De la constitution d'un corpus de SMS

Comment gérer un flux de données personnelles sensibles

Yosra GHLISS, Université Paul Valery- Montpellier, **Praxiling UMR 5267**
Frédéric ANDRÉ, Université Paris-Sorbonne, **EA 4509, STIH**

Rennes, le 23 Octobre 2015

Etudier les SMS

- Caractéristique principale
  - Pratique faisant partie de l'ensemble plus large que constitue la CMR (*Communication Médiée par les Réseaux*)

- Problèmes relatifs à leur étude
  - Réunir assez de données pour permettre l'exemplification de phénomènes linguistiques spécifiques
    - ⇒ Corpus existants trop réducteurs
  - Etudier des données authentiques
    - ⇒ Implique la présence de données personnelles

Dimension juridique

- Extrait de l'article 2 de la loi n° 2004-801 du 6 août 2004, modifiant la loi du 6 janvier 1978, « Informatique et liberté »

  - Constitue une 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.

Dimension juridique

- Guide : *Promouvoir une recherche intègre et responsable, Comité d'éthique du CNRS*

- *Considérant 26 de la directive du parlement européen*

  […] pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne.

Problématique

- Quelle méthodologie mettre en place pour permettre l'étude de SMS authentiques, malgré ces contraintes ?

Projet SMS4SCIENCE

- Présentation générale :
  - Projet international :
    - Objectif : mettre en place une méthodologie permettant la constitution de grands corpus de SMS authentiques.
  - 2 phases :
    - Collecte
      - Médiatisation
      - Collecte (avec questionnaire optionnel)
    - Traitement
      - Anonymisation
      - Transcodage
      - Annotation (optionnelle)

Projet SMS4SCIENCE

- Bilan en 2015

Région	Durée de la collecte	Année de collecte	Nombre de SMS recueillis	Nombre de participants
Belgique	2 mois	2004	73 127	2 773
La Réunion	2 mois	2008	12 661	365
Suisse	2 mois	2009	23 987	1 311
Québec	6 mois	2010	7 274	298
France/ Rhône-Alpes	3 mois	2010	22 054	285
France/Languedoc-Roussillon	4 mois	2011	93 085	393

Projet SMS4SCIENCE

- Bilan en 2015

Région	Durée de la collecte	Année de collecte	Nombre de SMS recueillis	Nombre de participants
Belgique	2 mois	2004	73 127	2 773
La Réunion	2 mois	2008	12 661	365
Suisse	2 mois	2009	23 987	1 311
Québec	6 mois	2010	7 274	298
France/ Rhône-Alpes	3 mois	2010	22 054	285
France/Languedoc-Roussillon	**4 mois**	**2011**	**93 085**	**393**

Anonymisation

Anonymisation

- Logiciel développé
  - *Seek&Hide*, (Accorsi *et al.*, 2012)
    - Permettre l'anonymisation automatique
    - Faciliter l'anonymisation semi-automatique (phase de vérification et relecture)

⇒ *http://www.msh-m.tv/spip.php?article450*

Anonymisation

- Données à anonymiser :
  - Données personnelles
    - Nom => balise **<NOM>**
    - Prénom => balise **<PRE>**
    - Surnom => balise **<SUR>**
    - Adresse => balise **<ADR>**
    - Numéro de tel => balise **<TEL>**
    - Autre => balise **<AUT>**
    - Code (porte, carte bleue, etc.) => balise **<COD>**
    - Terme étranger => balise **<ETR>**

Anonymisation

- **Exemple :**

N°19 675 :

Non j y suis pas comme <PRE_8> <PRE_8> et <PRE_8> ne voulaient pas y aller et que parmi les TL personne ne voulait aller en cours-au lycée et au bar a coté oui, ms pas au cours de <NOM_7>.

Anonymisation

- Messages problématiques :
  - SMS chaînes, ou publicitaires
  - Doublons
  - Messages relatifs au projet
  - Messages faisant référence à des activités potentiellement illégales

    => suppression des SMS du corpus

Anonymisation

- Messages problématiques :
  - Dénigrement d'un tiers (service juridique à contacter)
    - Personnes
    - Organismes
    - Autres

      => suppression des SMS du corpus

Bilan

Raison d'anonymiser	Travail effectué
Identification de l'auteur	Mise en place de balises (PRE, ADR, COD, etc.)
Dénigrement d'un tiers (quel qu'il soit)	Suppression du message
Messages relatifs à des activités potentiellement illégales	Suppression du message
Chaînes, publicités	Suppression du message
Doublons	Suppression du message
Messages relatifs aux projet	Suppression du message

Conclusion

- Le corpus anonymisé peut maintenant être exploité
  ⇒ *http://88milSMS.huma-num.fr/*

- Limites : légère perte de données nécessaire à la publication des corpus
  ⇒ Passage de 93 085 à 88 522 messages, après anonymisation

Perspectives

- Le SMS : pratique relative à la CMR présentant le plus de contraintes, quant à leur exploitation en vue de recherches

=> Adapter la méthodologie et les outils (logiciels) pour constituer et traiter de nouveaux corpus (mails, messagerie instantanée, chat, SRSs, Muds, etc.)

Merci !!