Corpus « 88milSMS ».
© 2014 Panckhurst, Détrie, Lopez, Moïse, Roche, Verine.

Lecture des données du corpus 88milSMS :

Colonne A : « NUM_SMS », numéro du SMS. La numérotation s'étend de 1 à 93 085, mais le nombre réel de SMS au sein du corpus 88milSMS est de 88 522.
Colonne B : « DATETIME », date et heure de la réception du SMS par l'équipe de chercheurs.
Colonne C : « ID_NUM_TEL », identifiant du numéro de téléphone mobile. Cette colonne est à mettre en lien avec la colonne D du fichier des réponses au questionnaire sociolinguistique.
Colonne D : vide, pour faciliter la lecture de la colonne E.
Colonne E : « SMS_ANON », SMS brut anonymisé.
Au sein de la colonne E, dix balises ont été utilisées pour l'anonymisation :

1. <PRE_X> : prénom ; X = le nombre de caractères ;
2. <NOM_X> : nom ; X = le nombre de caractères ;
3. <SUR_X> : surnom ; X = le nombre de caractères ;
4. <ADR_X> : adresse ; X = le nombre de caractères ;
5. <LIE_X> : lieu ; X = le nombre de caractères ;
6. <TEL_X> : numéro de telephone ; X = le nombre de caractères ;
7. <COD_X> : codes divers (bancaires, colis postaux, portes, etc.) ; X = le nombre de caractères ;
8. <URL_X> : URL ; X = le nombre de caractères ;
9. <MAR_X> : marque ; X = le nombre de caractères ;
10. <MEL_X> : courriel ; X = le nombre de caractères.

Bref descriptif du projet :

Une équipe pluridisciplinaire de linguistes et d'informaticiens (Rachel Panckhurst*, Catherine Détrie*, Cédric Lopez*****, Claudine Moïse***, Mathieu Roche**/****, Bertrand Verine*, Praxiling*, Lirmm**, Lidilem***, Tetis****, Viseo*****) a recueilli plus de 88 000 SMS authentiques en français à Montpellier, en 2011. Cette collecte a été effectuée dans le cadre du projet *sud4science LR* (http://sud4science.org, *Sud4science Languedoc Roussillon. Mutation des pratiques scripturales en communication électronique médiée* (financement principal : MSH-M)), lui-même faisant partie du projet international *sms4science* (http://sms4science.org), coordonné par le CENTAL à l'Université catholique de Louvain (UCL) en Belgique. Lors du recueil des SMS, un questionnaire sociolinguistique a également été proposé aux participants. Les SMS du projet *sud4science LR* ont été ensuite *anonymisés* de manière semi-automatique (en collaboration avec des étudiants stagiaires et un juriste-CIL, Nicolas Hvoinsky, SAJI-Université Paul-Valéry), puis partiellement *transcodés* (en français standardisé) et *annotés* (*cf*. Panckhurst *et al*. 2013).

Les analyses portent sur l'écriture, la textualité numérique, les fonctionnements

langagiers, les pratiques et usages des donateurs de SMS et ont été menées dans le cadre de *sud4science LR* et de deux projets supplémentaires : *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS* (financement principal DGLFLF) et *Analyse contrastive des **é**motions **co**ntenues dans les **mess**ages courts* (PEPS CNRS ECOMESS (HuMaIn)).

Cette recherche pluridisciplinaire a permis de réaliser un très grand nombre de travaux et de publications et d'encadrer 8 stagiaires étudiants. Les résultats de recherche ont été diffusés en France et dans 7 pays étrangers (Belgique, Canada, Espagne, Finlande, Grèce, Maroc, Suisse). Les chercheurs ont pu offrir l'occasion aux étudiants-stagiaires de *présenter* leurs propres travaux et d'écrire, pour deux d'entre eux, en tant que *premier auteur*, un article dans une revue internationale (*cf*. Accorsi *et al*. 2012, Patel *et al*., 2013).

L'objectif du dépôt sur la grille HUMA-NUM est de mettre 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 mettent donc à disposition publique, via un téléchargement direct, le corpus intitulé « 88milSMS », deux échantillons (100 SMS annotés, 1000 SMS transcodés en français standardisé), ainsi qu'un questionnaire sociolinguistique et ses 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.)).

L'intérêt de la mise à disposition du corpus 88milSMS sur la grille HUMA-NUM est de 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, avec l'autorisation de 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. »

© 2014 Panckhurst, Détrie, Lopez, Moïse, Roche, Verine.

Références citées :

Accorsi P., Patel N., Lopez C., Panckhurst R., Roche M. (2012), "Seek&Hide : Anonymising a French SMS corpus using natural language processing techniques", Lingvisticæ Investigationes, Special Issue : "SMS Communication : A Linguistic Approach", John Benjamins, 35:2, 163-180.

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.

Patel N., Accorsi P., Inkpen D., Lopez C., Roche M. (2013) "Approaches of anonymisation of an SMS corpus", Proceedings of CICLING (Conference on Intelligent Text Processing and Computational Linguistics), LNCS, Springer Verlag, March 24–30, 2013, University of the Aegean, Samos, Greece, p. 77-88, http://www.cicling.org/2013/

Site Web : http://sud4science.org
Vidéos : http://msh-m.tv/spip.php?rubrique138
Presse écrite, radio, télévision :
http://www.sud4science.org/?q=fr/node/5

Coordonnées et logos :

Praxiling* UMR 5267 CNRS,
Université Paul-Valéry
Montpellier 3,
Route de Mende,
34199 Montpellier cedex 5, France
88milsms@univ-montp3.fr
rachel.panckhurst@univ-montp3.fr
catherine.detrie@univ-montp3.fr
bertrand.verine@univ-montp3.fr

Tetis****
MTD (Maison de la Télédétection)
UMR TETIS
500, rue J.F. Breton
34093 Montpellier cedex 5, France
mathieu.roche@cirad.fr

Viseo***
Le Pulsar
4, av. du Doyen Louis Weil
38000 Grenoble
clopez@objetdirect.com

Lidilem***
Université Stendhal Grenoble 3
Laboratoire Lidilem,
BP 25, 38040, Grenoble cedex, France
claudine.moise@u-grenoble3.fr

Lirmm**
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier,
161 Rue Ada,
34090 Montpellier
mathieu.roche@lirmm.fr

© 2014 Panckhurst, Détrie, Lopez, Moïse, Roche, Verine.