Partage d'expériences

Un grand corpus de SMS en français : 88milSMS

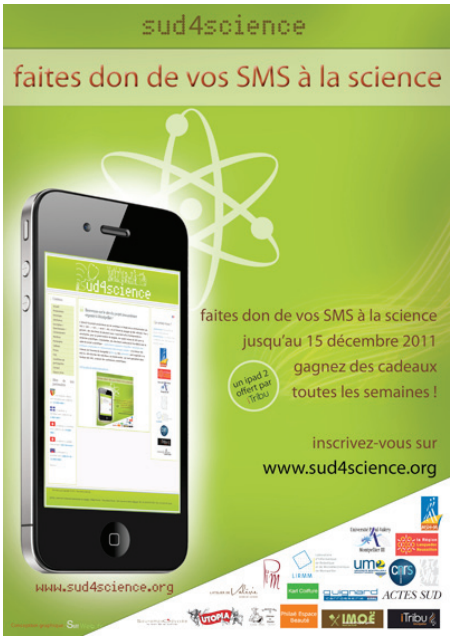
Contexte historique

Pour mieux comprendre comment la langue évolue, le linguiste recueille des données, les observe et les analyse. L'informaticien peut lui venir en aide en minimisant les traitements manuels ; la rencontre entre linguistique et informatique — ou le traitement (partiellement ou entièrement) automatique de la langue (TAL) — a lieu. Depuis plusieurs décennies, l'accès facilité aux données numériques permet la constitution aisée de corpus informatisés. Les données nativement numériques sont de plus en plus nombreuses et les recherches en humanités numériques foisonnent : le discours écrit des courriels, forums, *chats*, blogs, réseaux sociaux, etc. est passé au peigne fin par les chercheurs, car il révèle des usages émergents, novateurs, différents, spécifiques. Pour mieux comprendre les pratiques et les usages, les chercheurs s'intéressent aussi à la communication et aux discours qui circulent entre les individus se servant d'ordinateurs — la *communication médiée par ordinateur* ainsi que le *discours électronique médié* (Panckhurst, 1997, 2006[1]). Avec l'arrivée du téléphone portable, puis des SMS (services de messages succincts) au début des années 1990, un autre outil de médiation ouvrait aux chercheurs un grand terrain d'enquête, mais une difficulté *a priori* insurmontable persistait : comment recueillir ces données authentiques — écrites spontanément en situation réelle — sans qu'elles subissent aucune re-saisie au moment du recueil ?

Constitution d'une grande base de données mondiale de SMS

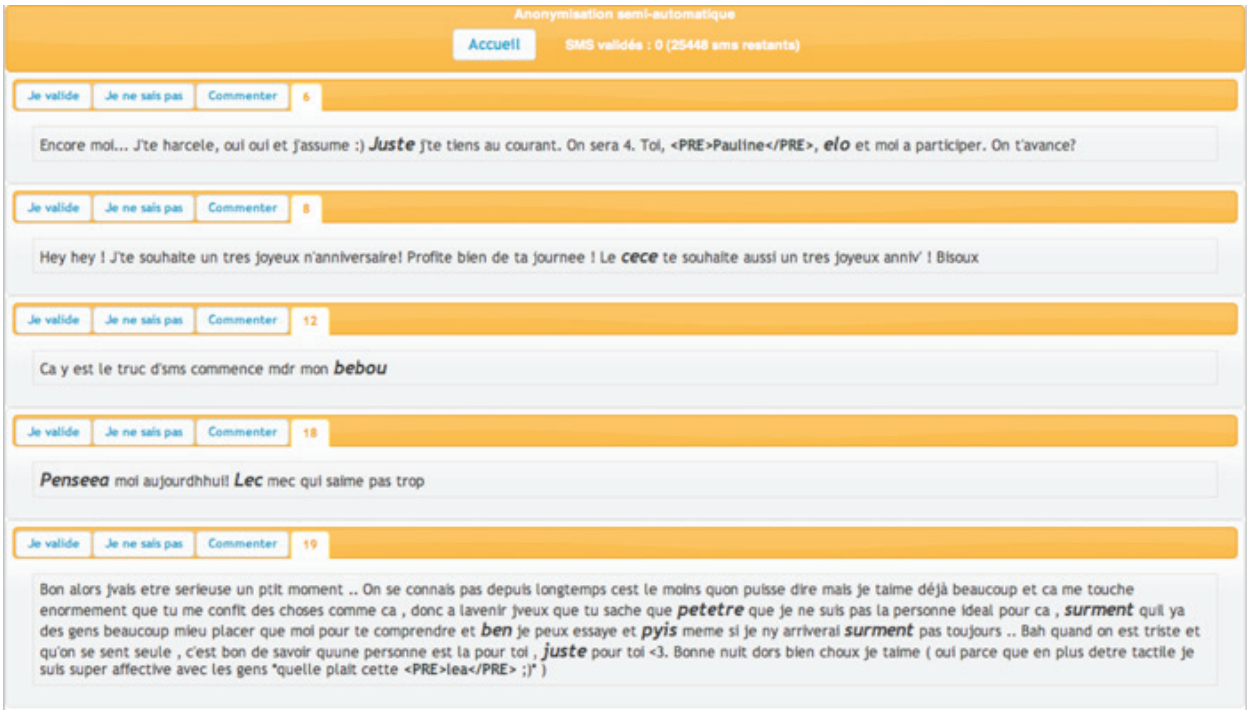
Jusqu'au début des années 2000, les analyses de SMS étaient restreintes, faute de collectes de quantité significative. En 2004, un groupe d'universitaires belges a lancé un projet international, intitulé *sms4science*, afin de recueillir, organiser (en une base de données mondiale) et analyser des SMS authentiques (Fairon et al. 2006, Cougnon, 2014[2]). S'en sont suivies d'autres collectes de SMS et l'initiative la plus récente pour le français est le projet *sud4science LR*. En trois mois, à l'automne 2011, plus de 93 000 SMS authentiques ont été recueillis auprès du grand public par un groupe de chercheurs dans la région Languedoc-Roussillon (Panckhurst et al. (2013), Panckhurst & Moïse (2014)[3]. Plus de 88 000 SMS seront finalement conservés et mis à disposition après divers prétraitements décrits ci-dessous.

Le téléphone comme outil de recueil



Le projet montpelliérain s'est distingué des collectes précédentes par la méthode de récolte. Après inscription et consentement légal en ligne, les participants donateurs de SMS, au moment de l'envoi de leur texto à autrui, pouvaient l'envoyer en copie aux chercheurs. Il était également possible de réexpédier des SMS (précédemment envoyés et contenus dans la mémoire du téléphone du scripteur) aux chercheurs. Le moyen utilisé ? Un smartphone[4], grâce auquel l'ensemble des textos a été recueilli pendant 13 semaines. Ce dispositif a été un véritable pari technique, car personne ne savait à l'avance si l'iPhone allait permettre un recueil de SMS très important, sans défaillir. En définitive, aucun problème n'est survenu. Chaque semaine, les SMS ont été copiés sur un disque dur externe, déconnecté d'Internet (pour des raisons juridiques). La grande base de

1. Panckhurst R. (1997), « La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? », Terminologies nouvelles, 17, 56–58.
Panckhurst R. (2006), « Le discours électronique médié : bilan et perspectives », in A. Piolat (Éd.). *Lire, écrire, communiquer et apprendre avec Internet*. Marseille : Éditions Solal, p. 345-366.
2. Fairon C., Klein J.-R., Paumier S., (2006), *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve, Manuel+CD-Rom.
Cougnon L.-A. (à paraître, 2014) *Langage et sms. Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.
3. Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., et Verine B. (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 R. et Moïse C., (2014), « French text messages. From SMS data collection to preliminary analysis », in SMS Communication. A Linguistic Approach, éd L.-A. Cougnon, C. Fairon, John Benjamins : Amsterdam/Philadelphia, p. 141-168.
4. L'entreprise iTribu, très enthousiaste à l'idée de participer à un projet de recherche universitaire, a prêté un iPhone aux chercheurs pour la durée de la collecte.



Interface de *Seek&Hide*

données (BD) en constitution devait rester dans son intégralité sur le téléphone également jusqu'à la fin de la récolte, afin d'assurer que la BD soit entière et homogène, avant transfert final. Depuis le début des collectes de SMS en 2004, il était très important que la méthode de recueil ne passe pas par une (re)saisie des données : seul ce type de transfert technologique était possible afin d'assurer que les données demeurent réellement authentiques.

Un logiciel pour l'anonymisation

Après l'étape de la collecte, en raison des aspects juridiques liés à la protection de la vie privée, tous les SMS du corpus « 88milSMS » incluant des prénoms, noms, surnoms, adresses, lieux, numéros de téléphone, codes, URL, marques, courriels, etc. ont été anonymisés de manière semi-automatique, en plusieurs étapes. Le logiciel *Seek&Hide* (Accorsi et al., 2014[5]), ayant pour double tâche d'anonymiser le corpus et de fournir aux annotateurs humains une interface en ligne agréable à utiliser, a été élaboré par des étudiants.

Ce logiciel s'appuie sur des méthodes de traitement automatique des langues. 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 :

▶ Phase automatique : traitement automatique des données (mots) qui ne présentent *a priori* aucune ambiguïté quant à leur interprétation (à anonymiser ou non). Par exemple, un prénom comme *Cédric* serait automatiquement anonymisé ; un nom commun comme *crayon* serait automatiquement écarté de l'anonymisation. Notons que des approches d'apprentissage automatique ont également été proposées (Patel et al. 2013[6]).

▶ Phase semi-automatique : traitement manuel de l'information nécessaire pour les SMS qui présentent des mots ambigus (*Pierre*=prénom, *pierre*=nom commun) ou inconnus (*Namrata* = prénom inconnu du dictionnaire utilisé). Ceci s'effectue à travers un système qui met en relief les éléments nécessitant une expertise. Cette mise en valeur facilite significativement le travail de l'annotateur.

▶ Phase de validation : relecture et validation des SMS anonymisés automatiquement ou modification d'une anonymisation appliquée par l'outil lors de la phase automatique (cf. cas 1 à 3 ci-dessous).

Seek&Hide a automatiquement anonymisé 72% du corpus « 88milSMS » ; les 28 % restants ont été soumis à une phase semi-automatique. Le logiciel propose une interface web sécurisée permettant aux annotateurs-experts linguistes de mener à bien la phase suivante, qui permet de désambiguïser les SMS et de décider si l'anonymisation doit ou non être effectuée. Par exemple, un prénom comme *Pierre* qui est également un nom commun, en minuscules, serait traité pendant cette deuxième phase, pour qu'un annotateur humain puisse décider, en fonction du cotexte/contexte si l'occurrence doit être ou non anonymisée.

La troisième phase de validation consiste en la lecture de tous les SMS (72% du corpus) anonymisés de manière automatique par *Seek&Hide*, afin de vérifier si tous les textos l'ont bel et bien été correctement. Trois cas de modification éventuelle ont été repérés par les annotateurs.

5. Accorsi P., Patel N., Lopez C., Panckhurst R., Roche M. (2014), « Seek&Hide : Anonymising a French SMS corpus using natural language processing techniques », in SMS Communication. A Linguistic Approach, éd L.-A. Cougnon, C. Fairon, John Benjamins : Amsterdam/Philadelphia, p. 11-28.
6. Patel N., Accorsi P., Inkpen D., Lopez C., Roche M. (2013) "Approaches of anonymisation of an SMS corpus", in Computational Linguistics and Intelligent Text Processing, pp. 77-88, Springer Verlag, Berlin, Heidelberg.

Cas 1 : anonymisation automatique à enlever

grace a lui on comprend trop bien franchement ke kiffe la physique cette *anne* meme si cest bien dur

Dans cet exemple, *grace* et *anne* ont été anonymisés, mais ce n'est pas une erreur du logiciel. Si le scripteur avait ajouté l'accent circonflexe, *Seek&Hide* n'aurait pas procédé à l'anonymisation en prénom pour *grâce* ; l'autre occurrence est *anne* au lieu d'*année*, qui n'est donc pas un prénom dans ce contexte.

Cas 2 : anonymisation manquante à insérer

Excuse pour c texto si tard c'était pour t dire q *mat* a u l permis bisous bisous

Mat est absent du dictionnaire de prénoms, puisqu'il s'agit d'un diminutif (e.g. *Mathieu*), et le nom commun (le *mat*) existe. Le logiciel avait donc ignoré cet élément.

Cas 3 : balises d'anonymisation à remplacer

Une *clio* noir phase 2 vendue par une amie dune collègue de boulot.

Clio est ici un nom de voiture de la marque Renault, et non un prénom. Il faut donc changer la balise de l'anonymisation.

Les annotateurs humains peuvent donc retirer, ajouter, modifier les étiquettes précédemment insérées de manière automatique par le logiciel[7]. À ce stade, ils peuvent également décider de noter certains SMS comme devant être supprimés du corpus si ceux-ci contiennent des propos inacceptables au regard de la loi.

L'opération totale d'anonymisation a nécessité 21 mois et a été réalisée grâce au travail de nombreux étudiants stagiaires. Si le travail avait été mené de manière entièrement manuelle, la périodicité aurait été augmentée de manière significative et le travail accompli aurait été très certainement moins fiable, tant la lecture humaine des SMS inflige une réelle surcharge cognitive.

Transcodage, alignement, annotation

Une fois l'anonymisation terminée, les SMS sont prêts à être transcodés en français standardisé afin de permettre d'éventuels traitements ultérieurs en linguistique-informatique (incluant des analyseurs morpho-syntaxiques). L'idée est de restituer l'orthographe et la grammaire afin de faciliter la compréhension, mais non d'« injecter » des éléments supplémentaires (cf. l'exemple 1 ci-dessous). Tous les SMS bruts anonymisés, un échantillon de 1000 SMS transcodés et un échantillon de 100 SMS annotés sont téléchargeables. Le transcodage est 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 de procéder à un transcodage qui convienne à tous, car les interprétations sont nombreuses et variées.

Exemple 1 : passage du SMS brut anonymisé au SMS transcodé

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 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 ?

Dans l'exemple 1, 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 pour qu'un analyseur morpho-syntaxique soit à même de traiter automatiquement la phrase. Dans cet exemple, l'apocope[8] « fac » demeure telle quelle dans la version transcodée, car les chercheurs ont décidé de valider le transcodage en lien avec les informations apparaissant au sein du *Petit Robert* en ligne 2014 (PR14) : 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 PR14, 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 l'espace absente avant le point d'interrogation final.

Ces choix ne conviennent pas nécessairement à tous, d'où l'importance de maintenir le lien visible entre la consultation du SMS brut anonymisé et le SMS anonymisé.

Des étudiants ont également travaillé sur le transcodage et l'alignement en explorant la faisabilité d'une méthode d'alignement des SMS pour faciliter le passage du SMS brut anonymisé au SMS transcodé en français standardisé. Ils ont ainsi proposé un modèle pour une interface en ligne afin d'aider le travail de l'annotateur humain. Le modèle d'alignement incluant une interface s'intitule AlignSMS (cf. Lopez et al. 2014[9]).

Les chercheurs ont renoncé à effectuer le transcodage sur l'ensemble du corpus pour deux raisons :
1) le temps très important exigé par cette tâche ;
2) la façon d'effectuer le transcodage susciterait vraisemblablement des désaccords théoriques.

7. Sur un échantillon de 20 000 SMS, seules 358 modifications ont dû être effectuées : 66 % (cas 1), 29 % (cas 2), 5 % (cas 3).
8. Chute d'un ou de plusieurs phonèmes à la fin du mot par suite d'une évolution phonétique ou d'un abrègement.
9. Lopez C., Bestandji R., Roche M., Panckhurst R. (2014) « Towards Electronic SMS Dictionary Construction: An Alignment-based Approach », Actes du colloque LREC, Reykjavik, Islande, 26-31 mai, p. 2833-2838.

Par la suite, un extrait de 100 SMS du corpus « 88milSMS » a été annoté, à l'aide de 8 balises : TYPographie, MODification, GRAmmaire, BINettes, ABSence, LANgue, ORThographe, DIVers. Il en ressort que les phénomènes de *typographie* sont les plus saillants, suivis par les *modifications* (substitutions, réductions, ajouts, etc.). La balise qui concerne la *grammaire* arrive en troisième position, suivie, dans l'ordre, par les *binettes*, l'*absence*, la *langue*, l'*orthographe* et la balise *divers*. Il est également intéressant de constater que 70 % de l'extrait des 100 SMS n'a subi aucune modification.

Comme pour le transcodage, il est extrêmement difficile de proposer une annotation standardisée. Lors du projet *sud4science LR*, les chercheurs ont 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 a ensuite permis aux chercheurs *sud4science* de réduire le nombre de balises précédemment utilisées, afin d'envisager le balisage éventuel du corpus 88milSMS. Tout bien pesé, ils ont décidé de fournir un petit échantillon d'annotation de 100 SMS. Mais ils ont renoncé à l'annotation de l'ensemble du corpus 88milSMS car, d'une part, la tâche aurait été gigantesque et, d'autre part, les chercheurs n'auraient pas été nécessairement en accord avec le choix des balises. Si cet échantillon permet de fournir des pistes de recherche, il nous est apparu au final que le plus important est d'abord de mettre à disposition le corpus anonymisé de telle sorte que chacun puisse le catégoriser et l'annoter en fonction de sa propre problématique de recherche.

Diffusion : mission de service public

L'objectif du dépôt sur la grille de services d'Huma-Num est de mettre sans tarder à la disposition de la communauté scientifique et, plus largement, de tous ceux qui sont intéressés par les mutations sociales, comme les responsables des politiques publiques en matière d'éducation et d'intégration sociale, une base de données directement téléchargeable. Les chercheurs du projet proposent donc au public, via un téléchargement direct, le corpus intitulé « 88milSMS » dans son intégralité, deux échantillons (100 SMS annotés, 1 000 SMS transcodés en français standardisé), ainsi qu'un questionnaire sociolinguistique soumis aux donateurs, et leurs réponses.
Les chercheurs ont débuté l'*observation*, la *fouille*, la *description*, le *traitement* et l'*analyse* du grand corpus « 88milSMS », mais beaucoup de recherches doivent encore être menées. Le corpus de SMS pourra être exploité afin d'élaborer des applications informatiques variées (par exemple, élaboration de lexiques transcodés français standardisé -> SMS ou SMS -> français standardisé consultables en ligne, mise en place de systèmes de vocalisation des SMS à l'usage de personnes déficientes visuelles ou de personnes momentanément empêchées de consulter leur écran de téléphone – en situation de conduite, etc.). Par ailleurs, il serait envisageable de rendre l'outil d'anonymisation disponible afin qu'il soit réutilisé dans d'autres projets de recherche, voire dans des activités professionnelles soumises à la confidentialité.

L'intérêt de la mise à disposition du corpus 88milSMS sur la grille de services d'Huma-Num (et un éventuel archivage au Centre Informatique National de l'Enseignement Supérieur) relève d'une véritable mission de service public : permettre à un grand nombre de chercheurs et d'étudiants, toutes disciplines confondues, ainsi qu'à des personnes du grand public de tous horizons, de fouiller, d'analyser, d'approfondir nos connaissances à propos des pratiques contemporaines de la textualité numérique pendant de nombreuses années.

▶ Référence officielle du corpus
"88milSMS. A corpus of authentic text messages in French" Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014), 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.

Rachel Panckhurst, Catherine Détrie, Bertrand Verine
Praxiling (UMR 5267 CNRS / Université Paul-Valéry Montpellier 3
▶ rachel.panckhurst@univ-montp3.fr
▶ catherine.detrie@univ-montp3.fr
▶ bertrand.verine@univ-montp3.fr

Cédric Lopez
Objet Direct - VISEO
▶ cedric.lopez@viseo.com

Claudine Moïse
Lidilem, Université Stendhal Grenoble 3
▶ claudine.moise@u-grenoble3.fr

Mathieu Roche
Tetis. MTD (Maison de la Télédétection) UMR TETIS
▶ mathieu.roche@cirad.fr

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 (SAJI, Université Paul-Valéry Montpellier 3) de nous avoir accompagnés et conseillés sur le plan juridique, 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 2), qui ont développé le système informatisé *Seek&Hide*, permettant d'anonymiser le corpus ; Michel Otell, Camille Lagarde-Belleville, Frédéric André et Yosra Ghliss (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 : Aghiles Lounes, Tarik Zaknoun, Zakaria Mokrani, Reda Bestandji, Takfarinas Sider, Ahmed Loudah, (Master I Informatique, Spécialité : « Informatique pour les sciences », Université Montpellier 2) qui ont travaillé sur un système de transcodage automatique.