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, n° 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, nº 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, nº 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* 24e *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`