Évaluation d'un système de transcription de SMS

Émilie Guimier De Neef[1], Sébastien Fessard[2]
FTR&D/TECH/EASY − France Telecom R&D, DELIC − Université de Provence

Abstract

We present the software TiLT for SMS transcription and we evaluate its performances on *SMS pour la Science* corpus with Jaccard distance and Blue score. The presentation of results is followed by a qualitative analysis of the system and its limits.

Nous présentons le logiciel TiLT pour la transcription des SMS et évaluons ses performances sur le corpus de *SMS pour la Science*. 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.

Keywords : SMS, corpus *SMS pour la Science*, correction orthographique, TiLT, évaluation.

1. Introduction

Nous avons présenté dans Guimier De Neef *et al.* (2007) une première évaluation des performances du logiciel TiLT correcteur, ou plutôt transcripteur, de SMS. Ce logiciel a été mis au point à FTR&D dans le cadre du développement d'une solution de vocalisation des messages SMS destinés aux téléphones fixes. Cette évaluation s'est faite sur un corpus de près de 10 000 SMS compilés et transcrits par le laboratoire DELIC (Hocq 2006). Les performances de TiLT, mesurées par les métriques Jaccard et BLEU, montrent que le logiciel corrige correctement à 75 % sur ce corpus. Nous terminions ce travail en projetant une évaluation supplémentaire sur le corpus *SMS pour la Science* développé au CENTAL par Fairon *et al.* (2006). Les premiers résultats de cette seconde évaluation sont exposés dans cet article.

Dans une première section, nous rappelons brièvement le fonctionnement du logiciel TiLT en insistant sur la constitution des données linguistiques utilisées pour la transcription des SMS. Dans une seconde section, nous rappelons les caractéristiques du corpus *SMS pour la Science* et son adaptation à notre évaluation. Les résultats chiffrés de l'évaluation sont fournis en troisième section ainsi qu'une première analyse qualitative des résultats obtenus.

2. TiLT « correcteur » de SMS

2.1. Aperçu du fonctionnement du logiciel

[1] FTR&D/TECH/EASY – France Telecom R&D 2, avenue Pierre Marzin 22 300 Lannion Cedex, France, emilie.guimierdeneef@orange-ftgroup.com

[2] DELIC – Université de Provence 29 av. Robert Schuman – 13 100 Aix-en-Provence, sebastien.fessard@free.fr

Utilisé pour la transcription des SMS, le logiciel TiLT fait intervenir séquentiellement trois briques :

1) un module de segmentation qui assure le découpage et le typage des différentes unités du message SMS : SMILEY, MOT (incluant ou non des chiffres), etc.,
2) un module d'analyse lexicale avec ou sans méthodes correctives qui permet d'associer chaque segment du message SMS avec une ou plusieurs entrées du lexique utilisé ; ce sont les méthodes correctives étendues au contexte SMS qui permettent dynamiquement de retrouver les orthographes standard à partir des formes SMS (*pl1 = plein, jappelle = j'appelle, SSSSUUUUUUPPPPEEER = super* etc.),
3) un module d'analyse syntaxique légère réalisant un regroupement des éléments lexicaux en chunks qui permet de choisir l'hypothèse corrective la plus adaptée.

Plus de détails sur les modules du logiciel peuvent être trouvés dans (Guimier De Neef *et al.* 2007).

TiLT fonctionne avec des données linguistiques qui sont principalement un lexique et une grammaire de chunking. En standard, pour l'analyse du français, TiLT utilise un lexique d'environ 100 000 entrées incluant une base de mots-composés. L'analyse en chunking se fait avec une grammaire hors contexte d'à peu près 2000 règles. Ces règles ont été mises au point manuellement à partir d'observations sur corpus. Elles expriment les séquences possibles de catégories grammaticales en français sous forme de regroupements en chunks. Des contraintes d'accord internes aux chunks valident les groupes. Des contraintes de séquentialité entre les chunks peuvent être définies.

Plusieurs mécanismes permettent de forcer l'analyse à être déterministe, c'est-à-dire à ne rendre qu'une seule solution par énoncé. Comme chez Abney (1991), les groupes les plus longs de la gauche vers la droite sont privilégiés. Un score peut être associé à chaque chunk en fonction des catégories de ses composants et/ou des unités lexicales en présence. Enfin, rendre la première hypothèse est l'heuristique finale en cas d'ambiguïté non résolue.

2.2. Données linguistiques pour le contexte SMS

Les données lexicales et grammaticales utilisées pour la correction des SMS sont les données standard du logiciel enrichies de données spécifiques principalement mises au point à partir du corpus de SMS du DELIC.

Un lexique de près de 2000 abréviations a été compilé à partir de différentes sources provenant du web ainsi que de relevés issus du corpus du DELIC. La fréquence d'apparition des lemmes dans la transcription du corpus a permis de délimiter le vocabulaire propre aux SMS pour guider le choix des transcriptions. Cette délimitation permet de garantir que la graphie phonétique *om* sera corrigée par *homme* plutôt que par *heaume* ou *ohm* par exemple.

La grammaire utilisée dans le logiciel a été mise au point pour l'analyse de textes issus de la presse écrite. L'écriture SMS bouleverse très peu la syntaxe du français mais présente des caractéristiques propres à la communication interpersonnelle. Des extractions de n-grammes, par exemple, montrent la fréquence des constructions présentatives (*c'est moi, c'est ta copine* etc.), des interrogatives (*keske tu dev1 ?*), des tournures impératives (*passe une bonne journée*), etc. Quelques règles de grammaire mises au point manuellement d'après observations ont été ajoutées pour traiter finement ces constructions. De même, certains déclencheurs de noms propres typiques de l'écriture SMS tels que les formules de salutation (*coucou, bisous* etc.) ont été identifiés et associés à un traitement adapté dans la grammaire.

La fréquence des mots se réduisant à un caractère est l'un des éléments qui oppose écriture SMS et écriture standard. Ces mono-caractères ont souvent pour équivalent des mots outils dont le rôle syntaxique est déterminant pour notre analyse en chunk. Ils sont, de surcroît, très souvent ambigus : *k* vaut pour *que* ou *qui*, *c* remplace *c'est, ces, s'est, sais, sait* etc. Un soin particulier leur a été consacré afin de pouvoir les désambiguïser aussi bien que possible.

D'un point de vue méthodologique, mise à part la délimitation du vocabulaire du français utilisé dans les SMS, les données de TiLT n'ont pas été apprises automatiquement à partir du corpus du DELIC. Aucune modification de ces données n'a été faite pour l'évaluation sur le corpus *SMS pour la Science*.

3. Préparation du corpus d'évaluation

3.1. Présentation du corpus source

Dans le cadre de l'opération « Faites don de vos SMS à la Science » menée entre octobre et décembre 2004, l'université catholique de Louvain a collecté un corpus de 75 000 SMS auprès d'un public diversifié et volontaire. Suite à cette collecte, environ 30 000 de ces messages ont été transcrits constituant ainsi une base alignée SMS / français standard. Les détails concernant la sélection des messages et le protocole de transcription peuvent être trouvés dans Fairon *et al.* (2006).

FTR&D a récemment acquis une version de cette base livrée sous format excel contenant différentes versions de la transcription des SMS vers le français. Ce corpus peut être caractérisé par quelques chiffres. Il contient 30 000 SMS ; leur taille moyenne est de 21,8 mots et 104 caractères ; 52 214 mots[3] différents ont été dénombrés[4]. Il est intéressant de noter que la taille moyenne des SMS du corpus du DELIC est de 14,5 mots pour 66,7 caractères. La taille moyenne des mots en écriture SMS (rapport du nombre de caractères / nombre de mots) est relativement stable : 4,6 dans le corpus du DELIC pour 4,75 caractères par mot en moyenne dans le corpus du CENTAL.

3.2. Extraction des données

Pour notre évaluation, nous avons retenu du corpus original les colonnes comportant le message SMS normalisé et sa transcription normalisée sans tag. Le tableau suivant fournit un extrait de ce corpus.

De rien. Jesper ktu va bien et kta passé une bonne journé. Kiss	De rien. J'espère que tu vas bien et que tu as passé une bonne journée. Kiss.
Jespèr ktu va ten rmetr ma puce...é jesper ktu men vx pa dpa tlavoir di mè voila,rep si tu px.gros bisouxxx	J'espère que tu vas t'en remettre ma puce...et j'espère que tu m'en veux pas de pas te l'avoir dit mais voilà,réponds si tu peux.Gros bisous
Oui pa d prob m pèr t prendra ver 7h-7h15.ça 2v1 bi1 t new house!Bizz	Oui pas de problème mon père te prendra vers 7h-7h15.Ça devient bien ta new house!Biz

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

[3] Le terme *mot* est synonyme ici du terme *type* utilisé dans Fairon *et al*. (2006 : 44).

[4] Les comptages obtenus varient en fonction de la méthode utilisée pour le dénombrement. Ici, un mot est une séquence de caractère trouvée entre des espaces ou des marques de ponctuation.

Le filtrage des doublons dans les données nous a laissé un corpus d'expérimentation de 29 802 SMS composés de 655 542 tokens.

3.3. Harmonisation des transcriptions

Des divergences sur les conventions de transcription de TiLT et du corpus du CENTAL ont nécessité quelques aménagements.

TiLT ayant été développé pour le pré-traitement des SMS avant synthèse vocale, les abréviations usuelles comme *lol, jtm, jtd, stp, svp*, etc., difficilement vocalisables en l'état, sont étendues : *je me marre, je t'aime, je t'adore, s'il te plait*, etc. au contraire de ce qui est fait dans le corpus du CENTAL[5]. Plus rarement, le cas inverse a été rencontré : TiLT n'étend pas systématiquement *min* en *minutes* par exemple.

De même, les mots d'origine étrangère tels que *today, my* etc. fréquemment utilisés en SMS mais présentant des difficultés de vocalisation avec une synthèse française sont traduits en français par TiLT. La préservation du mot étranger est préférée dans le corpus du CENTAL[6].

L'ensemble de ces opérations d'harmonisation se fait par un script qui opère sur les transcriptions manuelles du corpus du CENTAL garantissant ainsi la cohérence avec les transcriptions de TiLT.

4. Évaluation

Les objectifs de cette évaluation sont multiples. Il s'agit d'une part de savoir si la capacité corrective du système TiLT est stable d'un corpus à l'autre, en particulier sur un corpus de SMS français de Belgique. Il s'agit également, par l'analyse des erreurs, de cerner les phénomènes de l'écriture SMS résistants pour TiLT afin d'étudier des méthodes alternatives de traitement.

4.1. Évaluation objective

Pour évaluer les performances de TiLT sur le corpus du DELIC, nous avons choisi dans Guimier De Neef *et al*. (2007) les mesures 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 considère la phrase comme un sac de mots, tandis que la mesure BLEU prend en compte les n-grams et pénalise les corrections qui divergent quant à l'ordre des mots. La mesure BLEU standard étant très pénalisante pour des énoncés courts car elle prend en compte des 4-grams, nous avons proposé un aménagement de cette mesure qui permet d'ajuster la taille des n-grams calculés en fonction de la taille du SMS (cf Guimier De Neef *et al*. (2007)).

Le Tableau 2 donne les résultats obtenus sur les 29 802 SMS du corpus du CENTAL avec ces deux mesures, ainsi que les précédents résultats sur le corpus SMS du DELIC. Précisons que la casse et les signes de ponctuation ont été ignorés pour le calcul.

---

[5] La raison pour laquelle ces abréviations, sigles etc. ne sont pas étendus dans les transcriptions du corpus *SMS pour la Science* n'est pas explicitée dans Fairon *et al*. (2007) et semble assez discutable.

[6] Ce choix de préservation peut être discuté car l'emprunt permet une simplification de l'écriture comme pour *today* vs *aujourd'hui* et constitue à ce titre un mécanisme d'abréviation.

CORPUS	Jaccard	BLEU pondéré
DELIC (9 575 SMS)	0,769	0,712
CENTAL (29 802 SMS)	0,785	0,736

Tableau 2 : Scores BLEU et Jaccard obtenus sur le corpus SMS du CENTAL.

Les résultats obtenus sur les deux corpus sont donc tout à fait comparables, même légèrement meilleurs sur le corpus du CENTAL.

4.2. Évaluation qualitative

L'application de la mesure Jaccard entre le SMS et sa transcription manuelle permet de quantifier la distance entre les deux modes d'écriture, nous donnant une image de la difficulté a priori de la tâche de transcription automatique. Le score Jaccard moyen obtenu pour la difficulté a priori est de 0,61. Le transcripteur TiLT a donc permis de gagner 0.17 points[7].

La Figure 1 nous donne la répartition des SMS selon la difficulté de correction a priori (barres hachurées) et selon le score obtenu après la transcription automatique (barres unies). Cette figure nous montre très nettement qu'après transcription automatique très peu de messages sont notés en dessous de 0,5.

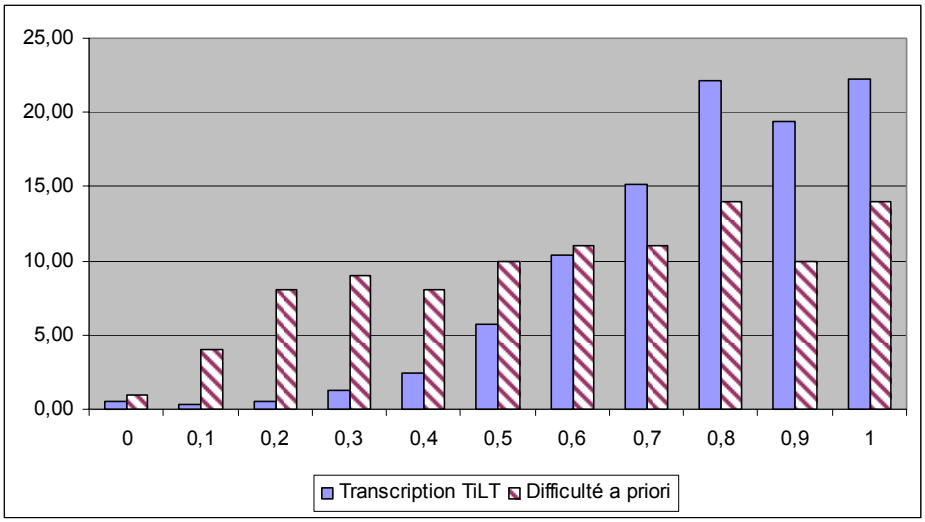

Figure 1 : Répartition des SMS en fonction du score Jaccard

On note que sur les 6 623 SMS qui obtiennent un score maximal de 1, 2 649 sont parfaitement orthographiés à l'origine. Parmi les SMS bien transcris par TiLT restants certains présentent des particularités d'écriture très classique du français (abréviations courantes et non ambiguës, phonétisations simples etc.). À l'opposé, des messages assez typiques du langage SMS sont aussi correctement transcrits comme le montre le Tableau 3.

[7] Le score de difficulté a priori obtenu sur le corpus DELIC est de 0,436. Sur ce corpus, TiLT permet de gagner 0,33 points.

Voilà, chui ds mon lit ac le catalogue 2 la redoute! J'espèr vais bien dormir! Bonne nuit a toi mon prince et bonne chance pr 2m1. Gros bisous tt plein	Voilà, je suis dans mon lit avec le catalogue de la redoute ! J'espère vais bien dormir ! Bonne nuit à toi mon prince et bonne chance pour demain . Gros bisous tout plein
Oui pq pas. C où et c Koi?C gentil de penser à moi.	Oui pourquoi pas . c'est où et c'est quoi ? c'est gentil de penser à moi .
pour ne pa mentir, j'ai ke 21 ans, c'est pour ca ke je ne voulai pa parlé au téléfone. j'atten ta réaction.	pour ne pas mentir, j'ai que 21 ans, c'est pour ça que je ne voulais pas parler au téléphone . j'attends ta réaction .
Pq ta fai sa	pourquoi t'as fait ça
Pq g pa couru! C a cose du non ke...	pourquoi j'ai pas couru ! c'est à cause du non que ...

Tableau 3: SMS dont la correction reçoit un score de 1

Les SMS recevant un score de 0 représentent 0,52 % des résultats. L'essentiel de ces messages présente une absence de séparateur généralisée, c'est-à-dire qu'aucun séparateur habituel n'est utilisé pour segmenter les mots. Ce problème, qui avait déjà été identifié lors du dépouillement des résultats obtenus sur le corpus du DELIC, est une des limites du système TiLT actuellement. Quelques exemples sont donnés dans le tableau suivant :

SMS	**Transcription manuelle**	**Correction TiLT**
MNANGEJETEDITDEMAINJETAIMEMNAMOUR TON LOULOU	Mon ange je te dis demain je t'aime mon amour ton loulou	MNANGEJETEDITDEMAINJETAIMEMNAMOUR TON LOULOU
Comencavamoigvb1a+	Comment ça va moi je vais bien à plus	Comencavamoigvb1a plus
.J.ESPERE.KE.VOS.FETES.C.SONT.BIEN.PASSEE.JE.PENSE.TJRS.A.TOI..MEME.SI.ON.NE.C.VOIT.+.BIZ	J'espère que vos fêtes se sont bien passées je pense toujours à toi même si on ne se voit plus biz	.JESPEREKEVOSFETESCSONTBIENPASSEEJEPENSETJRSATOI. MEME . SI . ON . NE . CVOIT plus . bisou
HelloJeRecoisSeulementTonSms..JétaisChezMaSoeurOuYaPasDeRéseau..CaSeraPourUneAutreFoisCoupezVousBienDam	Hello je reçois seulement ton SMS..J'étais chez ma soeur où y a pas de réseau..Ça sera pour une autre fois coupez-vous bien Dam	HelloJeRecoisSeulementTonSms .. JétaisChezMaSoeurOuYaPasDeRéseau .. CaSeraPourUneAutreFoisCoupezVousBienDam

Tableau 4 : SMS présentant une absence de séparateur

De même, la combinaison de plusieurs mécanismes d'écriture SMS sur un même mot, agglutination et écriture phonétique, par exemple, comme dans presque tous les mots du SMS suivant : *Jti1 atwa koma pRson dotr!Tmemank!,* ne trouve pas de solution complètement satisfaisante aujourd'hui.

La stratégie corrective choisie ne corrige que les mots inconnus du lexique français. Les formes existantes mais mal orthographiées ne sont pas corrigées. Les résultats montrent donc de nombreux cas d'homophones hétérographes non corrigés :

SMS	Transcription manuelle	Correction TiLT
On ta **fais** signe dans le break rouge	On t'a **fait** signe dans le break rouge	On t'a **fais** signe dans le break rouge
T'es où? t'es **partis** quand d'ici?	T'es où? T'es **parti** quand d'ici?	T'es où ? t'es **partis** quand d'ici ?

Tableau 5 : homophones hétérographes non corrigés

Concernant les belgicismes relevés dans Fairon *et al*. (2006 : 112), seuls les belgicismes lexicaux posent problème à TiLT qui les transforme au lieu d'en préserver l'orthographe :

SMS	Transcription manuelle	Correction TiLT
C gentil. Dis-lui comme çà, elle sera **binêsse**!	C'est gentil. Dis-lui comme ça, elle sera **binêsse**!	c'est gentil . Dis-lui comme ça, elle sera **bineuse** !
Je peux passer prendre les quelques **brols** que j'ai laissé dans la course vendredi? Tu seras lààmmmmh?	Je peux passer prendre les quelques **brols** que j'ai laissés dans la course vendredi? Tu seras là mmmh?	Je peux passer prendre les quelques **bols** que j'ai laissé dans la course vendredi ? Tu seras Lààmmmmh ?
Ben si...mè non on ira lè metr o **kot**. Biz	Ben si...mais non on ira les mettre au **kot**. Biz	Ben si ... mes non on ira l'ai Metr où **cotent** . bisou
C toi qui a repris **l'escabelle** bleue?	C'est toi qui as repris **l'escabelle** bleue?	c'est toi qui a repris **l'escabelle** bleue ?
AT,c'est toi? Que deviens-tu, **fieu**? A bientot, j'espère! Guibert	AT,c'est toi? Que deviens-tu, **fieu**? À bientôt, j'espère! Guibert	AS, c'est toi ? Que deviens-tu, **Fieu** ? à bientôt, j'espère ! Guibert

Tableau 6 : Quelques exemples de Belgicisme

Du point de vue des capacités du logiciel TiLT, les résultats obtenus confirment largement les limites déjà observées dans Guimier De Neef *et al*. (2007).

5. Bilan et perspectives

Le but de cette expérimentation était de valider la stabilité de TiLT transcripteur de SMS en testant ses performances sur un second corpus sans aucune préparation préalable. Les mesures effectuées montrent une stabilité des scores. Cet encourageant résultat n'était pas escompté a priori étant données les différences sociologiques et géographiques des deux corpus.

Les causes d'erreurs TiLT déjà repérées dans Guimier De Neef *et al*. (2007) se confirment : pas de correction des homophones hétérographes, pas de segmentation des messages sans séparateur, pas de mode correctif hybride etc. Aucune limite nouvelle spécifique ou non au corpus du CENTAL n'a été mise à jour, à part quelques difficultés liées aux belgiscismes lexicaux qu'il suffirait d'introduire dans les données lexicales de TiLT pour un traitement optimal.

Nous notons cependant que sur le corpus du DELIC, TiLT permet de gagner 0,33 points en score Jaccard par rapport à la difficulté de correction a priori. Il ne permet de gagner que 0,17 points sur le corpus du CENTAL. Les performances du logiciel semblent donc plafonner. De nouvelles méthodes de parsing seront à explorer pour repousser les limites actuelles.

Remerciements

Nous remercions Jean Véronis pour ses conseils.

Références

ABNEY, S. (1991), « Parsing by chunks », In. Berwick, Abney, Tenny (Eds.) *Principle-based parsing*. Kluwer Academic Publishers, Amsterdam.

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

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

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., DEBEURME, A., PARK, J., (2007). « TiLT correcteur de SMS : évaluation et bilan qualitatif », *Actes de TALN 2007*.

GUIMIER DE NEEF, É., VERONIS, 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.

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