Pour citer ce document :
Ledegen, G. (2014). Manuel du corpus *smslareunion (cmr-smslareunion-tei-v1-manuel.pdf)* .In Ledegen, G., Grand corpus de sms smslareunion. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-smslareunion-tei-v1]

Affiche de l'opération par Shovel (2008)

Grand corpus de sms : smslareunion

Le programme sms4science

Ce grand corpus de sms a été constitué dans le cadre de l'opération sms4science (Fairon 2006a, b)[1] ; ce programme de recherche lancé en 2004 par le CENTAL, Centre de Traitement Automatique du Langage de l'Université catholique de Louvain en Belgique (http://www.sms4science.org/ ; http://www.lareunion4science.org/), devient rapidement international : le Québec, la Suisse, différentes régions françaises (Grenoble, Montpellier, …), mais aussi l'Espagne, la Grèce, …

Mené à La Réunion, premier terrain français, le projet a permis de réunir 21 694 sms durant la période allant d'avril à juin 2008, provenant de 1 744 usagers, donnant 12 622 sms finalisés.

Le protocole

Afin d'obtenir des sms d'un échantillon de la population le plus représentatif possible des usagers du sms réels, nous avons mis en place des appels à participation diversifiés (télévision, radio, presse écrite, internet[2]) et permis une participation gratuite des usagers via l'envoi vers un numéro court (3699).

Nous avons par ailleurs indiqué vouloir obtenir des messages envoyés lors de véritables échanges communicationnels par envoi ultérieur ou parallèle des messages sur le numéro court.

---

[1] Fairon, C., Klein, J.R. & Paumier, S., 2006a, *Le langage SMS*, Louvain-la-Neuve, P.U.Louvain, Cahiers du Cental, 3.1. Fairon, C., Klein, J.R. & Paumier, S., 2006b, *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation*, CD-Rom, Louvain-la-Neuve, P.U.Louvain, Cahiers du Cental, 3.2.

[2] http://www.lareunion4science.org/?q=fr/node/5

Ainsi, la procédure de copie automatique des messages a permis d'éviter les biais obtenus lors du recopiage manuel qui était auparavant pratiqué habituellement pour ce type d'enquête.
Une fois collecté, le corpus est traité afin qu'il réponde aux exigences juridiques et scientifiques : nettoyage du corpus (élimination de messages qui étaient adressés à l'équipe, de messages publicitaires ou circulaires (ex. : chaîne de l'amour, …)), anonymisation de toutes les données personnelles (nom, prénom, adresse-mail, site-web, lieu, date, …) et transcription du corpus en français standard, pour permettre l'exploration logicielle (après balisages et annotations (cf. *ReadMe*)).

Financement

Ce programme de recherche a bénéficié du mécénat de SFR (5000€) et du financement de 10.000€ de la Région Réunion. SFR prenait par ailleurs en charge les frais d'envoi et de réception des sms pour les numéros SFR (pour l'autre fournisseur d'accès, Orange, les financements ont permis de couvrir les frais[3]), ainsi que les cadeaux offerts durant les 10 semaines de l'opération (un téléphone dernier cri et 5 cartes de téléphone / semaine).
Ces cadeaux incitaient les participants à répondre au questionnaire sociolinguistique en ligne et à envoyer au moins 5 messages sur une semaine : un tirage au sort était organisé chaque semaine pour les participants réunissant ces 2 conditions, et les noms des gagnants figuraient dans le journal *Le Quotidien*, autre partenaire de l'opération (annonce du lancement de l'opération et publication régulière d'articles donnant les premiers résultats de la récolte).

L'équipe

- Gudrun Ledegen : responsable scientifique du programme réunionnais : nettoyage du corpus, transorthographisation, balisage, annotation, exploration logicielle.
- Étudiants ayant participé au nettoyage du corpus et à la transcription du corpus en français standard : Gauthier Caron, Gaëlle Corré, Marie-Caroline Guillemain.

Bibliographie portant sur le corpus smslaréunion

Les études et publications auxquelles ce corpus a pu donner lieu s'organise en deux volets : le premier analyse la variation de l'écrit-sms dans une optique comparatiste sur le terrain francophone et franco-créolophone (références 1, 2, 8, 10), avec un éclairage portant spécifiquement sur les graphies du créole réunionnais (7 & 9). Le second volet se situe dans le cadre d'une comparaison avec des corpus complémentaires réunissant des corpus recueillis auprès de personnes sourdes, à La Réunion ou en Normandie (références 3, 4, 5, 6) :

[1] Ledegen, G., 2010, « Contact de langues à La Réunion : « On ne débouche pas des cadeaux. Ben i fé qoué alors ? » », *Langues et Cité*, 'Langues en contact', n° 16, 9-10, http://www.dglf.culture.gouv.fr/Langues_et_cite/LC16.pdf Disponible à : http://hal.archives-ouvertes.fr/hal-00879323.

[3] L'entreprise Cirrus Informatique récoltait les sms en provenance des deux opérateurs, envoyait les messages de confirmation (les participants s'inscrivaient à l'opération en envoyant 'sms' au 3699, manifestant ainsi leur volonté de participer, indispensable étape sur un plan juridique pour de telles opérations (cf. annexe 1. *Conditions de participation* et annexe 2. *Foire aux questions*) et les mettait à notre disposition en temps réel, sous format .xls.

[2] Cougnon, A. & Ledegen, G., 2010, « Une étude comparatiste des variétés du français dans l'*écrit-sms* (Réunion-Belgique) », in Abécassis, M. & Ledegen, G., *Les voix des Français : en parlant, en écrivant*, Peter Lang, 39-58.

[3] Ledegen, G., 2011, « Résonance SMS. « Jc c koi mé javé pa rèalizé sur le coup! » », *LINX*, n° 57, Gadet, F. & Guérin, E. (Dirs), 'Français parlé/français hors de France/créoles à base française d'un point de vue syntaxique', 101-112. Disponible à : http://hal.archives-ouvertes.fr/hal-00879319

[4] Ledegen, G., M. Blondel, J. Gonac'h & J. Seeli, 2011 « Contacts de langues dans les SMS 'sourds' », *Langues et cité Bulletin de l'observatoire des pratiques linguistiques*, n° 19, 'Parler (avec) plusieurs langues : l'alternance codique', 10. Disponible à : http://www.dglf.culture.gouv.fr/Langues_et_cite/LC19.pdf et http://hal.archives-ouvertes.fr/hal-00879337

[5] Blondel M., J. Gonac'h, G. Ledegen et J. Seeli, 2011, « Ecriture-sms en Métropole et à La Réunion : 'Zones instables et flottantes' du français ordinaire et spécificités du contexte de surdité », in Gilles Col et Sylvester N. Osu (dir.), *Transcrire, Écrire, Formaliser - (1) Travaux linguistiques du CERLICO*, n° 24, 51-70. Disponible à : http://hal.archives-ouvertes.fr/hal-00879181

[6] Ledegen G, J. Seeli, M. Blondel M. et J. Gonac'h, 2011, « 'Tu pense quoi mieux ?' De la Normandie à La Réunion, les interrogatives en question dans les SMS en contexte de surdité », in Liénard, F. et Zlitni, S. (éds), *La communication électronique : enjeux de langues,* Limoges, Lambert-Lucas, 223-234. Disponible à : http://hal.archives-ouvertes.fr/hal-00879194

[7] Gonac'h, J., Seeli, J., Ledegen, G. & Blondel, M., 2012, « Les contacts du français, du créole et de la LSF dans les écrits-sms", *CLAIX*, n° 24, Kriegel, S. & Véronique, D. (Dirs), 'Contacts de langues, langues en contact', 171-186. Disponible à : http://hal.archives-ouvertes.fr/hal-00879338.

[8] Ledegen, G., 2012, « Ecrits ordinaires du créole réunionnais dans les sms : « T kwa la fai ? » », in Colloque *Éclairages pluridisciplinaires pour une orthographe fonctionnelle et consensuelle du créole réunionnais* (27, 28, 29 mai 2009 – Université de La Réunion, Lofis la Lang Kréol La Rénion), http://lofis.unblog.fr/files/2011/11/gledegencommunication.pdf

[9] Ledegen, G., 2012, « Tchat – sms – oral : les « petits mots » des jeunes à La Réunion », in Bulot, T. & Feussi, V., *Normes, urbanités et émergences plurilingues: Parlers (de) jeunes francophones*, Paris, L'Harmattan, 45-66.

[10] Ledegen, à paraître, « L'« écrit réunionnais » dans les SMS. 'Ma fi vi conè pa komen!' », in Ledegen, G., Gkoskou, P. & Gauvin, A. (Eds), *Éclairages pluridisciplinaires pour l'aménagement des langues créoles, langues en situation de contact inégalitaire*, Paris, L'Harmattan.

[11] Ledegen, G. & Lyche, C., à paraître, « La particule négative *ne* dans les français d'Afrique et de l'Océan Indien : convergences et divergences », in Abécassis, M. & Ledegen, G. (Eds), *De la genèse de la langue à Internet*, Peter Lang.

Annexe 1. Conditions de participation

Objectif du projet

Le projet vise à rassembler un grand nombre de SMS afin de constituer un vaste "corpus" électronique qui puisse servir de base à des activités de recherche et développement (principalement en linguistique et ingénierie linguistique). Le corpus constitué fera l'objet d'une diffusion, car ce type de données intéresse une large communauté scientifique. Pour réaliser cet objectif, les promoteurs lancent un appel à participation et cherchent des volontaires disposés à "faire don" de leurs SMS en envoyant ceux-ci vers un numéro gratuit. Les participants sont également invités à répondre à une courte enquête au travers d'un formulaire Web. Les données anonymes recueillies dans cette enquête sont destinées à être diffusées avec le corpus électronique.

Droits d'auteur

1. Le participant déclare être l'auteur des messages qu'il envoie.
2. Le participant autorise le LCF à inclure les messages qu'il envoie dans le corpus destiné à la recherche scientifique.
3. Le participant cède gracieusement au LCF les droits de reproduction, de modification, d'adaptation et de communication sur les textes envoyés, et ce pour toute la durée de la propriétaire littéraire et artistique et pour le monde entier.
4. Le participant donne au LCF le droit de diffuser avec le corpus les informations du questionnaire sociolinguistique qu'il remplit (cf. ci-dessous "protection de la vie privée").
5. Le participant fait don de ses SMS sans aucune contrepartie.

Protection de la vie privée

INTERMEDIAIRE TECHNIQUE

La récolte des SMS nécessite l'intervention d'un partenaire technique (**Cirrus Informatique**) qui prend en charge la réception des SMS au 3699 et leur transfert au LCF. La société **Cirrus Informatique** s'est engagée à transmettre avec la plus grande confidentialité les SMS au LCF qui en est seul propriétaire.**Cirrus Informatique** gère également les interactions avec les participants. Le LCF demande à **Cirrus Informatique** de transmettre automatiquement par e-mail, à ces personnes qui ont marqué leur accord, un message d'information et de remerciement. Ce message invite les participants à compléter, sur le site Web du projet, un questionnaire portant sur les pratiques du SMS. Pour la bonne gestion de cette opération, **Cirrus Informatique** conserve, dans le respect de la législation sur la protection vie privée, votre numéro de gsm et votre adresse e-mail. Pour supprimer votre numéro de GSM et votre adresse, envoyez STOP par SMS au numéro gratuit 3699.

ENGAGEMENT DU LCF

Le LCF s'engage à ne pas diffuser et à protéger les informations personnelles des participants qui sont en sa possession (numéro de téléphone, email).Les informations du questionnaire sociolinguistique (nationalité, âge, langue maternelle, habitudes d'utilisation des SMS, etc.) sont destinées à être diffusées avec le corpus de manière strictement anonyme.Avant d'être intégrés au corpus, les SMS seront anonymisés (remplacement des noms de personnes par des noms génériques).Les données récoltées seront gérées par le laboratoire **LCF de l'Université de la Réunion (LCF-UMR 8143 du CNRS, 15 av. R. Cassin, 97715 Saint Denis – 02.62.93.85.77) et** le Centre de traitement automatique du langage de l'UCL (Place Blaise Pascal, 1 à 1348 Louvain-la-Neuve). Cette base de données a fait l'objet d'une déclaration à la Commission Nationale de l'Informatique et des Libertés (n°1251736 du 10/03/08).

Conformément à la loi du 8 décembre 1992 sur le respect de la vie privée, les participants peuvent exercer leur droit de regard sur les données les concernant et, le cas échéant, les faire modifier ou supprimer.

Durée du projet

La première phase de la collecte des SMS commencera le 15 mars 2008 et se poursuivra jusqu'au 15 juin 2008. Le LCF garde le droit d'interrompre ce projet à tout moment.

Règlement du tirage au sort

Un tirage au sort sera effectué chaque semaine par le LCF. Les personnes sélectionnées recevront un cadeau Proximus (cartes prépayées, etc.) en remerciement pour leur participation.Le tirage au sort aura lieu tous les jeudis (sauf la première semaine) sous le contrôle de Madame Gudrun Ledegen (LCF, Université de la Réunion). Seuls les participants ayant d'une part envoyé 5 SMS ou plus et d'autre part rempli le questionnaire disponible sur ce site participent au tirage au sort. Le tirage au sort est réalisé par programme informatique dans la base de données de messages reçus. Le nom des gagnants sera annoncé sur le site Web et par SMS (envoyé aux gagnants) ainsi que dans la journal *Le Quotidien*. Les gagnants seront invités par SMS à contacter le LCF pour obtenir leur cadeau. Les cadeaux non réclamés au moment du tirage au sort suivant sont remis en jeu. Les gagnants ont donc une semaine pour se manifester.

Annexe 2. Foire aux questions

Peut-on participer quel que soit notre opérateur?

Le numéro 3699 est gratuit et valable pour les opérateurs SFR et Orange.

Comment participer?

Rien de plus facile. Pour les détails, suivez ce lien

Quelqu'un saura-t-il que je suis l'auteur de ces SMS?

Personne! L'anonymat est garanti.

Y a-t-il des cadeaux à gagner?

En participant, vous pouvez gagner de nombreux cadeaux (chaque semaine les gagnants seront désignés par tirage au sort parmi ceux qui auront envoyé 5 SMS dans la semaine et par ailleurs rempli le questionnaire). Mais le plus important... c'est que vous faites progresser la science ;-)

Mes SMS ne sont pas très originaux et ils sont courts, puis-je aussi les envoyer?

Oui, ce qui nous intéresse principalement, c'est l'authenticité: nous ne cherchons pas à récolter les messages les plus originaux qui soient, mais bien à rassembler des messages qui ont été échangés dans le cadre normal de la communication entre correspondants.

Quels sont les messages qui nous intéressent?

Les messages qui nous intéressent sont dans la mémoire de votre téléphone: ce sont les messages que vous avez écrits à vos correspondants dans le cadre habituel de vos communications. Nous vous invitons à envoyer uniquement les messages dont vous êtes l'auteur.

Puis-je vous envoyer des messages que j'ai reçus?

NON, vous devez être l'auteur des messages que vous nous envoyez.

Je n'ai pas conservé de message en mémoire dans mon GSM, puis-je vous écrire directement un message?

NON, nous souhaitons collecter des messages qui ont réellement été échangés dans le cadre habituel de la communication. Il ne s'agit donc pas de nous écrire, mais de nous faire suivre des messages que vous avez écrits à d'autres personnes.

S'il y a des noms, adresses ou numéros de téléphone dans les messages que j'envoie, est-ce un problème?

Non, ce n'est pas un problème. Les messages reçus seront anonymisés : les données privées (nom, adresse, n° de téléphone) seront remplacées par des données génériques.