TiLT correcteur de SMS : évaluation et bilan qualitatif

Émilie GUIMIER DE NEEF, Arnaud DEBEURME, Jungyeul PARK
FTR&D/TECH/EASY – France Telecom R&D
2, avenue Pierre Marzin, 22300 Lannion Cedex, France
{emilie.guimierdeneef,arnaud.debeurme,jungyeul.park}
@orange-ftgroup.com

Résumé. Nous présentons le logiciel TiLT pour la correction des SMS et évaluons ses performances sur le corpus de SMS du DELIC. L'évaluation utilise la distance de Jaccard et la mesure BLEU. La présentation des résultats est suivie d'une analyse qualitative du système et de ses limites.

Abstract. This paper presents TiLT system which allows us to correct spelling errors in SMS messages to standard French. We perform Jaccard and Bleu metrics for its evaluation using the DELIC SMS corpus as a reference. We discuss qualitative analyses of system and its limits.

Mots-clés : SMS, SMS corpus, correction orthographique, TiLT, evaluation.

Keywords: SMS, SMS corpus, spelling correction, TiLT, evaluation.

1 Introduction

Les nouvelles formes de communication écrite (blogs, SMS, chats etc.) se caractérisent par de nombreux écarts vis-à-vis des conventions orthographiques standards. Ces écarts recensés par Anis (2002), confirmés par des études de corpus réels Bove (2005) et très brièvement rappelés ci-dessous, offrent un nouveau défi aux outils de correction automatique. En effet, comme l'ont montré Véronis et Guimier De Neef (2006), un simple recensement de couples graphie SMS / graphie standard ne suffit pas à répondre à la productivité et à la combinatoire des différents procédés d'écriture.

- Graphies phonétisantes et rébus : *g ht du kfé a+ (j'ai acheté du café à plus)*
- Abréviations diverses : *slt k f tu (salut que fais-tu ?)*
- Étirements graphiques : *ssuuuppperr ! hhhhhuuuuummmm !*
- Agglutinations : *g ésayé 2tapelé pl1 2foi (j'ai essayé de t'appeler plein de fois)*

Différentes motivations peuvent justifier la correction, ou plutôt la normalisation, de l'écriture SMS : l'extraction d'information, l'indexation, l'analyse de blogs, de wikis etc. La vocalisation des SMS destinés aux téléphones fixes a été l'occasion pour FTR&D d'adapter son logiciel TiLT à ce contexte. L'architecture globale du service est schématisée Figure 1.

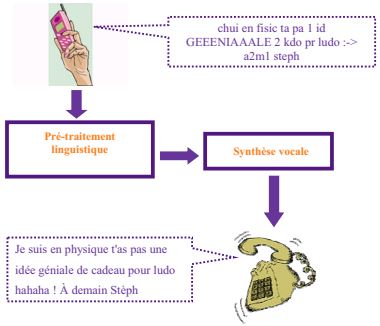

Figure 1 : correction des SMS avant vocalisation

Dans cet article, nous présentons une évaluation du correcteur de SMS TiLT en mesurant ses performances sur le corpus de SMS du laboratoire DELIC. Dans une première section, nous décrivons le logiciel et ses adaptations au contexte SMS. Nous présentons le corpus du DELIC en seconde section. Les mesures utilisées et les résultats de l'évaluation sont fournis en troisième section ainsi qu'une analyse qualitative des résultats obtenus. Nous terminons l'article par différentes perspectives de recherche.

2 TiLT correcteur de SMS

Le logiciel TiLT est une solution d'analyse linguistique modulaire permettant différents types de traitements automatiques comme la correction orthographique, le pré-traitement linguistique de documents avant indexation, l'extraction d'informations, la traduction automatique etc. Nous présentons ci-dessous l'architecture et les particularités de la solution TiLT pour la correction orthographique dont l'application SMS est une instance particulière.

Le logiciel fait intervenir séquentiellement trois briques : (i) un module de segmentation, (ii) un module d'analyse lexicale avec ou sans méthodes correctives, (iii) un module d'analyse en chunking. Des données linguistiques spécifiques au contexte SMS ont été développées à partir d'observations linguistiques et de tests sur corpus. L'ensemble des données symboliques utilisées par le logiciel est développé à partir d'une expertise humaine.

2.1 Segmentation

La segmentation permet le découpage et le typage des différents segments du message en entrée. Le typage des segments permet de différencier ensuite le traitement qu'il convient de

leur associer : seuls les segments de type MOT sont envoyés à l'analyse lexicale. La description des segments se fait au moyen d'expressions régulières compilées par FLEX[1].

Pour le traitement des SMS, une modification importante des données de segmentation est rendue nécessaire principalement pour l'identification des smileys et pour l'inclusion massive des chiffres et des symboles dans les mots.

```
Phrase analysée : tu vi1 2m1 a+ ;-)

0  1 : "tu"  : MOT
1  2 : "vi1" : MOT_AVECCHIFFRE
2  3 : "2m1" : MOT_AVECCHIFFRE
3  4 : "a+"  : MOT_AVECSYMB
4  5 : ";-)" : SMILEY
```

Figure 2 : segmentation TiLT

2.2 Lexique

Le lexique du français utilisé par TiLT comporte environ 100 000 entrées incluant une base de mots-composés. Pour le traitement des SMS, ce lexique a été enrichi d'une base d'un millier d'abréviations recueillies sur le web ou compilées d'après relevés sur corpus. Parmi ces abréviations, on recense des sigles (*atd = à ta disposition*, *tvb = tout va bien*…), des squelettes consonantiques (*slt a vs ts = salut à vous tous*), des troncations (*adr = adresse*) etc. Le lexique SMS inclut également une base de prénoms de 3 000 entrées environ.

Un apprentissage sur corpus a permis de délimiter la liste des mots les plus fréquemment utilisés dans un contexte SMS afin de favoriser ces formes par rapport aux autres mots de la base lexicale.

2.3 Les méthodes correctives

Au cours de l'analyse lexicale, le lexique est consulté et renvoie l'ensemble des informations disponibles : lemme, orthographe standard, catégorie grammaticale, traits morpho-syntaxiques. Cette consultation se fait selon différents modes dont des modes correctifs, certains ayant été spécialement adaptés au contexte SMS.

1. **Correction phonétique** : le module de correction phonétique, basé sur un transducteur appris par alignement des formes phonétiques et graphiques des mots d'une langue, a été enrichi de règles permettant la phonétisation des symboles, chiffres et lettres utilisés pour leur valeur phonétique en écriture SMS.

2. **Correction par découpage morpho-syntaxique** : des observations sur corpus (Bove 2005) ont montré que l'agglutination de mots non généralisée n'intervient pas au hasard mais concerne de façon privilégiée les séquences avec clitiques (*jtrappelle*, *gspère qtu va bien*…), avec préposition (*g ésayé 2tapelé pl1 2foi*), les séquences déterminant/nom (*c le foot ki te mé ds 7éta?*) ou les formes lexicales complexes

[1] http://fr.wikipedia.org/wiki/Flex_%28GNU%29.

(*Keske tu deviens?*) etc. Le module de correction par découpage a été associé à des données spécifiques pour permettre l'expansion de ces formes compactées.

3. **Tolérance à la répétition** : l'une des particularités de l'écriture SMS est la présence de marques expressives dans l'écriture. Ces marques incluent les smileys, des jeux sur les signes de ponctuation (*on est sur la plage!!!!!!!!!!*), l'utilisation de la casse (*Encore un grand MERCI à tous les deux*) mais aussi l'étirement graphique (*c la foliiiiiiiie !!*). Une méthode corrective dite de tolérance à la répétition a été spécialement développée pour restituer les orthographes non étirées.

Ces méthodes calculent dynamiquement les corrections possibles dans les SMS, ce qui répond à la créativité orthographique des utilisateurs de SMS.

2.4 La grammaire de chunking

L'analyse en chunking pratiquée par TiLT utilise une grammaire hors contexte de 2 000 règles environ. Son rôle est de permettre de choisir la correction adaptée pour un mot étant donné son contexte syntaxique local.

En SMS, deux particularités augmentent la difficulté du chunking :

1. l'ambiguïté en général et celle des mots outils en particulier, comme dans le paradigme suivant où le caractère *c* se normalise de 4 façons différentes :

Voilà c fini ca c bin passé	=>	*voilà c'est fini ça s'est bien passé*
Je c pa ki c	=>	*je sais pas qui c'est*
Jespere k vou descendé c soir	=>	*j'espère que vous descendez ce soir*

2. la ponctuation souvent absente qui cesse de jouer son rôle de césure :

alor t soulagé moi la jaten 1h ca me soule j menui a mourir biz
G un empechement previen sophie gros poutou à demain

Pour contrer ces difficultés, la grammaire a été enrichie de règles prenant en compte des structures particulièrement fréquentes en contexte SMS (interrogatives, formes présentatives introduites par *c'est*..., constructions modales etc.). En particulier, une grammaire locale des formules de politesse, s'appuyant sur des déclencheurs tels que *salut*, *bisous* etc. s'est avérée indispensable pour aider à la détection des noms propres (*HELLO RÉJANE, A BIENTO JOHAN, coucou Nénette, a+ Reno* etc.)

3 Le corpus de SMS du DELIC

3.1 Présentation du corpus

Dans le cas d'un contrat collaboratif avec France Télécom R&D, le laboratoire DELIC de l'université d'Aix en Provence a collecté, avec le concours de ses étudiants, un corpus d'environ 9 700 messages SMS correspondant à un peu plus de 132 000 mots. Ces messages ont été corrigés semi-automatiquement puis révisés manuellement pour constituer une base alignée de SMS et de transcriptions. Le Tableau 1 fournit un extrait de ce corpus.

tu pe tokup du cha 2m1?	Tu peux t'occuper du chat demain ?
Je sui dvt ché toi	Je sui devant chez toi.
Jarriv!!tnev pa!!	J'arrive ! [ne] t'énerve pas !
Je conte sur toi 2m1 a la danse!	Je compte sur toi demain à la danse
Dsl pour ier..enormes bisous	Désolé(e) pour hier… énormes bisous.
pti coucou d vacs!! C terrible c bo !! je vs racontré	Petit coucou des vacances ! C'est terrible, c'est beau ! Je vous raconterai !

Tableau 1 : Extrait du corpus aligné SMS / français standard du DELIC

Une description fine des conventions de correction utilisées dans la transcription du corpus peut être trouvée dans Hocq (2006). Globalement, les orthographes phonétiques, erronées, abrégées sont corrigées. Il en va de même des marques de ponctuation qui sont ramenées à l'usage standard. Des alternances d'accord sont proposées en cas d'ambiguïté (*dsl = désolé ou désolée*). Certains mots manquants, *ne* négatif, pronom sujet principalement, sont notés entre crochets. Les sigles et abréviations sont étendus (*mdr = mort de rire, dvt = devant* etc.). Des données personnelles ont été rendues anonymes. Les numéros de téléphone ont été remplacés par 01 02 03 04 05. La balise <NOM> se substitue aux noms de personne identifiables : Michel Dupont est remplacé par <NOM>. Par contre, les prénoms isolés ont été conservés.

Ce corpus peut être caractérisé par quelques chiffres. La taille moyenne des messages SMS est de 14,5 mots. Le nombre moyen de caractères par message est de 66,7 caractères[2]. Le rapport entre le nombre de caractères utilisés dans la correction et le nombre de caractères présents dans le SMS source, nous donne un taux de compression de l'écriture SMS de l'ordre de 20%.

3.2 Post-traitements pour l'évaluation

La correction fournie par TiLT ne suivant pas toujours les mêmes normes que la correction manuelle proposée dans le corpus du DELIC, une phase de normalisation des sorties de TiLT et d'enrichissement du corpus du DELIC a été effectuée. Les principales divergences portent sur la normalisation des heures (10h30 vs 10 h 30), des nombres notés en lettres ou en chiffres, des unités de mesure (km vs kilomètre), de certaines abréviations conservées ou étendues dans les corrections automatiques ou dans le corpus du DELIC. Certains choix de restitution TiLT en rapport avec l'application de vocalisation ont été revus : la restitution des smileys sous forme de balise par exemple (le smiley ;-) devient <SMILEY>).

Après post-traitement du corpus du DELIC, à chaque SMS correspond en moyenne 1,2 transcription standardisée. C'est cette version post-traitée du corpus qui sert de référence pour notre évaluation. Le Tableau 2 montre un extrait du corpus après les post-traitements.

[2] Le nombre maximum de caractères autorisés pour la saisie d'un SMS sur téléphone mobile est de 160 caractères. Certains téléphones autorisent la saisie de messages plus longs qui sont alors découpés avant d'être envoyés au destinataire.

tu t es planté en math t a eu 1 sal note	Tu t'es planté en math, t'as eu une sale note. Tu t'es planté en math, tu as eu une sale note.
tu pe venir me prendre a aix ver 2h30? merci ta couzine bizz	Tu peux venir me prendre à Aix vers 2 h 30 ? Merci. Ta cousine. Bise. Tu peux venir me prendre à Aix vers 2h30 ? Merci. Ta cousine. Bise. Tu peux venir me prendre à Aix vers 2 h 30 ? Merci. Ta cousine. Bisou. Tu peux venir me prendre à Aix vers 2h30 ? Merci. Ta cousine. Bisou.
TU PE VENIR	Tu peux venir.

Tableau 2 : Extrait du corpus aligné SMS / français standard après post-traitements

4 Évaluation

Les objectifs de cette évaluation sont doubles. Il s'agit d'une part de trouver des indicateurs permettant de quantifier objectivement les performances pour pouvoir en suivre les évolutions dans le temps. Il s'agit également de repérer les phénomènes de l'écriture SMS résistants pour mieux cerner les limites de notre approche et prévoir des extensions. En conséquence, la partie évaluation objective est suivie d'une analyse qualitative des résultats obtenus.

4.1 Évaluation objective

Parmi les mesures fréquemment utilisées pour mesurer les performances des traducteurs statistiques ou des systèmes de reconnaissance vocale, nous en avons retenu deux pour l'évaluation de notre correcteur de SMS : la mesure BLEU (Papineni et al. 2002) et le coefficient de Jaccard. La prise en compte ou pas de l'ordre des mots distingue les deux types de mesure. Le coefficient de Jaccard[3] considère la phrase comme un sac de mots, tandis que la mesure BLEU[4] prend en compte les n-grams et pénalise les corrections qui divergent quant à l'ordre des mots.

---

[3] Coefficient de Jaccard = $\frac{\text{nb mots dans l'intersection}}{\text{nb mots dans l'union}}$

nb mots dans l'intersection = nombre de mots communs entre la solution et la référence la plus proche.

nb mots dans l'union = nombre de mots de la solution + nombre de mots de la référence – nombre de mots communs. S'il y a un seul mot en commun, cette mesure n'est pas nulle.

[4] $BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{\log(\text{nb occurrences n - gram / nb n - grams})}{N}\right)$

nb occurrences n-gram = nombre de n-gram commun avec au moins une référence.

nb n-gram = nombre de n-gram dans la phrase à évaluer = taille de la phrase – (n – 1)

BP = Brevity Penalty = min (1, exp(1 – Min Nb Mots Reférence / Nb Mots Solution))

Si la solution a au moins le même nombre de mots que la plus petite des références, BP = 1 (pas de pénalité).

Pour la mesure BLEU standard N = 4. La mesure BLEU est donc une moyenne logarithmique (=géométrique) entre les taux de 1-gram, 2-gram, etc, en commun avec les références. Cette moyenne est pondérée par un facteur entre 0 et 1 : Brevity Penalty. Plus il y a de références, meilleur sera le score. En mesure BLEU standard

La solution de correction TiLT, à la différence d'un système de traduction automatique, ne reformule pas le message SMS. Elle corrige les mots dans l'ordre dans lequel ils sont formulés. Les risques d'erreurs liées à l'ordre des mots sont donc assez faibles. Néanmoins, l'écriture SMS présentant de nombreux cas d'agglutinations à étendre, l'utilisation de la métrique BLEU nous a semblé appropriée.

En standard, la métrique BLEU prend en compte les n-grams jusqu'au 4-grams. Une seule erreur de correction située au milieu d'un message SMS bref, dont la taille est de 5 ou 6 mots, est très fortement sanctionnée par BLEU. Le tableau suivant montre quelques exemples de ce type, fortement pénalisés par BLEU alors que plutôt bien notés par Jaccard.

SMS source	Correction TiLT	Correction DELIC	Jaccard	BLEU
CT koi ldebu du mess?	c'était quoi début du message ?	C'était quoi le début du message ?	0,85	0
COMEN FAI T ON ALOR?	COMMENT fait-t-ON alors ?	Comment fait-on alors ?	0,8	0
Lé fete toute seul c cool	Les fêtes toute seul c'est cool	Les fêtes toute seule, c'est cool	0,75	0

Tableau 3 : Comparaison de résultats obtenus avec les métriques Jaccard et avec BLEU

Considérant BLEU comme peu informatif sur les messages brefs très fréquents en SMS, un paramètre a été ajouté au calcul de la mesure BLEU permettant de faire varier *n* dans le calcul des n-grams en fonction de la taille du message. Pour un message de 4 ou 5 mots, BLEU s'arrêtera aux bi-grams, pour un message de 6 ou 7 mots, on s'arrête aux tri-grams etc. Avec ce paramètre, la mesure BLEU fait d'avantage sens pour les corpus SMS et la cohérence entre Jaccard et BLEU est plus importante.

Le Tableau 4 donne les résultats obtenus sur les 9 575 SMS du corpus DELIC avec ces trois mesures. Précisons que la casse et les signes de ponctuation ont été ignorés pour le calcul. En cas de référence multiple pour un SMS (cf. Tableau 2), le score retenu est le meilleur score obtenu sur l'ensemble des références possibles.

Jaccard	BLEU standard	BLEU pondéré
0,769	0,681	0,712

Tableau 4 : Scores BLEU et Jaccard obtenus sur le corpus SMS du DELIC

4.2 Évaluation qualitative

Une étude des scores obtenus SMS par SMS permet de voir qu'environ 25% du corpus reçoit le score maximal de 1. Parmi les SMS source, on trouve des exemples parfaitement ou quasi-parfaitement orthographiés et non dégradés par la correction automatique :

(N = 4), si aucun tétragramme est commun avec une référence, le score est 0. Si la solution est courte, la moindre différence avec les références donnera un score de 0. Il est donc préférable d'ajuster le paramètre N en fonction de la taille de la solution. N = min (4 , Nb Mots Solution / 2) par exemple.

Soliel et piscine tout va bien bisous
Sommes à arles maman
Souper à la maison ce soir pour feter le début de mes vacances

mais également des exemples à l'écriture typiquement SMS très bien corrigés :

ojrd8 jv ala pi6n tu ve vnir?	Aujourd'hui je vais à la piscine, tu veux venir ?
noubli pa ke jseré tjs la pr toi	N'oublie pas que je serai toujours là pour toi
bjr, vs avé le tps pr 1 kf ? avt 15h.G du taf	Bonjour, vous avez le temps pour un café ? Avant 15h. J'ai du taf.

Tableau 5 : SMS dont la correction reçoit un score de 1

L'analyse des mauvais scores montre trois grands types de limites dans notre système actuel. La première tourne autour de la primitive "mot" prise comme point de départ pour effectuer ses hypothèses correctives. On remarque, en effet, que les plus mauvais scores sont obtenus sur les SMS présentant une absence de séparateur généralisée ou avec un séparateur peu classique de type casse ou symbole particulier. Ce type de phénomène est présent dans 1 à 2% des messages SMS.

SMS	**Correction manuelle**	**Correction TiLT**
TuTcouchéto!Cbi1! moijaVpEr2pareusir adormir.onsphon...	Tu t'es couché tôt ! C'est bien ! Moi j'avais peur de pas réussir à dormir. On se phone...	TuTcouchéto ! c'est bien ! MoijaVpEr2pareusir dormir . Onspho
B o n n e f e t e p r o f i t e s b i e n d e v o t r e d e r n i e r j o u r d e v a c a n c e	Bonne fête, profite bien de votre dernier jour de vacances.	Bonnefeteprof il t'est biendevotrede rnierjourdeva cance

Tableau 6 : SMS présentant une absence de séparateur

De même, sur les méthodes correctives, notre approche trouve ses limites quand un même segment cumule différents procédés d'écriture : phonétique et agglutination (*je ne pep a mpaC dtoi => je ne peux pas me passer de toi*), étirement et phonétique : (*G haaaaaaateuh => j'ai hâte)* etc.

La solution à ces deux difficultés est sans doute à aller chercher du côté des méthodes employées en reconnaissance de la parole pour segmenter le signal acoustique. Le problème de l'absence de séparateur n'est pas sans rappeler une langue comme le chinois pour laquelle des algorithmes de segmentation ont été développés.

La seconde limite vient de l'absence d'apprentissage dans le développement des données. L'indisponibilité d'un corpus aligné lors de la mise au point de la solution en est la cause. On peut espérer pouvoir maintenant tenter des expériences pour extraire les n-grams fréquents et enrichir les lexiques d'expressions récurrentes telles que *rien de spécial*, etc. Apprendre des séquences fréquentes de mots pour pouvoir affiner les scores attribués aux différentes hypothèses etc.

Cet apprentissage permettrait sans doute d'ouvrir l'espace des corrections aux mots dont l'orthographe est connue du lexique[5]. En effet, la stratégie utilisée actuellement a l'avantage de ne pas dégrader les messages bien écrits mais a l'inconvénient de laisser de nombreux homophones hétérographes non corrigés :

SMS	**Correction manuelle**	**Correction TiLT**
Tu me racontera dis	Tu me raconteras, dis	Tu me racontera dis
ns avons dormis	Nous avons dormi	Nous avons dormis
T ou au moi daout?	T'es où au mois d'août	T'es où au moi d'août

Tableau 7 : homophones hétérographes non corrigés

La troisième lacune concerne la trop grande localité des règles de grammaire. Une grammaire vérifiant les principales contraintes de sous-catégorisation aiderait à orienter certains choix de correction. Dans l'exemple *jme langui tro dy aller*, TiLT échoue à corriger faute de lien entre *languir* et son dépendant *d'y aller*. De même, l'accord sujet verbe n'est pas vérifié dans *tes vacs se pass bi1* qui est corrigé en *tes vacances se passe bien*. L'utilisation d'une grammaire vérifiant des contraintes entre tête et dépendant serait à expérimenter.

Enfin, des travaux autour des noms propres restent également à faire même si étant donné leur anonymisation dans ce corpus, il est difficile de tirer des conclusions définitives sur ce point. Néanmoins, on remarque sans surprise que l'identification des prénoms sans contexte déclencheur connu des lexiques est générateur de nombreuses erreurs : *Julie c jb ap moi => Julie ce gibet après moi*, *gros bisous à vous tous caro => gros bisous à vous tous carreau*.

5 Bilan et perspectives

Les mesures BLEU et Jaccard pratiquées pour l'évaluation des performances de TiLT correcteur de SMS montrent que la solution est efficace à 75% environ. L'adéquation des mesures utilisées reste bien entendu à discuter. En particulier, il pourrait être intéressant de pondérer les scores obtenus par la difficulté a priori de la correction ; cette difficulté étant quantifiable par une distance entre le SMS et sa transcription. Il faut noter également que pour une application de correction de SMS avant synthèse vocale, il faudrait prévoir une évaluation par l'usage : les erreurs sur les homophones seraient sans doute moins pénalisantes.

La correction pratiquée par TiLT n'exploite pas de données apprises sur corpus. Une évaluation sur un autre corpus permettrait de s'assurer de la stabilité de la solution. C'est pourquoi, nous espérons pouvoir faire cette même évaluation sur le corpus de Fairon et al. (2006).

Les 25% à 30% de mauvaises corrections montrent les limites de l'approche corrective actuelle : pas de correction des homophones hétérographes, pas de segmentation des messages sans séparateur, pas de mode correctif hybride etc. autant de phénomènes qui nous montrent

[5] Sauf dans certains cas très fréquents d'erreur comme la confusion participe passé en *–é* et infinitif en *-er*.

la parenté entre les messages SMS et la langue orale et qui nous invitent à reconsidérer la place centrale de la notion de mot dans notre traitement.

Remerciements

Nous remercions Olivier Collin et Jean Véronis pour leurs conseils et Sabrina Hocq pour son travail de collecte et de transcription du corpus.

Références

ANIS, J., (1999). Chats et usages graphiques. *Internet, communication et langue française.* In Anis J. (éd.), Paris : Hermès, 71-90.

ANIS, J., (2001). *Parlez-vous texto ? Guide des nouveaux langages du réseau.*, Paris : Le cherche-midi éditeur.

ANIS, J., (2002). Communication électronique scripturale et formes langagières : chats et SMS., *Actes des journées « S'écrire avec les outils d'aujourd'hui »*, Université de Poitiers.

BOVE, R., (2005)., Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS, *Actes de RÉCITAL 2005*, Dourdan, 625-634.

FAIRON, C., KLEIN, J., PAUMIER, S., (2006)., *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'*, Presses universitaires de Louvain, Louvain-la-Neuve.

GUIMIER DE NEEF, É., VÉRONIS, J., (2004). 1 pw1 sr la kestion ;-), *Papier présenté à la Journée d'Étude de l'ATALA "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.),* Paris.

HOCQ, S., (2006). Étude des SMS en français : constitution et exploitation d'un corpus aligné SMS – langue standard. *Rapport de Master II "Industries des Langues"*, Aix-en-Provence.

PAPINENI, K., ROUKOS, S., WARD, T., ZHU, W. J., (2002), BLEU: a method for automatic evaluation of machine translation, in *ACL-2002 : 40th Annual meeting of the Association for Computational Linguistics*, 311-318.

VÉRONIS, J., GUIMIER DE NEEF, É., (2006). Le traitement des nouvelles formes de communication écrite. In Sabah, G. (Éd.), *Compréhension automatique des langues et interaction*, 227-248, Paris: Hermès Science.