# De la collecte à la normalisation des SMS : linguistique de corpus et traitements automatiques par apprentissage

Richard Beaufort
Université catholique de Louvain – CENTAL
richard.beaufort@uclouvain.be

Le *Short Message Service* (SMS), qui offre depuis quelques années la possibilité d'échanger des messages écrits entre téléphones mobiles, a été rapidement adopté par les utilisateurs. Une caractéristique notoire des SMS est leur tendance à s'éloigner fortement des conventions orthographiques. Selon les spécialistes (Thurlow and Brown, 2003; Fairon et al., 2006; Bieswanger, 2007), cette variabilité est le fruit de l'utilisation simultanée de plusieurs stratégies de codage comme les jeux phonétiques (*2m1* lu *demain*), les transcriptions phonétiques (*kom* à la place de *comme*), les squelettes consonantiques (*tjrs* pour *toujours*), les séparateurs incorrects, manquants ou superflus (*j esper* pour *j'espère* ; *j'croibi1k*, pour *je crois bien que*), etc. Ces déviations sont dues à trois facteurs principaux : la taille réduite des messages autorisés par le service (140 octets, donc environ 140 caractères), la difficulté de manipuler le petit clavier à neuf touches des mobiles standard et, *last but not least*, le fait que les gens envoient des SMS principalement à leurs proches et leurs amis, dans un registre informel. Quelles qu'en soient les causes, ces déviations compliquent considérablement la tâche de tout système standard de Traitement Automatique du Langage (TAL), qui ne sait comment réagir face à tant de mots hors vocabulaire. C'est la raison pour laquelle, comme le remarquent Sproat *et al.* (2001), un SMS doit être normalisé *avant* qu'un autre traitement, plus conventionnel, ne puisse lui être appliqué. Yvon (2008) définit la normalisation d'un SMS comme "le procédé qui consiste à réécrire un SMS en utilisant une orthographe plus conventionnelle, afin de rendre ce SMS plus facile à lire pour l'humain et pour la machine." [1]

Dans le cadre général d'un système de synthèse de la parole (cf. le projet Vocalise, http://cental.fltr.ucl.ac.be/team/projects/vocalise/), nous avons développé un **module de normalisation des SMS** basé exclusivement sur des **modèles appris sur corpus**. La normalisation se déroule en trois étapes. Premièrement, nous consultons un **dictionnaire dédié aux SMS**, afin de distinguer, dans une séquence SMS bruitée, les parties connues des parties inconnues. Deuxièmement, nous appliquons des **modèles de réécriture pondérés** sur les différentes parties de la séquence bruitée, ce qui produit un treillis de solutions pondérées. Les modèles appliqués diffèrent selon que la partie à réécrire est connue ou non. Troisièmement, nous combinons le treillis de solutions avec un **modèle de langue**, afin de tenir compte du contexte dans le choix de la meilleure retranscription. Nos modèles de normalisation ont été entrainés sur un corpus de 30 000 SMS écrits en français, qui ont été récoltés en Belgique, anonymisés semi-automatiquement et retranscrits manuellement par des membres du Centre de Traitement Automatique du Langage (CENTAL) de l'Université catholique de Louvain (Fairon and Paumier, 2006). Ensemble, le corpus et sa retranscription constituent des *corpus parallèles*, alignés *au niveau du message*.

La présentation de ce vendredi 22 octobre 2010 sera organisée comme suit. Premièrement, nous nous intéresserons à la manière dont les collectes de SMS ont été et sont encore réalisées dans le cadre des projets « Faites don de vos SMS à la Science » et « sms4science » (cf.

[1] Traduit de l'anglais.

http://www.sms4science.org/), dont l'objectif général est de rassembler des corpus SMS dans le plus grand nombre de langues possibles. Nous en profiterons pour nous arrêter quelques instants sur certains constats linguistiques et statistiques obtenus par les chercheurs à partir de ces corpus. Deuxièmement, nous décrirons l'approche que nous avons implémentée pour aligner le corpus SMS et sa transcription *au niveau du caractère*, une étape nécessaire dans l'optique d'apprendre des modèles de normalisation à partir de ces corpus. Troisièmement, nous détaillerons le processus de normalisation mis en place et, dans un même élan, la manière dont les modèles de normalisation ont été appris. Cette présentation se conclura par une évaluation de l'approche, suivie d'une petite démonstration du système complet de synthèse de la parole à partir de SMS, *text-it/voice-it*, dont un prototype est déjà disponible pour les smartphones utilisant le système d'exploitation Android.

**Références**


Markus Bieswanger. 2007. abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space and time-saving strategies in English and German text messages. In *Texas Linguistics Forum*, volume 50.

Cédrick Fairon and Sébastien Paumier. 2006. A translated corpus of 30,000 French SMS. In *Proc. LREC 2006*, May.

Cécrick. Fairon, Jean R. Klein, and Sébastien Paumier. 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. 136 pages.

Richard Sproat, A.W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Crispin Thurlow and Alex Brown. 2003. Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1).

François Yvon. 2008. *Reorthography of SMS messages*. Technical Report 2008, LIMSI/CNRS, Orsay, France.