Étudier l'écrit SMS –

Un objectif du projet *sms4science*

Louise-Amélie Cougnon (UCLouvain, ILC, Cental) & Thomas François (Aspirant FNRS, UCLouvain, ILC, Cental)

Abstract

This paper details an international project called *sms4science* that aims to collect text message corpora (hereafter referred to as "SMS corpora") from across the globe for scientific research. The project already has ten participating regions, including Belgium, Réunion, Switzerland and Quebec. This article first presents the initial corpora collected from these four areas (resulting in a combined total of 116'000 text messages) and the accompanying methodology. It then exposes the research possibilities related to it: the corpus-based studies pertain as much to linguistics and sociolinguistics as they do to natural language processing and statistics. A specific statistical study is thus presented here and its possible conclusions outline the differences in SMS practices between regions, notably when you consider abbreviation rate or message length. Finally, the paper delineates the project obstacles and correspondingly proposes fresh perspectives for the ongoing year (2011).

1 Introduction

L'essor des technologies de l'information et de la communication (TIC) a marqué de façon décisive le XX^{e} siècle: par la radiodiffusion, la télévision, la téléphonie ou Internet, tant l'information de masse que la communication interpersonnelle ont subi une évolution – voire une mutation – sans précédent. Le développement de ces technologies, notamment dans le domaine de l'informatique, a donné naissance à de nouvelles formes de communication entre locuteurs, par le biais des machines[1]. La notion de *communication médiatisée par ordinateur*, ou *CMO*,[2] est apparue au début des années 80 au sein du milieu scientifique international; nous lui préférons l'acronyme *CéMO*, pour *communication écrite médiée par ordinateur*, qui semble plus précis. Grâce à la CéMO, la pratique de l'écrit est plus importante aujourd'hui qu'elle ne l'a jamais été. La communication par SMS tient une place importante au sein de cet ensemble de CéMO. À titre indicatif, 2.300 milliards de SMS et MMS ont été échangés en 2008 à travers le monde[3]. Selon le rapport *Mediappro* (2006), basé sur les pratiques communicatives de 9 pays européens[4], 95 % des jeunes[5] possèdent un téléphone portable, et parmi eux, 79 % préfèrent envoyer un SMS qu'initier un appel téléphonique.

[1] Notons que ce développement n'est pas géographiquement ni socialement uniforme. Comme l'explique Thurlow (2003: 1): «which is not to say that this technology is properly global; worldwide patterns of [media development] necessarily follow the socioeconomic contours of which distinguish the "media rich" and "media poor" generally».

[2] En anglais on dira *Computer Mediated Communication* ou *CMC*.

[3] Ces données proviennent du *Journal du Net*. Source: http://www.journaldunet.com/chiffres-cles.shtml.

[4] Ces 9 pays sont: la Belgique, le Danemark, l'Estonie, la France, la Grèce, l'Italie, la Pologne, le Portugal et le Royaume Uni.

[5] Ces jeunes sont âgés de 12 à 18 ans.

Une liste finie de termes équivalents à *SMS* est employée dans la littérature francophone: *texto* et *mini-message*[6] sont ceux qui apparaissent le plus souvent. Ces termes renvoient à une même réalité, «une communication interindividuelle entre des partenaires qui se connaissent préalablement et ont un certain niveau d'intimité. Le régime temporel est le différé, mais à échéance rapide: une quasi-immédiateté est visée pour la lecture du message et une réponse est attendue dans de brefs délais» (cf. Anis 2002).

Cougnon/Ledegen (2008), dans une étude contrastive dédiée aux SMS francophones de Belgique et de La Réunion, ont montré une préférence pour la notion d'*écrit* ou d'*écriture SMS* (abréviée *eSMS*) par rapport à celle de *langage SMS*. Non seulement les variations présentes pour un seul et même mot (quelquefois chez le même auteur et destinateur) amènent à penser qu'il n'existe pas une seule variété de *SMS*, c'est-à-dire une série de pratiques linguistiques liées à l'utilisation du SMS, mais plutôt un ensemble de phénomènes applicables ou non, cumulables ou non. De plus, dans le cas du SMS, il semble s'agir plus précisément d'une nouvelle manière de transcrire la langue : nous aurions ainsi un *code SMS* qui utilise, tout comme la langue écrite standard, les lettres de l'alphabet latin et les chiffres arabes, en les employant de manière traditionnelle ou en leur attribuant une valeur différente de l'orthographe standard (phonétisante, symbolique, etc.).

La fréquence de la pratique du SMS ainsi que l'originalité de ce nouveau code posent question en de nombreux points quant aux probables répercussions de l'emploi du SMS sur les compétences linguistiques générales de chacun. De nombreuses études se sont déjà penchées sur cette nouvelle forme de communication, mais les connaissances acquises sont partielles et lacunaires. Ceci s'explique par le fait que la recherche ne dispose pas encore d'une quantité suffisante de données (de corpus) pour que soient menées à bien des études objectives.

C'est de cette lacune qu'est né le projet *sms4science*[7]. À travers cet article, nous présenterons le corpus de 116.000 SMS collecté dans 4 aires géographiques différentes: nous préciserons la méthodologie rigoureuse employée afin d'obtenir des données les plus authentiques possibles. Nous exposerons ensuite le panorama des études possibles à partir d'un tel corpus: des domaines tels que la sociolinguistique, la lexicologie et le traitement automatique des langues seront concernés. Nous mettrons en avant un exemple de recherche sur les messages: les statistiques du SMS. Enfin, nous livrerons les perspectives d'avenir du projet *sms4science*.

2 Un corpus international de 116.000 SMS

Comme expliqué précédemment, le projet *sms4science* est né en réponse à un manque de données SMS authentiques et conséquentes en nombre. En effet, le projet a pour but premier la constitution méthodique, pour un grand nombre de langues, de vastes corpus de SMS pour la recherche scientifique. C'est le Cental[8], associé à des partenaires scientifiques[9] et à des sponsors[10], qui a mis sur pied le projet en 2004. Depuis lors, *sms4science* a permis la collaboration de 10 régions[11] et de 15 universités.

Afin d'obtenir des corpus dits *comparables* dans les différentes régions, la méthodologie de collecte et de traitement des messages doit être précise et rigoureusement respectée. Cette

[6] Le Petit Robert propose, pour *texto*, une définition semblable à celle de *SMS*, à ceci près qu'il ne reprend pas le sens de «service». En ce sens, *texto* serait donc moins ambigu. *Mini-message* n'apparaît pas dans le dictionnaire; en outre, il est principalement employé en France.

[7] Pour plus d'information au sujet du projet : www.sms4science.org.

[8] Il s'agit d'un centre de traitement automatique du langage créé à la fin des années 90 à l'UCL (Belgique). Pour plus d'information au sujet du Cental: www.uclouvain.be/cental.

[9] Jean Klein (Celexrom, UCL, Belgique) et Sébastien Paumier (Université de Marne-la-Vallée, France).

[10] Proximus, Ogilvy et NEWAy.

[11] Belgique, Canada, Espagne, Alpes françaises, La Réunion, Grèce, Italie, Roumanie, Royaume-Uni et Suisse.

méthodologie comporte 4 niveaux distincts résumés dans la Fig. 1. La première étape concerne les aspects médiatique et technique de la collecte elle-même. La stratégie utilisée permet d'obtenir des copies de messages véritablement échangés et de les sauvegarder automatiquement, de manière à éviter un recopiage manuel entraînant des erreurs inévitables; elle permet également de sauvegarder des profils sociolinguistiques concernant les émetteurs des messages. Ces données sont donc collectées automatiquement, mais aussi anonymisées avant d'être récupérées par l'université en charge, afin de protéger les informations confidentielles des participants (étape 2). Lors de la troisième étape, les SMS sont triés (suppression de messages non conformes, etc.) et transcrits en graphie standard. Un corpus de SMS transcrits est ainsi aligné, au message près, au corpus de SMS bruts. La dernière étape consiste en une annotation[12] des messages, notamment pour ce qui est de la langue en usage.

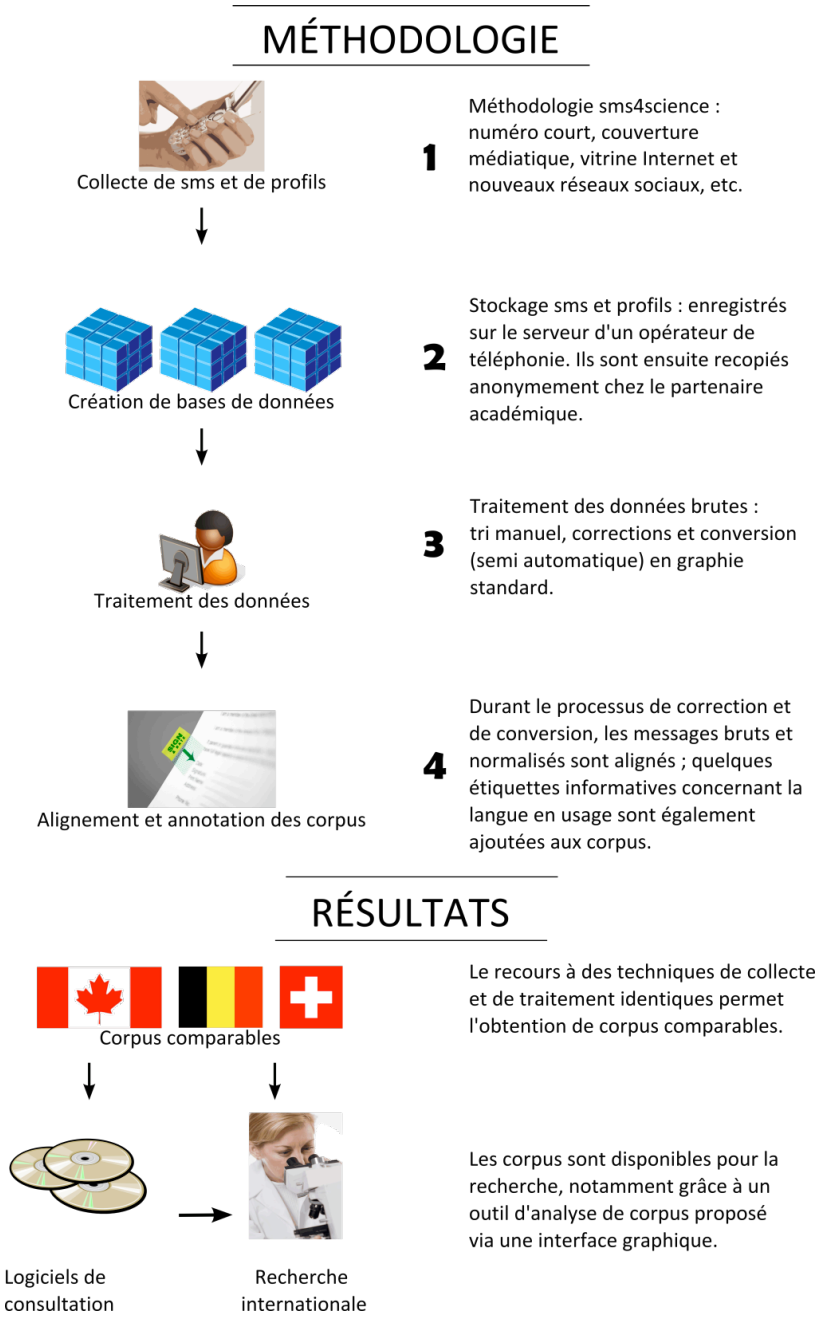

Figure 1: Méthodologie et type de résultats du projet *sms4science*

[12] Sous la forme d'étiquettes comprenant des informations textuelles et supratextuelles.

Des précautions légales doivent être prises pratiquement à chaque niveau du projet. Ainsi, avant même de participer, chaque usager du SMS doit valider un consentement de participation qui stipule notamment que «les sms sont transmis volontairement et librement par des utilisateurs témoins qui déclarent être les auteurs des messages transmis, connaître les objectifs de cette collecte»[13], etc. Lors du passage de la base de données de l'opérateur privé à celle de l'université en charge, la confidentialité est assurée: les numéros de téléphone ne sont pas divulgués. Au moment de la troisième étape, une procédure d'anonymisation du contenu des messages est prise en charge par un logiciel de reconnaissance de données privées, puis, dans un deuxième temps, manuellement: les noms, adresses, numéros de téléphone et autres données à caractère personnel incluses dans les SMS sont remplacés par des symboles neutres. Enfin, après diffusion des données pour la recherche, le projet assure encore le droit des usagers en leur offrant la possibilité de consulter la base de données et de supprimer éventuellement les messages dérangeants.

Depuis 2004, la Belgique[14], la Réunion[15], la Suisse[16] et le Québec[17] ont d'ores et déjà mené à bien une collecte dans leur région. Ainsi, nous disposons à ce jour d'une base de plus de 116.000 messages[18] rédigés en 3 langues principales (plus les langues régionales) et de 3.708 profils sociolinguistiques décrivant les auteurs de ces messages. Les Fig. 2 et 3 détaillent le résultat de ces collectes.

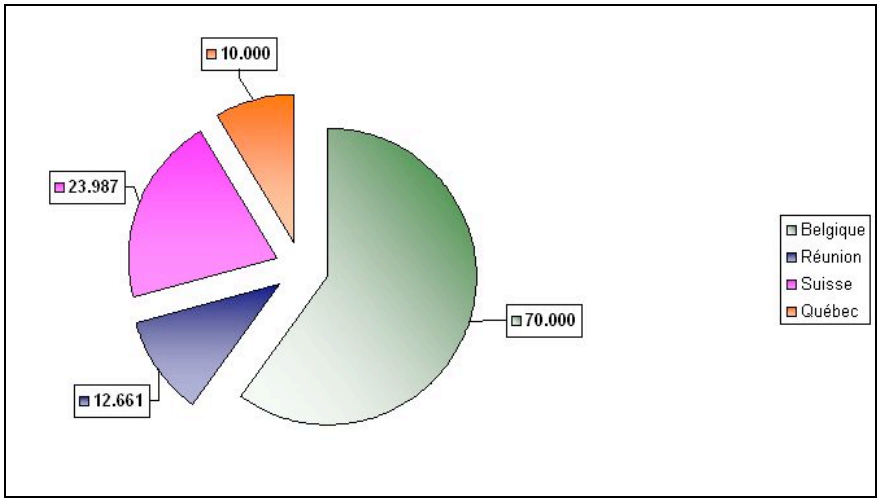

Figure 2: Nombre de SMS collectés dans chaque région participant au projet *sms4science*

[13] Extrait du contrat de collaboration établi entre l'Université catholique de Louvain et ses partenaires dans le cadre du projet *sms4science*.

[14] Information sur le projet en Belgique: http://www.smspourlascience.be/.

[15] Information sur le projet à la Réunion: www.lareunion4science.org.

[16] Information sur le projet en Suisse: www.sms4science.ch/.

[17] Information sur le projet au Québec: www.texto4science.ca/.

[18] Notons que pour des raisons techniques et légales (anonymisations non clôturées, SMS segmentés, etc.), les données exploitables dans des recherches telles que celles présentées dans ce papier ont été restreintes à 71.703 SMS, comme suit : 30.001 SMS en Belgique, 12.661 SMS à la Réunion, 23.987 SMS en Suisse et 5.054 SMS au Québec. Le nombre proposé plus haut représente l'ensemble des SMS exploitables après traitement complet des corpus, d'ici septembre 2011.

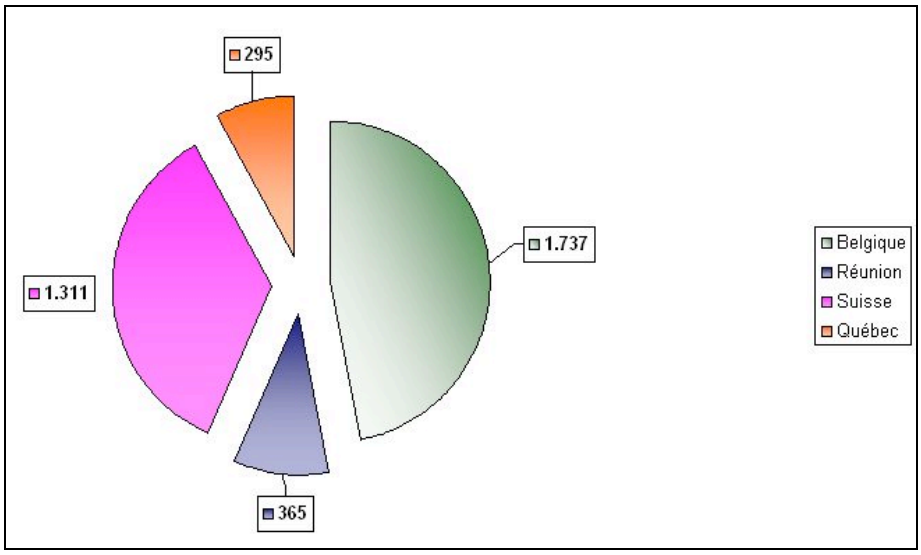

Figure 3: Nombre de profils collectés dans chaque région participant au projet *sms4science*

Le consortium scientifique est, à l'heure actuelle, le plus important dans le domaine de la collecte de SMS: il est à la fois international, plurilingue et pluridisciplinaire. Les centres de recherche participants relèvent de disciplines aussi variées que le traitement automatique du langage, l'ingénierie, la sociologie, la linguistique, la sociolinguistique, l'analyse interactionnelle, la pédagogie, la neuropsychologie, etc. Pour chacune de ces spécialités, les corpus de SMS et de profils constituent une base de recherche précieuse.

3 Quelques disciplines concernées par l'étude des SMS

3.1 La linguistique

La linguistique des SMS peut s'intéresser au corpus de *sms4science* et ce, en travaillant sur des données variées. En ce moment, la linguistique attache une importance particulière à la graphie dans les messages *bruts*[19]. Ainsi, elle analyse les types d'abréviation (cf. Fairon et al. 2006c) et s'attache à la récurrence de certaines néographies qui pourraient devenir avec le temps des variantes orthographiques. Elle se concentre aussi sur les messages transcrits et adopte souvent une visée lexicologique: elle tente par exemple d'identifier et de classifier un certain nombre de créations lexicales (par ex. le sigle *mdr*) qui apparaissent en contexte SMS. La linguistique étudie également la place des emprunts et les règles de l'alternance de code en contexte SMS (cf. Cougnon 2011). Dans l'avenir proche, ces études graphiques et lexicologiques vont être complémentés par d'analyses grammaticales à proprement parler, c'est-à-dire des analyses de morphologie et de syntaxe dans les SMS (marquage d'accord, de cas, ordre de mots, ellipses etc.).

Afin de faciliter ces analyses linguistiques, le projet *sms4science* propose un outil de recherche textuelle[20], autorisant des requêtes dans les SMS bruts autant que dans les SMS transcrits en graphie standard. La Fig. 4 présente l'interface de cet outil.

[19] C'est-à-dire non transcrits en graphie standard.

[20] Notons que cet outil permet également des recherches d'expressions régulières.

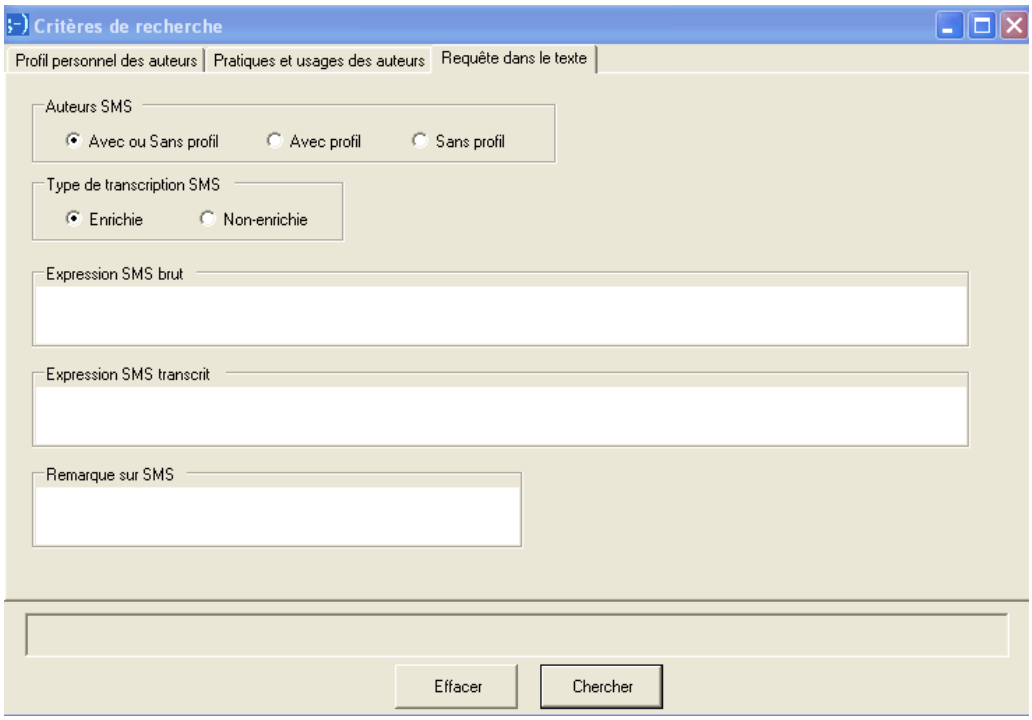

Figure 4: Interface de recherche textuelle proposée par le logiciel *sms4science*[21]

3.2 La sociolinguistique

En sociolinguistique, on voit apparaître un intérêt grandissant pour le corpus SMS, notamment en ce qu'il permet une série d'études sur la variation (diatopique, diaphasique et diastratique): les variétés d'une même langue, les registres différents en fonction du destinataire, les divers styles, l'usage de l'argot, le recours à un langage *ordinaire*, les marques de l'oral (cf. Fairon et al. 2006a, Cougnon/Ledegen 2008), etc. Cet intérêt englobe donc également le sujet des langues minoritaires, des dialectes, des régionalismes, des créoles, etc. La variation s'observe encore à travers les conséquences des contacts entre les langues: *code-switching*, *code-mixing* et emprunts font désormais l'objet des études sociolinguistiques sur les SMS (cf. Dürscheid/Stark 2010, Cougnon 2011).

Un deuxième courant de la sociolinguistique se penche sur les pratiques linguistiques en fonction de données démographiques, sociales ou psychologiques. Ainsi, grâce au corpus *sms4science* (voir Fig.s 5 et 6), les chercheurs peuvent restreindre leurs requêtes textuelles (par ex. «rechercher tous les messages qui comportent la forme "*que* + infinitif"») à une population précise («et émis par les hommes âgés de plus de 30 ans»). Ils peuvent également interroger la base de données à propos des pratiques électroniques interpersonnelles: en effet, certaines questions posées aux participants, telles que «à qui envoyez-vous le plus souvent des SMS ?», permettent d'en savoir plus sur le sujet.

Enfin, dans chaque aire géographique où la collecte de SMS et de profils a eu lieu, une zone de commentaire libre était proposée aux participants. Les réponses enregistrées dans cette zone offrent des informations très précieuses sur les représentations linguistiques et, en les comparant aux productions effectives qui apparaissent dans les messages, elles apportent

[21] On y distingue trois zones de recherche principales: recherche dans les messages bruts, recherche dans les messages transcrits et recherche dans les remarques émises par les transcripteurs. Les recherches peuvent porter sur des unités lexicales, syntaxiques ou des expressions régulières.

également des clés sur l'insécurité linguistique[22] des locuteurs. Exemple: *ben en fait ça dépend de la personne à qui j'écris et surtout de la circonstance. Exemple si j'écris à un copain j'écris souvent en mélangé mais à mes vieux c'est en français plus ou moins correct et si c'est un poème par exemple ben là c'est en bon français ;)*. En outre, cette notion d'insécurité linguistique est particulièrement importante et ne touche pas seulement en général les productions dans leurs écarts par rapport à la norme orthographique, mais concerne plus spécifiquement aussi les écarts par rapport à la norme dans les zones périphériques francophones.

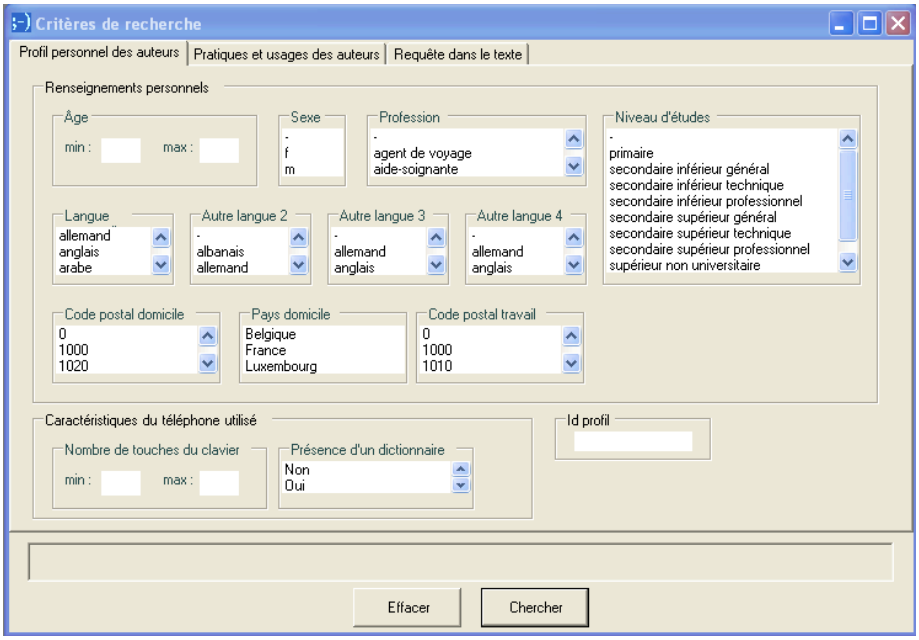

Figure 5: Interface de recherche sur les profils des participants au projet *sms4science*

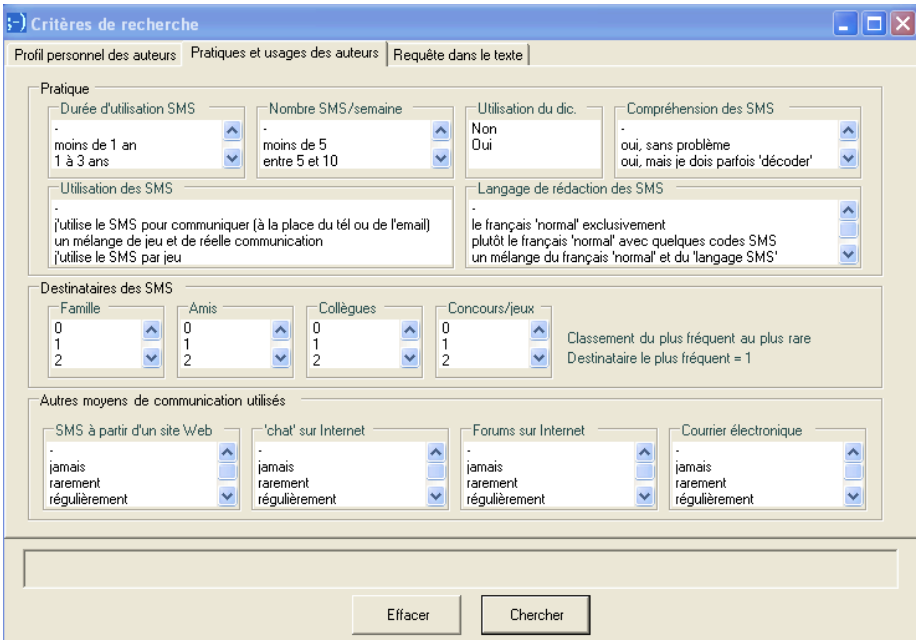

Figure 6: Interface de recherche sur les pratiques des participants au projet *sms4science*

[22] Francard (1997: 172) explique: «les locuteurs dans une situation d'insécurité linguistique mesurent la distance entre la norme dont ils ont hérité et la norme dominant le marché linguistique. L'état de sécurité linguistique, par contre, caractérise les locuteurs qui estiment que leurs pratiques linguistiques coïncident avec les pratiques légitimes, soit parce qu'ils sont effectivement les détenteurs de la légitimité, soit parce qu'ils n'ont pas conscience de la distance qui les sépare de cette légitimité».

3.3 Le traitement automatique du langage

Le traitement automatique du langage, plus fréquemment nommé sous son acronyme TAL, est une discipline qui travaille très souvent sur la base de corpus. En effet, un corpus de mots, de messages ou de textes, comporte des structures récurrentes, des logiques internes qui peuvent être formalisées, en vue d'une meilleure compréhension, d'une génération de nouveaux corpus ou d'autres utilisations spécifiques. Les corpus de SMS n'y font pas exception. En effet, ces corpus sont originaux, car la langue écrite en usage montre un écart par rapport à la norme orthographique (ex. «je veux 1 cal1» pour «je veux un câlin»). Ainsi, le TAL se propose tout naturellement de normaliser cet écrit au moyen d'un outil de transcription automatique en graphie standard. Cet outil peut par exemple être intégré dans le système interne d'un téléphone portable, pour des individus non habitués à la lecture de ces phénomènes scripturaux.

Le TAL se propose également d'utiliser les corpus de SMS afin de construire des outils de vocalisation. La vocalisation d'un SMS reçu sur un téléphone portable est utile à de nombreux groupes d'individus. Elle sert en premier lieu à des personnes visuellement déficientes, qui se servent d'un téléphone mobile. Ainsi, au moment de recevoir un SMS, celui-ci est directement vocalisé par l'outil en question. La vocalisation aide de la même manière les clients de téléphonie fixe, qui ne disposent pas d'un écran pour visualiser le message du destinateur. Enfin, un système de vocalisation peut être utile aux automobilistes, dont l'attention ne peut pas être, légalement et pratiquement, déviée de la conduite elle-même. Le système de vocalisation des SMS fonctionnerait ainsi comme les appels en kit main-libre. La Fig. 7 présente le logiciel de transcription et de vocalisation des SMS mis au point par le Cental (cf. Beaufort et al. 2010a).

Figure 7: Outil de transcription et de vocalisation des SMS. © 2010, Cental

4 Un exemple d'analyse: les statistiques du SMS

Les études statistiques constituent un outil précieux pour l'exploration et l'analyse de corpus SMS. On peut les décliner selon deux axes différents, suivant que l'approche se limite à décrire les données ou qu'elle vise à inférer des tendances globales à partir d'un corpus (considéré alors comme un échantillon représentatif d'une population plus large). Chacune de ces deux techniques possèdent des avantages, mais également des limites méthodologiques.

La première démarche possible est **descriptive**. Elle permet d'explorer un corpus de manière systématique et fournit des indices résumant diverses caractéristiques du corpus. Ces statistiques, qu'elles soient de localisation ou de dispersion, restent sommaires: elles doivent être conçues comme un simple moyen d'épauler l'analyse qualitative du linguiste, que ce soit en l'informant d'une problématique qui lui aurait échappé ou en lui fournissant une mesure précise d'un phénomène qu'il aurait repéré. Ainsi, on remarque dans le corpus belge de nombreux emprunts à des langues étrangères. Après une exploration systématique des données, on observe que 83% des emprunts proviennent de l'anglais. Reste au linguiste à détailler, voire à expliquer ce phénomène.

La seconde démarche, **inférentielle**, est plus ambitieuse, mais également plus risquée. Il s'agit de chercher à expliquer des phénomènes linguistiques en fonction d'une série de variables explicatives supposées entretenir une relation de corrélation[23] avec le phénomène d'intérêt. Étant donné que les variables à disposition pour ce type d'analyse sont par exemple de nature socio-démographique (sexe, âge, région d'origine, langues pratiquées...), on adopte alors une perspective sociolinguistique. Une telle démarche a déjà été mise en pratique dans Cougnon/François (2010).

4.1 La question méthodologique

Les études descriptives et inférentielles sont toutes deux confrontées à une première problématique: les *outliers* (ou «données aberrantes»). En effet, la préparation des corpus[24] requiert des interventions manuelles qui engendrent un certain nombre d'erreurs inévitables[25]. Ainsi, lorsque l'analyse révèle des outliers, il est primordial de déterminer si ceux-ci sont effectivement des données aberrantes issues de ces erreurs ou simplement des messages hors normes à conserver.

Deuxièmement, le traitement de corpus de SMS pose la question de la représentativité des données. Nous avons montré précédemment (Cougnon/François 2010) qu'il était probable que les collectes basées sur un échantillonnage par volontaires ne produisent pas des échantillons représentatifs des populations qu'elles ciblent. Nous avons pris le corpus belge en exemple et nous ne sommes pas parvenus à apporter une réponse définitive à cette question. En effet, il est normalement possible de confirmer ou infirmer la représentativité d'un jeu de données en lui appliquant un test khi-carré d'ajustement (Agresti 2002: 22), qui compare la distribution empirique observée dans le corpus à la distribution théorique de la population. Cependant, dans le cas des études sur les SMS, les opérateurs de téléphonie mobile ne sont pas autorisés à transmettre des informations précises sur les caractéristiques de leurs clients, afin de garantir le respect de la vie privée. Il est dès lors impossible de connaître précisément les attributs de cette population d'utilisateurs de SMS. Comme expliqué précédemment, Cougnon/François (2010) ont effectué un tel test d'ajustement sur le corpus belge en prenant comme population

[23] Rappelons qu'une corrélation entre des variables A et B peut correspondre à une relation de cause à effet (A engendre B), ce qui intéresse le chercheur. Toutefois, la démarche statistique seule n'autorise pas cette conclusion, car A et B peuvent tous deux être liés à la même cause C et varier simultanément.

[24] Ceci inclut la transcription en graphie standard ainsi que l'anonymisation.

[25] Par exemple, on note une irrégularité dans le processus d'anonymisation, tel que l'anonymisation d'un numéro de téléphone tantôt via la balise {TEL}, tantôt via {NoTEL}.

de référence les habitants de Wallonie et Bruxelles. D'importantes divergences ont été observées entre cette population et les participants à la collecte de SMS en Belgique francophone. Malheureusement, rien ne permet de décider s'il faut interpréter ces différences comme le résultat d'un biais dans l'échantillonnage ou comme le fait que les utilisateurs de SMS ne recouvrent pas cette population de Wallonie et Bruxelles.

Ainsi, le chercheur utilisant des techniques issues de la statistique inférentielle pour l'étude des SMS se doit de rester très prudent dans ses conclusions tant que ces questions de représentativité ne sont pas tranchées. En effet, si cette condition n'est pas satisfaite, tout résultat d'un test d'inférence statistique effectué sur la population d'intérêt risque d'être biaisé.

La question de la représentativité touche également les messages du corpus. En effet, dans un corpus issu d'une méthodologie telle que celle de *sms4science*, où la participation n'est ni encouragée ni limitée, certains participants sont à l'origine d'un très grand nombre de messages dans le corpus, tandis que d'autres peuvent n'avoir envoyé qu'un seul message. Un même phénomène peut donc paraître récurrent, alors qu'il se trouve en réalité être le fait d'un seul locuteur et devrait être considéré comme un hapax ou un trait idiosyncratique. Par conséquent, il est important de bien distinguer le niveau auquel l'analyse s'applique: celui des locuteurs, des messages ou des phénomènes linguistiques.

4.2 Longueur des SMS et taux d'abréviation

Dans la seconde partie de cette section, nous allons proposer quelques exemples d'analyses statistiques du corpus. Nous nous intéresserons ainsi à la distribution du nombre de caractères par messages en fonction de la région où ces messages ont été envoyés, puis au taux d'abréviation entre le nombre de caractères des messages bruts et celui de leur transcription en graphie normalisée.

La longueur des messages, mesurée en nombre de caractères, fournit une première voie d'exploration du corpus. La Fig. 8 résume les caractéristiques du corpus total, ainsi que celles de chaque aire géographique. On y observe que les messages récoltés au Québec sont en moyenne plus courts, et qu'ils sont suivis par ceux de la Réunion. À l'opposé, les SMS collectés en Suisse et en Belgique sont nettement plus longs. Les médianes trahissent un phénomène intéressant: si certains messages suisses sont très longs (dont un message-fleuve de plus de 2200 caractères), ce qui fait croître la moyenne, il y a un plus grand nombre de messages belges qui, tout en dépassant les 100 caractères, restent de longueur modérée.

Région	Nombre de données[26]	Médiane	Moyenne	Écart-type	Message le plus long
Réunion	12139	57	72,7	60,8	842
Belgique	26581	107	105,4	59	1046
Suisse	23987	98	110,9	83	2209
Québec	4698	42	56,9	54,5	908
Total	67405	89	98,1	70,79	2209

Figure 8: Longueur des messages en nombre de caractères

[26] Nous spécifions à chaque fois le nombre de données utilisées pour les analyses. En effet, certains messages doivent encore être transcrits. Par ailleurs, selon les régions, il a été nécessaire d'écarter un certain nombre «d'outliers». Il s'agit soit d'erreurs (messages vides, sans traduction ou qui ont été répétés), soit de messages en créole ou en anglais dont la transcription n'a pas été effectuée.

Ces premières observations se voient confirmées par l'étude des quatre distributions[27] suivantes parmi lesquelles on observe deux tendances distinctes. La première tendance, qui regroupe les corpus du Québec (Fig. 9) et de la Réunion (Fig. 10), montre une quantité importante de messages courts (moins de 50 caractères) et une longue queue de distribution sur la droite ; tandis que la seconde tendance, propre aux corpus belge (Fig. 11) et suisse (Fig. 12), se caractérise par une distribution plus ou moins uniforme jusqu'aux alentours des 160 caractères, où apparaît un pic, suivi d'une longue queue de distribution à droite.

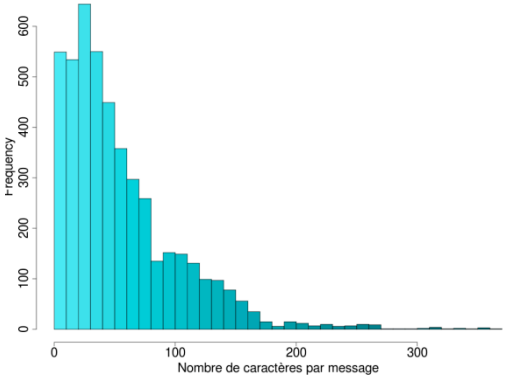

Figure 9: Nombre de caractères/SMS – Québec

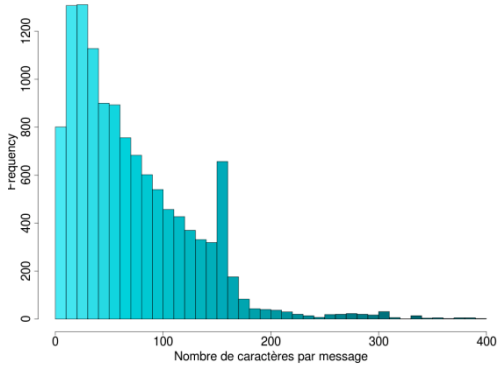

Figure 10: Nombre de caractères/SMS – Réunion

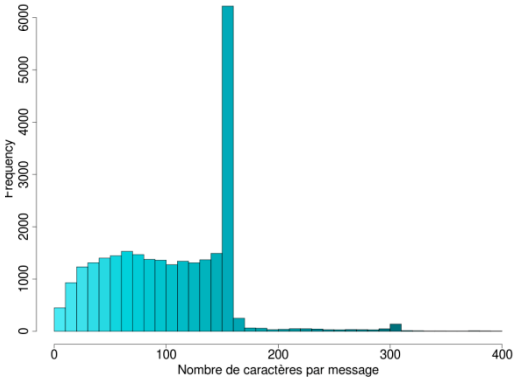

Figure 11: Nombre de caractères/SMS – Belgique

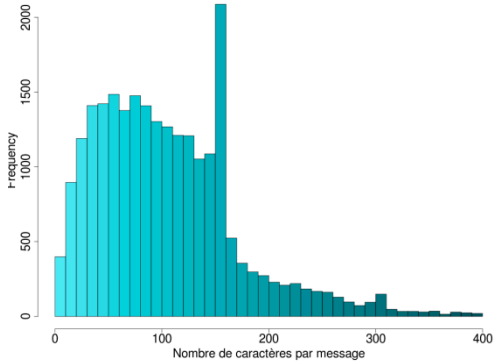

Figure 12: Nombre de caractères/SMS – Suisse

Ce pic, également présent dans les données réunionnaises, est particulièrement intéressant. Il s'explique par la limite imposée par les opérateurs de la téléphonie mobile[28] sur la taille des SMS et par le fait que le dépassement de ce seuil entraîne un surcoût. On remarque toutefois des comportements différents face à cette contrainte. Les messages belges trahissent un souci extrême de se limiter à un seul message (de 160 caractères) par envoi. Nous trouvons plus de 6.221 messages comportant de 151 à 160 caractères, ce qui correspond à près d'un quart du corpus total. Par ailleurs, nous observons un phénomène similaire, à une moindre échelle, pour la limite des 320 caractères. Le corpus suisse révèle, dans une moindre mesure, une tendance similaire avec un pic de plus de 2.000 messages autour des 160 caractères. À l'opposé, la distribution du Québec ne révèle aucune préoccupation des utilisateurs par rapport à cette limite. Cela s'explique sans doute par le fait que les messages sont nettement plus

[27] Pour des raisons de lisibilité des graphiques, les histogrammes présentés ici ne considèrent pas les messages dont la longueur dépasse les 400 caractères.

[28] Cette limite régit l'ensemble du marché mondial de la téléphonie mobile, à l'exception de quelques pays, tels que le Japon. Les régions participant au projet *sms4science* sont toutes concernées par cette contrainte.

courts et que le seuil est rarement atteint. Le cas de la Réunion, dont le profil est assez similaire à celui du Québec, est toutefois caractérisé par un tel pic, même si son ampleur (657 messages de 151 à 160 caractères) n'a rien à avoir avec la situation belge.

Au terme de ces premières observations, on constate une nette divergence dans les comportements des usagers du SMS issus des quatre régions. Il apparaît également que la limite de taille des messages influe sur les comportements des utilisateurs du SMS, particulièrement dans le corpus belge francophone. La question qui se pose alors est la suivante : puisque certains utilisateurs essaient d'optimiser la quantité d'information[29] qui transitent par SMS, n'auront-ils pas davantage tendance à abrévier leurs messages lorsque ceux-ci approchent du seuil de surcoût?

Afin de tenter de répondre à cette question, nous avons décidé de tirer profit d'une nouvelle variable pour chaque message: le taux d'abréviation. Il est défini comme la différence en nombre de caractères entre le SMS brut et sa transcription en graphie standard; cette différence est ensuite normalisée par le nombre de caractères de la transcription, afin de rendre le taux d'abréviation indépendant de la longueur des messages. La Fig. 13 résume l'essentiel des mesures calculées pour cette nouvelle variable.

Région	Nombre de données	Médiane	Moyenne	Écart-type	Proportion de SMS non abréviés	Proportion de SMS plus longs que leur transcription
Réunion	12139	0,188	0,175	0,26	5,8%	4,7%
Belgique	26581	0,074	0,095	0,098	18,7%	2,5%
Suisse[30]	/	/	/	/	/	/
Québec	4698	0,12	0,146	0,163	8,7%	5,3%
Total	43418	0,106	0,124	0,17	14%	3,4%

Figure 13: Taux d'abréviation des messages

Les moyennes de la Fig. 13 montrent que l'on rencontre les messages les plus abréviés dans le corpus de la Réunion, avec un taux moyen presque deux fois supérieur à ceux issus de Belgique francophone. Le phénomène est également assez important au Québec. Or, rappelons que les messages de Réunion et du Québec sont aussi ceux qui comportent le moins de caractères. On peut donc postuler deux hypothèses:

- ce sont les messages les plus courts qui sont les plus abréviés.
- les utilisateurs de SMS du Québec et de la Réunion, dont les messages sont plus courts en moyenne et qui semblent davantage abréviés, n'useraient pas du sms pour les mêmes fonctionsque les utilisateurs belges (et peut-être suisses), lesquels l'utilisent pour échanger une quantité d'informations plus grande.

Si la seconde hypothèse nécessite d'être explorée dans un futur travail selon une perspective plus qualitative, la première peut être examinée à l'aide de quelques outils statistiques. La Fig. 14 présente, pour le corpus du Québec, un nuage de points qui résume la relation entre la longueur des SMS et leur taux d'abréviation. On y voit que ce sont les messages les plus courts qui atteignent les plus hauts taux d'abréviation[31]. De plus, la corrélation entre ces deux

[29] Nous considérons ici la notion d'information dans une acception proche de celle de Shannon (1948).

[30] Les données suisses n'ayant pas encore été transcrites, nous n'avons pas pu appliquer cette analyse à cette partie du corpus.

[31] Notons qu'il est probable que la longueur des messages conserve une certaine influence sur le taux de réduction, en ce sens qu'il est plus facile d'abrévier quelques mots (ex.: «rendez-vous là à 9 heures» qui devient

variables est significative, même si elle reste faible (r = -0,28; p < 0.001). Les SMS les plus courts tendent donc à être plus abréviés. La situation est relativement similaire pour la Réunion. Par contre, le profil du corpus belge (Fig. 15) diffère. Non seulement, il semble que plus les messages sont longs, plus ils sont abréviés (r = 0,17; p < 0,001), mais on observe aussi un effet du seuil des 160 (et 320) caractères sur le niveau d'abréviation.

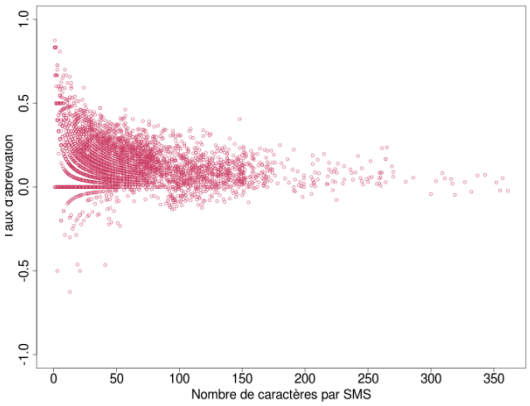

Figure 14: Taux d'abréviation du nombre de caractères par SMS – Québec

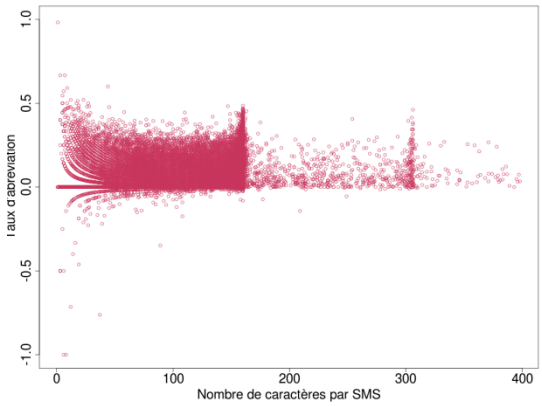

Figure 15: Taux d'abréviation du nombre de caractères par SMS – Belgique

Ainsi, nous clôturons notre étude en revenant sur la question posée à la fin de la section précédente, à savoir si le taux d'abréviation est significativement plus élevé dans les messages qui comportent près de 160 caractères. Pour ce faire, nous avons isolé les messages comportant de 151 à 160 caractères des autres SMS et avons appliqué un test T de Student à ces deux échantillons distincts.

De manière attendue, on observe une différence largement significative (t = 29,32 ; p < 0,001) pour le corpus belge. Pour peu que celui-ci soit représentatif de la population des utilisateurs francophones du SMS, on peut en conclure que les messages belges d'environ 160 caractères présentent bien un taux supérieur d'abréviation aux autres SMS (13,3% contre 9%). Dans le corpus québécois, la différence est significative (t = -6,21; p < 0,001), mais inversée. Le niveau d'abréviation est plus faible pour les messages d'environ 160 caractères que pour les autres (9% contre 15,6%). Par contre, les résultats du test sont non significatifs pour la Réunion (t = 1,84; p = 0,07) où il n'y a guère de différence (19% pour les messages d'environ 160 caractères contre 17,4% pour les autres).

En conclusion, les analyses ont révélé des comportements singuliers dans les différents lieux de collecte. Les informateurs belges (et, dans une moindre mesure, suisses) tendent à rédiger des messages plus longs qu'ils adaptent à la limite des 160 caractères, alors que les utilisateurs de la Réunion et du Québec rédigent des messages plus courts et se soucient dès lors moins de cette limite. Malgré l'absence de contrainte liée au seuil, les messages de ces derniers sont généralement plus abréviés que ceux des usagers belges et suisses. Il semblerait donc que l'abréviation ne s'explique pas, pour ces deux régions, par une contrainte d'espace/de coût, mais réponde à une autre motivation qui pourrait faire l'objet d'une étude ultérieure.

5 Obstacles et perspectives

Le but du projet *sms4science*, qui est de collecter des corpus de SMS dans un grand nombre de régions pour la recherche scientifique, rencontre un certain nombre d'obstacles. La

«rdl 9h») qu'une longue phrase dont certains termes sont moins courants et ne comportent pas de formes abréviées aussi évidentes.

première étape de préparation de la collecte, telle que schématisée dans la Fig. 1, consiste à rassembler une équipe et à assurer un financement et un fonctionnement technique. Cette première étape n'est pas toujours simple : ainsi, trouver des académiques motivés dans une région où l'on a déjà un sponsor (Israël), constituer une équipe dans une région où l'on a déjà un académique motivé (Espagne) ou rassembler des fonds dans une région où l'on a déjà une équipe constituée (Grèce) sont quelques-uns des problèmes rencontrés. La deuxième étape de la Fig. 1 pose également certains problèmes: chaque région montre ses particularités légales et techniques. Dans certains pays, la gratuité de participation de la population au projet est impossible pour des raisons techniques (France). Dans d'autres, la protection de la vie privée restreint le type de question que l'on peut poser via le formulaire sociolinguistique (Québec).

Les quatre collectes réalisées depuis le lancement du projet international ont montré elles-mêmes quelques limites qui servent d'exemples pour le futur. Ainsi, on regrettera le faible taux de participation de la communauté francophone de Suisse, peut-être dû à la non-gratuité des envois pour les personnes relevant d'un autre opérateur que la Swisscom; de la même façon, on remarquera la participation relativement faible de la population au Québec où la non gratuité était de mise pour tous les usagers de la téléphonie mobile. Pour ce qui est de la participation au questionnaire sociolinguistique, la population a réagi de deux manières différentes: au Québec, et malgré toutes les précautions prises, on a constaté une certaine crainte du non respect de la vie privée; en Suisse romande, la procédure d'inscription a été perçue comme légèrement trop compliquée.

Au niveau des résultats de ces diverses collectes, comme l'ont notamment montré les analyses statistiques du corpus, les chercheurs rencontrent un problème de représentativité de la population étudiée: comment atteindre une représentativité statistique de la population? Comment obtenir les données concernant la population cible qui nous permettraient de réajuster ces corpus?

Enfin, le traitement et l'étiquetage des données montrent une réelle diversité en fonction des régions et surtout des équipes: par exemple, dans les laboratoires où les aspects sociaux prévalent, un prénom tel que *Abdel* peut être anonymisé en le remplaçant par *Akim*, ce qui permet de maintenir la référence ethnique; une équipe de statisticiens ou de talistes préférera remplacer le prénom par une suite de symboles abstraits qui respecte le nombre d'unités, tel que [XXXXX]. Une équipe pluridisciplinaire proposera *Ahmed* à la place d'*Abdel*, solution qui permet de maintenir la référence ethnique et le nombre d'unités. La diversité se manifeste également au niveau de la transcription des messages en graphie standard: certaines régions réalisent une transcription manuelle (c'est-à-dire lente, avec erreurs de distraction), d'autres se penchent sur la transcription automatique; cette transcription automatique pose notamment la question des lexies non françaises, qui ne sont pas reconnues par les outils actuels (par exemple à La Réunion, avec le créole). Cette diversité apparaît comme un obstacle à l'uniformisation des corpus qui permettrait d'obtenir des corpus comparables.

Les perspectives du projet *sms4science* sont nombreuses et pertinentes: un corpus de 100.000 SMS francophones est pratiquement constitué et un corpus de 100.000 SMS anglophones est visé. 3.708 profils sociolinguistiques sont déjà disponibles pour les études en tout genre sur les populations. Depuis octobre 2010, une nouvelle collecte est lancée en France métropolitaine[32]. Des collectes en Roumanie, au Canada, à Chypre, en Grèce et en Italie sont annoncées pour 2011. Si elles sont effectivement réalisées, elles offriraient un ensemble de 10 corpus dans 7 langues et beaucoup de dialectes: français, allemand, italien, romanche, anglais, roumain et grec. Cette perspective apparaît comme une richesse évidente pour la recherche scientifique future.

[32] Elle est organisée par l'université Stendhal Grenoble 3 et le Conseil général des Hautes Alpes. Pour plus de renseignement: www.france4science.org.

Bibliographie

Agresti, Alan (^{2}2002): *Categorical Data Analysis*. New York: Wiley-Interscience.

Anis, Jacques (2002): *Communication électronique scripturale et formes langagières: chats et SMS*. http://edel.univ-poitiers.fr/rhrt/document.php?id=547 <27/10/2009>.

Beaufort, Richard et al. (à paraître, a): "Une approche hybride traduction/correction pour la normalisation des SMS". In: *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10), actes électroniques, Montréal, Canada, juillet 2010*.

Beaufort, Richard et al. (à paraître, b): "A hybrid rule/model-based finite-state framework for normalizing SMS messages". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 11-16 July 2010, Uppsala, Sweden*: 770–779.

Cougnon, Louise-Amélie (2011): "Tu te prends pour the king of the world? Language contact in text messaging context". In: Hasselblatt, Cornelius/Houtzagers, Peter/van Pareren, Remco (eds.): *Language Contact in Times of Globalization*. Amsterdam/New York, Rodopi: 45-59.

Cougnon, Louise-Amélie (2008a): "La néologie dans 'l'écrit spontané'. Etude d'un corpus de SMS en Belgique francophone". In: *Actes du Congrès International de la néologie dans les langues romanes. Barcelone*: 1139-1154. (=*Sèrie Activitats* 22).

Cougnon, Louise-Amélie (2008b): "Le français de Belgique dans 'l'écrit spontané'. Approche d'un corpus de 30.000 SMS". In: *Travaux du Cercle Belge de Linguistique*. http://webh01.ua.ac.be/linguist/SBKL/sbkl2008/cou2008.pdf <10/06/2010>.

Cougnon, Louise-Amélie/Ledegen, Gudrun (2008): "*c'est écrire comme je parle*. Une étude comparatiste des variétés du français dans l'écrit sms". In: *Actes du Congrès annuel de L'AFLS. Oxford. 3-5 septembre 2008*. Bern et al., Peter Lang: 39-57.

Cougnon, Louise-Amélie/Beaufort, Richard (2010): "SSLD: a French SMS to Standard Language Dictionary". In: Granger, Sylviane/Paquot, Magali (eds.): *eLexicography in the 21st century: New applications, new challenges. Proceedings of eLEX2009*. Louvain-la-Neuve, Presses universitaires de Louvain: 33-42. (=*Cahiers du Cental* 7).

Cougnon, Louise-Amélie/François, Thomas (2010): "Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS". In: *Actes du colloque JADT 2010, volume 1*. Edizioni Universitarie di Lettere Economia Diritto: 619-630.

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

Fairon, Cédrick/Paumier, Sébastien (2006): "A translated corpus of 30,000 french sms". In: *Proceedings of LREC 2006, Genova*. http://hnk.ffzg.hr/bibl/lrec2006/ <4/12/2010>.

Fairon, Cédrick/Klein, Jean René/Paumier, Sébastien (2006a): *Le langage SMS*. Louvain-la-Neuve: Presses universitaires de Louvain. (=*Cahiers du Cental* 3/1).

Fairon, Cédrick/Klein, Jean René/Paumier, Sébastien (2006b): *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation*. CD-Rom. Louvain-la-Neuve: Presses universitaires de Louvain. (=*Cahiers du Cental* 3/2).

Fairon, Cédrick/Klein, Jean René/Paumier, Sébastien (2006c): "Le langage sms: révélateur d'1compétence". In: Didier, Jean-Jacques et al. (eds): *Le français m'a tuer. Actes du colloque "L'orthographe française à l'épreuve du supérieur"*. Louvain-la-Neuve, Presse universitaire de Louvain: 33-42. (=*Cahiers du Cental* 1).

Francard, Michel (1997): "Insécurité linguistique". In: Moreau, Marie-Louise (ed.): *Sociolinguistique, les concepts de base*. Bruxelles, Mardaga: 170-176.

Mediappro (2006): *A European Research project: The Appropriation of New Media by Youth. Rapport final, Commission européenne*. http://www.mediappro.org/publications/finalreport.pdf <2/07/2010>.

Shannon, Claude (1948): "A Mathematical Theory of Communication". *Bell System Technical Journal* 27: 379-423/623-656.