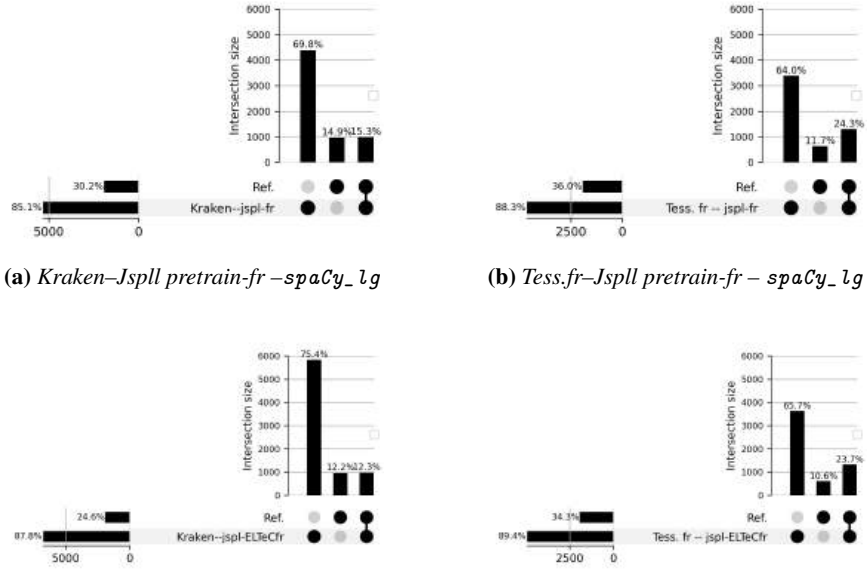
– on remarque que la baisse du nombre d'EN types, et donc des hapax est plus significative sur les textes de version Kraken que Tess., ce qui laisse penser qu'il y aurait un effet de seuil concernant la qualité des versions de ROC au-delà de laquelle la correction automatique serait moins efficace. Autrement 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 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'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 Jspll (2a-2b), et le modèle entraîné avec le corpus ELTeC français (2c-2d). La figure 2b montre que la configuration Tess.fr–Jspll *pretrain*-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 Kraken préentraîné de Jspll 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 Jspll, 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. « Morlincourtl » qui reste « Morlincourtl »), n'améliorent pas les résultats obtenus par calcul des intersections. Pour les corpus *small*-ELTeC-{fr, en, pt} et *small*-TGB-fr, il



(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 préentraîné et modèle ELTeCfr,* `spaCy_lg` *sur le corpus small-ELTec-fr*

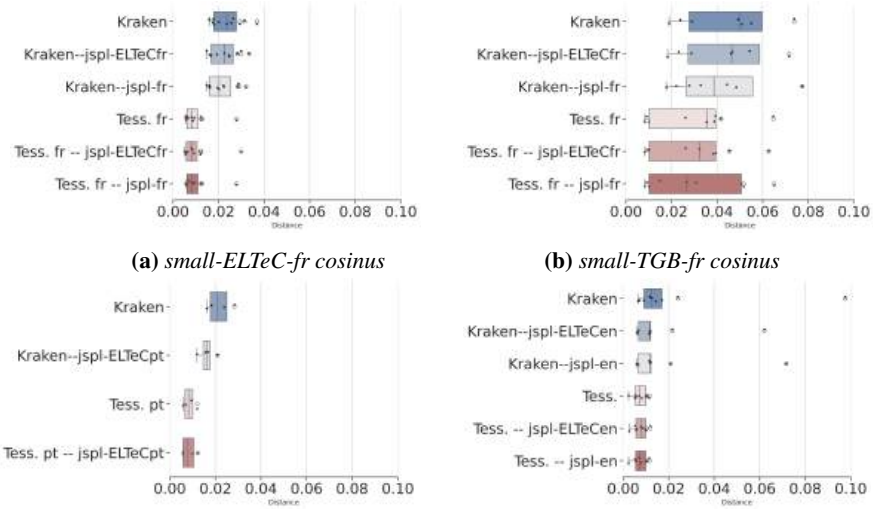
semble que la correction automatique avec le modèle entraîné sur une partie de chaque corpus ELTeC fasse perdre 5 % des EN dans l'intersection avec Kraken et 10 % avec Tesseract, alors que concernant les modèles préentraînés, on perd 3 % avec Kraken et 9 % avec Tesseract. Cette expérience est l'occasion 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 caractères[36].



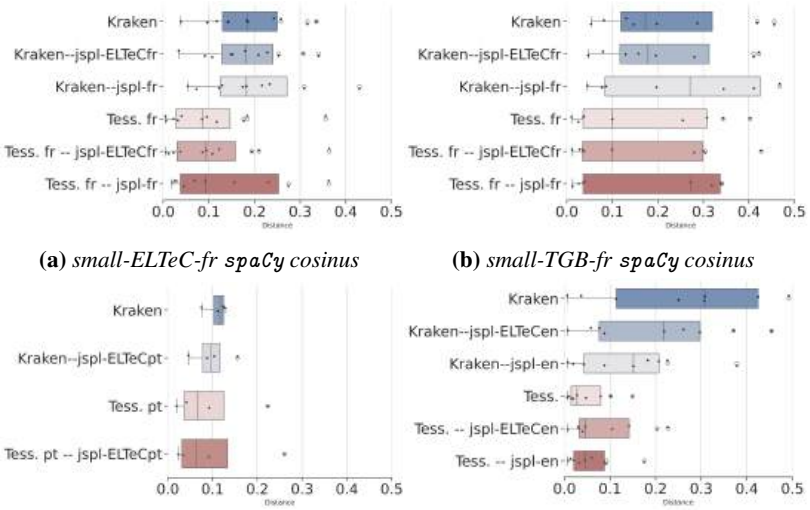
(a) *small-ELTeC-fr cosinus* **(b)** *small-TGB-fr cosinus*

(c) *small-ELTeC-pt cosinus* **(d)** *small-ELTeC-en cosinus*

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 Jspll-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`.



(a) *small-ELTeC-fr* `spaCy` *cosinus* **(b)** *small-TGB-fr* `spaCy` *cosinus*

(c) *small-ELTeC-pt* `spaCy` *cosinus* **(d)** *small-ELTeC-en* `spaCy` *cosinus*

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 Jspll-fr provoque beaucoup de variations, des surcorrections, autant sur Kraken que sur Tess. car les boîtes sont plus larges, et le modèle Jspll-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 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'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 évoquions dans la partie 5.3.1.

	# Entités		Évaluation par NERVAL			
Version	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 + Jspll-fr	1 027	744	471	0,46	0,63	**0,53** ⇓
Tess.fr + Jspll-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*

		# 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, 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 signifiants 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 appuierons 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.

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, 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/teklia/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., 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.

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.