L'architecture d'un modèle hybride pour la normalisation de SMS

Eleni Kogkitsidou Georges Antoniadis
Laboratoire LiDiLEM, Université Grenoble Alpes, Bâtiment Stendhal BP25 38040 Grenoble Cedex 9, France
{eleni.kogkitsidou,georges.antoniadis}@univ-grenoble-alpes.fr

RÉSUMÉ

La communication par SMS (*Short Message Service*), aussi bien que tout autre type de communication virtuelle sous forme de textes courts (mails, microblogs, tweets, etc.), présente certaines particularités spécifiques (syntaxe irrégulière, fusionnement et phonétisation de mots, formes abrégées, etc.). A cause de ces caractéristiques, l'application d'outils en Traitement Automatique du Langage (TAL) rend difficile l'exploitation d'informations utiles contenues dans des messages bruités. Nous proposons un modèle de normalisation en deux étapes fondé sur une approche symbolique et statistique. La première partie vise à produire une représentation intermédiaire du message SMS par l'application des grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base de règles pour convertir la représentation intermédiaire vers une forme standard.

ABSTRACT

A hybrid model architecture for SMS normalization

SMS (Short Message Service) communication, as well as other types of computer mediated communication (e-mails, tweets, blogs, chats, etc.) presents several specific irregularities such as phonetic spelling, irregular syntax, abbreviations, etc. The application of standard Natural Language Processing (NLP) tools seems to be unable to process this kind of noisy data. In this paper, we describe a two-steps hybrid model for SMS normalization based on symbolic and statistic approaches. We first aim to produce an intermediate representation of SMS text via the application of local grammars, whereas second step uses a machine translation (MT) system based on rules in order to convert this representation into standard French.

MOTS-CLÉS : normalisation de SMS, traduction automatique, grammaires locales, traitement automatique du langage.

KEYWORDS: SMS normalization, machine translation, local grammars, natural language processing.

1 Introduction

Les grands développements techniques des années 90 ont été marqués par le rapprochement du domaine de l'informatique à celui des télécommunications. Ce dernier a vu naitre une nouvelle forme de communication, les SMS (*Short Message Service*), qui a été notamment décrite grâce à l'apparition de termes tels que : communication médiatisée ou médiée par ordinateur (Panckhurst, 1997; Marcoccia, 2000), communication écrite médiée par ordinateur (Cougnon & François, 2011),

cybercommunication, netspeak (Crystal, 2001) etc. L'intérêt d'étudier la communication par SMS réside surtout dans les particularités que nous observons dans ce langage.

La plupart des messages courts présentent des différences significatives en comparaison avec les messages du langage standard puisqu'ils doivent contenir au maximum 160 caractères et leurs auteurs utilisent diverses formes pour abréger les mots dans l'objectif de gagner du temps tout en réduisant l'effort fourni. Comme Barasa & Mous (2009) le mentionnent, le langage utilisé dans la communication par SMS est particulièrement caractérisé par la création d'une nouvelle orthographe et d'une richesse créative qui échappe aux conventions. Les approches fondées sur l'étiquetage morphosyntaxique de textes standards atteignent de hauts niveaux de précision dans des tâches liées au traitement du langage naturel. Cependant, les résultats sont significativement mitigés lorsque l'étiquetage est appliqué sur des textes courts contenant du bruit (Gadde *et al.*, 2011). Une des contraintes que nous devons affronter avec les systèmes de Traitement Automatique du Langage (TAL) est la graphie particulière des mots SMS (fusionnement et phonétisation de mots, formes abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc.). L'étiquetage morphosyntaxique constitue une étape fondamentale afin de pouvoir traiter davantage de données textuelles, comme dans la reconnaissance d'entités nommées, la traduction automatique, les systèmes de questions-réponses, l'extraction d'information etc. (Yvon, 2010). La normalisation d'un texte devient une étape de prétraitement indispensable, de même, Sproat *et al.* (2001) indiquent notamment l'importance d'appliquer ce processus de normalisation avant tout autre traitement basique issu du TAL. La normalisation de SMS consiste à réécrire un message SMS en utilisant des formes lexicales plus conventionnelles afin de rendre ce message plus lisible aussi bien pour l'homme que pour la machine (Jose & Raj, 2014).

2 Travaux Antérieurs

Afin de surmonter le problème des particularités lexicales des SMS, plusieurs approches pour la normalisation lexicale des SMS ont vu le jour au cours des dernières années. Étant donné un texte $T = S_1, S_2, ..., S_n$, la tâche de normalisation lexicale est de trouver pour chaque token S_i hors-vocabulaire (*OOV : Out-Of-Vocabulary*) un token correspondant en forme standard (*IV : in-vocabulary*), par exemple : $cmb \rightarrow combien$. Pour mieux visualiser le problème de normalisation, nous empruntons la formulation initialement introduite par Shannon (1948) afin de décrire le processus de communication à travers un canal de communication bruité : Supposons que le texte malformé est T et que sa forme standard est S, l'approche de normalisation vise à trouver $\operatorname{argmax} P(S|T)$ en calculant $\operatorname{argmax} P(T|S)P(S)$, dont $P(S)$ est généralement un modèle de langage et $P(T|S)$ est un modèle d'erreurs (Han & Baldwin, 2011).

Les méthodes à base de canaux bruités (*noisy channel model*) étaient les premières à être appliquées dans la Communication Médiée par Ordinateur (CMO). Toutanova & Moore (2002) utilisent les similarités de prononciation entre les mots dans le but d'améliorer la correction orthographique. Choudhury *et al.* (2007) proposent un modèle au niveau du mot pour chaque mot en langue standard et produisent, à partir d'un modèle de Markov caché, toutes les variations possibles en langage SMS. De leur côté, Han & Baldwin (2011) exploitent une méthode à base de cascades qui détecte les mots malformés et génère des candidats à partir de similitudes morphophonémiques. Dans le cadre de la normalisation de tweets en anglais, Kaufmann (2010) utilise un modèle qui fait passer les messages par une phase de prétraitement afin d'enlever le bruit et par la suite alimente un modèle de traduction

automatique pour le convertir en anglais standard. Aw *et al.* (2006) considèrent la normalisation comme un problème de traduction et proposent un outil de traduction automatique au niveau des phrases. La technique de la reconnaissance de la parole a été appliquée par Kobus *et al.* (2008), dans une première étape, pour décoder la représentation phonétique d'un mot en forme écrite. Avec une autre approche, Beaufort *et al.* (2010) utilisent une méthode fondée sur les automates finis tout en combinant la traduction automatique et les canaux bruités. Un des travaux les plus récents est le modèle qui a été proposé par Jose & Raj (2014) basé sur trois types de canaux (abréviations, graphèmes, phonétiques). Le modèle consiste à passer une phrase par quatre bases de données différentes pour identifier et corriger les mots inconnus en donnant à l'utilisateur la possibilité de choisir parmi les meilleurs candidats.

Nous distinguons deux approches pour la transcription automatique de SMS en français (Kobus *et al.*, 2008; Beaufort *et al.*, 2010). Comme Beaufort *et al.* (2010) le mentionnent, les analyses de normalisation produites par leur système démontrent que lors de la normalisation le système n'est pas capable de traiter les erreurs liées au contexte qui concernent le genre (choix féminin ou masculin), le nombre (singulier ou pluriel), la personne (pronom personnel) ou le temps du verbe. Ils soulignent, d'ailleurs, que selon Kobus *et al.* (2008) les modèles basés sur les *n*-grammes ne sont pas capables de traiter ce type d'erreurs. Le modèle hybride que nous proposons permet, à l'aide de règles de transfert couplées avec les dictionnaires morphologiques, de résoudre ce type d'erreurs.

3 Normalisation

Notre démarche est inspirée des méthodes de traduction automatique, à la différence des travaux précédents (Bangalore *et al.*, 2002; Aw *et al.*, 2006; Raghunathan & Krawczyk, 2009; Kaufmann, 2010), qui sont limités lors du traitement des variations lexicales et par conséquent ont besoin de grandes quantités de données d'apprentissage étiquetées. Nous proposons un modèle de normalisation en deux étapes à l'aide des approches symboliques et statistiques : la première partie vise à produire une représentation intermédiaire du message SMS par l'application des grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base de règles pour convertir la représentation intermédiaire vers une forme standard (tableau 1).

SMS brut	Coucouuuuu! Où t es? Kfé? :)) Bizz
Représentation intermédiaire	**Coucou** ! Où t es ? Kfé ? *****EMOTICON*** Biz**
Traduction automatique	Coucou ! Où **tu** es ? **Café** ? ***EMOTICON*** **Bise**

TABLEAU 1 – Exemple de normalisation

3.1 Représentation intermédiaire

L'observation de particularités liées aux SMS nécessite l'utilisation de matériaux authentiques dans le but d'obtenir un point de vue plus objectif. Notre étude a comme point de départ le corpus de SMS, collecté et traité semi-automatiquement, dans le cadre du projet *alpes4science*. La collecte s'est déroulée du 1er octobre 2010 au 31 janvier 2011 dans les Hautes-Alpes et l'Isère. Au total, 22 054 SMS authentiques de 359 personnes ont été recueillis. Sur la base de données nous trouvons les SMS (anonymisés, alignés et transcrits en langue standard), le lexique (mots SMS avec leurs traductions en

langue standard et leurs fréquences) et des informations variées sur les expéditeurs (âge, sexe, niveau d'études, langue maternelle, etc.).

La représentation intermédiaire du message se divise en deux processus et se base sur l'analyse typographique et néologique des SMS menée par Anis *et al.* (2004); Panckhurst (2009); Fairon *et al.* (2006); Veronis & Guimier De Neef (2006). Tout d'abord, nous avons le processus de normalisation structurelle qui prend en charge la normalisation des séparateurs et le traitement des symboles de ponctuation ; pour accomplir cette partie nous faisons appel à l'utilisation d'heuristiques développées à partir de notre corpus d'apprentissage. Une deuxième étape a pour objectif de reconnaître, et par la suite, traiter les unités reconnues à l'aide des grammaires locales (Paumier, 2006). Cette étape est responsable du processus de normalisation pour la résolution des formes non ambiguës, les abréviations, la détection des émoticônes et des mots hors-vocabulaire, ainsi que le découpage en unités lexicales. Pour ceci nous avons conçu de réseaux de transition récursives (RTN : *Recursive Transition Networks*) appliqués en combinaison avec des dictionnaires électroniques du français standard et une base de connaissance regroupant des informations spécifiques aux mots SMS (figure 1). Nous présentons comme exemple deux types des traitements effectués.

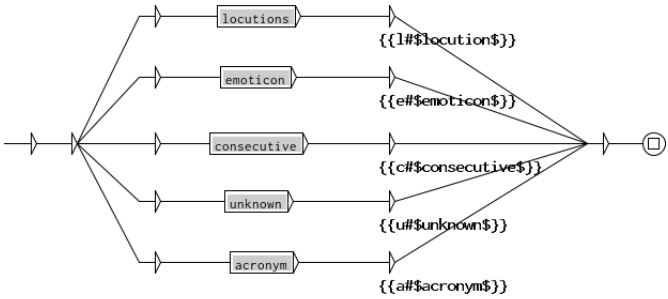

FIGURE 1 – Graphe principale de reconnaissance

3.1.1 Expression graphique des sentiments

L'expression graphique des sentiments (émoticônes) est une chaîne de caractères produite à l'aide d'un clavier qui peut être proposée pour imiter un visage exprimant une émotion particulière (Danesi, 2009). L'utilisation des émoticônes dans les interactions CMO représente une alternative à la forme verbale et constitue une représentation plus au moins imagée de la signification. Ces caractéristiques linguistiques sont codifiées sous forme de pictogrammes ou d'émoticônes pour représenter, dans la plupart des cas, des gestes, des expressions faciales et des éléments prosodiques puisque la CMO ne comporte pas ces caractéristiques (Amaghlobeli, 2012). Afin de pouvoir reconnaître toute forme d'émoticônes contenues dans les SMS, nous avons créé un dictionnaire électronique contenant plus de 300 entrées (par exemple : `o_o,.Emoticon+Meaning=surprise`). Ce dictionnaire a été utilisé pour développer une grammaire locale qui identifie et multiplie la capacité de reconnaissance des émoticônes contenues dans les SMS, par exemple : *:)* *:))* *:) :D* *;) :) :D ;) :)) :D*.

3.1.2 Répétition de caractères

La répétition de caractères constitue un phénomène très courant dans la CMO dans le but de souligner les sentiments exprimés (Brody & Diakopoulos, 2011). De leur côté Kalman & Gergle (2014) en

étudiant la fréquence et l'utilisation de répétitions de lettres dans la communication par courriel, remarquent que les répétitions tendent à imiter un morphème étendu dans la conversation parlée, à donner de l'emphase ou à imiter des sons. La reconnaissance d'unités lexicales hors-vocabulaire contenant de caractères répétées (par exemple : *cooool→cool*), s'effectue à l'aide d'une grammaire locale qui identifie les unités lexicales inconnues et d'un dictionnaire ($\approx$ 50 000 entrées) constitué par les mots les plus fréquents qui ont une ou plusieurs extensions de lettres. Ce dictionnaire a été élaboré à partir d'un corpus de résumés courts en français issu de DBpedia [1].

3.2 Traduction Automatique

La transformation de la représentation intermédiaire vers la forme standard du message s'effectue à l'aide d'un système de traduction libre basé sur des règles de transfert lexical. Apertium est une plateforme libre de traduction automatique, connue initialement pour traiter des paires de langues assez proches. La conception de cette plateforme est fondée sur la traduction mot à mot avec une désambiguïsation lexicale (étiquettes morphosyntaxiques), un traitement lexical robuste et structuré basé sur des règles simples bien formulées (Forcada *et al.*, 2009).

Le système nécessite deux types de ressources linguistiques : a) deux dictionnaires morphologiques monolingues (*smsfra, fra*) et b) un dictionnaire bilingue (*smsfra-fra*). Le dictionnaire *smsfra* est une extension du dictionnaire français (*fra*) fourni par Apertium avec des ajouts supplémentaires, par exemple : *<e lm="impec">impec<par n="impec__adj"/></e>*.

Pour construire le dictionnaire bilingue *smsfra-fra* nous avons utilisé deux corpus constitués chacun de 14 200 SMS [2]. Le premier corpus contient les SMS qui ont été transcrits manuellement, tandis que le deuxième est la représentation intermédiaire de SMS en format brut. Les deux corpus ont été ensuite étiquetés morphosyntaxiquement avec Apertium [3] et alignés à l'aide de GIZA++ (Och & Ney, 2003), un outil combinant des modèles statistiques et heuristiques pour fournir un outil d'alignement de mots. Nous avons ensuite induit une première version du dictionnaire bilingue avec ReTraTos (Caseli *et al.*, 2006), un outil capable construire des dictionnaires bilingues à partir d'un corpus parallèle aligné. Finalement, nous avons effectué une validation manuelle pour corriger d'éventuelles erreurs lors de la phase d'alignement.

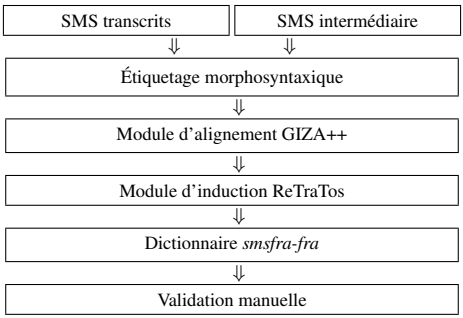

FIGURE 2 – Schéma d'induction du dictionnaire bilingue

1. short_abstracts_fr $\approx$ 50 millions de mots, projet DBpedia http ://fr.dbpedia.org
2. Ceci correspond aux deux tiers du corpus alpes4science.
3. Nous avons utilisé les dictionnaires monolingues *smsfra* et *fra* avec le modèle de probabilité pour l'étiquetage morphosyntaxique du français fourni par Apertium.

Le schéma de la figure 2 synthétise les étapes de traitement pour l'induction du dictionnaire bilingue *smsfra-fra*. Exemple d'entrée dans le dictionnaire *smsfra-fra* : *<e> <l>impec<s n="adj"/></l><r>impeccable<s n="adj"/></r> </e>*.

Afin de transformer la représentation intermédiaire d'un SMS vers sa forme standard, la chaîne de traitement d'Apertium réalise, dans un premier temps, une analyse morphologique en utilisant le dictionnaire SMS (*smsfra*) et un étiquetage morphosyntaxique grâce au modèle de probabilité (français) fourni par le système. Par la suite, une phase de transfert s'effectue à l'aide du dictionnaire bilingue (*smsfra-fra*) pour arriver ainsi à la génération morphologique en utilisant le dictionnaire français (*fra*). Finalement, l'ajout des corrections est produit dans la phase de la post-génération (pas de ressources requises) [4].

4 Evaluation

L'évaluation de l'approche a été effectuée sur les 7 000 SMS bruts (*corpus de test*) tirés du corpus alpes4science qui n'ont pas été utilisés pour la construction du dictionnaire bilingue. Le *gold standard* est constitué par la transcription manuelle des SMS du corpus de test. Les métriques que nous avons appliquées sont le BLEU et le NIST score [5], deux mesures couramment utilisées pour l'évaluation des transcriptions en CMO (Aw *et al.*, 2006; Han & Baldwin, 2011; Kaufmann, 2010; Beaufort *et al.*, 2010; Sidarenka *et al.*, 2013). Le BLEU score (Papineni *et al.*, 2002) est un outil destiné à évaluer la précision du processus de traduction d'une langue à l'autre. Cette évaluation nécessite préalablement un gold standard, autrement dit, une représentation de la traduction produite manuellement. Le gold standard sera par la suite comparé avec la traduction produite automatiquement afin d'établir un score entre *0* et *1*, où *1* signifie l'exactitude absolue entre les deux traductions et *0* que les deux traductions ne présentent aucune similarité (Kaufmann, 2010). Nous utilisons également le NIST score, une autre métrique alternative dérivée du BLEU score qui diffère sur le degré informatif qu'un *n*-gramme particulier peut avoir. Si un *n*-gramme rare a été correctement trouvé, il lui sera attribué plus de poids (Doddington, 2002). Les résultats de l'évaluation figurant sur la tableau 2 nous permettent

Technique	**BLEU score**	**NIST score**
Approche de référence	0.60	11.78
Représentation intermédiaire (*RI*)	0.62	11.92
Traduction automatique (*TA*)	0.72	13.45
Modèle hybride (*RI+TA*)	**0.76**	**14.00**
Gold standard	1	16.38

TABLEAU 2 – Résultats d'évaluation

d'observer l'écart de 0.16 (26.67%) en BLEU score entre l'approche de référence (*baseline*) [6] et la transcription obtenue par l'approche hybride. Par ailleurs, les résultats pour le NIST score nous confirment, tout comme l'écart de 2 points entre le gold standard et l'approche hybride, la bonne qualité de la production, en comparaison avec d'autres systèmes qui fournissent des informations complètes sur ce type de métriques (Kaufmann, 2010; Beaufort *et al.*, 2010).

4. Pour plus de détails concernant la chaîne de traitement d'Apertium, voir Forcada *et al.* (2008).
5. Les résultats d'évaluation ont été produits à l'aide du logiciel MTEval toolkit : https ://github.com/odashi/mteval
6. Évaluation effectuée entre les SMS bruts (sans aucune normalisation) et leurs transcriptions.

5 Conclusion

Dans cet article, nous avons proposé l'architecture d'un modèle hybride pour la normalisation de textes courts. L'approche est fondée sur l'analyse typologique et néographique à l'aide des grammaires locales, couplée avec un système de traduction automatique. Nos résultats nous montrent que l'approche proposée améliore la qualité, la lisibilité et l'opérationnalité des SMS. Ces résultats sont encourageants et nous incitent à explorer de nouveaux modes d'hybridation (traitement des emprunts, phonétisation, calcul de distances d'édition, mémoires de traduction). Nous envisageons également d'effectuer l'évaluation de notre système sur le corpus de SMS *88milSMS*[7] récolté à Montpellier, en 2011. Nous estimons que cette approche pourrait être appliquée à d'autres types de textes qui contiennent du bruit : chats, e-mails, forums de discussion, réseaux sociaux.

Remerciements

Ce travail a été soutenu par la Région Rhône Alpes.

Références

AMAGHLOBELI N. (2012). Linguistic Features of Typographic Emoticons in SMS Discourse. *Theory and Practice in Language Studies*, **2**(2), 348–354.

ANIS J., DE FORNEL M. & FRAENKEL B. (2004). La communication électronique : Approches linguistiques et anthropologiques. In *Colloque international, EHESS*.

AW A., ZHANG M., XIAO J. & SU J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, number July, p. 33–40, Sydney : 2006 Association for Computational Linguistics.

BANGALORE S., MURDOCK V. & RICCARDI G. (2002). Bootstrapping Bilingual Data Using Consensus Translation For A Multilingual Instant Messaging System. *Coling-2002*.

BARASA S. & MOUS M. (2009). The oral and written interface on SMS : Technology Mediated Communication in Kenya. In *Low-educated adult second language and literacy acquisition*, p. 234–242. Kenya.

BEAUFORT R., ROEKHAUT S., COUGNON L.-A. & FAIRON C. (2010). A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 770–779 : Association for Computational Linguistics.

BRODY S. & DIAKOPOULOS N. (2011). Cooooooooooooooollllllllllllll ! ! ! ! ! ! ! ! ! ! ! ! ! ! Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings Empirical Methods in Natural Language Processing (EMNLP)*, number 2010, p. 562–570, Edinburgh, UK.

CASELI H. M., NUNES M. D. G. V. & FORCADA M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora : Application to shallow-transfer machine translation. *Machine Translation*, **20**(4), 227–245.

7. http ://88milsms.huma-num.fr

CHOUDHURY M., SARAF R., JAIN V., MUKHERJEE A., SARKAR S. & BASU A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, **10**(3-4), 157–174.

COUGNON L.-A. & FRANÇOIS T. (2011). Etudier l'écrit SMS - Un objectif du projet sms4science. In *Linguistik online 48*.

CRYSTAL D. (2001). *Language and the Internet*. Cambridge University Press.

DANESI M. (2009). *Dictionary of media and communications*. New york & edition.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, p. 138–145 : Morgan Kaufmann Publishers Inc.

FAIRON C., KLEIN J.-R. & PAUMIER S. (2006). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête "Faites don de vos SMS à la science"*. Cahiers du Cental (Louvain-la-Neuve), ISSN 1783-2845. Louvain-la-Neuve, Belgique : UCL Presses universitaires de Louvain, DL 2006.

FORCADA M., BONEV B., ROJAS S., ORTIZ J., SÁNCHEZ G., MARTÍNEZ F., ARMENTANO-OLLER C., MONTAVA M. & TYERS F. (2008). Documentation of the open-source shallow-transfer machine translation platform Apertium.

FORCADA M. L., TYERS F. M. & RAMÍREZ-SÁNCHEZ G. (2009). The Apertium machine translation platform : Five years on. *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, (November), 3–10.

GADDE P., SUBRAMANIAM L. & TANVEER A. F. (2011). Adapting a WSJ trained Part-of-Speech tagger to Noisy Text : Preliminary Results. In *2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data (MOCR_AND '11)*, New York, USA : ACM.

HAN B. & BALDWIN T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a # twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, p. 368–378, Portland, Oregon : Association for Computational Linguistics.

JOSE G. & RAJ N. S. (2014). Lexical normalization model for noisy sms text. In *Computational Systems and Communications (ICCSC), 2014 First International Conference on*, p. 57–62.

KALMAN Y. M. & GERGLE D. (2014). Letter repetitions in computer-mediated communication : A unique link between spoken and online language. *Computers in Human Behavior*, **34**, 187–193.

KAUFMANN M. (2010). Syntactic Normalization of Twitter Messages. p. 1–7.

KOBUS C., YVON F. & DAMNATI G. (2008). Normalizing SMS : are two metaphors better than one ? *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, (August), 441–448.

MARCOCCIA M. (2000). La communication écrite médiatisée par ordinateur : faire du face à face avec de l ' écrit. *Journée d'étude de l'ATALA « Le traitement autom atique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, p. 1–4.

OCH F. J. & NEY H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1), 19–51.

PANCKHURST R. (1997). La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? *Terminologies nouvelles*, (17), 56–58.

PANCKHURST R. (2009). « Short Message Service (SMS) : typologie et problématiques futures », in Arnavielle T. (coord.),. p. 33–52.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

PAUMIER S. (2006). Unitex 1.2 - Manuel d'utilisation.

RAGHUNATHAN K. & KRAWCZYK S. (2009). CS224N : Investigating SMS Text Normalization using Statistical Machine Translation. *Www-Nlp.Stanford.Edu*.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(July 1928), 379–423.

SIDARENKA U., SCHEFFLER T. & STEDE M. (2013). Rule-based normalization of german twitter messages. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*.

SPROAT R., BLACK A. W., CHEN S., KUMAR S., OSTENDORF M. & RICHARDS C. (2001). Normalization of non-standard words. *Computer Speech & Language*, **15**(3), 287–333.

TOUTANOVA K. & MOORE R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. In *Proceedings of the 40th Annual Meeting on Associ- ation for Computational Linguistics*, number July, p. 144–151, Philadelphia, USA.

VERONIS J. & GUIMIER DE NEEF E. (2006). *Le traitement des nouvelles formes de communication écrite*. Paris : Hermès Science.

YVON F. (2010). Rewriting the orthography of SMS messages. *Natural Language Engineering*, **16**(02), 133.