Histoire Épistémologie Langage
38/2 (2016), p. 73-85

DOI : 10.1051/hel/2016380206

Disponible en ligne sur :
www.hel-journal.org

De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire

Rachel Panckhurst*, Mathieu Roche/***, Cédric Lopez****, Bertrand Verine*, Catherine Détrie*, Claudine Moïse*******

*Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3,

**UMR TETIS (Cirad, CNRS, Irstea, AgroParisTech), Montpellier,

***LIRMM (Univ. Montpellier, CNRS), Montpellier,

****R&D Viseo Technologies, Grenoble,

*****Lidilem, Université Grenoble Alpes

Résumé

Nous présentons notre approche fondée sur les données authentiques, en nous concentrant sur des recherches récentes, portant sur le recueil, le traitement et l'analyse d'un grand corpus de SMS en français, intitulé *88milSMS* (http://88milsms.huma-num.fr/, Panckhurst, Détrie, Lopez, Moïse, Roche, Verine, 2014), incluant un questionnaire sociolinguistique soumis aux donateurs au moment de la collecte ainsi que leurs réponses. Puis nous expliquons pourquoi, dans une démarche pluridisciplinaire (située entre sciences du langage, informatique et traitement automatique du langage naturel), nous avons décidé de fournir à la communauté scientifique et au grand public le corpus de SMS.

Mots-clés

Corpus, SMS, pluridisciplinarité, données authentiques, traitement automatique du langage naturel (TALN), discours électronique médié, logiciel d'anonymisation, dictionnaires électroniques, alignement

Abstract

This article highlights an approach based on authentic data, by focusing on recent research related to collection, processing and analysis of a large French text-message corpus, entitled *88milSMS* (http://88milsms.huma-num.fr/, Panckhurst, Détrie, Lopez, Moïse, Roche, Verine, 2014), including a sociolinguistic questionnaire submitted to donors (with their answers). The authors, using a pluridisciplinary approach (linguistics/language sciences, computer science, Natural Language Processing), explain why they chose to give the scientific community and the general public access to the SMS corpus.

Keywords

Corpus, SMS, pluridisciplinarity, authentic data, natural language processing (NLP), mediated electronic discourse, anonymisation software, electronic dictionaries, alignment

À Augusta Mela,
en mémoire de son œuvre interdisciplinaire entre linguistique et informatique

1. INTRODUCTION

En deux décennies, la constitution de corpus numérisés ou nativement numériques est devenue monnaie courante, et cette accessibilité massive constitue, en soi, une nouveauté. Les données authentiques existent, entre autres, sous la forme de courriels, forums, clavardages, blogues et autres réseaux sociaux, et, plus récemment de SMS. L'avantage est qu'elles sont facilement exploitables par les chercheurs, et qu'elles permettent à la fois l'observation, la fouille et l'analyse des pratiques et des usages (novateurs ou non) des scripteurs (*cf.* Cori *et al*. 2008, pour un débat sur la place des corpus en sciences du langage).

Dans le cadre de cet article, nous présenterons notre approche fondée sur les données authentiques, en nous concentrant sur des recherches récentes, portant sur le recueil, le traitement et l'analyse d'un grand corpus de SMS en français, intitulé *88milSMS*, incluant un questionnaire sociolinguistique soumis aux donateurs au moment de la collecte ainsi que leurs réponses. Puis, nous expliquerons pourquoi, dans une démarche pluridisciplinaire, nous avons décidé de fournir à la communauté scientifique et au grand public le corpus de SMS.

2. COLLECTE ET ANONYMISATION DE DONNÉES AUTHENTIQUES

En 2004, des universitaires belges ont lancé un projet international, *sms4science* (http://www.sms4science.org, Fairon *et al.,* 2006), afin de recueillir, organiser en une grande base de données mondiale et analyser des SMS authentiques, pour mieux comprendre comment la langue évolue : pour saisir dans toutes ses nuances le dynamisme linguistique, les processus de création lexicale, grammaticale, scripturale, notamment, très nombreux dans les SMS. Pourquoi un grand corpus de SMS ? Parce que seule une observation longitudinale d'un corpus de cette taille permet de faire émerger ce qui fonde l'échange tel qu'il est pensé dans les SMS. L'empathie et la complicité avec son autre, pensé comme un même, quels que soient les textoteurs, sont repérables à de nombreux marqueurs, tels les désignants affectifs, les jeux phoniques constants, les jeux sur les registres spécifiques, etc. Ceux-ci laissent affleurer une dynamique intersubjective privilégiant la consonance et la synchronisation des points de vue, soit une façon d'être spécifique qui privilégie le clan à l'autonomie individuelle. Ces observables, *in fine,* dessinent une photographie assez précise des relations humaines entre pairs dans les années 2010 (Cougnon et Fairon, 2014, Cougnon, 2015, Panckhurst *et al*. 2013, 2014a).

Cinq collectes ont eu lieu[1], puis, à l'automne 2011, plus de 93 000 SMS authentiques ont été recueillis auprès du grand public par notre groupe de chercheurs dans la Région Languedoc-Roussillon (projet *sud4science LR*, http://www.sud4science.org, Panckhurst *et al*. 2013). Plus de 88 000 SMS seront finalement conservés[2] et mis à disposition après divers prétraitements. Le corpus intégral, intitulé *88milSMS,* deux échantillons (100 SMS annotés, 1 000 SMS transcodés en français standardisé), un questionnaire sociolinguistique soumis aux donateurs, et leurs réponses, sont désormais téléchargeables (sur la grille de services d'Huma-Num : http://88milsms.huma-num.fr/, Panckhurst *et al*. 2014a).

L'élaboration de notre corpus a nécessité le développement d'un logiciel d'anonymisation semi-automatique, *Seek&Hide* (Accorsi *et al*. 2014, Patel *et al*., 2013), que nous décrivons ci-dessous.

Après l'étape de la collecte (voir Panckhurst *et al*., 2014b, pour plus de détails), en raison des aspects juridiques liés à la protection de la vie privée des personnes, tous les SMS du corpus *88milSMS* ont été anonymisés[3] de manière semi-automatique, en plusieurs étapes. Le logiciel *Seek&Hide* avait pour double tâche d'anonymiser le corpus et de fournir aux annotateurs humains une interface en ligne agréable à utiliser. Il s'appuie sur des méthodes de TALN, incluant des approches d'apprentissage automatique, et il propose une page web sécurisée accessible pour les annotateurs. Le but du logiciel est de faciliter l'expertise et de traiter une quantité importante de données.

L'approche développée se décline en trois phases :

1. une phase automatique, pendant laquelle tous les mots du corpus ont été automatiquement confrontés à un lexique électronique[4], ceux qui ne présentaient aucune ambiguïté étant anonymisés : par exemple, un prénom comme *Cédric* est automatiquement anonymisé ; un nom commun comme *crayon* ne l'est pas ;

1 Par ordre chronologique : Cental, Université catholique de Louvain, Belgique (75 000 SMS, corpus final : 30 000 SMS, 2004, http://www.sms4science.org, Fairon *et al*., 2006, Cougnon et Fairon, 2014, Cougnon, 2015), Île de la Réunion (20 000 SMS, 2008, http://www.lareunion4science.org/, Cougnon et Ledegen, 2010), Suisse (24 000 SMS, 2009-2010, http://www.sms4science.ch/, Dürscheid et Stark, 2011), Quebec (5 000 SMS, 2010, http://www.texto4science.ca/, Langlais *et al.,* 2012), region Rhône-Alpes (22 000 SMS, 2010, http://www.alpes4science.org/, Antoniadis *et al,* 2011)

2 Les SMS supprimés incluent : des doublons, des messages envoyés par la responsable du projet aux participants, des SMS automatiques envoyés par l'opérateur téléphonique au téléphone portable utilisé pour la collecte (prêté aux chercheurs par l'entreprise iTribu : http://www.itribustore.fr/), des textos en provenance de l'étranger, des messages envoyés par des personnes n'ayant pas rempli le formulaire de consentement, des messages publicitaires.

3 Les 10 étiquettes d'anonymisation, par ordre de fréquence d'utilisation dans le corpus sont : PREnom (10 905 étiquettes), SURnom (1 042), NOM (785), TELéphone (123), LIEu (102), ADResse (85), MARque (58), CODe (50), MEL (adresse électronique, 27), URL (13).

4 Le lexique électronique utilisé était le LEFFF (*Lexique Électronique des Formes Fléchies du Français*, Sagot, 2010).

2. une phase semi-automatique pour les mots ambigus (*Pierre*[5]=prénom, *pierre*=nom commun) ou inconnus (*Namrata*=prénom inconnu du lexique utilisé). Ceci s'effectue à travers une interface web sécurisée qui met en relief les éléments nécessitant une expertise. Cette mise en valeur facilite significativement le travail de l'annotateur ;
3. une phase de validation, consistant en une acceptation des SMS anonymisés automatiquement ou une modification d'une anonymisation appliquée par l'outil lors de la phase automatique.

Seek&Hide a automatiquement anonymisé 72 % du corpus *88milSMS* ; les 28 % restants ont été soumis à la phase semi-automatique. La phase de validation consiste à lire tous les SMS anonymisés de manière automatique par *Seek&Hide*, afin de vérifier s'ils ont été correctement anonymisés. Trois cas de modification éventuelle ont été identifiés par les annotateurs :

1. anonymisation automatique à enlever ;
2. anonymisation manquante à insérer ;
3. balises d'anonymisation à remplacer.

Les annotateurs humains peuvent donc retirer, ajouter, modifier les étiquettes précédemment insérées de manière automatique par le logiciel. Afin d'évaluer la répartition des différents cas, nous avons isolé un échantillon de 20 000 SMS. Sur ces données, seules 358 modifications (1,79 % du corpus) – réparties de la manière suivante : 66 % (cas 1), 29 % (cas 2), 5 % (cas 3) – ont été apportées. Notons enfin qu'à ce stade, les experts peuvent décider de noter certains SMS comme devant être supprimés du corpus si ceux-ci contiennent des propos inacceptables au regard de la loi. L'étape suivante du processus est liée au transcodage, c'est-à-dire au passage d'un SMS 'brut' anonymisé à un SMS en français 'standardisé', avec une éventuelle annotation linguistique.

3 Transcodage et annotation

Suite à l'anonymisation, les SMS sont prêts à être transcodés en français standardisé afin de permettre d'éventuels traitements ultérieurs en linguistique-informatique (incluant des analyseurs morphosyntaxiques). L'idée est de restituer l'orthographe et la grammaire afin de faciliter la fouille et la compréhension, mais non d'« injecter » des éléments supplémentaires.

Le transcodage peut être utile pour le grand public, ou pour ceux qui veulent lire et comparer rapidement les SMS bruts anonymisés et transcodés, à des fins de recherche. Cependant, d'un point de vue linguistique, il est extrêmement difficile

5 Dans l'écriture SMS, les majuscules sont fréquemment remplacées par des minuscules, d'où le problème d'ambiguïté pour le traitement automatique.

de procéder à un transcodage qui convienne à tous, car les interprétations sont nombreuses et variées. Prenons un exemple pour illustrer ce propos.

SMS brut anonymisé (n° 22446 du corpus 88milSMS) :
« En fait c rien de spécial, jprends juste un peu de recul et jcomprends pas ce que jfous là, fac, psycho, montpellier, pourquoi simplement je vis, enfin bref rien de grave. Qu'est ce qui cloche chez toi ? »
SMS anonymisé et transcodé en français standardisé :
En fait **c'est** rien de spécial, **je** prends juste un peu de recul et **je** comprends pas ce que **je** fous là, fac, **psychologie**, **Montpellier**, pourquoi simplement je vis, enfin bref rien de grave. Qu'est-ce qui cloche chez toi **?**

Exemple : transcodage

Dans l'exemple ci-dessus, on n'ajoutera pas la particule de négation, *ne*/*n'*. On n'« injectera » pas non plus des éléments prépositionnels ou des déterminants (« *à la* fac », « *en* psychologie », « *à* Montpellier »), car le traitement automatisé demeure possible sans ces informations. En revanche, pour des formes abrégées, agglutinées, etc., on transcode en français standardisé ('c' => 'c'est' : ici, il s'agit d'une *abréviation sémantisée*, lorsqu'un mot est réduit à l'initiale et seul le co(n)texte permet de déterminer de quel mot il s'agit ; agglutinations : *jprends, jcomprends, jfous*) pour qu'un analyseur morphosyntaxique soit à même de traiter automatiquement la phrase. La forme en apocope « fac » demeure telle quelle dans la version transcodée, car nous avons décidé de valider le transcodage en lien avec les informations apparaissant au sein du *Petit Robert* en ligne (PR) : si une entrée dictionnairique existe, elle n'est pas transcodée dans sa forme entière (« fac » demeure intact, mais « psycho » sera transcodé en « psychologie », car si l'élément « psycho- » existe effectivement dans le PR, l'apocope qui renvoie à « psychologie » n'y figure pas). Par ailleurs, lorsque la ponctuation est présente, les normes typographiques sont rétablies pour le français, ici sont réintroduites la lettre 'M' majuscule pour le nom de ville Montpellier et l'espace absent avant le point d'interrogation final. Cet exemple de transcodage donne un aperçu de la difficulté de cette opération.

Six étudiants (en Master d'informatique) ont travaillé sur le transcodage. Ils ont étudié la faisabilité d'une méthode d'alignement des SMS pour faciliter le passage du SMS brut anonymisé au SMS transcodé en français standardisé, et ils ont proposé un modèle pour une interface en ligne afin de faciliter le travail de l'annotateur humain. Le modèle d'alignement incluant une interface s'intitule AlignSMS (*cf.* Lopez *et al*. 2014). Lorsque des étudiants (en Master de Sciences du Langage) ont procédé à un test de transcodage sur un échantillon de 1000 SMS, ils ont conclu que l'écriture SMS étudiée à travers le travail de transcodage « contient

une réelle forme de créativité […] imprégnée des personnalités des émetteurs et récepteurs des SMS (de leur passé, de leur humour, de leur quotidien), des effets de modes, des contextes actuels. » (Dalle *et al*., 2013, Dossier étudiant).

La suite du processus mis en place concerne l'annotation linguistique, qui consiste à apposer des balises associées à certains mots contenus dans un SMS, permettant de retrouver facilement une information. Cette annotation peut être utile pour des chercheurs qui souhaitent étudier certains aspects, en accédant rapidement aux données qui les intéressent. Lors du projet *sud4science LR*, nous avons invité les acteurs des collectes précédentes, dans le cadre de *SMS4science*, à présenter les balises utilisées pour l'annotation de leurs corpus de SMS. Une harmonisation générale nous a ensuite permis de réduire le nombre de balises précédemment utilisées. Huit balises ont été retenues : TYPographie, MODification, GRAmmaire, BINettes, ABSence, LANgue, ORThographe, DIVers. Les exemples figurant dans le tableau 1 correspondent chacun à une balise spécifique qui peut être appliquée manuellement. Nous n'indiquons pas l'annotation complète pour la totalité du SMS, mais simplement la balise concernée, afin d'éclairer la lecture.

TABLEAU 1
Exemples de SMS annotés

SMS	BALISES	AVANT BALISAGE	APRÈS BALISAGE
n° 6 885	TYP[6]	Zorro est arrive, sans s'presse […]	Zorro est <TYP_arrivé> arrive, sans s'presse […]
n° 4 360	MOD[7]	[…] Oui, j sui zalé ! […]	[…] Oui, <MOD_j'y> j <MOD_suis> sui <MOD_allé> zalé ! […]
n° 5 536	GRA[8]	Cc tu va mieux. Mam ma dis ke tèté retmbè malade. Et bb ? Bisx	Cc tu <GRA_vas> va mieux. Mam ma <GRA_dit> dis ke tèté retmbè malade. Et bb ? Bisx
n° 6 887	BIN[9]	Au dos d'son beau tornado J elle est trop bien cette prof, chui amoureux d'elle ^^	Au dos d'son beau tornado <BIN> J elle est trop bien cette prof, chui amoureux d'elle <BIN> ^^
n° 19 621	ABS[10]	[…] je met tout ça de coté et peux tout encaisser juste pour toi . […]	[…] je met tout ça de coté et <ABS_je> peux tout encaisser juste pour toi . […]

6 <TYP> typographie : ponctuation, symboles mathématiques, signes diacritiques (accents, etc.), nombres, format des heures, ponctuations ou symboles inattendus, signe (&, chevrons, parenthèses), respect de la casse (majuscules/minuscules), mise en page.

7 <MOD> modification (soit en réduction, soit en augmentation, soit en remplacement de caractères, abrègements et abréviations, acronymes, sigles et abréviations, répétition de lettres, transformations phonétiques, interjections et onomatopées…) : *ht* (acheter), *pr* (pour), *c* (s'est, c'est, ces…), *dcd* (décidé)…

8 <GRA> accords : *il viens* (il vient), syntaxe : *si j'aurais su, je serais pas venu* (si j'avais su, je ne serais pas venu), etc.

9 <BIN> binettes/frimousses/émoticônes/smileys :) ^^ :p ;) :d <3 :-) xd :(:/

10 <ABS> absence/ellipse : négation, pronoms, éléments manquants faciles/difficiles à identifier, etc.

n° 43 133	LAN [11]	if(ce_soir == film) {get_commande;} else {set_tagueule;} return "bisous"	<LAN> if(ce_soir == film) { <LAN> get_commande;} <LAN> else { <LAN> set_tagueule;} <LAN> return "bisous"
n° 19 621	ORT [12]	[…] notre couple sera tel un rosau à jamais se casser . […]	[…] notre couple sera tel un <ORT_rosseau> rosau à jamais se casser […]
n° 4 671	DIV [13]	Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe	<DIV> Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe

Toutes ces étapes (acquisition, anonymisation, transcodage, annotation) ont nécessité des choix théoriques qui ont été effectués dans une démarche pluridisciplinaire.

4. Choix théoriques dans une démarche pluridisciplinaire

Faut-il fournir à la communauté un corpus brut uniquement anonymisé ou doit-on aussi le transcoder et l'annoter entièrement ? Y a-t-il un consensus entre les chercheurs ? Le transcodage et l'annotation sont-ils neutres ? Ou sont-ils liés à un cadre interprétatif ? Comme indiqué précédemment, nous avons eu une discussion scientifique avec tous les partenaires de *sms4science* afin d'arrêter les balises ensemble. Par ailleurs, nos séminaires scientifiques pluridisciplinaires ont soulevé quantité de questionnements par rapport à ces deux étapes.

Par exemple, dans le tableau 2 ci-dessous, les psychologues Goumi et Bernicot (2011) montrent des exemples de messages originaux et leurs transcriptions :

Tableau 2
Transcription selon Goumi et Bernicot (2011)

Âge du participant	Genre	Message Original	Transcription conservant les formes de l'oral	Transcription rétablissant l'écrit académique
11;2	F	Cc alor ta dmende pour ce soir	Coucou alors t'as d'mandé pour ce soir	Coucou alors as-tu demandé pour ce soir ?
11;11	M	Jen sai rien chui pa aller ché le toubib	J'en sais rien j'suis pas allé chez le toubib	Je n'en sais rien je ne suis pas allé chez le toubib.
13;5	M	Ouai jen ai jamai vu des vivan mes kan il sn mor ca mderange pa	Ouais j'en ai jamais vu des vivants mais quand ils sont morts ça m'dérange pas	Ouais je n'en ai jamais vu des vivants mais quand ils sont morts ça ne me dérange pas.
13;8	F	Lèa t c se kil i a fair en techno	Léa t'sais ce qu'il y a à faire en techno	Léa sais-tu ce qu'il y a à faire en technologie ? [17]

11 <LAN> (contact de langues, emprunts, régionalismes, néologismes, verlan, argot, etc.)
12 <ORT> (uniquement l'orthographe lexicale : erreurs de saisie, interversion de lettres, etc.)
13 <DIV> (dans le cas où aucune autre balise ne semble convenir).

Si ces transcriptions peuvent convenir pour des psychologues, elles soulèveraient sans doute un débat pour des linguistes, notamment la transformation de formes de modalité non marquée en formes interrogatives incluant des inversions pronom sujet + verbe, l'ajout de la particule « ne » et la transformation de formes contractées/élidées/agglutinées : « ta => t'as => tu as », « Jen => J'en => Je n'en », « chui => j'suis => je ne suis ».

Pour prendre un autre exemple, au sein du consortium *Corpus Écrits*, Chanier *et al*. (2014) ont clairement stipulé qu'il faut nécessairement rendre disponible un corpus en deux versions : v1) le corpus brut sans segmentation et sans annotation et v2) le corpus avec ses annotations[14].

À travers les exemples du § 3, nous constatons qu'il est extrêmement difficile, voire impossible, de proposer un transcodage et une annotation linguistique standardisés consensuels. Mais ce n'est pas parce que cela prendrait un temps conséquent que nous avons décidé d'y renoncer. Pour nous, il s'agit plutôt d'une position théorique. Le transcodage et l'annotation suscitent des désaccords théoriques, que ce soit au sein d'une même discipline, ou de manière interpluridisciplinaire. Nous considérons qu'annoter n'est pas une opération descriptive neutre. Elle relève nécessairement d'un cadre interprétatif. On comprend alors pourquoi elle peut ne pas faire consensus, parce qu'il y a des cadres théoriques différents, des démarches pluridisciplinaires distinctes, des questionnements scientifiques variés, etc. Nous pensons qu'il est préférable que les chercheurs prennent en charge leurs transcodage et annotation en fonction de leur(s) propre(s) questionnement(s).

Depuis le début du projet *sud4science LR*, notre démarche se veut résolument pluridisciplinaire : chercheurs et étudiants en sciences du langage et en informatique (TALN) travaillent main dans la main, et en coordination avec un juriste institutionnel. Cette recherche n'aurait pu être menée à bien sans cette approche pluridisciplinaire – qui, selon nous, reste trop rare dans la recherche publique française actuelle. Elle nous a permis d'impliquer des étudiants (inscrits en Master à l'époque), dès la collecte des SMS, puis pendant tout le déroulement du projet. Quant aux 6 chercheurs participant directement au projet (4 linguistes et 2 informaticiens), nos spécialités pluridisciplinaires recouvrent les thématiques suivantes : analyse des discours oraux, écrits et médiés par les technologies, TALN, fouille de textes, recherche d'information.

Dans le cadre de *sud4science LR*, nous avons organisé deux années de séminaires de recherche et des journées d'étude, à la MSH-M (http://www.msh-m.tv/spip.php?rubrique138), pour lesquels nous avons invité des collègues émanant

14 "Dissemination will take two different forms: one version of a corpus with the 'raw' text without any tokenization and annotation (v1), and a second version of the same corpus with the annotations (v2)." (Chanier *et al*. 2014, p. 2)

de différentes disciplines (sciences du langage, informatique, psychologie, information-communication, notamment) à exposer leurs travaux de recherche à Montpellier. Grâce à cette collaboration fructueuse et à la confrontation systématique des différents points de vue – très précieuse pour ce projet – nous avons abouti à des choix théoriques déterminants.

Notons enfin que les méthodes informatiques mises en place sont clairement semi-automatiques, c'est-à-dire qu'elles nécessitent d'intégrer les connaissances expertes, en l'occurrence en sciences du langage, dans le processus. C'est ainsi que les différents logiciels développés ont été conçus de concert entre informaticiens et linguistes. En effet, pour une utilisation adaptée, ces derniers ne doivent pas seulement maîtriser les questions purement techniques mais ils doivent également appréhender les principes des algorithmes utilisés ou conçus et les différents paramètres associés.

5. Conclusion

In fine, les chercheurs du projet proposent au public, via un téléchargement direct (sur la grille de services d'Huma-Num : http://88milsms.huma-num.fr/, Panckhurst *et al*. 2014a), le corpus intégral, intitulé *88milSMS* entièrement anonymisé (format .ods), deux échantillons (SMS annotés, 1 000 SMS transcodés en français standardisé), le questionnaire sociolinguistique soumis aux donateurs, et leurs réponses. Nous avons proposé en 2015 une version du corpus encodé en XML dans le cadre d'une contribution Dariah[15]. Cela est déterminant pour un archivage à long terme au CINES[16]. Fin 2016, notre corpus 88milSMS – dans une nouvelle version structurée en XML/TEI – a intégré la plateforme Ortolang (Panckhurst et al. 2016 ; cmr-88milsms-tei-v1; https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1).

Outre les étapes d'anonymisation, de transcodage, d'annotation fondées, en partie, sur des techniques de TAL, d'autres applications informatiques sont envisageables : élaboration de lexiques transcodés français standardisé => SMS ou vice versa, consultables en ligne ; mise en place de systèmes de vocalisation des SMS à l'usage de personnes aveugles ou de personnes momentanément empêchées de consulter leur écran de téléphone – en situation de conduite, etc.

Par ailleurs, les analyses en sciences du langage fusent depuis la collecte des SMS authentiques. Celles-ci entrent dans un cadre « guidé par corpus » ('corpus driven') (Tableau 3) ou « fondé sur corpus » ('corpus based') (Tableau 4) :

15 *Digital Research Infrastructure for the Arts and Humanities* : Dariah-fr, http://www.dariah.fr/
16 *Centre Informatique National de l'Enseignement Supérieur* : https://www.cines.fr/

TABLEAU 3
Guidé par corpus/corpus driven

SMS brut anonymisé	N° du SMS
Wesh ma vache :-) je lol	13 213
On va rater les bandes annonces espèce de nazgul en tongue lol	902
Wesh gros ! Et bien je sais pas si je pourrai parce que jai ptetre cours, enfin jte dirai ca ce soir ^^	38 593
Espèce de gloutonne des validations d'acquis^^	48 178
Lol, non j't-ai pas oublié !	59 947
Wesh trkl tkt ;) tu fou quoi ?	692

Détrie et Verine (2015) ont découvert un phénomène qu'ils ont dénommé « insultes-mots doux » : des SMS incluant des « insultes » (« ma vache », « espèce de nazgul en tongue », « gros », « gloutonne »), radoucies par d'autres éléments textuels (« je lol »), ou des binettes/emojis (« ^^ », « :-) »), etc. Voyant l'utilisation de « wesh », « trkl », « tkt », « lol », etc. dans différents contextes, Moïse (2013) a décidé d'approfondir l'étude sociolinguistique des notions de « norme » et de « faute » au sein de l'écriture SMS.

TABLEAU 4
Fondé sur corpus/corpus-based

	SMS brut anonymisé	N° du SMS
« er » au lieu de « é »	Moi non plus comment ça se fait que tu a changer de Num ?	86 936
	Ehh la j'en peux plus mdrr tu sais je sais plus si je te l'avais dit mais on a enfin acheter la machine a laver aujourd'hui et donc je dis : elle est belle ? Quel est blanche ? elle marche bien ? Et la mon pere me dit : elle est verte !! Nan mais VERTE quoi ! MDRRRR	20 475
« é » au lieu de « er »	J'arive ps a tlavoué mé jsuis tbr sr tn charme	8 530
	Oki :) jvais chercher ca, jte tiens au jus si j'ai réussi a réservé	63 150
mélange des deux	Je me suis levé, j'ai mangé, j'ai révisé mes cours de Jsp, j'ai jouer sur l'ordi, là, je cherche des applis a télécharger sur mon iPod, et toi?	13 001
	Ma mere a retrouver la housse :) t as retrouve tes rido?? Ta commencer a tt regroupe ? :)	41 203

Panckhurst s'est interrogée sur le remplacement de l'accent aigu « é » par « er » ou vice versa : la fouille du corpus a révélé effectivement un grand nombre de formes infinitives en remplacement de formes participiales. Le tableau 4 montre les résultats de recherches « fondées sur corpus ». L'exemple suivant, trouvé au hasard d'une lecture, qui constituerait donc plutôt une recherche « guidée par le corpus », semble faire pencher la balance en direction d'une réponse ergonomique (l'accent étant le résultat d'un appui long sur la touche en question) : « Ok t a coter d ki? » (n° 71 634).

Les exemples des tableaux 3 et 4 montrent l'importance d'effectuer un va-et-vient constant entre les hypothèses et l'observation des données. Cela constitue le point essentiel de notre démarche et rejoint la position de Tagg (2009) et de Cougnon (2015).

Le corpus *88milSMS* et les collectes *SMS4science* constituent effectivement des « clichés » d'une époque (2004-2011) déjà révolue, tant les technologies mutent constamment. La déclinaison perpétuelle des supports (SMS et messages courts/instantanés sur tablettes, ordinateurs, etc.) et des modes rédactionnels (écriture intuitive, correcteurs intégrés, etc.) feront très certainement apparaître dans les futures collectes des phénomènes inédits et de nouvelles modes sociétales.

Aux journalistes qui nous demandent souvent si les SMS constituent un phénomène de mode, nous répondons que ces derniers ont encore un bel avenir devant eux : de la créativité lexicale/scripturale à la survie (démontrée par les attentats de Charlie Hebdo en janvier 2015). Nous, linguistes et informaticiens, unis dans une démarche pluridisciplinaire fructueuse, continuerons à nous passionner pour le corpus *88milSMS* de longues années encore.

Éléments bibliographiques

Accorsi, Pierre, Patel, Namrata, Lopez, Cédric, Panckhurst, Rachel, Roche, Mathieu, 2014. "Seek&Hide : Anonymising a French SMS corpus using natural language processing techniques", in Cougnon, Louise-Amélie, Fairon Cédric (eds), *SMS Communication. A Linguistic Approach*, Amsterdam/Philadelphia, John Benjamins, 11-28.

Antoniadis, Georges, Chabert, Gaëlle, Zampa, Virginie, 2011. « Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine », Communication, 79e Colloque Acfas, Sherbrooke, 9-10 mai 2011.

Chanier, Thierry, Poudat, Céline, Sagot, Benoît, Antoniadis, Georges, Wigham, Ciara R., Hriba, Linda, Longhi, Julien & Seddah, Djamé, 2014. "The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres", *Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics*, *JLCL* (*Journal of Language Technology and Computational Linguistics*), 1-31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

Cori, Marcel, David, Sophie, Léon, Jacqueline, (dir.), 2008. *Construction des faits en linguistique : la place des corpus*, *Langages*, 171, Paris, Larousse, septembre 2008.

Cougnon, Louise-Amélie, 2015. *Langage et sms. Une étude internationale des pratiques actuelles. Cahiers du Cental*, 8, Louvain-la-Neuve, Presses universitaires de Louvain.

Cougnon, Louise-Amélie, Fairon, Cédric, (éd.), 2014. *SMS Communication. A linguistic approach*, Amsterdam/Philadelphia, John Benjamins.

Cougnon, Louise-Amélie, Ledegen, Gudrun, 2010. « "c'est écrire comme je parle". Une étude comparatiste de variétés de français dans l'"écrit sms" », *Les voix des Français. Modern French Identities*, 2, 94, 39–57.

Dalle, Laurine, Faisant, Joséphine, Jaffal, Marie, de Martino, Véronique, 2013. « Transcodage de 1 0001 SMS », Dossier étudiant, Master 1, Sciences du Langage parcours DiMIP (EAD), Université Paul-Valéry Montpellier 3.

Détrie, Catherine, Verine, Bertrand, 2015. « Quand l'insulte se fait mot doux : la violence verbale dans les SMS », Tuomarla, Ulla *et al*., (eds.), *Dialogic Language use 3. Dimensions du dialogisme 3. Dialogischer Sprachgebrauch* 3, Helsinki, Société néophilologique, 59-71.

Dürscheid, Christa, Stark, Elisabeth, 2011. "sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland", Thurlow, Crispin, Mroczek, Kristine (eds), *Digital Discourse. Language in the New Media,* Oxford, Oxford University Press, 299-320.

Fairon, Cédric, Klein, Jean René, Paumier, Sébastien, 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*, Louvain-la-Neuve, Presses universitaires de Louvain, Manuel+CD-Rom, http://www.smspourlascience.be/

Goumi, Antonine, Bernicot, Josie, 2011. « Un corpus de SMS produits par de jeunes adolescents : méthode de recueil et premières données », séminaire invité, projet *sud4science LR*, MSH-M, 15/3/2011.

Langlais, Philippe, Drouin, Patrick, Paulus, Amélie, Rompré Brodeur, Eugénie, Cottin, Florent, 2012. "Texto4Science: a Quebec French Database of Annotated Short Text Messages" Proceedings, International conference on *Language Resources and Evaluation* (*LREC*), 1047-1054, http://www.lrec-conf.org/proceedings/lrec2012/pdf/413_Paper.pdf

Lopez, Cédric, Bestandji, Reda, Roche, Mathieu, Panckhurst, Rachel, 2014. "Towards Electronic SMS Dictionary Construction : An Alignment-based Approach", Proceedings, International conference on *Language Resources and Evaluation* (*LREC*), 2833-2838, www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf

Moïse, Claudine, 2013. « "Lol non tkt on ta pas oublié" Rapports à la norme et valeurs de la « faute » dans l'écriture SMS (projet et corpus *Sud4science*). Réflexions sociolinguistiques. », Conférence plénière, Colloque « Si j'aurais su, j'aurais pas venu ! Linguistique des formes exclues : description, genre, épistémologie », Université Libre de Bruxelles, 20-22 juin.

Panckhurst, Rachel, Détrie, Catherine, Lopez, Cédric, Moïse, Claudine, Roche, Mathieu, Verine, Bertrand, 2013. « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS », *Épistémè* – revue internationale de sciences sociales appliquées, 9 : *Des usages numériques aux pratiques scripturales électroniques*, 107-138.

Panckhurst, Rachel, Détrie, Catherine, Lopez, Cédric, Moïse, Claudine, Roche, Mathieu, Verine, Bertrand, 2014a. "88milSMS. A corpus of authentic text messages in French" produit par l'Université Paul-Valéry Montpellier 3 et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8.

Panckhurst, Rachel, Détrie, Catherine, Lopez, Cédric, Moïse, Claudine, Roche, Mathieu, Verine, Bertrand, 2014b. « Un grand corpus de SMS en français : 88milSMS », *La lettre de l'InSHS*, *La Tribune d'Huma-Num*, pages 22-25, septembre 2014, http://www.cnrs.fr/inshs/Lettres-information-INSHS/lettre_infoinshs_31hd.pdf

Panckhurst, Rachel, Détrie, Catherine, Lopez, Cédric, Moïse, Claudine, Roche, Mathieu, Verine, Bertrand, 2016. "88milSMS. A corpus of authentic text messages in French"

(new version of the ISLRN 024-713-187-947-8 corpus). In Chanier, Thierry, (ed), Banque de corpus CoMeRe. Nancy: Ortolang. [cmr-88milsms-tei-v1; https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1].

Patel, Namrata, Accorsi, Pierre, Inkpen, Diana, Lopez, Cédric, Roche, Mathieu, 2013. "Approaches of anonymisation of an SMS corpus", in *Computational Linguistics and Intelligent Text Processing*, 77-88, Berlin, Heidelberg, Springer Verlag.

Sagot, Benoît, 2010. "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French", Proceedings, International conference on *Language Resources and Evaluation* (*LREC*), Valletta, Malta, http://hal.inria.fr/inria-00521242/

Tagg, Caroline, 2012. *The discourse of text messaging: analysis of SMS communication*, New York, Continuum.

Soutiens/remerciements

Nous remercions la MSH-M (Maison des Sciences de l'Homme de Montpellier), la DGLFLF (Délégation générale à la langue française et aux langues de France) et le CNRS (PEPS ECOMESS, HuMaIn) qui ont soutenu ce travail. Nous remercions chaleureusement le correspondant informatique et libertés, CIL, Nicolas Hvoinsky de nous avoir accompagnés et conseillés sur le plan juridique, ainsi que sa directrice, Stéphanie Delaunay (DAJI, Université Paul-Valéry Montpellier) tout au long de notre projet. Nous remercions vivement nos étudiants stagiaires : Anthony Stifani (étudiant en Master Information et Communication à l'Université Paul-Valéry Montpellier 3), qui a manuellement analysé une partie des SMS, permettant ainsi d'évaluer le système d'anonymisation ; Pierre Accorsi et Namrata Patel (étudiants en Master d'Informatique à l'Université de Montpellier), qui ont développé le système informatisé *Seek&Hide*, permettant d'anonymiser le corpus ; Frédéric André, Yosra Ghliss, Camille Lagarde-Belleville et Michel Otell (étudiants en Master de Sciences du Langage à l'Université Paul-Valéry Montpellier 3) qui ont procédé à l'anonymisation manuelle en ligne à l'aide de *Seek&Hide* et à la vérification de l'anonymisation automatique du corpus ; Reda Bestandji, Ahmed Loudah, Aghiles Lounes, Zakaria Mokrani, Takfarinas Sider, Tarik Zaknoun (Master I Informatique, Spécialité : « Informatique pour les sciences », Université Montpellier) qui ont travaillé sur un système de transcodage automatique.