HAL
open science

Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88MILSMS

Yosra Ghliss, Frédéric André

To cite this version:

Yosra Ghliss, Frédéric André. Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88MILSMS. Ciara R. Wigham, Gudrun Ledegen. Corpus de communication médiée par les réseaux, L'Harmattan , 2017, 978-2-343-11212-1. hal-01722169

HAL Id: hal-01722169

https://hal.science/hal-01722169

Submitted on 3 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88MILSMS

Yosra GHLISS, PRAXILING (UMR 5267 – CNRS) – Université Paul-Valéry Montpellier 3
Frédéric ANDRÉ, STIH (EA 4509) – Université Paris-Sorbonne

Cet article propose une réflexion sur le comportement éthique et les pratiques méthodologiques qui se présentent lors de la constitution de corpus qui relèvent de la communication médiée par les réseaux (désormais CMR). En effet, compte tenu de sa très rapide expansion, au cours de ces deux dernières décennies et au vu de la multiplication des recherches dans ce domaine, il semble primordial de poser les bases d'une réflexion commune au sujet des enjeux sous-jacents de la constitution de corpus. Car s'il est relativement aisé d'acquérir des corpus relevant de la CMR, ces derniers demeurent néanmoins problématiques, et imposent d'être traités avec beaucoup de précaution ; il ne s'agit plus de discours littéraires ou d'archives ouvertement accessibles, mais de productions discursives personnelles. Il sera ici question d'un type en particulier de CMR : celui de la communication par SMS. En effet, la constitution de ce type de corpus pose un ensemble de contraintes que nous allons tenter d'exposer aussi exhaustivement que possible en première partie, avant de présenter par la suite les choix retenus par l'équipe du projet *sud4science* (http://sud4science.org/), lors de la constitution du corpus *88milSMS* (http://88milsms.huma-num.fr/).

Premièrement, nous exposerons les différentes contraintes éthiques et juridiques qui régissent la constitution de corpus électroniques (notamment de type SMS), ces corpus qui exigent consentement des donateurs car ils font apparaître de nombreuses données personnelles. Dans un second temps, nous présenterons brièvement les principales étapes de la méthodologie que propose le projet international *sms4science*, permettant la constitution de grands corpus de SMS. Enfin, dans une troisième partie, nous nous intéresserons plus en détails à la phase d'anonymisation du corpus 88milSMS, avant de chercher à en connaître les vrais enjeux éthiques, mais aussi juridiques, et de poser les bases d'une réflexion commune.

Directement confrontés à ces problématiques dans le cadre de notre participation au projet *sud4science*, lors de la phase d'anonymisation, nous n'avons cessé de nous questionner quant à la mise en place d'une méthode permettant de conserver un maximum de données malgré les contraintes, et il nous paraissait donc essentiel de partager cette expérience dans le cadre de cette rencontre internationale sur la CMR.

1. Cadre théorique

1.1. De la linguistique de corpus SMS

Avec la diversification rapide des médias et des pratiques communicationnelles associées, la linguistique de corpus est depuis quelques années au cœur des débats scientifiques (cf. Rastier : 2005, Cori *et al*, 2008). Pour bon nombre de linguistes, il était certain que la recherche devait être ancrée dans la réalité, dépassant ainsi la réflexion sur de simples énoncés préfabriqués, pour s'appuyer sur des énoncés authentiques, issus de situations communicationnelles attestées. S'est ainsi vu développée ce que l'on appelle la linguistique de/des corpus, avec notamment les travaux de John Sinclair (cf. Sinclair, 1991). Nous renvoyons par ailleurs aux travaux de Damon Mayaffre (voir notamment 2005) où il

présente une discussion intéressante quant à la nécessité de penser le « corpus » dans la réflexion linguistique.

Les débats se sont depuis succédé, afin de définir la notion de « corpus », notion aussi vague qu'imprécise. Pour Sylvie Mellet (2002 : 2), la notion de corpus est bien ancrée dans certaines traditions des sciences humaines et sociales (domaine juridique par exemple).Elle renvoie à un recueil formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline. Or, dans le champ linguistique, Mellet explique que la notion s'est complexifiée au cours des dernières décennies en fonction de la diversité des approches et des objectifs assignés à la constitution et à l'exploitation des corpus. Ce caractère flou de la notion de corpus a généré de nombreux travaux dont le but était d'en identifier les critères. Ainsi, du côté des analystes de discours par exemple, le corpus est appréhendé sous une acception générale, intégrant même les éléments iconographiques.

« Le corpus peut être naturel (ou attesté) : les occurrences sont prélevées dans des interactions sociales ; il peut être plus ou moins artefactuel : sa production est provoquée par le chercheur qui fait réaliser les occurrences à une population cible …» (Détrie & al., 2001)

D'autres linguistes, s'inscrivant dans une approche du traitement automatique des langues (désormais TAL), soulignent quant à eux l'importance de l'exploitation informatique du corpus. Pour cette étude, nous nous référerons, à la suite de Cougnon (2015) à la définition de Leech (1991 : 9) : « un ensemble conséquent de textes exploitables informatiquement ». Cette définition, aussi large soit-elle, permet en effet de ne pas se poser de barrières théoriques qui limiteraient l'exploitation dudit corpus, d'autant plus que celui-ci, dans notre cas, est constitué de SMS, et peut donc faire l'objet d'analyses diverses selon des axes et thématiques qui dépassent les frontières de la linguistique.

D'une manière générale, un corpus fait de données authentiques et attestées est souvent un gage de scientificité, car il permet, dans une certaine mesure[1], d'arriver à des résultats fidèles aux pratiques réelles. Selon ce principe, la qualité d'un corpus peut donc être déterminée en fonction de son authenticité. Dans ce cadre, un corpus SMS remplit largement le contrat de l'authenticité puisqu'il donne des données écrites attestées. Mais, de par la qualité des données (qui par nature font apparaître des informations personnelles), les SMS imposent des exigences strictes en matière de traitement.

L'authenticité du corpus, si elle valorise le travail scientifique, impose en contrepartie aux chercheurs de faire preuve de responsabilité et de rigueur dans la conception de toute démarche méthodologique à adopter. Or, cette démarche doit bien entendu répondre à une injonction éthique, mais également juridique, régissant la recherche scientifique dans son ensemble. Il serait intéressant alors de se demander qui décide des règles en matière de comportement éthique, devant le traitement de différents corpus ? Quelles sont les injonctions qui pèsent réellement sur la constitution de ces *digital corpora* ?

1.2. Les SMS, un corpus soumis à l'injonction juridique

Dans un numéro des *Cahiers de Praxématique* portant sur les « Corpus sensibles » (Paveau & Perea, 2015[2012]), de nombreux chercheurs questionnent le caractère dit « sensible » de leurs données et l'impact que celui-ci peut engendrer sur la recherche menée mais aussi sur le chercheur. L'ensemble des contributions du numéro partent des instructions émises par les textes juridiques pour les confronter à leurs données et discuter par la suite, les contraintes que ces textes posent à la recherche scientifique. Situant ce travail dans cette même perspective, le corpus SMS semble aussi répondre aux caractéristiques d'un corpus sensible. De fait, la constitution d'un corpus SMS est une phase très problématique, et ce, pour deux raisons : tout d'abord, les SMS renvoient à ce que les textes juridiques classent comme des *données sensibles* (notion que nous expliciterons en dessous) ; et ensuite, comme ils relèvent d'un genre écrit de l'intime au sens de Détrie (2015), ils sont ainsi susceptibles de contenir

[1] Un corpus constitué de données authentiques permet bien d'analyser des pratiques existantes, mais n'est pas forcément un gage de représentativité de la pratique au niveau global : il convient donc de clairement définir, en amont de toute constitution de corpus, la finalité de celle-ci, afin d'adapter notamment les méthodes d'échantillonnage.

des informations qui relèvent de la vie privée d'une personne. Ces deux points sont formellement bien explicités par la Commission nationale informatique et libertés (CNIL) qui définit ainsi :

Les « données sensibles » :

« Données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci » (Art. 8 de la loi Informatique et libertés, dans CNIL, Guide : la sécurité des données personnelles).

La « donnée à caractère personnel » :

« Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne. » (Art. 2 loi I&L).

Ainsi, suivant la logique des textes juridiques, tout SMS faisant apparaître une allégation raciste, stigmatisant un tiers, ou incitant au meurtre, au viol, à la prise de drogue, etc. (tout propos hors la loi), fait apparaître des données sensibles, ou à caractère personnel. La constitution de corpus SMS s'inscrit donc dans ce que le CNIL nomme le « Traitement de données à caractère personnel » :

« Toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. » (Art. 2 loi I&L) »

Ces différentes définitions, bien que très éclairantes, ne constituent néanmoins pas de réelle marche à suivre, aussi la CNIL stipule également tout un ensemble de recommandations à l'attention des chercheurs, listées ci-dessous.

- Préserver la vie privée.
- Ne pas mettre en danger les individus avec des données les concernant.
- Rassembler toutes les informations qui aident à la protection et à la sécurité des personnes.
- Assurer la transparence des opérations afin de permettre la reproductibilité des expériences.
- Satisfaire les contraintes de traçabilité et de conservation pour les générations futures.

(COMET[2] : 9)

Ainsi, tout chercheur doit respecter l'ensemble de ces règles, s'il souhaite constituer, puis étudier un corpus issu de la CMR. En ce qui concerne les SMS, une méthodologie claire, aidant à la constitution de corpus comparables entre eux, a heureusement été mise au point en 2004, grâce au projet international *sms4science*, puis fut reprise et adaptée au fil des années, au sein d'équipes régionales, telles *sud4science*.

2. Du projet sm4science à sud4science

Le projet international *sms4science*[3] (« SMS *for* science ») a permis de regrouper, à l'initiative du CENTAL (Centre de Traitement Automatique du Langage de l'université catholique de Louvain, en Belgique) des chercheurs de 19 universités et fondations, issues de 10 pays, dans le but de constituer des corpus de SMS les plus larges possibles. Son ambition est donc, encore aujourd'hui, la collection d'un maximum de sous-corpus géographiquement affiliés, qui présenteraient des caractéristiques

[2] www.cnrs.fr/comets/IMG/pdf/guide_promouvoir_une_recherche_inte_gre_et_responsable_8septembre2014.pdf

[3] Plus d'informations sur http://www.sms4science.org/

communes. La première étape consistait à mettre en place une méthodologie claire, permettant de constituer de grands corpus de données authentiques. Le projet a démarré en 2004 par une campagne de collecte de SMS limitée à la Belgique francophone, intitulée « Faites don de vos SMS à la science ». Ce premier pas constitue la première phase d'un protocole qui en comporte quatre : la collecte, puis les phases d'anonymisation, de transcodage, et, enfin, d'annotation.

Comme son titre le suggère, la collecte des SMS se base sur le volontariat ; les participants acceptent, après avoir rempli, s'ils le souhaitent, un questionnaire permettant de définir leur profil, de transmettre aux chercheurs les messages écrits de leur main enregistrés sur leur téléphone, et dans certains cas, toujours selon leur consentement, ceux qu'ils émettent en temps réel[4]. Les SMS sont alors enregistrés, puis traités selon les différentes phases décrites ci-après.

La phase de collecte terminée, commence alors l'étape d'anonymisation des SMS. Cette étape primordiale, que nous décrirons plus en détails *infra*, consiste à écarter toutes formes de données personnelles ou indésirables, permettant ainsi la publication du corpus.

Une fois l'anonymisation de l'ensemble des SMS collectés effectuée, commence éventuellement alors l'étape de transcodage : chaque SMS est transcrit sous une forme standardisée de la langue de référence. Il ne s'agit pas de réécrire un message, mais bien de le transcrire, en alignant une version *brute* et une version *standardisée*. *L'idée est de restituer l'orthographe et la grammaire afin d'aider la compréhension et la fouille automatisée, mais non pas d'« injecter » des éléments supplémentaires* (Panckhurst *et al.*, 2013).

Enfin, avec ou sans transcodage, une étape d'annotation optionnelle peut alors être effectuée : des étiquettes sont ajoutées aux SMS afin de faciliter la recherche et l'analyse de phénomènes précis. Huit étiquettes ont été utilisées dans le corpus *88milSMS* : *ABS* (pour *absence*, en cas d'omission d'un élément de la phrase standard), *BIN* (pour *binette*, indiquant l'ajout d'un smiley/émoticône), *GRA* (pour *grammaire*, étiquetant les variations graphiques d'ordre grammatical), *LAN* (pour *langue*, pour indiquer les cas de code-switching, code-mixing, ou la présence dans un message d'une langue autre que le français), *ORT* (pour *orthographe*, indiquant les variations d'ordre orthographique), *TYP* (pour *typographie*, indiquant les ajouts/suppression d'éléments typographiques), *MOD* (pour toute autre *modifications* par rapport à la graphie standardisée), et enfin *DIV* (pour *divers*, indiquant finalement tout autre élément intéressant) (Panckhurst *et al.*, 2013). Auparavant, le trop grand nombre de balises (elles sont par exemple au nombre de 18 dans le corpus traité par l'équipe québécoise), exigeait un temps de traitement très important, aussi, en 2011, deux journées avec l'ensemble des partenaires du projet ont permis de fixer les étiquettes employées.

En mettant en place cette méthodologie, le CENTAL a donc développé une manière rigoureuse de constituer des corpus de SMS géographiquement affiliés, qui regroupent les usages d'une large population, sur une période donnée. Ainsi, depuis 2004, plusieurs grands corpus ont été constitués selon cette méthode, dont les principales caractéristiques sont regroupées dans le tableau 1, placé ci-après :

[4] Donc, uniquement des messages qu'ils ont eux-mêmes rédigés, et non pas ceux qu'ils ont reçus, ces derniers ne pouvant pas être légalement étudiés.

Région	Durée de la collecte	Année de la collecte	Nombre de SMS recueillis (avant traitement)	Nombre de participants	Participants ayant répondu au questionnaire
Belgique	2 mois	2004	73 127	3 200	86,65 %
La Réunion	2 mois	2008	12 661	884	40,61 %
Suisse	2 mois	2009	23 987	2 784	47,27 %
Québec	6 mois	2010	7 274	297	100 %
France/Région Rhône-Alpes	3 mois	2010	22 054	359	79,39 %
France/Région Languedoc-Roussillon	3 mois	2011	93 085	410	95,80 %

Tableau 1 : sms4science - Bilan en 2015
sources : cf. www.sms4science.org

Afin de présenter plus en détail la phase d'anonymisation et ses enjeux, nous nous appuierons finalement sur le dernier corpus régional, intitulé *88milSMS*, dans la région Languedoc-Roussillon, en 2011. Bien que n'en définissant pas les règles, nous avons en effet directement participé à l'anonymisation de ce corpus, au sein de l'équipe du projet *sud4science*, à qui nous avons notamment signalé les cas problématiques.

3. Anonymisation du corpus *88milSMS*

3.1. Méthode d'anonymisation

Une fois la collecte arrivée à son terme, les chercheurs se sont trouvés en possession d'une grande base de données « brutes ». Ces milliers de SMS portaient parfois des renseignements d'ordre privé, explicitant des informations personnelles sur les donateurs (nom, adresse, etc.), ou demeuraient juridiquement ambigus, ne pouvant donc pas être directement exploités. La phase d'anonymisation allait finalement permettre de réduire ce nombre, en masquant les données personnelles, et écartant les messages problématiques.

Afin de suivre au mieux les recommandations de la CNIL indiquées *supra*, nous avons fait le choix d'établir une feuille de route, nous permettant de filtrer le contenu du corpus brut en fonction de règles clairement définies:

- Ne laisser aucune trace permettant d'identifier un scripteur ou un tiers.
- Ne pas publier les SMS dont le contenu relève de la *donnée sensible.*
- Éliminer les SMS chaines (voir *infra*).
- Éliminer les SMS publicitaires (des messages reçus et non envoyés).

De par ces différents choix, l'anonymisation, compte tenu des contraintes décrites ci-avant, permet donc non seulement de respecter la charte signée avec les donateurs, mais également de rendre le corpus conforme à la législation, autorisant sa publication ou celle de travaux issus de son exploitation. Mais la phase d'anonymisation ne répond pas uniquement à une injonction juridique : elle vient aussi faciliter le travail qui se fera après, en procédant à une sorte d'épuration des SMS récoltés. Ainsi, a-t-on fait le choix de supprimer tous les SMS publicitaires et autres messages qui ne rendaient pas compte d'une

pratique effective du SMS (SMS chaînes, ou messages envoyés par les organisateurs du projet *sud4science* aux participants).

La toute première étape de l'anonymisation consiste à attribuer un numéro d'identification à chaque participant au projet, afin que son profil, anonyme, mais présentant de très utiles informations (âge, sexe, langue maternelle et/ou seconde, niveau d'étude, type de téléphone utilisé, etc.) puisse être mis en relation avec les messages produits. Dans un second temps, les SMS bruts regroupés suite à la phase de collecte, faisant apparaître très régulièrement des informations personnelles sur l'usager (noms, adresses, numéros de téléphone, numéros de carte bancaire, ou autre), doivent impérativement être passés au crible afin d'effacer, comme nous l'avons vu *supra*, toutes données sensibles. Néanmoins, il s'agit également de conserver le type de données anonymisées, ce qui est rendu possible par l'intermédiaire de balises associées à un type de données précis. Ainsi, dans le cadre du traitement du corpus *88milSMS*, 12 balises ont été mises en place : <PRE> pour un prénom, <NOM> pour un nom, <SUR>pour un surnom, <TEL> pour un numéro de téléphone, <ADR> pour une adresse, <LIE> pour un lieu, <URL> pour l'adresse d'un site internet, <MAR> pour une référence à une marque déposée, <MEL> pour une adresse mail, <COD> pour un code, <AUT> pour tout autre type de données, et plus tard, <ETR>, balise placée au début des messages rédigés entièrement dans une langue autre que le français. Une fois le traitement d'anonymisation effectué, ces balises apparaissent directement dans le message anonymisé, accompagnées d'un chiffre indiquant le nombre de caractères effacés, comme dans l'exemple ci-après :

SMS n°19723 : Recherche beau vampire musclé comme <PRE_6> et <PRE_5> ou indien souriant sympa et... Musclé! Ca ferait une belle petite annonce ;-)

Dans cet exemple, deux prénoms, de respectivement 6 et 5 caractères, ont donc pu être anonymisés, sans que le fond du message n'en souffre. Nous avons par ailleurs dû baliser d'autres informations privées, comme des numéros de téléphone (<TEL>) et de cartes bleues (<COD>), ou encore des lieux (<LIE>), et des marques déposées (<MAR>). L'ensemble des balises sélectionnées masquent ainsi les données privées, mais donnent une indication sur le type de donnée anonymisée. Cette méthode d'anonymisation permet de conserver, d'une part, les SMS qui contiennent des données privées, mais également des informations sur l'élément masqué, de telle sorte que le sens véhiculé par le message n'est pas perdu ; le corpus finalement mis à la disposition des chercheurs regroupe ainsi une quantité maximale de messages, dont la qualité n'a pas pâti.

Dans le cadre plus large du projet international *sms4science*, d'autres choix ont pu être fait, en matière d'anonymisation des prénoms par exemple. L'équipe suisse a par exemple choisi de les remplacer par d'autres, permettant de *conserver les informations de genre*, de *faciliter la lecture*, et d'i*dentifier les conversations* (Cougnon, 2015). D'autres encore ont choisi de les laisser apparent, jugeant qu'ils ne constituaient pas d'élément suffisant pour permettre l'identification de l'auteur. Dans notre cas, nous avons fait le choix de tous les masquer, tout en sachant que cela allait engendrer une légère perte de données, dans l'optique de publier un corpus le plus « propre » possible vis-à-vis de la législation française actuelle, et qui le restera même si la loi est amenée à évoluer.

Outre l'étiquetage des données, la phase d'anonymisation nous a confrontés à des SMS relativement problématiques. Leur contenu s'inscrit en effet dans ce que le CNIL nomme des « données sensibles », aussi nous avons supprimé tous les SMS qui stigmatisent un tiers ou un groupe de personnes. Ces SMS qui renvoient à des discours racistes, xénophobes, sexistes ou autres, sont tout à fait intéressants dans l'optique de réaliser des recherches sur ces sujets, mais leur publication restent formellement interdite, notamment en tant que corpus. Dans ce cas de figure, le seul choix possible était donc d'écarter les messages du corpus, afin de pouvoir mettre en ligne la base de données récoltée.

L'étape d'anonymisation nous a également permis d'écarter d'autres SMS qui posent problème, comme les SMS-chaînes du type : « […] *Envoie ce sms aux personnes que tu aimes* […] *Si 5 sms te reviennent, demain 1 personne que tu aimes te fera une surprise.* », les doublons, les messages d'erreurs, ceux à caractère pornographique, en lien avec toutes formes de trafics illégaux, ou encore, en fonction

des recommandations d'un service juridique, les SMS dénigrant ouvertement un tiers, conformément aux recommandations du CNIL.

En suivant l'ensemble de ces directives, le corpus, ainsi balisé et épuré de ces messages indésirables, est donc prêt à être éventuellement transcodé, annoté, mais aussi, publié[5], fournissant aux chercheurs actuels, et aux générations futures, un témoignage authentique d'écrits éphémères de ce début de XXIème siècle.

3.2. Résultats

Le tableau 2, ci-dessous, présente finalement un bilan de l'ensemble de l'étape d'anonymisation effectuée dans le cadre du traitement du corpus *88milSMS*.

Type de modifications	Ajout de balises												Suppression d'un message
	PRE	NOM	SUR	TEL	ADR	LIE	URL	MAR	MEL	COD	AUT	ETR	
Nombre total	8647	682	963	117	72	93	13	52	27	47	14	349	4 563

Tableau 2 : Bilan de la phase d'anonymisation du corpus *88milSMS*

Cette phase du traitement a conduit à la mise en place de 10 889 balises, dont une très grande majorité fait référence à des prénoms, et à la suppression de 4 563 messages indésirables. Le corpus final se compose donc de 88 522 SMS, produits par un total de 410 scripteurs différents, dont 393 ont répondu à un questionnaire sociolinguistique.

3.3. Les enjeux de l'anonymisation d'un corpus de SMS

Il apparaît que la méthode décrite *supra* a bien fait ses preuves, en permettant la publication du corpus *88milSMS*, mais elle conduit invariablement à la réduction des possibilités d'exploitation de ces données. Elle présente donc des limites non négligeables, qui font toute la lumière sur les enjeux de la phase d'anonymisation. Cette étape ne doit pas se faire au fur et à mesure du traitement, mais être construite très en amont du projet, en fonction de problématiques qui n'ont pas encore vu le jour, tout en tenant compte de la finalité même de la constitution du corpus. Ménager la chèvre et le chou n'est néanmoins pas chose aisée, lorsqu'on cherche à produire un corpus pour le mettre à la disposition du grand public. En effet, dans un cadre comme celui des SMS, ou de toute autre pratique en lien avec la CMR, de nombreux axes de recherches pourraient être développés sur la base de ces données anonymisées et autres messages malheureusement écartés. Nous avons, avec l'équipe du projet *sud4science*, proposé une méthode réduisant les pertes de données potentiellement intéressantes, par l'intermédiaire d'un nombre conséquent de balises, mais celle-ci pourrait certainement être améliorée, à l'avenir, tout en restant conforme à la législation. Travailler ensemble sur ces questions nous semble aujourd'hui essentiel, dans la perspective d'un monde de la CMR toujours plus multitâche, qui intéresse des chercheurs de disciplines de plus en plus diverses.

4. Conclusion

La phase d'anonymisation, comme nous venons de le voir, consiste finalement à éliminer les données problématiques d'un corpus, soit en masquant une partie des SMS (celle où figurent les données personnelles des donateurs), soit en supprimant l'ensemble d'un message, si son contenu présente des données sensibles (incitation à la haine, racisme, etc.). L'anonymisation donc, si elle donne l'apparence

5 Le corpus 88milSMS est disponible en ligne à l'adresse suivante : http://88milSMS.huma-num.fr/

d'être une simple mesure de précaution, se trouve être une étape indispensable pour la constitution de corpus et ce pour deux raisons :

- Elle permet la publication des travaux qui ont utilisé le corpus.[6]
- Elle permet la mise en ligne du corpus lui-même, sa diffusion dans la communauté scientifique afin qu'il puisse être exploitable par différents chercheurs.

De fait, l'anonymisation entraînant immanquablement une perte de données, le linguiste est alors appelé à limiter au maximum ces pertes par des choix méthodologiques adaptés. Dans le cadre de la constitution du corpus *88milSMS*, la démarche était celle de masquer les données personnelles avec des balises qui catégorisent la nature de la donnée, ce qui permet de préserver le sens véhiculé par le message, tout en conservant des informations sur le nombre de caractères cachés. Or, la loi *Informatique et Liberté* ne ciblant pas exclusivement les SMS, on peut donc maintenant s'interroger : cette méthode pourrait-elle être réemployée afin d'anonymiser des corpus d'un autre type, regroupant par exemple des courriels ou des conversations issues de réseaux sociaux, et comment l'améliorer pour conserver encore plus d'informations ?

Avec la multiplication des types de corpus et des approches possibles pour les analyser, la phase d'anonymisation impose aux chercheurs de renégocier en permanence *les bonnes et les mauvaises pratiques* à adopter (Perea, 2015[2012]) face à des corpus juridiquement « sensibles ». Or, au vue des initiatives déjà prises et des solutions proposées, poser les bases d'une réflexion d'ordre épistémologique, questionnant les textes de loi pour les confronter avec les problématiques de la recherche scientifique, parait être une nécessité à laquelle nous devrions travailler collectivement.

[6] Beaucoup de docteurs se trouvent dans l'impossibilité de publier leur thèse et ce, parce que celle-ci a exploité un corpus relevant de données personnelles non anonymisées (corpus forum, chats,etc.)

REFERENCES BIBLIOGRAPHIQUES

Cori, Marcel & David, Sophie (2008) : « Les corpus fondent-ils une nouvelle linguistique? », *Langages* 3, 111-129.

Cougnon, Louise Amélie (2015) : *Langage et SMS: Une étude internationale des pratiques actuelles*, Louvain : Presses universitaires de Louvain.

Détrie, Catherine & al. (2001) : *Termes et concepts pour l'analyse du discours : Une approche praxématique*, Paris : Honoré Champion.

Détrie, Catherine (2015) : « Gentlemanminette d'amour, ma chou, colocounette et autres formes nominales d'adresse dans les SMS: de quelques spécificités liées au genre » in D. Ablali & al. (dir) : *En tous genres. Normes, textes, médiations*. Louvain-la-Neuve : L'Harmattan, 43-57.

Leech, Geoffrey (1991) : « The state of the art in corpus linguistics », in Aijmer K. & Altenberg B. (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London : Longman, 8-29.

Mayaffre, Damon (2005) : « Rôle et place du corpus en linguistique. Réflexions introductives » in Vergely, P (eds) : *Actes du colloque JETOU'2005*. Toulouse, Université de Toulouse-Le Mirail, 5-17.

Mellet, Sylvie (2002) : « Corpus et recherches linguistiques », *Corpus* [En ligne], 1 |mis en ligne le 15 décembre 2003.consulté le 27 mai 2016. URL : http://corpus.revues.org/7

Paveau Marie-Anne & Perea, François (2012), (dir.) « Corpus sensibles », *Cahiers de Praxématique* 59. Montpellier : Pulm

Panckhurst, Rachel & al. (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, 107-138.

Perea, François (2012) : « Bonnes et mauvaises pratiques dans les corpus consacrés aux sexualités numériques. Discussion et exemples autour du consentement et de l'anonymisation », *Cahiers de praxématique* 59. Montpellier : Pulm, 91-108

Rastier, François (2005) : *Enjeux épistémologiques de la linguistique de corpus. La linguistique de corpu*s, Grenoble : Presses Universitaires de Grenoble, 31-46.

Sinclair, John (1991) : *Corpus, concordance, collocation*, Oxford : Oxford University Press.