Degache, C. & Garbarino, S. (Ed.) (2012). *Actes du colloque IC2012. Intercompréhension : compétences plurielles, corpus, intégration.* Université Stendhal Grenoble 3 (France), 21-22-23 juin 2012.

Stratégies pour l'anonymisation systématique d'un corpus d'interactions plurilingues

Christophe REFFAY[1], *François-Marie BLONDEL*[1], *Emmanuel GIGUET*[2]
[1] STEF – ENS-Cachan, IFÉ – ENS-Lyon
[2] GREYC, Université Caen Basse-Normandie, CNRS
christophe.reffay@ens-cachan.fr, francois-marie.blondel@ens-cachan.fr, emmanuel.giguet@unicaen.fr

RESUME.

Dans le champ de l'analyse des interactions textuelles, les chercheurs désirant partager leurs corpus font face à de grandes difficultés pour en éliminer les marques d'identification des personnes physiques. La loi européenne suggère pourtant que ces marques soient soigneusement retirées avant toute publication. Dans la suite des outils de la plateforme Calico dédiée à l'analyse des interactions en ligne, nous proposons ici un procédé interactif d'anonymisation systématique, fonctionnant sans dictionnaire a priori et donc applicable à toute langue. Ce procédé a été appliqué à un premier corpus plurilingue issu de Galanet. Cet article souligne les difficultés de l'anonymisation et présente les premiers résultats de cette expérience. Au-delà de la transformation elle-même, nous proposons deux stratégies de fouille afin de détecter de nouvelles graphies pouvant révéler des données personnelles.

ABSTRACT

Strategy for systematic anonymisation of multi-lingual interaction corpora

Considering the textual interaction analysis field, researchers who want to share their corpus are facing many difficulties when they try to remove the marks identifying physical persons from their corpus. The European law suggests that such marks may be removed before any publication of the corpus. Many tools dedicated to online discussion analysis have been already developed in the Calico platform. They are language independent. Following this way, we propose here an interactive and systematic anonymisation process working without dictionary and being then available for any language. This process has been applied to a first multi-lingual corpus coming from the Galanet project. This paper emphasises the difficulties arising during this anonymisation process. We present the results of this experience. Beyond the substitution of identity marks, we propose two mining strategies that help to detect new lexical forms that may reveal personal information.

MOTS-CLES : Anonymisation, Données personnelles, Corpus plurilingue, Partage de données.
Keywords: Anonymisation, personal information, multi-lingual Corpus, Data sharing.

Avec d'autres chercheurs (Henri et Charlier, 2005 ; King, 2007 ; Reffay et al., 2008), nous pensons que le partage des données de recherche est une étape indispensable dans le processus de comparaisons des résultats de la recherche en sciences humaines et particulièrement dans les recherches portant sur l'analyse des interactions. Dans le cas d'une expérience non reproductible telle qu'une session de formation ou d'échange, ce partage permet à d'autres chercheurs de répliquer le processus d'analyse, de le vérifier et de le discuter en profondeur. En accédant aux données, il devient plus aisé de repérer les contours des domaines de validité des résultats, ouvrant ainsi de meilleures perspectives pour la comparaison des résultats issus de contextes différents.

Sur le plan opérationnel, nous nous sommes déjà engagés dans le développement de méthodes et la réalisation d'outils pour favoriser le partage des données, en particulier :

- ☑ Calico (2012) : plateforme plurilingue pour l'analyse et le partage de corpus de discussions en ligne (Giguet et al., 2009 ; Blondel et Giguet, 2011) ;
- ☑ Mulce (2012) : plateforme de partage de corpus multimodaux de situations d'enseignement et d'apprentissage. (Reffay et al., 2008, Reffay et al., in press).

À partir de l'expérience de ces plateformes, nous avons pu constater que le coût d'anonymisation d'un corpus constituait un des obstacles qui limite le partage. C'est ce constat qui a motivé la recherche d'un processus d'aide à l'anonymisation systématique de corpus textuels d'interactions. Notre objectif est de proposer une méthode et des outils pour rendre l'anonymisation plus efficace, en réduisant le temps de travail du chercheur et en améliorant la qualité du résultat.

Après une brève présentation du contexte et des enjeux de l'anonymisation en section 1, et notre positionnement dans l'état de l'art en section 2, nous sensibiliserons le lecteur à la problématique (section 3) en nous appuyant sur des exemples tirés de notre corpus d'étude, issu de la plate-forme Galanet. La section 4 est consacrée à l'explicitation de la méthode d'anonymisation proposée. L'article détaille en section 5 les techniques de fouille avant de fournir les premiers résultats de l'application de la méthode sur le corpus d'étude en section 6. Nous concluons (section 7) par une discussion et précisons les suites envisagées.

1. Contexte et enjeu

L'anonymisation des corpus s'impose d'abord pour des raisons juridiques :

> *« Les applications informatiques à des fins pédagogiques et éducatives mobilisent des données permettant d'identifier directement mais aussi indirectement les personnes physiques. Une attention particulière doit être portée sur la collecte de données sensibles ainsi que sur les procédés d'anonymisation des données. »*
>
> (Mallet-Poujol 2004: p 21)

Dans son guide sur la protection de la vie privée et des données personnelles, Mallet-Pujol rappelle aux enseignants et aux chercheurs les règles juridiques qui s'appliquent à tout recueil ou traitement de données automatisé lié à une action éducative.

Dans le cas d'une activité qui fait intervenir des participants de plusieurs pays, une attention particulière doit être portée aux différentes réglementations nationales. En Europe, la protection des personnes physiques à l'égard du traitement des données à caractère personnel est essentiellement encadrée par trois directives : 95/46/CE, 97/66/CE et 2002/58/CE.

L'anonymisation des corpus s'impose aussi pour des raisons éthiques. Les relations qui se sont établies entre le chercheur et les personnes avec lesquelles il travaille et qui deviennent dans certains cas ses « sujets » d'étude, supposent une confiance réciproque et engagent la responsabilité du chercheur ; les documents qu'il va communiquer publiquement doivent par conséquent respecter certaines règles, souvent implicites, de non divulgation de données à caractère personnel. Au Canada, c'est le « Groupe consultatif interagences en éthique de la recherche », par le biais de son *Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres humains*, qui balise ces enjeux.

L'anonymisation a pour fonction principale de supprimer ou de masquer les données qui permettent d'identifier une personne soit directement, par son nom, son prénom ou un

identifiant unique, soit indirectement, par une adresse, un numéro de téléphone ou tout autre indice révélateur.

Encore faut-il être en mesure de repérer toutes ces données à caractère personnel.

Le cas des corpus de textes issus de sessions d'interaction en ligne est particulièrement important pour tous les chercheurs qui s'intéressent aux dispositifs de formation à distance. Pour les interactions qui ne mobilisent qu'une seule langue, le repérage des données peut reposer en partie sur des indices construits à partir d'une connaissance profonde de la langue en question. Si, comme dans le cas qui nous intéresse, les interactions ont lieu dans plusieurs langues simultanément, et si, de surcroit, les locuteurs ne maîtrisent qu'imparfaitement les langues qu'ils utilisent, une connaissance a priori ne paraît pas suffisante pour effectuer ce repérage des données personnelles.

Notre proposition repose sur l'idée qu'une méthode faisant appel aux données mêmes du corpus et basées sur l'exploitation de régularités contextuelles peut faciliter le repérage des données personnelles dans de grands corpus d'interactions multilingues.

1.1. Quels sont les risques et les difficultés ?

Au même titre que les applications informatiques citées dans (Mallet-Poujol 2004: p 21), les corpus d'interactions plurilingues contiennent *des données permettant d'identifier directement mais aussi indirectement les personnes physiques*, à savoir tous les acteurs ayant participé ou simplement ayant été cités dans ces interactions. Parmi les marques d'identification, on peut citer : les patronymes, les lieux de vie ou de travail, les institutions, les URLs de blogs ou de pages personnelles, les identifiants numériques (courriel, téléphone, MSN, Facebook, Twitter, …) L'objet de l'anonymisation est de détecter ces données personnelles, et, si nécessaire, de les éliminer afin d'empêcher toute personne extérieure d'identifier les acteurs, de les localiser ou d'entrer en contact avec eux.

Ce brouillage est nécessaire lorsque l'on rend les données accessibles à des tiers (hors contrat de recherche). En effet, dès lors que nous ne pouvons garantir les intentions des personnes pouvant accéder à un corpus, ce dernier doit être épuré des marques identificatrices. Les données de localisation que l'on trouve dans les récits autobiographiques sont autant d'informations utilisables par des personnes pour le kidnapping, le vol au domicile, le vol d'identité, etc.

Comme il a déjà été mentionné dans (Thomson et al, 2005 ; Reffay et Teutsch, 2007), ce brouillage doit être opéré avec suffisamment de précision pour conserver certaines marques culturelles ou maintenir la cohérence du contenu. Ainsi, dans un messsage de notre corpus d'étude, une élève se présente de la manière suivante : « [...] o meu é uma homenagem a uma de minhas tias e minha avó que se chamam **Ana** e ao resto de minhas tias que se chamam **Maria**. Daí, **Mariana** » Si dans ce message, on remplace **Mariana** par **Melita** sans changer **Ana** et **Maria**, le texte perd toute sa cohérence sémantique.

2. Etat de l'art et positionnement

2.1. Les marques d'identification introduites par le système : une tâche automatisable

Une partie des marques d'identification situées dans les dispositifs numériques sont introduites par le système technique lui-même pour permettre de rendre un service aux utilisateurs : date de dépôt d'un message, auteur, lecture d'un message, date de lecture, identifiant du fil de discussion ou de l'espace, en réponse à un autre message ou premier message, etc. Toutes ces marques étant générées de façon automatique, elles apparaissent nécessairement sous une forme et dans un contexte parfaitement identifiables. Elles sont donc relativement faciles à modifier en cas de besoin. Ces aspects peuvent être traités par les techniques d'anonymisation développées pour les bases de données et dont on trouve une synthèse dans le livre blanc (Net ltd, 2000).

2.2. Les marques laissées par les acteurs : un problème difficile

Les marques d'identification laissées par les acteurs eux-mêmes dans le contenu des interactions peuvent prendre des formes très variées et surtout non systématiques. C'est à l'anonymisation de cette catégorie de marques que nous nous intéressons dans cet article. Dans l'état de l'art, des travaux connexes portent sur le traitement des dossiers médicaux : il s'agit d'anonymiser des retranscriptions d'entretiens avec les patients. Au cours des entretiens médicaux, les noms des établissements et des professionnels de santé peuvent être évoqués ainsi que ceux des patients ou de leurs proches. Les données de santé pouvant être particulièrement sensibles, les données personnelles qui y sont attachées sont à masquer impérativement avant toute diffusion. Des techniques (Aramaki et al., 2006 ; Grouin et al., 2009 ; Gardner et al., 2010 ; Mestre et al., 2010) et outils tels que Scrub (Sweeney, 1996), ainsi qu'une méthodologie, PIPE (Neubauer & Heurix, 2011) ont été proposés : ils se basent sur des dictionnaires d'une langue particulière et d'un réseau spécifique (santé).

2.3. Notre positionnement

Nous souhaitons nous inspirer des travaux portant sur l'anonymisation des dossiers médicaux pour mettre au point une méthode adaptée à l'analyse des interactions en ligne, où l'orthographe et la grammaire sont instables, et où les contenus peuvent être plurilingues.

Conscients des difficultés qui peuvent apparaître lors d'une analyse effectuée sur un corpus anonymisé (Thomson et al., 2005), nous pensons que la tâche d'anonymisation ne peut être entièrement automatisée. C'est pour cette raison que notre travail s'oriente vers la construction d'un *processus interactif* qui nécessite l'intervention du chercheur pour identifier les données personnelles et de sa décision pour y substituer des données anonymes. En revanche, compte tenu de la taille de certains corpus (plusieurs centaines, parfois plusieurs milliers de messages), il nous semble qu'un outil pour assister le chercheur dans cette tâche d'anonymisation systématique est indispensable.

3. Illustrations de la problématique de l'anonymisation

3.1. Présentation du corpus d'étude

Le corpus que nous avons utilisé pour élaborer la méthode d'anonymisation proposée dans cet article est issu d'une session réalisée sur la plateforme Galanet (2012) et utilise donc l'intercompréhension pour l'apprentissage des langues romanes. Il s'intitule « Nômades... nomadi... nómades... des langues » et a été dirigé par Sandrine Deprez. Les interventions dans le forum s'échelonnent sur trois mois : d'octobre 2011 à février 2012. Ce forum contient 915 messages rédigés par 83 auteurs (sur 103 inscrits) en 5 langues romanes : français, italien, catalan, castillan et portugais (du Brésil). Les élèves sont inscrits par l'intermédiaire de leur lycée et sont âgés de 16 ans environ. L'objectif des échanges est de permettre à chacun de mieux comprendre la langue et la culture de l'autre. Dans le contrat didactique de Galanet, chacun doit s'exprimer prioritairement dans sa langue. Mais il est fréquent que les élèves utilisent d'autres langues que leur langue maternelle et certains ont même recours à plusieurs langues dans un même message.

Ce corpus de 212 902 caractères comprend 47 740 graphies au total. Le lexique permet d'identifier 9 653 graphies différentes avec des fréquences allant de 1 à 2163 occurrences. Les dix graphies les plus fréquentes sont : ‘,’(2163), ‘.’(1305), ‘de’(1015), ‘que’(965), ‘!’(859), ‘la’(673), ‘a’(556), ‘:’(515), ‘-’(451) et ‘)’(446). À l'autre extrême, 6074 graphies n'apparaissent qu'une seule fois dans tout le corpus.

Sur la plateforme Galanet, un espace privé intitulé « Qui est qui » permet aux participants d'échanger leurs données personnelles (adresse MSN, courriel, page Facebook, etc.) L'accès à cet espace est strictement réservé aux participants de la session. Comme nous avons pu le confirmer au cours de notre étude, l'existence de cet espace privé permet de limiter l'apparition de données personnelles à l'intérieur du forum. Dans la méthode proposée, le contenu de cet espace privé est utilisé dans le *processus* d'anonymisation. Mais il ne peut évidemment pas faire partie du *résultat visé*, i.e. : le contenu du forum anonymisé partageable.

En annexe 1, nous présentons quelques messages du corpus. Le lecteur intéressé par le contenu trouvera plus d'informations dans (Deprez, 2012). L'accord des participants ayant été obtenu au préalable, le corpus original est en accès public sur les plateformes Galanet et Calico. C'est pourquoi nous avons pu exceptionnellement dans cet article, présenter des cas réels avant anonymisation. Une version anonymisée de ce corpus est également disponible sur la plateforme Calico.

3.2. Illustrations sur des exemples extraits du corpus d'étude

Les quatre messages cités en exemple dans cette section sont donnés en annexe 1.

Dans le premier message, posté par Gabibr, malgré la barrière de la langue, nous identifions aisément qu'elle indique son prénom (Gabriela) et son patronyme (Medeiros) en précisant de plus qu'ils identifient son compte Facebook. Ces informations suffisent à l'identifier et à la contacter via sa page Facebook (parmi les 10 comptes ouverts sous ce nom). Dans ce cas précis, et pour qu'on puisse l'identifier, il suffirait de remplacer son patronyme

par un autre en indiquant (par un marquage spécifique) aux chercheurs (qui feraient une seconde analyse de ce corpus, une fois anonymisé) que cette graphie représentait un patronyme et qu'elle a été modifiée durant le processus d'anonymisation. Le cas échéant, on remplacera toutes les occurrences de cette graphie par la même graphie de substitution.

Dans le deuxième, Miche indique, sans aucune précaution, son adresse de messagerie électronique. Rien n'est plus facile pour un robot de détecter des adresses de courriel qui contiennent le caractère '@'. Cette adresse électronique permet bien évidemment de contacter directement la personne et dès lors, bien des scénarios peuvent être à redouter, allant de la simple farce aux pièges les plus tordus. C'est pourquoi, cette information ne peut rester indemne dans le corpus. Elle peut être remplacée par une adresse inventée pourvu que l'adresse de substitution n'existe pas et qu'elle soit suffisamment éloignée syntaxiquement et sémantiquement de l'adresse originale. Par exemple *codabscons@hotmail.it* qui conserve la marque 'it' identifiant le pays et la langue, pourrait convenir.

Dans les deux exemples précédents, les marques d'identification sont évidentes et permettent d'identifier ou de contacter la personne directement, ce qui est tout à fait naturel puisque les auteurs de ces deux messages les ont écrits dans l'intention d'être contactés par les destinataires du message. Arrêtons-nous un instant sur cette remarque. En publiant le corpus non anonymisé, les lecteurs potentiels du corpus ne sont plus les destinataires des messages originaux. C'est pourquoi il est si important que les acteurs des expérimentations (dans notre cas, les élèves) soient informés par les chercheurs (via un formulaire de consentement éclairé) du devenir des traces de leurs interactions dans le dispositif, de leur traitement et du public qui pourra y avoir accès ; c'est précisément l'esprit de la directive européenne 95/46/CE.

Le troisième message montre un des problèmes les plus difficiles à détecter et à résoudre dans la problématique de l'anonymisation. En effet, nous aurions pu penser, comme cela est fréquemment accepté par les chercheurs, que le seul prénom (soit-il composé) d'un participant n'est pas une information suffisante pour l'identifier de façon formelle. Dans cet exemple PBS nous explique elle-même (à qui comprend un peu l'espagnol) que nous aurions eu tort. En effet son prénom comprenant un prénom thaïlandais (rare en Thaïlande) et un prénom italien (rare en Italie), leur composition le rend unique dans une recherche sur Google ou Facebook. De plus, sa recherche sur le web permet de trouver immédiatement son patronyme, et, via une page référencée du site de son lycée, le nom et le site web de son lycée, le nom de sa classe et même celui de sa camarade de classe, primée avec elle lors d'un concours local. Dès lors, avec le nom de la classe, du lycée et la date des messages, il devient possible, grâce aux différents prénoms ou pseudos de son équipe, de retrouver l'identité de tous les élèves de la classe de PBS. Ce constat rend obligatoire le remplacement des graphies de ses deux prénoms au cours du processus d'anonymisation alors que cela n'était pas nécessairement envisagé au départ. Ce remplacement pourrait ne pas suffire puisque PBS explique clairement l'étymologie de chacun des composants de son prénom de sorte qu'on ne peut transformer l'un des composants sans modifier aussi son étymologie. Pour faire le travail plus finement, il faudrait proposer une étymologie différente (correspondant à d'autres prénoms) et recomposer ainsi un nouveau prénom rare (voire inexistant) pour que le contenu du message reste cohérent. Pour finir sur cet exemple, et pour conserver la cohérence des prénoms avec le pseudo (PBS), il faudra aussi envisager de modifier les deux premières initiales en

conséquence. Une alternative plus rapide et efficace serait de choisir deux autres prénoms, si possible thaïlandais et italiens pour conserver leurs attributs culturels, puis de remplacer les étymologies par « XXXX » indiquant qu'elles ont dû être masquées pour décourager toute tentative de ré-identification.

Les trois premiers exemples nous ont montré qu'une seule information (compte Facebook, adresse de courriel, prénom rare) pouvait suffire pour l'identification d'une personne. Dans le message posté par KellyM, chaque information, prise séparément, n'est pas suffisante pour identifier l'élève. En revanche, le faisceau d'information incluant le prénom, la spécialité de la classe (S), le nom du lycée et la ville est amplement suffisant pour identifier les participants de ce lycée à cette session Galanet à partir de leur prénom. Ainsi, pour anonymiser un corpus contenant un tel faisceau de données, il faut décider, par exemple, de renommer, ou d'effacer, tous les noms de lycée et de changer tous les noms de ville ayant une population réduite. Une version anonymisée du message de KellyM est donnée en annexe 2. Les experts au jeu des différences noteront très vite que parmi les graphies repérées en gras en annexe 1, seule Perpignan (120 000 habitants) est conservée ; toutes les autres graphies ont été transformées en tenant compte de leurs particularités lexicales. Dans ce résultat, le lecteur (chercheur) sait que cette information est bien l'information originale. Il conserve donc l'idée que c'est une élève de la région de Perpignan, et donc proche de la mer et de la Catalogne.

4. La méthode proposée

Cette section constitue le cœur de notre proposition. Elle présente la méthode en partant des hypothèses fondatrices. Nous illustrons ensuite la méthode générale sur la figure 1 pour montrer l'articulation des trois processus principaux : le *marquage*, la *fouille* et la *substitution*. Après avoir présenté la phase d'amorce, nous détaillons ces trois processus en les illustrant sur des exemples issus du corpus test. Dans chacune des étapes, nous nous efforcerons de préciser la part de la méthode automatisable (i.e. : ce que devrait apporter l'outil), et les choix ou actions qui resteraient à la charge de l'utilisateur (i.e. : le chercheur souhaitant anonymiser son corpus).

4.1. Hypothèses et caractéristiques fondant la méthode

Nos hypothèses suivantes de travail sont les suivantes :

- ☑ Une méthode d'anonymisation entièrement automatique et applicable à tout corpus n'est pas envisageable car de nombreuses décisions sont du ressort de l'utilisateur, pas de l'outil. C'est donc une méthode interactive ;
- ☑ Elle doit être applicable à toute langue (alphabets, sens d'écriture, …) ;
- ☑ Elle doit pouvoir s'enrichir et s'affiner par l'expérience sur des corpus similaires (par la langue, le genre, etc.)
- ☑ La partie automatique de l'outil doit assurer l'application systématique du processus d'anonymisation et assister l'utilisateur dans sa recherche des graphies susceptibles de poser un problème d'identification des personnes ;
- ☑ Le processus d'anonymisation doit être irréversible pour qui ne possède pas la table de remplacement des graphies.

4.2. Présentation générale de la méthode

La méthode que nous proposons, a pour objectif de transformer un corpus textuel représentant un ensemble d'interactions (ici un forum de discussion). Nous démarrons donc avec ce « Corpus à anonymiser » pour terminer avec un « Corpus anonymisé » (Fig. 1). Nous supposons que certaines données personnelles sont connues a priori : il peut s'agir a minima des marques d'identification introduites automatiquement par le système. Dans la situation d'une session de Galanet, des données personnelles sont également disponibles dans un espace privé dédié à cet effet : « Qui est qui ? ».

Nous renseignons la version initiale de la « liste des entités nommées » à partir de ces données. Nous trouverons dans cette liste les graphies dans leur version normalisée des entités à repérer dans le corpus : noms des participants, des institutions, des lieux précis, etc. Quelques exemples sont donnés dans la table 1. Pour chacune de ces graphies, nous indiquons l'entité (identifiée par un code) qu'elle représente, la catégorie (Prénom, Nom, Lieu, etc.) ainsi que les caractéristiques de la graphie (casse, accents et si l'on s'écarte de la graphie normalisée de référence).

Code	Original form	Category	Case	Accent	Misspelled?
F058	Kelly	First Name	Capitalized	Normal	No
I03	Rosa Luxembourg	Institution	Capitalized	Normal	No
P007	Canet	Small City	Capitalized	Normal	No
P004	Perpignan	City	Capitalized	Normal	No
C086	mikinessi@yahoo.it	Email	Lower	Normal	No
F039	Gabriela	First Name	Capitalized	Normal	No
L039	Medeiros	Last Name	Capitalized	Normal	No

Table 1 : Extrait de la liste initiale des entités nommées

En utilisant cette liste, une première recherche des graphies dans le corpus initial permet de *marquer* toutes leurs occurrences dans le corpus. Pour chaque occurrence trouvée (automatiquement) par le concordancier, le chercheur peut voir le contexte dans lequel elle apparaît et décider de l'associer à l'entité nommée correspondante. Bien sûr, on pourra décider de marquer et d'associer automatiquement toutes les occurrences d'une même graphie à une entité donnée, mais c'est courir le risque d'associer à tort une graphie à une entité nommée. Par exemple, le corpus de test contient une occurrence de « Kelly » qui ne correspond pas au prénom de l'un des participants, mais au patronyme du danseur « Gene Kelly ». Dans ce cas, le contexte permet immédiatement au chercheur de distinguer cette occurrence de « Kelly » en définissant une nouvelle entité nommée « Gene Kelly » qui fait référence à une personne publique et non à l'un des participants. Au cours du processus de marquage, le chercheur peut donc être amené à ajouter des entités nommées (homonymes) désignées par des graphies (définies dans la liste à l'étape précédente) figurant dans le corpus.

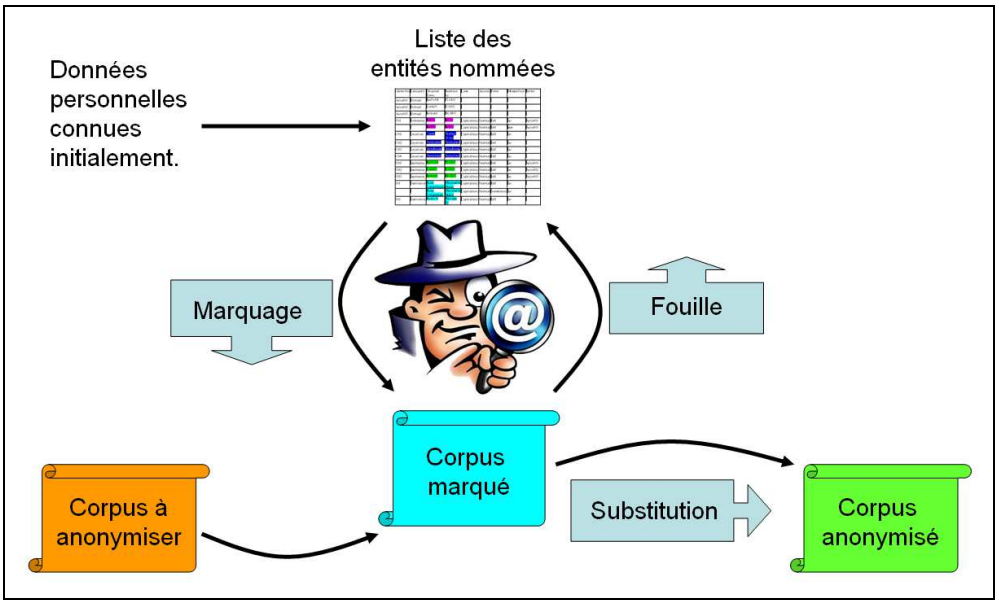

Figure 1 : Vue générale sur la méthode d'anonymisation

Toutes les occurrences des graphies recensées ayant été marquées dans le corpus, l'étape suivante, dite de *fouille*, consiste à rechercher des graphies nouvelles, pouvant désigner elles-aussi les catégories d'entités déjà identifiées et marquées. Dans cette étape, nous considérons deux cas :

1) Le cas où les participants ont utilisé des graphies altérées (introduisant des modifications linguistiques ou typographiques : casse, accents, orthographe, etc.)

2) Le cas où les participants ont introduit des références à de nouveaux lieux ou à des personnes pouvant être considérées comme des données personnelles (ex : personnes de la famille, lieux ou institutions très spécifiques liés à un participant ou à ses proches, etc.).

Dans les deux cas, nous proposons une technique pour découvrir de nouvelles graphies candidates. Le chercheur retient, parmi les nouvelles graphies proposées, celles qui représentent des données personnelles et rejette les autres. Il associe chaque graphie retenue à une entité nommée existante ou à une nouvelle entité nommée qu'il créée à cette occasion.

La liste des graphies et entités nommées ayant été mise à jour, on peut effectuer le *marquage* des nouvelles graphies repérées. Les étapes de *fouille* et de *marquage* sont répétées tant que le chercheur retient au moins une nouvelle graphie après la fouille.

Quand plus aucune nouvelle graphie n'émerge de ce processus itératif, on considère que toutes les graphies personnelles ont été repérées et leurs occurrences marquées dans le corpus. C'est à ce moment que le chercheur décide quelles graphies marquées doivent être remplacées, et par quelles graphies de substitution. Un outil de vérification de la cohérence devrait aider le chercheur à choisir des graphies de remplacement. En effet, ces graphies de substitution ne doivent pas entrer en collision entre elles ou avec des graphies déjà existantes dans le corpus. Par ailleurs certaines situations ambigües dans le corpus d'origine devraient rester tout aussi ambiguës dans le corpus ainsi transformé (homonymie, synonymie). Le

résultat de ce travail est consigné dans la table de remplacement dont un extrait est proposé dans la table 2.

Code	Original form	Modified	Substitution form	Category	Misspelled?
S058	KellyM	No		Login Name	No
F058	Kelly	No		First Name	No
F058	Kellly	No		First Name	Yes
P007	Canet	Yes	Aigues-Vives	Small City	No
P004	Perpignan	No		City	No
C086	mikinessi@yahoo.it	Yes	codabscons@hotmail.it	Email	No
I002	Rosa Luxembourg	Yes	Marguerite Duras	Institution	No
I002	Rosa Luxemburg	Yes	Margherita Duras	Institution	Yes
F014a	Peimikà	Yes	Chailai	First Name	No
F014b	Bibiana	Yes	Imelda	First Name	No
S014	PBS	Yes	CIS	Login Name	No
S039	Gabibr	No		Login Name	No
F039	GABRIELA	No		First Name	Yes
L039	MEDEIROS	Yes	CARLITOS	Last Name	No

Table 2 : Extrait de la table de remplacement

Pour chaque ligne de la table 2, la graphie d'origine est inscrite dans la colonne "*Original form*", la décision de remplacement est donnée dans la colonne "*Modified*". Lorsque la valeur de cette colonne est "*Yes*", une graphie de remplacement doit être proposée dans la colonne "*Substitution form*".

Dès que la table de remplacement est consolidée (après vérifications), le processus de *substitution* peut être exécuté : il s'agit de rechercher les graphies marquées et de les remplacer par les graphies de substitution. Cette opération termine le processus global d'anonymisation. Son résultat premier est le corpus anonymisé devenu partageable au sens des textes légaux. Mais le chercheur ayant réalisé l'opération peut aussi conserver (un certain temps) la table de remplacement ainsi que le corpus marqué (avant remplacement) pour pouvoir (sur demande spécifique d'un chercheur) modifier légèrement certaines substitutions et ainsi rendre possible des analyses que la précédente anonymisation aurait empêchées.

Après avoir présenté les grandes lignes de cette méthode, nous précisons dans la section suivante les deux stratégies de fouille permettant de détecter de nouvelles graphies.

5. Recherche de nouvelles graphies représentant des données personnelles

Ce qui rend le processus d'anonymisation difficile dans un grand corpus, ce n'est pas tant la taille du corpus ou le nombre d'occurrences à remplacer, mais plutôt le caractère imprévisible et donc difficile à détecter de certaines graphies. Parmi elles, très peu peuvent suffire pour fragiliser l'anonymat d'un pan entier du corpus (cf. section 2.3). C'est pour cette raison qu'il nous semble essentiel de proposer au chercheur des stratégies et des outils pour détecter de telles graphies. Dans cette section, nous présentons les idées principales de deux stratégies. La première est basée sur des règles de *variations lexicales* tandis que la deuxième

utilise les *contextes* trouvés dans le corpus lui-même pour détecter de nouvelles formes. Le résultat de leur application sur notre corpus de test est donné dans la section 6.

5.1. Stratégie des règles de variation lexicale

En contexte d'apprentissage en général et de communication spontanée en particulier, l'orthographe et les règles habituelles de typographie ne sont pas nécessairement connues ou adoptées par les participants. Ceci est d'autant plus vrai que les participants sont jeunes (du point de vue de leur apprentissage de la langue) ou qu'ils s'expriment dans une autre langue que leur langue maternelle. Les variantes possibles d'une graphie peuvent être nombreuses et leur nombre croît de manière importante avec sa longueur.

Par exemple, combien de graphies peuvent être considérées comme lexicalement proches de « Kelly » ? Dans les petits calculs d'illustration suivants, on se restreint à l'alphabet français et on ne compte ni les différences d'accents ni les différences de casse) ?

- ☑ 4 ayant une lettre manquante : elly, Klly, Kely, Kell ;
- ☑ 26×6, soit 156 ayant une lettre en trop : AKelly, … ZKelly, Kaelly, …, Kellyz ;
- ☑ 25×5, soit 125 ayant une lettre incorrecte : Aelly, …, Zelly, …, Kellz ;
- ☑ ...

Si l'on s'arrête ici, on compte déjà près de 300 formes considérées proches de « Kelly ». Si l'on ne considère que des prénoms, le nombre de graphies proches des 103 prénoms initiaux est de l'ordre 30 000 alors que le lexique complet de notre corpus ne comporte que 9 653 graphies différentes.

Nous proposons deux techniques complémentaires pour rechercher dans le lexique la présence de variations lexicales d'une graphie donnée. La première technique consiste à ne tenir compte ni de la casse (majuscule vs minuscule) ni des accents (signes diacritiques) lors de la recherche, par exemple *adriana* pour *Adriana*, et *Alexia* pour *Alèxia*.

La seconde technique repose sur la mesure de la ressemblance entre deux graphies. La distance de Levenshtein, en l'occurrence, compte le nombre de transformations qu'il faut effectuer pour passer de l'une à l'autre par suppression, insertion ou remplacement d'un caractère (Levenshtein, 1966). La mesure permet de limiter la recherche aux variantes obtenues avec une seule transformation pour les graphies courtes et deux transformations au plus pour les graphies plus longues.

Si l'avantage de cette approche réside dans le fait qu'elle ne propose que des graphies effectivement présentes dans le corpus, elle suppose cependant de connaître a priori toutes les formes canoniques des entités nommées à repérer. Pour pallier cette limite de l'approche lexicale, nous proposons une stratégie complémentaire qui repose sur l'étude des contextes.

5.2. Stratégie des règles de contextes

Si la première stratégie repose sur l'idée que des graphies peuvent apparaître avec des variantes lexicales, la seconde stratégie cherche à exploiter le fait que les graphies représentant une catégorie (prénom, surnom, patronyme, lieux, institutions, etc.) peuvent apparaître dans des contextes similaires. Typiquement, dans une session de Galanet, dans l'activité « Briser la glace » où il s'agit de se présenter (discussion « Qui sommes-nous »), plusieurs élèves font précéder leur prénom (en français) par « je m'appelle » et le nom de leur

ville par « j'habite à ». En utilisant, pour chaque catégorie, les graphies identifiées dans la liste des entités nommées, on peut repérer les contextes dans lesquels leurs occurrences apparaissent.

Illustration de la méthode sur les contextes de prénoms :

On recherche toutes les graphies (mot ou caractère isolé) qui apparaissent immédiatement à gauche d'un prénom. Pour chacune de ces graphies, on va travailler différemment suivant leur fréquence dans le corpus.

Si cette fréquence est faible, on recherchera systématiquement toutes les graphies qui la suivent pour y repérer de nouvelles entités nommées. Par exemple, on trouve « hombre » dans le contexte gauche immédiat de Marc (« Hombre Marc ») ; hombre est présent 5 fois dans tout le corpus (fréquence faible) ; les mots qui suivent hombre sont : Marc, es, siempre, sigue, valeroso. En dehors de Marc, déjà identifié, le contexte « hombre » ne conduit pas à la découverte de nouvelles entités nommées.

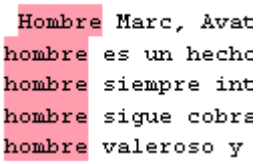

Figure 2. Les graphies qui suivent « hombre » dans le corpus de test

Si la fréquence est élevée, on va s'intéresser à la graphie située à gauche de la première, et on recherchera systématiquement toutes les occurrences de ce couple de graphies dans le corpus. Si leur fréquence est faible, on regarde ce qui les suit, comme on l'a fait pour une seule graphie. Si leur fréquence est élevée[1], on va s'intéresser à la graphie suitée à gauche pour construire un triplet de graphies et réitérer la méthode. Par exemple, on trouve « l' » dans le contexte gauche immédiat de plusieurs prénoms ; « l' » est présent 195 fois dans le corpus ; on va donc s'intéresser aux graphies situées à gauche du « l' » dans les contextes gauches des prénoms ; il s'en trouve quatre : « amb l' », « com l' », « diu l' » et « i l' » ; la fréquence des ces couples est faible et la recherche des graphies qui les suivent est rapide ; ainsi la visualisation des graphies qui suivent « amb l' » permet de découvrir une nouvelle entité nommée, « Asenjo ».

Cette méthode appliquée au contexte gauche d'une entité nommée connue, s'applique de la même manière au contexte droit, en recherchant d'abord la première graphie située immédiatement à droite puis les suivantes si nécessaire.

Cette deuxième stratégie fait émerger des contextes révélateurs de graphies représentant des données personnelles (dans une langue donnée). Nous pensons que ces contextes (incluant leurs erreurs lexicales) sont capitalisables dans chaque langue (et inter-langue) et pour chaque catégorie de graphies. Ils devraient donc pouvoir servir à la détection de graphies de la même catégorie dans d'autres corpus de la même famille, ici par exemple : les sessions Galanet.

[1] La limite : « trop grand » ou « raisonnable » doit être modifiable par le chercheur. Elle peut être initialisée à partir du nombre d'occurrences maximum d'une graphie de la liste des entités.

6. Les premiers résultats

Dans cette section, nous ne développons que de la recherche des graphies dans le contenu des messages et nous concentrons notre analyse sur les prénoms.

La même méthode de recherche est applicable aux noms d'institution, noms de lieux, noms d'utilisateur (Facebook, MySpace, Hi5, Soundcloud, Badoo, Bebo, Friendster, Netlog, skype, Twitter, etc.) et surnoms pouvant apparaître dans les messages.

Pour rechercher d'autres informations spécifiques, des techniques systématiques simples peuvent être employées.

- ☑ les adresses et les numéros de téléphone sont recherchés par les graphies composées uniquement de chiffres ; dans notre corpus : 229 occurrences trouvées, mais aucune donnée personnelle ;
- ☑ les adresses de courriel sont recherchées par la présence du caractère '@' ; une seule occurrence trouvée correspondant effectivement à une adresse de courriel (voir annexe 1).
- ☑ Les URL de pages personnelles ou de blogs sont recherchées par la présence de « http: » ; 35 occurrences trouvées dans le corpus dont aucune ne correspond à des informations personnelles.

6.1. Application des règles de variation lexicale

En partant des 103 prénoms différents répertoriés dans l'espace « Qui est qui » de la session et en les comparant aux 9 653 graphies du lexique du corpus de test, la recherche de variantes lexicales renvoie 317 graphies vérifiant les critères présentés plus haut (section 5.1.)

Parmi celles-ci, 70 graphies peuvent être éliminées automatiquement :

- ☑ 16 graphies font déjà partie de la liste des graphies identifiées ;
- ☑ 34 graphies ont une fréquence trop grande pour être candidates ;
- ☑ 20 graphies représentent des noms de login de participants ;

Ensuite, 180 autres graphies sont assez rapidement évaluées comme des mots des langues en présence dans le corpus.

Les 67 dernières ont nécessité une vérification de contexte plus approfondie avant de décider ou non de les introduire comme de nouvelles graphies associées à des entités nommées. On a ainsi trouvé :

- ☑ 5 mots communs dont il a fallu vérifier le contexte pour décider ;
- ☑ 31 graphies représentant de vrais prénoms ou surnoms des participants ;
- ☑ 1 graphie représentant le prénom d'une personne proche d'un participant ;
- ☑ 30 graphies représentant des prénoms de personnes publiques.

Le résultat de ce processus est illustré sur la table 3 où l'on montre les prénoms de la liste d'origine qui ont trouvé des variantes proches retenues comme prénoms ou surnoms.

Prénoms connus dans la liste initiale	Adriana, Alèxia, Anthony, Baptiste, Cleissa, Eli, Elouise, Emmanuel, Federica, Ferran, Gabriela, Guillem, Iñigo, Jaqueline, Jean, José, Kelly, Léo, Mariana, Mary, Michela, Monica, Olalla, Oleguer
Prénoms ou surnoms découverts	adriana, Alexia, Antonhy, baptiste, Cleisa, Elô, Ely, ELY, Seli, Louise, MAnuel, Federiac, fran, Fran, GABRIELA, guillem, iñigo, Jacqueline, jean, Jose, Kellly, Leo, léo, MariAna, mary, May, Miche, michelina, moni, olalla, oleguer

Table 3 : Variations lexicales des prénoms retrouvées dans le corpus

Par construction, les nouvelles graphies détectées par cette méthode de fouille sont proches des graphies d'origine. Ce n'est pas nécessairement le cas pour la fouille contextuelle dont nous présentons les résultats ci-dessous.

6.2. Application des règles de contexte

En partant des mêmes 103 graphies représentant les prénoms connus initialement, un concordancier, celui de Calico par exemple, permet de trouver automatiquement les 145 occurrences de ces graphies dans le corpus.

L'examen des contextes gauches (graphies situées immédiatement à gauche) de ces prénoms permet de repérer :

- ☑ 72 contextes gauches différents en ne prenant en compte qu'une seule graphie. Parmi ces 72 graphies, 30 sont très fréquentes et ne sont donc pas utilisées. 9 graphies n'apparaissent qu'une seule fois et sont donc écartées car cela signifie qu'elles n'apparaissent qu'avec un prénom déjà connu. Quant aux 33 restantes, près de la moitié (17 sur 33) peuvent être conservées car elles peuvent faire apparaître de nouvelles graphies ;
- ☑ 123 contextes gauches différents constitués d'une séquence de 2 graphies. En éliminant ceux qui sont trop fréquents et ceux qui n'apparaissent qu'une seule fois, on est conduit à en retenir 31.

Par ailleurs, nous avons testé 9 règles générales. Elles concernent les énumérations de graphies d'une même catégorie, qu'elles soient séparées par un espace, une virgule, un « et » ou un « ou » dans chacune des 5 langues. Sur le corpus de test, seules 4 de ces règles sont retenues ; ce sont celles dont les prénoms sont séparés par : « , », « i », « et », et « e ». On note que le « y » ne fonctionne pas bien pour les espagnols (catalans) car ils utilisent souvent un article devant les prénoms. On peut noter aussi que le « ou » ne fonctionne dans aucune des 5 langues.

Au total, nous avons donc passé en revue environ 200 contextes et nous en avons retenu 51 règles qui ont permis de détecter 20 graphies nouvelles par rapport à la liste des 103 graphies initiales de prénoms. Ces graphies sont présentées dans la table 4, et séparées en deux sous-ensembles : les dix qui avaient déjà été détectées par les variations lexicales (nouvelles graphies retrouvées), et les dix autres qui n'ont été détectées que par cette méthode contextuelle (graphies entièrement nouvelles).

10 graphies déjà trouvées par variation lexicale	adriana, Antonhy, Federiac, fran, iñigo, jean, Kellly, Leo, léo, May,
10 graphies trouvées uniquement par les contextes	Asenjo, Belle, Bet, Beth, Christine, Fede, Line, Maria, Peimika, Regina.

Table 4 : Nouvelles graphies de prénoms détectées dans le corpus par les contextes

Après avoir détecté 31 nouvelles graphies avec la méthode lexicale, et 20 avec la méthode des contextes (dont 10 communes), nous mettons à jour la liste des graphies de prénoms recensées dans le corpus. Ce cycle s'achève en marquant toutes les occurrences de ces 41 nouvelles graphies.

6.3. Nouvelle itération du cycle de fouille et marquage

Chaque cycle comporte une phase de marquage (des nouvelles graphies), puis une phase de fouille utilisant les deux stratégies présentées dans les sections précédentes. Cette deuxième phase est d'autant plus coûteuse que le nombre de graphies nouvelles est élevé. On peut optimiser ce travail de fouille dans le cycle suivant en ne calculant les variations lexicales que pour les graphies nouvelles et en ne considérant que les contextes des nouvelles graphies obtenues par les variations lexicales.

Sur le corpus de test, au cours du deuxième cycle, toujours en ne considérant que les prénoms, voici ce que l'on obtient :

- ☑ Les variations lexicales appliquées aux 41 nouvelles graphies font ressortir du lexique : des prénoms publics et leurs variantes (Gabriel, Frank, Joan, Michel, Christie, juan, Juan, Lleo, mac, Mac, Manel), des identifiants utilisés dans le dispositif Galanet (MarcA, FranO, maryb, Maryb AdrianaL AlexiaT, FedeS, OlallaN) et 90 mots courants (noms communs, verbes…) des langues utilisées. Aucun nouveau prénom n'est découvert par cette méthode.
- ☑ La recherche des contextes gauches des 21 nouvelles graphies (celles qui ne proviennent pas des règles contextuelles du cycle précédent) fait apparaître 4 nouveaux contextes gauches intéressants de taille 1 : « amic », « chamada », « ja » et « també ». Aucun de ces contextes ne permet cependant de découvrir de nouveau prénom ou de nouvelle entité.

Au vue de la première application de notre méthode de fouille sur les prénoms, nous pouvons émettre 2 hypothèses :

- ☑ Soit la fouille est si efficace que la première itération suffit à trouver toutes les graphies représentant des prénoms ;
- ☑ Soit elle n'est efficace qu'à la première itération et ne permet pas de trouver les graphies restantes dans le corpus.

À l'aide de l'outil Colagora de Calico nous avons pu demander à Sandrine Deprez (détentrice du corpus) de faire une relecture générale pour repérer les graphies que la méthode aurait oubliées. Sur 269 graphies référencées, 117 effectivement présentes, et 279 occurrences marquées, Sandrine Deprez a détecté 7 graphies oubliées par la méthode : OleguerLI, P_Monemurro, LauraPa, CR Martins, Peimikà, Cléia et Reginaldo. Sauf le dernier (que la méthode n'aurait pas pu trouver), les 6 autres oublis sont attribués à des défaillances humaines

lors de l'application manuelle de la méthode. Nous espérons qu'un outil qui la systématise la rende plus robuste.

6.4. Premier bilan de l'application de la méthode

Au final, la méthode proposée montre les caractéristiques suivantes :

- ☑ Elle est itérative.
- ☑ La liste des graphies à repérer, et éventuellement à remplacer dans le corpus, se construit à partir d'une liste initiale.
- ☑ Cette liste s'étend à chaque itération du double processus de marquage et de fouille.
- ☑ Le nombre d'itérations est très réduit (ici : 2).
- ☑ Les règles sont construites à partir du corpus, sans nécessiter a priori de connaissances extérieures.

7. Discussion et perspectives

Afin d'obtenir une anonymisation de bonne qualité, il nous parait important de respecter les critères suivants pour le remplacement des entités :

- ☑ Le marquage doit être complet : toutes les occurrences des graphies repérées doivent être marquées, qu'elles nécessitent d'être finalement remplacées ou non ;
- ☑ Chaque information doit être suffisamment imprécise pour englober des centaines de personnes candidates, de sorte qu'un faisceau d'informations ne permette pas d'identifier formellement une personne ;
- ☑ La cohérence du texte doit être maintenue autant que possible ;
- ☑ Le processus d'anonymisation ne doit pas changer le niveau d'ambiguïté du corpus initial. Ainsi deux participants de même nom au départ devraient avoir le même nom à l'arrivée.

Comme le montre l'un des exemples de l'introduction, la puissance actuelle des moteurs de recherche est telle que la moindre marque suffisamment rare dans un corpus peut mener assez directement à un participant ou à une institution.

Les moteurs de recherche sont particulièrement efficaces lorsqu'il s'agit de retrouver des graphies rares ou des combinaisons de graphies plus communes (noms de personnes, institutions, etc.). Pour reprendre notre exemple, loin d'être unique, nous pouvons considérer que les prénoms seuls ne permettent pas, *en général*, d'identifier formellement une personne. Pourtant, sur les 103 personnes inscrites à la session Galanet, deux possèdent un prénom composé extrêmement rare, ce qui permet de les retrouver immédiatement par une simple requête sur un moteur de recherche. Comme nous l'avons fait pour ce corpus, nous pouvons donc envisager de ne modifier que certains prénoms du corpus. Mais pour décider qu'un prénom (ou une autre entité) est rare, nous suggérons de définir une métrique basée sur le nombre de liens trouvés par un moteur de recherche. On pourrait ainsi considérer qu'en dessous d'une centaine de liens, une telle entité peut être considérée comme rare.

La méthode décrite dans cet article a été appliquée à un corpus de test composé de messages assez courts, dont les marques d'identité étaient assez peu nombreuses. Ce peu de marques d'identité peut s'expliquer par l'existence d'un dispositif privé (i.e. : « Qui est qui ») qui informe les participants sur les données personnelles de chacun.

Nous envisageons de réappliquer cette méthode sur deux autres corpus. Le premier, plus important mais semblable car issu de Galanet nous permettra d'évaluer la réutilisabilité des règles de contexte retenues dans le premier corpus. Le second, de nature très différente, comporte des messages beaucoup plus longs en moyenne, produits par un groupe beaucoup plus restreint d'apprenants, dans un contexte monolingue (en langue maternelle) avec des écrits plus académiques. Il est possible que, les apprenants mis en confiance dans leur petit groupe, puisse susciter des échanges d'informations plus spécifiques au regard de leur contexte de travail, qui était en lien avec le sujet du cours.

Parallèlement, un outil d'anonymisation, modélisant la méthode interactive décrite dans cet article, sera spécifié, testé, et mis à disposition de la communauté.

Références

(95/46/CE) Directive du Parlement européen et du Conseil du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. Accessible en ligne. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:fr:HTML

(97/66/CE) Directive du Parlement européen et du Conseil du 15 décembre 1997 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des télécommunications. Accessible en ligne. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31997L0066:FR:HTML

(2002/58/CE) Directive du Parlement européen et du Conseil du 12 juillet 2002 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques. Accessible en ligne. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:fr:HTML

Aramaki, E., Imai, T., Miyo, K., & Ohe, K. (2006). Automatic Deidentification by using Sentence Features and Label Consistency. In *Workshop on Challenges in Natural Language Processing for Clinical Data*. http://luululu.com/profile/paper/2006-i2b2/i2b2-deid.pdf

Blondel , F.-M., & Giguet, E. (2011). Analyses et partages de corpus de discussions avec Calico - Leçons tirées d'une expérience récente. In Dejean, C., Mangenot, F., Soubrié, T. (Ed.). Actes du colloque Epal 2011 (Échanger pour apprendre en ligne), Université Stendhal – Grenoble 3. http://w3.u-grenoble3.fr/epal/dossier/06_act/pdf/epal2011-blondel-giguet.pdf

Calico (2012) : http://woops.crashdump.net/calicorss/ Plate-forme CALICO pour visualiser et analyser des forums de discussion, issue de l'ERTé CALICO (2006-2010).

Deprez, S. (2012). Analyse semi-automatique d'un corpus plurilingue. Degache, C. & Garbarino, S. (Ed.) (2012). Actes du colloque IC2012. Intercompréhension : compétences plurielles, corpus, intégration. Université Stendhal Grenoble 3 (France), 21-22-23 juin.

Galanet (2012) : www.galanet.be/ Plateforme de formation à l'intercompréhension en langues romanes, issue du projet européen Socrates Lingua (2001-2004).

Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J., & Grandison, T. (2010). An evaluation of feature sets and sampling techniques for de-identification of medical records (p. 183). ACM Press. doi:10.1145/1882992.1883019

Giguet E., Lucas N., Blondel F.-M. et Bruillard E. (2009). Share and explore discussion forum objects on the Calico website, CSCL 2009, Rhodes, ICLS. pp. 174-176 (nominated for best technical paper). http://www.stef.ens-cachan.fr/annur/blondel/cscl09_calico_share.pdf

Grouin, C., Rosier, A., Dameron, O., Zweigenbaum, P. (2009) Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. 13èmes Journées francophones

d'informatique médicale, Nice, 28-30 avril 2009. *Risques, Technologies de l'Information pour les Pratiques Médicales*, Fieschi, M., Staccini, P., Bouhaddou, O. et Lovis, C. (Eds), *Informatique et Santé*, Vol. 17, Springer, 2009.

Henri, F., Charlier, B. (2005). L'analyse des forums de discussion Pour sortir de l'im-passe. In Baron G-L., Bruillard E., Sidir M. (Dir.), *Symposium Symfonic. Formation et nouveaux instruments de communication*. Amiens, Université de Picardie, janvier. http://archive-edutice.ccsd.cnrs.fr/edutice-00000897

King, G (2007) 'An Introduction to the Dataverse Network as an Infrastructure for Data Sharing', *Sociological Methods and Research*, 32. pp. 173-199. Available at http://j.mp/iHJcAa

Levenshtein, V. (1966). Binary codes capable of correcting deletions insertions and reversals. Soviet Physics- Dokludyj 10:707~710, 1966.

Mallet-Poujol, N. (2004). Protection de la vie privée et des données personnelles. *Legamedia*, Février 2004. Disponible en ligne http://eduscol.education.fr/chrgt/guideViePrivee.pdf.

Meystre, S.M., Friedlin, F. J., South, B.R., Shen, S., Samore, M.H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010, 10:70, http://www.biomedcentral.com/1471-2288/10/70

Mulce (2012) : http://repository.mulce.org/ Plateforme de partage de corpus d'apprentissage multimodaux, issue du projet ANR Mulce (2007-2010).

Net ltd (2000). Data Masking: What You Really Need To Know Before You Begin. White paper. http://www.datamasker.com/DataMasking_WhatYouNeedToKnow.pdf

Neubauer, T., & Heurix, J. (2011). A methodology for the pseudonymization of medical data. International Journal of Medical Informatics, 80(3), 190-204. doi:10.1016/j.ijmedinf.2010.10.016

Reffay, C., Noras, M., Chanier, T., Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. *Numéro spécial EPAL de la revue Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation*, pages 185-219, Vol 15, 2008. http://edutice.archives-ouvertes.fr/edutice-00159733/fr/

Reffay, C., Betbeder, M.-L., Chanier, T. (In press). Multimodal Learning and Teaching Corpora Exchange: Lessons learned in five years by the Mulce project. Int. J. Technology Enhanced Learning, Special Issue "dataTEL - Datasets and Data Supported Learning in Technology-Enhanced Learning", Inderscience, In press.

Reffay, C. & Teutsch, P. (2007). Anonymisation de corpus réutilisables : masquer l'identité sans altérer l'analyse des interactions. Poster (2 pages) Actes de la conférence Environnements Informatiques pour l'Apprentissage Humain (EIAH 2007), Lausanne, Suisse, Juin 2007.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. Proceedings: A Conference of the American Medical Informatics Association. AMIA Annual Fall Symposium, 333-337.

Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T., & Estabrooks, C. A. (2005). Central Questions of Anonymization: A Case Study of Secondary Use of Qualitative Data. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6(1)*. Consulté de http://www.qualitative-research.net/index.php/fqs/article/view/511

ANNEXE 1 : Quelques exemples de messages non anonymisés

Message n°f331s2970m2 du 2011-11-30T19:24 posté par **Gabibr**
Re: Quelques informations ...
"Eu amo a língua Francesa! Quem sabe falar francês me adiconem no meu FACEBOOK;) J'aime parler français! Qui peut parler français? M'ajouter dans **FACEBOOK**;) Nom: GABRIELA MEDEIROS."

Message n°f333s3016m2 du 2011-12-27T09:25 posté par **Miche**
Re: Les stéréotypes culinaires
"inviate i vostri documenti alla mia mail **mikinessi@yahoo.it** grazie!!!;)"

Message n°f330s2914m8 du 2011-10-22T19:52 posté par **PBS**
Re: Por que me chamo assim?!
"Yo me llamo **Peimikà Bibiana**. Como mi madre es tailandesa y mi padre es italiano, mi primer nombre, Peimikà, es tailandés y significa " dueña del amor ", mientras mi según nombre, Bibiana, es italiano y procede del etrusco " vibius " que significa " vida ". Me gusta mucho tener dos nombres (en Italia es más usual tener un nombre) y sobre todo estoy orgullosa de los orígenes diferentes que tienen **y que hacen mi nombre aún más particular (además Peimikà no es muy difundido en tampoco en Tailandia y tampoco Bibiana en Italia**"

Message n°f330s2880m3 du 2011-10-17T08:22 posté par **KellyM**
Re: Qui sommes- nous?
"Bonjour, je m'appelle **Kellly**. J'ai 16 ans, je suis une élève en 1ère S dans le **lycée Rosa Luxemburg** à **Canet**, non loin de **Perpignan**. Je ne fais pas de sport, mais il m'arrive de regarder des matchs de foot avec des amis. J'écoute à peu près tout les styles de musique, sauf le classique. J'aime sortir en soirées, avec des amis, aller en ville, à la plage (quand il fait beau ^ ^). J'aime bien cette plateforme car on peux discuter avec d'autres élèves du même âge, malgrès qu'on vienne de pays et cultures différentes. "

ANNEXE 2 : Version anonymisée des messages de l'annexe 1

Message n°f331s2970m2 du 2011-11-30T19:24 posté par **Gabibr**
Re: Quelques informations ...
"Eu amo a língua Francesa! Quem sabe falar francês me adiconem no meu FACEBOOK;) J'aime parler français! Qui peut parler français? M'ajouter dans **FACEBOOK**;) Nom: **GABRIELA CARLITOS***."

Message n°f333s3016m2 du 2011-12-27T09:25 posté par **Miche**
Re: Les stéréotypes culinaires
"inviate i vostri documenti alla mia mail **codabscons@hotmail.it*** grazie!!!;)"

Message n°f330s2914m8 du 2011-10-22T19:52 posté par **CIS**
Re: Por que me chamo assim?!
"Yo me llamo **Chailai*** **Imelda***. Como mi madre es tailandesa y mi padre es italiano, mi primer nombre, **Chailai***, es tailandés y significa "bonita* ", mientras mi según nombre, Imelda*, es italiano y [...]* significa "batalla universal*". Me gusta mucho tener dos nombres (en Italia es más usual tener un nombre) y sobre todo estoy orgullosa de los orígenes diferentes que tienen **y que hacen mi nombre aún más particular (además Chailai* no es muy difundido en tampoco en Tailandia y tampoco Imelda* en Italia**"

Message n°f330s2880m3 du 2011-10-17T08:22 posté par **KellyM**
Re: Qui sommes- nous?
"Bonjour, je m'appelle **Kellly**. J'ai 16 ans, je suis une élève en 1ère S dans le lycée **Margherita Duras*** à **Aigues-Vives***, non loin de **Perpignan**. Je ne fais pas de sport, mais il m'arrive de regarder des matchs de foot avec des amis. J'écoute à peu près tout les styles de musique, sauf le classique. J'aime sortir en soirées, avec des amis, aller en ville, à la plage (quand il fait beau ^ ^). J'aime bien cette plateforme car on peux discuter avec d'autres élèves du même âge, malgrès qu'on vienne de pays et cultures différentes. "

ANNEXE 3 : Contextes (mono-graphie) testés et évalués

Contexte gauche	Nombre d'occurrences	Nb occ. devant un prénom connu	Nb occ. devant un nouveau prénom	Taux de réussite de la règle
anche	49	0		0%
avec	45	6	2	18%
come	45	1	0	2%
ao	21	1	0	5%
ça	15	1	0	7%
dir	11	1	0	9%
amicizia	11	1	0	9%
Bon	11	0	1	9%
sou	10	2	0	20%
appelle	9	4	1	56%
mãe	7	0	1	14%
Cara	7	1	1	29%
Ciao	6	1	0	17%
Merci	6	2	1	50%
grandi	6	0	0	0%
soy	5	2	0	40%
ara	5	0	0	0%
equipe	5	1	0	20%
Hombre	4	1	0	25%
Pedro	4			0%
costruito	4			0%
dicho	3	1	0	33%
llamo	3	2	1	100%
appel	3	1	0	33%
Ana	3			0%
raison	3	1	0	33%
Peimikà	3	1	0	33%
choix	3	1	0	33%
mangiano	3	0	0	0%
aiuto	2	0	0	0%
chamam	2	1	1	100%
tampoco	2	1	0	50%
vedo	2	0	0	0%

Pour chaque contexte gauche donné en première colonne (ex : avec), on indique en deuxième le nombre total de ses occurrences (ex : 45) dans le corpus. On précise alors le nombre ce ces occurrences précédant un prénom connu en troisième colonne (ex : 6) et un nouveau prénom détecté en quatrième colonne (ex : 2). Enfin, la dernière colonne calcule le rapport entre la somme des colonnes 3 et 4 et la colonne 2 (ex : (6+2)/45 = 0,178, soit environ 18%).

Attention :

Les contextes surlignés dans la première colonne étant des graphies identifiées dans la liste des données personnelles. Elles ne peuvent être capitalisées même si leur taux de réussite était haut (ce qui n'est pas le cas ici). Elles devront donc être supprimées à l'issue de l'anonymisation de ce corpus. Cette vérification devra être faite juste avant le remplacement pour prendre en compte toutes les graphies identifiées (ex : Peimikà).