Un corpus de SMS est-il un corpus comme les autres ?

Cédrick Fairon[1], Sébastien Paumier[2]
Université catholique de Louvain, Université de Marne-la-Vallée

Abstract

The development of communication technologies has contributed to the appearance of new forms in the written language that scientists have to study according to their peculiarities (typing or viewing constraints, synchronicity, etc). In the particular case of SMS (Short Message Service), studies are complicated by a lack of data, mainly due to technical constraints and privacy considerations.

In this paper, we discuss the main difficulties encountered while collecting the data and building the corpus (legal aspects, technical and practical issues). We also highlight the specificities of the resulting corpus as a language resource and explain why it is difficult to use these data that are, by several aspects, different from other kinds of corpora.

Keywords : SMS, texto, minimessages, corpus de SMS

1. Introduction

La communication par SMS (*textos, minimessages*) est devenue un véritable phénomène de société. Le succès commercial des SMS est spectaculaire : des millions de messages s'échangent chaque jour pour communiquer, mais aussi désormais, pour participer à des concours, obtenir des renseignements, payer une place de parking, faire un paiement chez un commerçant, etc. Les services sont innombrables et de nouvelles possibilités apparaissent sans cesse.

Comme on le sait, ces messages sont généralement courts (au-delà de 160 caractères, le message est expédié en plusieurs tranches et donc facturé plus cher), encodés à l'aide d'un clavier difficilement maniable (les lettres sont disposées sur plus ou moins 9 touches, en fonction des téléphones), parfois rédigés précipitamment entre deux autres activités, voire en même temps. Pour ces raisons, mais aussi pour le plaisir du jeu avec la langue ou encore pour répondre à des besoins expressifs, des habitudes de rédaction particulières sont dans certains cas adoptées par les utilisateurs. Pour étudier ces nouvelles pratiques de l'écrit, de vastes corpus sont nécessaires. Comme l'ont montré (Fairon *et al.* 2006a), les auteurs qui ont travaillé sans corpus, ont souvent travaillé... sans filet. Ainsi, quand on compare certains

[1] CENTAL, Université catholique de Louvain, 1 Place Blaise Pascal, 1348 Louvain-la-Neuve, Belgique, cedrick.fairon@uclouvain.be

[2] IGM, Université de Marne-la-Vallée, 5 bd Descartes, 77454 Champs-sur-Marne, France, paumier@univ-mlv.fr

résultats au corpus "SMS pour la science"[3] (Fairon *et al.* 2006b), la différence entre les prédictions et les observations réalisées sur des données authentiques est parfois très spectaculaire.

Dans cet article, nous proposerons une réflexion sur les particularités d'un corpus SMS : les difficultés techniques que constituent sa composition, ses particularités internes (taille des textes, nombre des auteurs, problème de la variété des formes, difficulté du dénombrement) ainsi que des difficultés d'exploitation d'un texte parfois "illisible".

2. Constitution d'un corpus

Certains auteurs (Anis 2001, 2003 ; Liénard 2005) estiment le langage SMS suffisamment proche de ce que l'on peut trouver dans des chats ou des forums pour pouvoir étudier ceux-ci et extrapoler leurs conclusions aux SMS. Ce regroupement est également rendu explicite par des appellations telles que *cyberlangue*, ou *cyberlangage* (Dejond 2002, 2006). En réalité, on observe des différences linguistiques, et donc même si les chats et les forums sont beaucoup plus faciles à collecter que les SMS, ils ne représentent pas un matériau parfait pour étudier le langage SMS.

2.1. Problèmes techniques

Contrairement aux chats et aux forums, la communication par SMS se caractérise par le fait que les messages sont privés, visibles uniquement des correspondants en contact. Il est donc techniquement impossible de les collecter aisément en téléchargeant des pages web, comme c'est généralement le cas pour les chats et les forums (notons que si c'est techniquement possible, cela ne règle pas la question de la légalité de cette opération).

En ce qui concerne les SMS, une première solution technique pour les obtenir serait de les intercepter, mais cela présente deux inconvénients majeurs. Tout d'abord, il faudrait s'assurer de la complicité d'au moins un opérateur de téléphonie mobile. Une complicité bien entendu incompatible avec la loi, qui protège la vie privée et le secret de la correspondance. Ensuite, il serait impossible d'associer les messages obtenus à des informations sur leurs auteurs à moins de donner franchement dans l'espionnage. Ce défaut est sérieux, car sans de telles informations, il devient impossible de faire des études sociolinguistiques sur ce matériau qui s'y prête pourtant admirablement.

Une seconde méthode de collecte est la copie manuelle des messages. C'est ce procédé qui a été adopté dans les rares travaux présentant des corpus de SMS (Véronis et Guimier de Neef 2006), mais il souffre également de deux défauts dont le premier est la fiabilité de la copie. En effet, il est très probable que de nombreuses erreurs viennent altérer le message d'origine. On pense bien sûr à des fautes de frappe, mais le plus grand risque se situe du côté de l'auto-

[3] Dans le cadre du projet « Faites don de vos SMS à la science » (Fairon *et al.* 2006a), l'UCL a constitué un corpus de 75 000 SMS. Pour mener à bien cette collecte, un appel a été lancé vers le grand public au travers des médias. Les messages ont été récoltés en deux mois, au travers d'un numéro gratuit vers lequel les participants pouvaient faire suivre une copie de messages « réellement envoyés dans le cadre habituel de leur correspondance ». Ceci dans le but de récolter des messages « authentiques » et non des messages rédigés pour la circonstance. Le dispositif technique de la collecte permettait de son côté de préserver l'intégrité des messages (pas d'erreur de recopiage, autocorrection, etc.). Parallèlement les participants étaient invités à remplir un questionnaires socio-linguistiques sur le Web. Un extrait de ces 75 000 SMS a été diffusé à l'attention de la communauté scientifique. Ce sous-corpus a été enrichi d'une transcription en français standardisé et de certaines annotations permettant de simplifier les opérations de recherche dans le corpus (Fairon *et al.* 2006b).

correction, car l'oeil humain a tendance à corriger beaucoup de choses inconsciemment. Sur du texte particulièrement « bruité » comme le SMS, la propension à rétablir des formes standard est très importante comme nous avons pu le constater lors du travail de transcription mené au CENTAL (Fairon *et al.* 2006a). Le deuxième problème lié à la copie concerne le public visé. Typiquement, un professeur va demander à ses étudiants de ramener une vingtaine de SMS chacun, et il obtiendra ainsi un corpus ne contenant que des SMS provenant de l'entourage immédiat des collecteurs, ce qui pose le problème de la représentativité des données (c'est un problème récurrent à chaque création de corpus, de SMS ou autre type de données).

Afin d'éviter les inconvénients de l'interception et de la copie manuelle, il reste la solution de se faire envoyer des SMS directement par leurs auteurs, solution qui a été retenue lors du projet « Faites don de vos SMS à la science ». Bien que simple, cette idée est difficile à mettre en œuvre. D'un point de vue technique, il faut pouvoir collecter sous forme numérique un grand nombre de messages, ce qui nécessite la mise en place d'un numéro de téléphone *ad hoc* par une entreprise spécialisée[4]. Il faut ensuite convaincre le public de participer doublement : en envoyant des messages d'une part, et en fournissant des informations sur leur profil d'autre part (par exemple, en remplissant un formulaire web). On peut ainsi obtenir un corpus de SMS assorti des profils sociolinguistiques des auteurs (Fairon *et al.* 2006a). Pour y parvenir, nous avons combiné la promotion du projet avec un concours récompensant les donataires.

Cette méthode souffre tout de même d'une limitation, car elle ne permet pas de saisir les différents tours de l'interaction. On ne récolte pas les questions et les réponses, mais uniquement les messages rédigés par un seul des correspondants. Même si les deux correspondants envoient leurs messages, il n'y a pas moyen de reconstituer leur conversation. Et si un correspondant transmet les messages qu'il a envoyés ainsi que ceux qu'il a reçus au cours d'une conversation, cela posera un double problème : d'une part, on ne sera probablement pas en mesure d'identifier son correspondant (on ne disposera donc pas d'information sociolinguistique à son sujet) et d'autre part, on risque de se retrouver face à un problème de droit d'auteur dans la mesure où les messages n'ont pas été directement transmis par l'auteur, seul ayant droit.

Un autre problème est que l'on n'échappe pas au paradoxe de l'observateur que l'on rencontre dans les enquêtes de terrain : pour obtenir les données les plus importantes pour la théorie linguistique, on doit observer comment parlent les gens quand ils ne sont pas observés, disait Labov. En d'autres termes, il faut reconnaître que les conditions de collecte peuvent introduire un biais. La différence avec la collecte de corpus oraux est cependant notable : une fois que l'enregistrement est lancé, le témoin se sait observé de A à Z, du début à la fin de l'interview. Dans notre système, les participants peuvent rédiger librement leur message sans se soucier de l'enquête, et ensuite, décider si oui ou non ils l'enverront. Cela peut être vu comme une situation plus favorable que celle de l'enregistrement dans la collecte de corpus oraux, car il y a moins de pression, puisque l'on pourra choisir plus tard les messages à donner ou, au contraire, comme une situation moins favorable, dans la mesure où cette sélection introduit nécessairement un biais.

2.2. Problèmes légaux

[4] Pour cette opération l'UCL a bénéficié de l'aide de deux sociétés belges de télécommunication, Proximus et NEWay, ainsi que d'un troisième partenaire privé, Ogilvy.

La méthode de collecte des SMS règle les problèmes légaux du droit d'auteur et de la vie privée. On peut en effet demander aux auteurs de renoncer à leurs droits sur les messages en vue de leur exploitation ultérieure. Cependant, si l'on a bien l'autorisation de l'auteur, on ne sait rien à propos du destinataire qui n'a peut-être pas envie qu'un message privé qui lui est adressé soit transmis à un tiers à son insu. La seule solution est donc d'anonymiser les messages de telle sorte qu'il soit impossible d'identifier qui que ce soit de privé. Cette procédure est très compliquée, car les SMS sont, par essence, en langue non standard et il est impossible d'appliquer des techniques automatiques de recherche de noms propres. Il faut donc parcourir manuellement tout le corpus pour repérer les indications personnelles (noms, numéros de téléphone, adresses, mails, sites web, blogs, etc.). Certaines données personnelles peuvent également apparaître sous des formes très variées, comme dans l'exemple imaginaire suivant : *J'habite rue de la pêcherie, la maison rouge avec un cochon doré sur la porte*. Ce travail est terriblement fastidieux, mais il est nécessaire si l'on veut disposer de données utilisables en toute légalité.

3. Exploitation des données

3.1. Lisibilité

Pour pouvoir être utilisé, un corpus de texte doit être lisible. Cela peut faire sourire à l'heure du tout électronique, mais il faut se rappeler que l'on peut rencontrer divers problèmes comme une langue inconnue (Linéaire A), des symboles illisibles (manuscrits de Balzac, ordonnances médicales) ou encore un cryptage volontaire (Enigma). Dans le cas des SMS, on se rapproche plutôt de cette dernière catégorie, bien que le but soit moins de protéger les messages des yeux indiscrets que de les décorer et les raccourcir. Quoi qu'il en soit, on peut avoir de réelles difficultés à les déchiffrer, par exemple si la séparation en mots n'est pas respectée :

```
LO???dsl    msGarticlàrendr    pr    vend&jEdi&jdoialéensal1fomètpEvnir2
2m1jusKvendyanoprobtmDrangpa???&noublipavnirbalDbleujsui1pEmaladmèonvafRf
ètken mm bisou
```

```
BijourMonAmourDiMoiJPeVnirChéToi?GRi1DOtrPrFerPaséLTpsJusk19hPuiGCorPasé1
SuperMatinéeJMeSuiFéTrétéDTtLèNomPuiMmSiCPrDodoCLeMèmMsGTroBzoinDTVoirEnf
1SiTuVeBi1?JTM
```

Les messages peuvent également comporter beaucoup de formes altérées :

```
Hep.cfè plésir dav 1msg dtoi.g u math l1di é ca abof éT.ier scienc é
Go,alèz.toi   oci   bon   merd.la   jaten   lbus   é   ca   gèl.mè   d
couch,bonè,gan,écharp...lol.bis a+
```

```
CdsCk là keCb1pratik davoir2num.jVtsoné ds10min+ou-.ok? (1ap4wi,2 4non)
```

Pour lire ces messages, il est nécessaire de maîtriser les codes, usages et habitudes des auteurs de SMS, qui ne sont pas forcément compréhensibles pour un néophyte. Il est donc quasi indispensable de disposer d'une transcription en forme standard pour pouvoir réellement exploiter de telles données.

3.2. Problèmes d'exploration des données

Le caractère aléatoire des variantes rencontrées pour un mot ou une expression fait qu'il est impossible de les lister *a priori*. Or, sans une telle liste, il est impossible de rechercher toutes les occurrences d'une séquence donnée à moins de passer manuellement en revue tous les SMS, ce qui est d'expérience assez fastidieux. En revanche, cela devient facile si l'on dispose de la transcription, puisqu'il suffit de rechercher toutes les occurrences de la séquence en français standardisé et d'examiner les SMS qui leur correspondent. On trouvera ainsi sans peine des variantes de la forme *soirée* :

```
Merci. Bisous, bonne soirée...
```

```
Bonne swarée et a+??? Bisouxx
```

```
Bizzøux bone soiré
```

La transcription des SMS pose de nombreux problèmes (cf. description du protocole adopté dans Fairon *et al.* 2006a), mais elle s'avère presque incontournable si l'on veut pouvoir explorer de telles données.

3.3. La notion de « mot »

Les SMS posent un autre problème, celui de la définition des mots. On peut trouver dans des SMS des séquences où la séparation en mots n'est pas faite (cf. ci-dessus, sous 3.1), ce qui nous rapproche des langues sans séparateurs comme le thaï. Parfois, le découpage est même impossible, car la séquence mélange intimement plusieurs mots comme c'est le cas pour `G` (*j'ai*) ou `CT` (*c'était*). Cet état de fait rend très difficile la mise en place de procédures automatiques de segmentation. Pour s'en convaincre, il suffit de constater que le découpage d'une séquence standard comme *aujourd'hui à 10h05* pose déjà des problèmes pour le choix des règles de segmentation :

```
aujourd/'/hui/    /à/ /10/h/05
aujourd'hui//à/ /10/h/05
aujourd/'/hui/    /à/ /10h05
aujourd'hui//à/ /10h05
aujourd/'/hui/    /à/ /1/0/h/0/5
aujourd'hui//à/ /1/0/h/0/5
```

Sur un exemple en SMS comme *0jourd'8 à 10h05*, on voit que ces règles devront être encore affinées pour tenir compte des chiffres faisant partie de mots.

Une autre difficulté est qu'il est presque impossible d'associer automatiquement un mot standard à un « mot » SMS à cause des procédés utilisés (abrègements, suppression de lettres muettes, etc.) qui augmentent beaucoup les ambiguïtés. Ainsi, difficile de déterminer automatiquement si *comm* apparaît pour *comme*, *comment* ou encore autre chose. Une procédure aussi classique que le décompte d'occurrences pose donc déjà des problèmes en soi. Au final, il apparaît qu'il est bien difficile d'établir des chiffres fiables sur les SMS.

À la question, pourtant triviale en linguistique de corpus, *de combien de mots se compose le corpus ?* il n'y a donc pas de réponse simple : tout dépendra des règles utilisées pour la segmentation des mots. A moins que l'on établisse le compte à partir de la version transcrite du corpus, ce qui ne manquerait pas non plus de soulever certains problèmes. Il ne sera pas

non plus facile de calculer une métrique telle que le "type/token ratio" (TTR). Souvent invoqué dans les études de linguistique de corpus, le TTR mesure une certaine forme de richesse lexicale [5] : par exemple, on ne sera pas étonné que le TTR d'un apprenant de l'anglais soit inférieur au TTR d'un natif de la langue. Comme le calcul de cette valeur passe par l'identification des formes différentes (*types*), l'opération est particulièrement compliquée.

3.4. Influence du média

À l'exception de messages envoyés depuis un portail web, les SMS sont toujours liés aux supports physiques que sont les téléphones portables. Ils sont soumis aux contraintes de coût, d'affichage et de saisie inhérentes à ces objets. Si le coût influence la taille des messages, le clavier d'un téléphone mobile impose une façon de composer le texte tout à fait particulière puisqu'il faut appuyer plusieurs fois sur une même touche afin d'atteindre une lettre donnée.

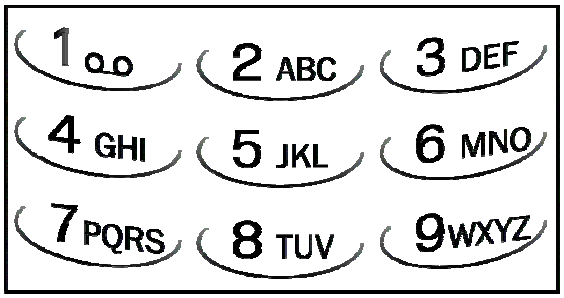

Fig 1. Répartition des lettres sur les touches d'un clavier

Une première conséquence en est que la longueur d'un SMS ne se compte pas qu'en nombre de caractères, mais également en nombre de touches pressées. Cela explique par exemple le fait qu'on trouve souvent *wa* à la place de *oi* comme dans *pourkwa*, car il faut presser 2 touches pour *wa* contre 6 pour *oi*.

Ce mode de composition fastidieux à naturellement conduit à la mise au point d'outils d'aide à la rédaction : les dictionnaires T9. Le principe est de prédire un mot à partir de la suite de touches nécessaire pour composer le mot, mais sans que l'auteur ait à préciser quelle lettre il veut sur chaque touche. Ainsi, pour obtenir le mot *avoir*, on appuiera sur 2, 8, 6, 4 et 7. Naturellement, il y a des ambiguïtés et l'on peut obtenir par erreur un autre mot qui correspond à la même suite de touches. Par exemple, on trouve dans certains SMS *et* à la place de *du* :

```
Je colle les photos et portugal
```

Ce phénomène peut fausser l'appréciation du corpus puisqu'il fait peut apparaître des séquences incorrectes là où l'auteur n'avait pas voulu taper cela. Ainsi le fait que l'on trouve *à* au lieu de *a* :

```
Salu si tu veu il y à escrime de 7h à 9h
```

[5] Nombre de *types* (formes différentes) divisé par nombre de *tokens* (nombre total d'occurrences dans le texte). Cette métrique très superficielle dépend directement de la taille du texte et ne permet donc pas de comparaison entre des textes de tailles différentes.

pourrait conduire à penser que l'auteur confond les deux, alors que c'est peut-être simplement une erreur du dictionnaire qu'il n'a pas vu ou pas pris la peine de corriger. On peut d'ailleurs se demander si ce genre d'erreurs du T9 ne pourrait pas avoir une influence sur les jeunes en cours d'acquisition de la langue.

4. Le discours

Bien que les SMS présentent des particularités de discours, il serait faux de croire qu'il s'agit d'une version destructurée de la langue. En effet, une fois que l'on a déchiffré un SMS, on constate la plupart du temps que la syntaxe est respectée et que le « langage SMS » n'est qu'une modification de surface d'un texte correct. Il arrive même que sous des dehors particulièrement cryptés se cachent des messages plutôt littéraires :

```
Tcom1drog+jtparl+Genvi2toi.tn rir mréveil,Tyx memport,tnSpri mgard éjSpR
ktn coer mlSra1jr bRCtn cor.jcouvriré ta douc po2 Kl1 .tusra plonG ds
1tendr bonher chéri
```

Le respect de la grammaire n'est pas très étonnant, car le but est d'être compris du destinaire, et un message qui serait agrammatical en plus d'être crypté aurait peu de chances d'être lu.

Les particularités du discours des SMS se situent ailleurs, notamment dans certaines formes de messages typiques, comme les réponses en rafale :

```
Je regarde la TV. Je veux bien ton massage... les magasins de Bruxelles.
Ça a l'air d'aller toi?
```

```
C'est turquoise je faisais dodo à demain Élo va prendre du chocolat à
plus
```

Il s'agit bien sûr de SMS répondant à des questions en rafale. On rencontre également un autre phénomène consistant à poser une question et à donner sa propre réponse :

```
ça va ? Moi oui.
```

```
T'as été à la foire finalement? Moi avec mon grand frère
```

```
Tu fais quoi de beau dimanche?Moi je suis censé dormir toute la journé
pour me reposer
```

Conclusion

Comme nous l'avons montré, les corpus de SMS posent des problèmes spécifiques depuis leur collecte jusqu'à leur exploitation. Matériau particulièrement difficile à obtenir légalement, les SMS présentent des problèmes de lisibilité qui sont difficilement surmontables sans une transcription en langue standardisée. Les contraintes liées aux téléphones mobiles ont favorisé l'apparition de nouvelles pratiques qui rendent ce mode de

communication très particulier, malgré sa ressemblance avec d'autres modes de communication électronique, ce qui donne à la communauté scientifique des corpus atypiques pleins de défis à relever.

Références

ANIS, J. (2001). *Parlez-vous texto ?*, Paris, Le Cherche-Midi.

ANIS J. (2003), « Communication électronique scripturale et formes langagières », in Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques, Poitiers, 31 mai et 1er juin 2002. Documents, Actes et Rapports pour l'Education, CNDP : 57-70 [sur le Web : http ://edel.univ-poitiers.fr/rhrt/document547.php, visité le 1/11/2006].

DEJOND A. (2002), *La cyberl@ngue française*, La Renaissance du Livre, Tournai.

FAIRON C., KLEIN J. et PAUMIER S. (2006a), *Le langage SMS. Etude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'*, in *Cahiers du Cental* 3/1, Presses universitaires de Louvain, Louvain-la-Neuve.

FAIRON C., KLEIN J. et PAUMIER S. (2006b), *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation*, CD-Rom, in *Cahiers du Cental* 3/2, Presses universitaires de Louvain, Louvain-la-Neuve.

LIENARD, F. (2005), « Langage texto et langage contrôlé, Description et problèmes », in *Lingvisticæ Investigationes*, 28/1 :49-60

VERONIS J. et GUIMIER de Neef É. (2006), « Le traitement des nouvelles formes de communication écrite », in G. Sabah (éd.) *Compréhension automatique des langues et interaction*, Hermès Science, Paris.

YIJUE H. et MIN-YEN K. (2005), « Optimizing predictive text entry for short message service on mobile phones », in *Proceedings of 11th International Conference on Human-Computer Interaction (HCII 05)*, Las Vegas, July 2005.