Anonymisation semi-automatique de corpus d'interactions éléments pour une méthode interactive

Christophe Reffay, François-Marie Blondel, Stéphane Allaire, Emmanuel Giguet

Anonymisation semi-automatique de corpus d'interactions éléments pour une méthode interactive

Christophe Reffay[1], François-Marie Blondel[1], Stéphane Allaire[2], Emmanuel Giguet[3]

[1] STEF, ENS-Cachan – IFÉ (ENS-Lyon)

[2] Université du Québec à Chicoutimi (UQAC), Canada

[3] GREYC, Université de Caen Basse-Normandie, CNRS

RÉSUMÉ. Pour faciliter le partage de données de recherche, et donc la comparaison, il est indispensable que les chercheurs puissent disposer de méthodes et d'outils permettant d'anonymiser les grands volumes d'interactions de leurs corpus. Nous rappelons le cadre légal et les enjeux de l'anonymisation avant de montrer les difficultés de son automatisation. La méthode proposée ici laisse au chercheur-utilisateur visé, le contrôle du processus de transformation de son corpus. C'est une méthode interactive, systématique et applicable à des corpus écrits en toutes langues. Elle est basée sur un cycle de marquage et de fouille détaillé dans cet article et appliqué à deux corpus de forums très différents. Les résultats de ces premières applications sont présentés et discutés pour envisager de nouvelles améliorations à cette méthode et une mise en œuvre comme outil de la plateforme Calico.

MOTS-CLÉS : Anonymisation, Corpus, Partage de données de recherche, Fouille de graphies personnelles.

ABSTRACT. In order to ease research data sharing and scientific comparison, researchers need appropriate methods and tools to anonymise their huge corpora of interaction. We first draw the legal context and present the stakes of anonymisation. We emphasise the subtleties that avoid complete automation of the anonymisation process. The target user-researcher will keep the control of the anonymisation process with the method we propose here. It is mainly based on two processes: marking and mining presented in details in this article. The whole method has been applied to two very different corpora. Preliminary evaluation of these first tests is given in the discussion and gives the floor to interesting perspectives for the method and its implementation on the Calico platform.

KEYWORDS: Anonymisation, Corpus, Research data sharing, lexical mining, identifying forms discovering.

1. Anonymiser : une nécessité légale pour le partage de données dans le respect de l'éthique

Pour faciliter la confrontation des points de vue sur des recherches construites à partir d'expériences difficilement reproductibles, des dispositifs comme Mulce (2012) (Reffay et al., 2008, Reffay et al., in press) et Calico (2012) (Giguet et al., 2009 ; Blondel et al, 2011) ont été développés pour permettre le partage de données entre chercheurs. Bien qu'elles soient encore peu nombreuses, on peut trouver une liste de plateformes de partage de données dans (Datacite, 2012). Par ailleurs, une étude de Nelson (2009) montre que la majorité des chercheurs de toutes les disciplines reconnaissent l'intérêt du partage des données pour améliorer la scientificité des travaux et la validité des résultats. Pourtant, les communautés pratiquant le partage ouvert constituent l'exception plutôt que la règle. Parmi les raisons invoquées en particulier dans les disciplines de Sciences Humaines et Sociales, nous trouvons l'obligation et la difficulté de rendre ces données anonymes pour en permettre le partage.

> *« 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)

L'OCDE et les directives 95/46/CE, 97/66/CE et 2002/58/CE de la Commission Européenne, ont énoncé les principes relatifs à la protection des données personnelles (avis, but, consentement, divulgation, accès et responsabilité). Celui relatif à la divulgation précise que les données personnelles ne devront pas être divulguées à une tierce partie sans le consentement de l'utilisateur. 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.

Dans un état de l'art du Royaume Uni, du Canada et des Etats-Unis concernant les aspects éthiques, légaux et pratiques de l'anonymisation dans le domaine de la santé, nous avons trouvé cette citation qui reflète bien le point de vue qualitatif (en SHS) de ce que doit faire l'anonymisation :

> *"[…] CORTI et al. (2000, para. 21.), who suggest removing "identifying details," such as proper names or street names, and replacing these with pseudonyms. In theory, then, the data remaining after anonymization tells us a story without telling whose story it is."*
>
> (Corti et al., 2000 : para. 21.) cité par (Thomson et al., 2005: p.6)

Que ce soit pour des raisons éthiques ou légales, les données permettant d'identifier des personnes physiques doivent être expurgées des corpus de données susceptibles d'être partagés entre chercheurs au-delà de contrats de recherche.

Dans le cas des données d'interaction, anonymiser ne se limite pas à supprimer ou masquer les identifiants des personnes. Il s'agit bien de procéder à une *dépersonnalisation totale* (AFNOR 2000) des données afin de garantir « qu'une personne, étrangère au dispositif, ne puisse pas identifier l'un des acteurs de la formation ». Plus précisément, cela implique notamment de faire fi de toute information sensible qui pourrait permettre de retracer un individu, sans toutefois perdre d'éléments sémantiques à travers les interactions.

Cette opération ne peut être réalisée directement sur la plateforme qui sert à la formation car les acteurs ont besoin de se reconnaître et de reconnaître leurs productions. Ainsi, tant que dure la session de formation, voire au-delà de sa durée prévue, les interactions restent inchangées et accessibles aux acteurs de la formation et aux éventuels observateurs (dûment autorisés). Dès que le partage des données d'interaction est nécessaire pour la recherche, il est indispensable de transformer ces données et de les isoler dans un espace différent pour y appliquer analyses et traitements spécifiques.

Le premier travail à accomplir est de retirer toute marque identifiant les acteurs physiques à l'intérieur des corpus. Cependant, les volumes à traiter sont tels que les procédures manuelles ne sont pas envisageables et il n'existe pas à l'heure actuelle d'outils automatiques permettant de réaliser ce travail sans craindre d'omettre la transformation de certains éléments ou encore de perdre des éléments contextuels nécessaires à une pleine compréhension des interactions. Un support informatique est indispensable pour garantir une transformation systématique et cohérente.

Après avoir rappelé son cadre légal et ses enjeux dans cette partie, nous présenterons les principaux obstacles à l'automatisation du processus d'anonymisation avant de proposer une méthode générale. Cette méthode d'anonymisation est interactive, semi-automatique, indépendante des langues utilisées dans le corpus et spécialement conçue pour s'appliquer à un corpus d'interactions. Après une présentation générale de la méthode, nous détaillerons deux des techniques de fouilles et les illustrerons sur deux corpus très différents. Les résultats de l'application de la méthode sur les deux corpus tests nous permettrons, après une première évaluation, d'ouvrir une discussion et des perspectives pour améliorer la méthode et orienter le développement de l'outil qui pourra l'implémenter.

2. Anonymiser des corpus d'interactions : une tâche difficile à automatiser

Dans le cas des interactions médiées par ordinateur, transformer les identifiants des auteurs de messages pour les rendre anonymes est une opération aisée et souvent automatisée. Le problème le plus délicat à traiter est celui de l'anonymisation à l'intérieur des messages. En effet, il est très fréquent que les acteurs de la formation (enseignants, tuteurs, apprenants) utilisent dans leurs messages des éléments qui permettent de les identifier aisément. Ce sont par exemple des noms, prénoms, adresses de courriel, sites web personnels, adresses postales, numéros de téléphone, identifiants dans des outils de communication externes (Skype, Twitter, Facebook), lieux de résidence, institutions de rattachement, espaces fréquentés, etc. Toutes ces données permettant d'identifier aisément les acteurs doivent être masquées pour assurer l'anonymisation du corpus et donc la protection des acteurs. Le choix des éléments à repérer est une étape préliminaire à notre méthode. Ce choix doit être guidé par la règle suivante :

Chaque information conservée dans le corpus, prise séparément, doit pouvoir correspondre à de nombreuses personnes pour leur assurer l'anonymat. Même combinées en un faisceau d'informations, elles ne doivent pas permettre l'identification d'une personne physique. Pour illustrer cette règle, voici un exemple extrait du corpus « Nomades » de (Desprez, 2012) :

> « Bonjour, je m'appelle **Kelly**. J'ai 16 ans, je suis une élève en **1ère S** dans le lycée **Rosa Luxemburg** à **Canet**, pas très loin de **Perpignan**. »

Kelly est un prénom fréquent (plus de 150 comptes facebook contenant cette graphie). Être en 1e S n'est pas rare en soit, mais préciser dans quel lycée et dans une ville de 3 000 habitants (qui n'a sans doute qu'un seul lycée) sont des informations bien trop précises qui permettent certainement de trouver la seule Kelly qui réponde à ces critères (en 2012). Dans ce cas précis, il faut impérativement masquer au moins le nom de l'institution et le nom de la (petite) ville.

Ce qui rend la tâche très difficile à automatiser entièrement relève de plusieurs aspects. Le premier vient du fait que ces marques (ex : « Paris ») sont susceptibles de variations, notamment syntaxiques, comme des erreurs (Pari, Parsi, paris, etc.) ou des altérations volontaires (Parigi en italien). Le deuxième provient de l'homonymie : une même graphie (i.e. : forme lexicale) peut représenter deux objets différents dans le monde physique. Exemple :

> « Sylvie **Paris** semble avoir développé une véritable addiction au PMU (**Paris** Mutuels Urbains). Elle fréquente assidument l'hippodrome de Longchamp à côté de **Paris**. »

Sur cet exemple fictif, il est facile de montrer combien il serait problématique de remplacer (à l'aveugle : automatiquement) toutes les occurrences de la graphie « Paris » par une même autre graphie (fusse-t-elle codée comme Dupond par exemple). D'abord, le texte perdrait de sa consistance comme dans l'explicitation du sigle : « PMU (Dupond Mutuels Urbains) ». Mais surtout, justement à cause de cette

explicitation d'un sigle par ailleurs bien connu, ou pour la précision géographique en fin de phrase, il serait très facile pour un français de déduire que la graphie substituée par Dupond était « Paris » et donc, de ré identifier le patronyme de Sylvie. Dans ce cas, il faudrait pouvoir distinguer les occurrences de « Paris » qui représentent le patronyme de Sylvie des autres occurrences (PMU et ville) et n'effectuer le remplacement (par un pseudo) que lorsque « Paris » est utilisé comme patronyme désignant Sylvie.

Cette règle de transformation cesse de fonctionner quand le contenu relie explicitement une graphie personnelle (ex : patronyme) à un homonyme ou à une description (étymologie, signification, etc.) Pour reprendre l'exemple précédent, considérons l'extrait : « Je m'appelle Sylvie *Paris* (comme la *capitale française*)… ». La description du patronyme donnée entre parenthèses annule tout effet d'anonymisation. Dans cette situation, on peut choisir de ne pas transformer le patronyme « Paris » (qui peut être considéré comme suffisamment commun, on changera alors le prénom pour brouiller les pistes). On obtiendrait alors « Je m'appelle Sandrine *Paris* (comme la *capitale française*)… ». Ou bien, il faut envisager de modifier la description en rapport avec la graphie de remplacement ainsi : « Je m'appelle Sylvie *Dublin* (comme la *capitale irlandaise*) … »

Enfin, comme cela a déjà été évoqué dans une première étude (Reffay & Teutsch, 2007), certaines analyses peuvent nécessiter des informations telles que le prénom, la localisation, la langue maternelle, ou la nationalité, et, à ce titre, doivent être conservées sous une forme intelligible pour le chercheur. À l'issue du projet canadien KUPI (*Knowledge Utilization and Policy Implementation*) dans le domaine de la santé, Denise Thomson et ses collègues montrent les difficultés d'une analyse qualitative seconde (après anonymisation).

Tous ces obstacles à une anonymisation entièrement automatisée nous ont conduits à envisager une méthode *interactive*, nécessitant l'intervention du chercheur, mais rendue *systématique* par l'assistance des outils de traitements qui s'y adossent.

3. Une méthode d'anonymisation semi-automatique

3.1. État de l'art

Meystre et al. (2010) comparent de nombreux outils anglophones d'anonymisation (dépersonnalisation) de dossiers de patients dans les hôpitaux. Ils classent 18 outils en 2 catégories principales : ceux utilisant les techniques d'apprentissage machine, et ceux basés sur des dictionnaires et des listes. Ils montrent que la plupart des outils s'appliquent à des textes prétraités (étiquetage morphosyntaxique) et que ceux de la deuxième catégorie sont fondés sur trois techniques fondamentales : a) des ressources extrinsèques au corpus à anonymiser : dictionnaires de noms communs, listes d'entités nommées (noms propres de personnes d'établissements ou de villes, états ou pays), b) des listes de déclencheurs (voisinage immédiat des entités nommées, ex : de, Mc, Van, Mister, lives in, resident of, etc.) et des expressions régulières (motifs lexicaux réguliers comme par exemple « .*@.* » ou « http.* » permettant de trouver des adresses de courriel ou des URL).

Après avoir essayé d'adapter l'outil De-Id (Neamatullah et al., 2008) au français, Grouin et al. (2009) ont développé Medina, un nouveau logiciel pour l'anonymisation de comptes-rendus d'hospitalisation en français, qui utilise les expressions régulières pour les entités numériques (ex : dates), des dictionnaires et listes pour les entités nommées, et réalisent une deuxième passe pour scruter le voisinage des termes déjà anonymisés.

Pour l'anonymisation de corpus d'interactions potentiellement multilingues, nous nous proposons de construire un outil adapté à ce type de données textuelles et répondant aux exigences de l'analyse de telles ressources par les chercheurs (a priori SHS). En proposant des techniques interactives de fouille et de marquage, nous visons une méthode interactive qui laisse le contrôle du niveau d'anonymisation au chercheur, mais qui le seconde dans les tâches complexes ou systématiques.

3.2. Présentation générale de la méthode

Dans notre méthode d'anonymisation, nous proposons de distinguer 7 étapes partant d'un corpus initial (à anonymiser) pour rendre, à la fin de la méthode, le corpus anonymisé :

1. Choisir les catégories d'informations (types d'entités) à identifier : noms, prénoms, institutions, villes, adresses de courriel, comptes facebook, MSN, numéros de téléphone, etc. ;
2. Catalogage : Lister toutes les graphies connues susceptibles de figurer dans le corpus et qu'il faut repérer et marquer (ex : prénom, patronyme, ville, courriel, et institution des participants) ;
3. Marquage : Marquer toutes les occurrences des graphies connues en les associant aux objets (entités réelles) qu'elles représentent (qui peuvent être différents en cas d'homonymie) ;
4. Fouille : Détection de nouvelles graphies ;
5. Définir pour chaque association (graphie - objet référencé) si la graphie doit être transformée au cours de la dernière étape (de substitution), et par quelle graphie de substitution ;
6. Vérifier la cohérence des graphies de substitution ;
7. Substitution de toutes les graphies (trop révélatrices) par leur graphie de substitution.

Les sept étapes de la méthode sont placées sur la figure 1 pour en expliciter l'enchaînement, les données nécessaires à chaque étape automatisable ou manuelle, ainsi que les données produites.

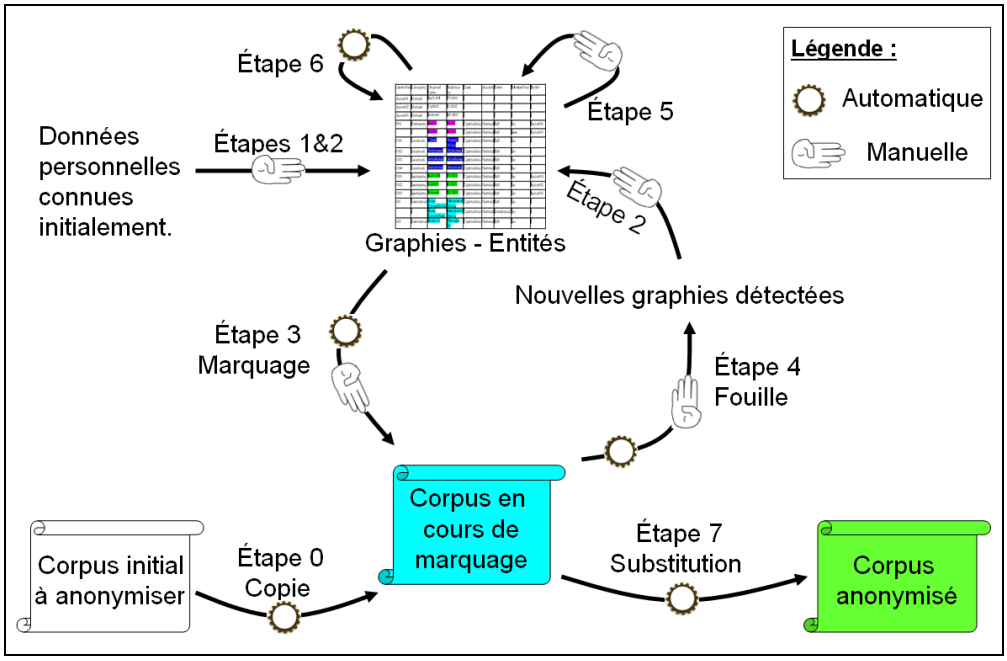

Figure 1. *Présentation générale de la méthode*

Les étapes 1 et 2 sont cruciales pour la qualité du résultat de l'anonymisation. Toutes les informations (externes au corpus) connues par le chercheur a priori sur les marques d'identification, sont autant d'indices permettant de trouver des références aux objets à repérer pour masquer. Plus ces indications sont précises et nombreuses, et plus large peut être la couverture des graphies repérées parmi celles susceptibles de ré-identifier les acteurs de la situation. Dans de nombreuses situations d'apprentissage, les chercheurs disposent d'une liste des participants avec quelques-unes de leurs caractéristiques : prénom, patronyme, institution, adresses, nationalité ou langue maternelle, etc. Ce sont précisément ces informations qu'il convient de recenser dans une première liste de graphies.

L'étape 3 de marquage peut être assistée d'outils (recherche dans un éditeur, concordancier) et doit permettre de retrouver à l'intérieur du corpus (dans leur contexte) toutes les occurrences des graphies listées au cours de l'étape 2. Pour chacune d'elles, le chercheur doit pouvoir choisir l'entité (objet du monde réel) à laquelle l'occurrence de cette graphie fait référence (ex : pour Paris : patronyme de Sylvie, PMU ou ville). Cette étape peut révéler de nouvelles entités (homonymes) à répertorier et à associer à la

graphie correspondante. À l'issue de cette étape, chacune des occurrences des graphies listées est marquée et associée à une entité définie.

L'étape 4 est celle que nous détaillerons dans la suite de cette communication. Elle est constituée de processus (partie automatique) de fouille proposant de nouvelles graphies (potentiellement personnelles) au chercheur qui doit (partie manuelle) les retenir ou les écarter.

Lorsque l'étape 4 est fructueuse, c'est-à-dire qu'elle a permis au chercheur de retenir de nouvelles graphies, il doit reprendre à l'étape 2 pour mettre à jour la liste des graphies dont il faudra marquer toutes les occurrences à l'étape 3. Si l'étape 4 ne détecte pas de nouvelles graphies, il passe à l'étape 5.

Ce n'est qu'à l'étape 5, quand le chercheur a une idée précise des entités représentées dans le corpus et des graphies (personnelles) recensées et marquées, qu'il est en mesure de décider avec finesse, quelles sont les graphies à remplacer pour rendre le corpus anonyme. Il doit alors choisir les pseudos qu'il peut utiliser pour ne pas dénaturer le corpus aux yeux des chercheurs susceptibles de l'analyser subséquemment.

L'étape 6 (entièrement automatique) doit vérifier certaines propriétés concernant les pseudos, et, le cas échéant, alerter le chercheur avant d'effectuer les substitutions (de l'étape 7) :

- ☑ Un même pseudo choisi pour deux graphies initialement différentes : ce qui risque d'engendrer des quiproquos qui n'avaient pas lieu d'être dans le corpus initial ;
- ☑ Un pseudo entre en collision avec une graphie (non modifiée) du corpus originel ;
- ☑ Une même graphie est remplacée par deux pseudos distincts : ce qui risque de rendre un quiproquo (dans le corpus d'origine) inconsistant dans le corpus anonymisé.

Le chercheur, ainsi alerté, peut bien sûr choisir de modifier certains pseudos en reprenant l'étape 5 ou d'ignorer les alertes et passer à l'étape 7 finale.

L'étape 7 de substitution est entièrement automatisable. Elle repose sur la table de transformation d'une part et le marquage de chaque occurrence d'autre part. La table précise pour chaque association (graphie – entité) par quel pseudo la graphie doit être remplacée. Dans le corpus marqué, chaque occurrence est associée à une entité. Ainsi, le processus parcourt le corpus (en une seule passe), et pour chaque occurrence marquée, vérifie dans la table si elle doit être remplacée. Dans l'affirmative, la graphie est substituée par le pseudo correspondant à l'association (graphie – entité) de cette occurrence.

4. Les techniques de fouille

Dans ce papier, nous mettons l'accent sur les techniques de fouille (i.e. : étape 4 de notre méthode) permettant de détecter des graphies non encore listées. Elles constituent à la fois l'originalité et le point fort de notre méthode. La détection de graphies pouvant révéler des informations personnelles est en effet l'étape la plus sensible de l'anonymisation. C'est elle qui détermine à la fois le temps nécessaire au traitement et la qualité du résultat de l'anonymisation. Cette étape peut être révisée en fonction des exigences légales et éthiques déterminées par le contexte. Nous nous plaçons dans des situations d'apprentissage où les interactions entre participants ne devraient pas dévoiler d'informations sensibles. Le risque que nous souhaitons réduire au minimum (idéalement supprimer), est celui d'une détection fortuite (par un internaute extérieur à la situation et utilisant un moteur de recherche) d'une graphie pouvant identifier un participant de la situation.

Compte tenu de la taille des corpus visés, nous cherchons des techniques de détection des graphies personnelles pour s'affranchir de la lecture exhaustive du corpus.

Une première approche basée sur des dictionnaires et testée en 2008, avait permis de repérer les graphies du corpus Simuligne (Chanier et al., 2007) qui ne faisaient pas partie des mots des dictionnaires français et anglais. Mais pour s'affranchir des langues, nous nous sommes imposés ici de ne pas nous

appuyer sur des dictionnaires (académiques) mais plutôt sur les graphies effectivement produites par les apprenants (éventuellement jeunes, et maîtrisant parfois mal les règles d'écriture de la langue : ce qui est particulièrement vrai dans le cas de situations d'apprentissage des langues).

Ainsi, dans l'approche développée ici, nous utilisons 3 techniques de fouille :

- ☑ **Expressions régulières** (ex : « http://.* », « .+ @ .+\..+ ») : pour détecter les adresses web, de messagerie, des dates ou toute forme suivant des règles précises conduisant à des motifs lexicaux réguliers ;

- ☑ **Variations lexicales** : Etant donnée une liste de graphies, nous recherchons celles qui figurent effectivement dans le corpus et qui peuvent être considérées comme (lexicalement) proches. Ces variations concernent des inversions de lettres, des erreurs de majuscule ou d'accents ou des caractères manquants ou supplémentaires par rapport à la graphie attendue ;

- ☑ **Règles contextuelles** : nous reprenons ici l'hypothèse de (Meystre et al., 2010) des contextes déclencheurs : i.e. : contextes potentiellement récurrents révélant les graphies recherchées. Nous systématisons la recherche pour proposer au chercheur un nombre limité de candidats.

L'utilisation des expressions régulières n'est pas détaillée ici, mais certains exemples pourront être présentés au moment de l'application sur les deux corpus. Cette section présente de façon assez formelle les deux autres techniques avant d'en présenter le résultat de leur application en section 6.

4.1. Le principe de propagation

Dans notre méthode, nous partons d'une liste de graphies connues (ex : noms des participants, lieux de résidence, institutions de référence, etc.), et nous utilisons deux techniques de fouille (variation lexicale et règles contextuelles) pour tenter de détecter de nouvelles graphies. La propagation utilise soit la forme lexicale des graphies connues pour trouver des graphies lexicalement proches, soit les contextes d'apparition des graphies connues pour rechercher (dans les autres occurrences de ces mêmes contextes) d'éventuelles nouvelles graphies. Les nouvelles graphies ouvrent de nouvelles variations lexicales et offrent de nouveaux contextes d'occurrences. La couverture des graphies repérées peut ainsi s'élargir par ce système de propagation miroir : contextes/graphies.

4.2. La technique des variations lexicales

Cette technique a pour but de comparer les graphies déjà référencées à l'ensemble du lexique du corpus : liste exhaustive de toutes les graphies (distinctes), et pour chacune d'elles, le nombre d'occurrences dans le corpus. Cette technique tente de palier au problème des erreurs typographiques que les participants peuvent commettre lorsqu'ils font références aux autres participants ou aux objets de leur monde. La taille du lexique est évidemment bien plus petite que le nombre total des graphies du corpus (incluant les doublons). Ce lexique peut être extrait par des outils tels que Calico (2012) ou TXM (2012).

Un outil (prototype sur tableur) compare automatiquement chaque graphie G_{ent} recensée (dans la liste des entités nommées) avec chaque graphie G_{lex} du lexique du corpus. Cet outil soumet au chercheur les graphies du lexique qui vérifient l'un des critères suivants :

- ☑ (Majuscule_et_sans_accent(G_{lex}) = Majuscule_et_sans_accent (G_{ent})) et ($G_{lex} \neq G_{ent}$) : Les graphies qui ne diffèrent que de la casse ou des accents, ou
- ☑ Levenshtein(G_{lex}, G_{ent}) = 1 et Longueur(G_{ent}) ≤ 5 : Les petites graphies (maximum 5 caractères) qui ne diffèrent que d'un caractère ou inversion, ou
- ☑ Levenshtein(G_{lex}, G_{ent}) ≤ 2 et ($G_{lex} \neq G_{ent}$) et Longueur(G_{ent}) > 5: Les grandes graphies (plus de 5 caractères) dont la distance de Levenshtein est égale à 1 ou 2.

Où les fonctions citées ci-avant sont définies par :

- ☑ Majuscule_et_sans_accent(G) : renvoie la graphie G' identique à G mais en transformant les lettres en majuscules et en supprimant les accents ;
- ☑ Levenshtein(G1, G2) : détermine le nombre d'ajouts, suppressions ou inversions de caractères pour passer de la graphie G1 à la graphie G2 (Levenshtein, 1966).
- ☑ Longueur(G) : renvoie le nombre de caractères composant la graphie G.

Comme nous le montrerons en section 6. , cette technique permet de repérer très efficacement de nombreuses variantes dans des corpus d'interactions entre participants qui s'accordent de grandes libertés quant à la façon d'utiliser la langue. Par construction, cette technique ne peut détecter que des graphies lexicalement proches. En revanche, la technique des règles contextuelles présentée ci-après peut révéler des graphies entièrement différentes.

4.3. La technique des règles contextuelles

Pour qui est habitué à l'analyse de corpus d'interactions en ligne, il n'est pas extravagant de penser que certaines catégories de graphies apparaissent dans des contextes récurrents, appelés contextes déclencheurs dans (Meystre et al., 2010). Par exemple, dans un corpus de courriels en français, les expressions « bonjour », « cordialement », « bien à vous », peuvent précéder un prénom ou un nom complet désignant l'auteur ou un interlocuteur. Pour implémenter cette idée sur les contextes gauches, nous proposons un algorithme en annexe 1. Le même algorithme est transposable aux contextes droits.

Partant d'une liste de graphies repérées (dans la liste des graphies et entités nommées), cet algorithme permet de recenser tous les contextes gauches capitalisables pour chacune de ces graphies. Au cours de son exécution, cet algorithme présente à l'utilisateur une liste (raisonnable) de contextes suivis de graphies potentiellement personnelles. C'est l'utilisateur qui décide de retenir (ou non) une nouvelle graphie et il l'associe à une entité nommée. Par conséquent, le résultat de cet algorithme est double : d'un côté, nous avons les nouvelles graphies détectées et de l'autre les contextes capitalisables. Ces derniers vérifient les propriétés suivantes :

1. Leur fréquence dans l'ensemble du corpus est inférieure à une limite F_{Max} ;
2. En proportion, le nombre d'occurrences de ce contexte suivi d'une graphie personnelle (ou nouvellement retenue) est supérieur à un taux minimum T_{Min}. Ce qui signifie que le contexte déclenche (à sa droite) des graphies personnelles dans une proportion supérieure à T_{Min}.
3. Un contexte ne peut pas contenir de graphie personnelle (i.e. : de la liste des graphies repérées)

La dernière propriété est nécessaire car nous ne pouvons pas nous permettre de conserver des contextes contenant des graphies personnelles, si ceux-ci sont réutilisés pour l'anonymisation d'autres corpus. Pour ne pas nous priver de ces contextes, nous proposons de les généraliser en remplaçant chaque graphie personnelle incluse par son type.

Exemple avec $F_{Max} = 50$ et $T_{Min} = 15\%$:

Supposons que nous ayons parmi les contextes gauches de la graphie « Paris », le prénom « Sylvie ». Si, dans le corpus, nous avons (critère 1) moins de 50 occurrences de « Sylvie », et que (critère 2) la proportion des « Sylvie » suivies de « Paris » est supérieure à 15%, alors on peut considérer que « Sylvie » est un bon contexte gauche déclencheur du patronyme « Paris ». Mais, si « Sylvie » est une graphie représentant le prénom d'une actrice (ex : une tutrice), elle figure donc dans la liste des graphies recensées et ce contexte gauche viole ainsi le critère 3. Il ne peut donc pas être capitalisable sous cette forme. Pour résoudre le problème, nous généralisons ce contexte en proposant que tous les prénoms (de la liste des graphies recensées) puissent servir de contexte gauche, révélateur d'une nouvelle graphie personnelle. Avec cette méthode, nous pouvons capitaliser des contextes gauches généralisés tels que : « <prénom>… », « <prénom> et … », « <prénom>, … », etc. De façon symétrique, nous obtenons des contextes droits généralisés comme « …<patronyme> », « … et <prénom> », « …, <prénom> », etc.

Au moment de la rédaction de cet article, cette technique a été appliquée sur deux corpus (présentés dans la section suivante). Nous n'avons considéré pour l'instant que les contextes immédiats des graphies connues. Mais pour des corpus structurés (tels que les forums de discussion), nous envisageons de prendre

en compte d'autres informations telles que : l'auteur du message, la liste des participants au fil de discussion, l'auteur du message auquel on répond, etc.

Les deux techniques de fouille que nous venons de présenter sont complémentaires : elles fonctionnent sur des principes indépendants (l'un lexicale, l'autre contextuel). L'utilisation d'expressions régulières repose sur un principe de régularité lexicale également indépendant. C'est pourquoi, la combinaison des trois techniques, associée au principe de propagation nous semble prometteuse.

Après avoir décrit dans cette section la méthode générale d'anonymisation et ses techniques particulières de fouille, nous présentons (dans la section 5.) les deux corpus sur lesquels cette méthode a été appliquée avant de donner (en section 6.) les résultats de cette application.

5. Présentation des corpus

Comme le montre le tableau 1 à la fin de cette section, les deux corpus ayant servi de test pour cette méthode diffèrent sur de nombreux points. Ils sont en effet issus de cadres très différents présentés dans les deux sous-sections suivantes.

5.1. Le cadre du corpus « Programme Court »

Dans le cadre de la maitrise en technologie éducative à l'Université Laval, le programme de formation concerné porte sur l'intervention dans les petites écoles et les classes multi-âges en réseau. Il s'inscrit dans le projet canadien de « l'École Éloignée en Réseau » (Laferrière et al., 2011). Ce programme comporte trois cours ; chacun d'eux étant sous la responsabilité d'une des trois universités partenaires (UQAC, UQAT, UQO). Puisqu'une partie importante des étudiants visés par ce programme de formation est susceptible d'être située en milieu rural, et donc éloignés d'un centre universitaire, les cours sont dispensés en réseau par l'entremise d'un forum électronique (Knowledge Forum : KF) et d'un outil de vidéoconférence (VIA) accessible à partir d'un ordinateur personnel. Le corpus qui nous concerne est constitué de l'ensemble des notes déposées par le groupe (le tuteur et les 7 étudiants) dans le Knowledge Forum entre janvier et mai 2011. Le scénario d'usage du KF dans ce contexte est inspiré des 12 principes de coélaboration de connaissances (knowledge building), initialement suggérés par Bereiter et Scardamalia (2003), remaniés en cinq dans (Allaire & Lusignan, 2011).

5.2. Le cadre du corpus « Nomades »

Ce deuxième corpus est issu d'une session réalisée sur la plateforme Galanet (2012) et utilise l'intercompréhension pour l'apprentissage des langues romanes : 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 devrait s'exprimer dans sa langue maternelle. Cette session a été dirigée par Sandrine Deprez qui en a fait une première analyse dans (Deprez, 2012). Les interventions dans le forum s'échelonnent sur trois mois : de novembre 2011 à janvier 2012. En dépit des consignes, il est fréquent que les élèves utilisent d'autres langues que la leur et certains ont même recours à plusieurs langues dans un même message.

5.3. Comparaison de la nature des deux corpus

Pour montrer combien la nature de ces deux corpus est différente, nous présentons quelques de leurs traits caractéristiques dans le tableau 1 ci-dessous.

Corpus	Niveau de formation, Pays	Inscrits/Participants	Langues	Taille
Programme Court (privé)	Master en technologie éducative, Université Laval au Québec (Canada)	1 tuteur, 7 étudiants. Tous participent	Français (du Québec) langue maternelle	Lexique : 4900 graphies Taille : 203 messages, soit : 41 317 graphies.
Nomades	Lycée (niveau Première) au Brésil, en Espagne (Catalogne),	2 tuteurs, 101 élèves :	Français, italien, catalan, castillan et portugais (du	Lexique : 9652 graphies Taille : 915 messages,

(public)	en France et en Italie	seuls 83 participent	Brésil)	soit : 46 825 graphies.

Tableau 1. *Comparaison des deux corpus*

Un rapide calcul à partir des données de la dernière colonne du tableau 1 nous montre que la taille moyenne d'un message est 4 fois plus importante dans le premier corpus (204 graphies/message) que dans le deuxième (51 graphies/message) tandis que la taille globale des deux corpus est très similaire (41 317 contre 46 825 graphies). Dans le deuxième corpus, il y a 83 auteurs écrivant en moyenne 11 messages chacun, tandis que dans le premier, les 8 auteurs ont écrit en moyenne 25 messages chacun.

Au-delà des chiffres, on constate que ces deux corpus diffèrent sur de nombreux aspects : la taille des groupes, leur niveau d'engagement (visible dans la taille, le nombre et la qualité des messages), le niveau de maîtrise de la langue et l'aspect plus ou moins formel des contenus attendus. Nous montrerons dans les résultats que ces différences ont probablement un impact sur la nature et la diversité des marques personnelles que l'on peut trouver dans ces deux corpus, et par conséquent sur l'efficacité de nos techniques de fouille.

6. Application de la méthode sur les deux corpus

Compte tenu du fait que l'accès au premier corpus (Programme court) est restreint aux personnes autorisées, nous ne pourrons pas donner d'exemples de graphies personnelles authentiques (avant anonymisation) pour ce corpus car, après avoir décrit ici l'institution et le contexte, compte tenu de la taille du groupe, chaque prénom suffit à identifier chaque participant. C'est pour cette raison que nous puiserons essentiellement nos exemples d'illustration sur le deuxième corpus, qui est accessible (sur demande) sur les plateformes Galanet et Calico.

Les données d'interaction ont été traitées à partir de leur import sur la plateforme Calico[1] (Blondel & Giguet, 2011). Les outils en ligne de Calico (lexique de Colagora, Concordagora) ont utilisés pour construire des listes de graphies à partir des lexiques issus des corpus et pour contrôler la présence ou l'absence ou même le nombre d'occurrences de graphies particulières dans les corpus anonymisés. Pour pouvoir réaliser certaines opérations hors ligne, nous avons aussi eu recours aux fonctionnalités équivalentes du logiciel TXM (2012).

6.1. Initialisation de la méthode : étapes 0, 1 et 2 (cf. section 3.2.)

Nous supposons que le corpus de messages a déjà été mis au format XMLForum et déposé sur la plateforme Calico. La méthode ayant été appliquée essentiellement à la main sur ces deux premiers corpus, nous avons réalisé les opérations de marquage et de substitution sur une copie locale du corpus dans un éditeur de texte (notepad++). À partir des données connues du chercheur, nous définissons les types d'entités à repérer dans le corpus et cataloguons les graphies connues. Pour les deux corpus, nous souhaitions a priori identifier pour chaque participant le prénom, le patronyme, la ville de résidence ou de travail, les adresses connues (postales, courriel, facebook, etc.), et l'institution de référence. Le tableau 2 illustre par quelques lignes, les informations rentrées dans la table à cette étape d'initialisation du processus pour le traitement du corpus « Nomades».

Graphie origine	Forme	Catégorie	Type Entité	Identifiant Entité
Kelly	Normale	Prénom	Participant	F058
Rosa Luxembourg	Normale	Nom	Institution	I03
Canet	Normale	Nom	Petite ville	P007
Barcelona	Normale	Nom	Ville	P002
…	Normale	…	…	…
Adrià	Normale	Prénom	Participant	F021
Medeiros	Normale	Patronyme	Participant	L039

[1] http://www.stef.ens-cachan.fr/calico/calico.htm

Tableau 2. *Extrait de la table des graphies et entités nommées pour le corpus « Nomades 2011-2012 » - Initialisation*

Pour le corpus « Nomades », ce sont près de 150 graphies répertoriées (103 prénoms, 44 patronymes, 4 institutions, 5 villes de résidence), que nous avons pu trouver dans un espace protégé de Galanet appelé « Qui est qui » attaché à cette session. Pour « Programme court », ce sont les 8 prénoms et patronymes qui permettent d'initialiser la table, ce qui, compte tenu des noms composés, représente 20 graphies distinctes. Comme le montre le tableau 2, il s'agit de préciser pour chaque information : la graphie (sous sa forme normale), sa catégorie (en colonne 3), le type d'entité à laquelle la graphie fait référence (i.e. : quel genre d'objet du monde réel la graphie représente ?) et l'entité précise (désignée par un identifiant) à laquelle cette graphie fait référence. Par exemple, sur la troisième ligne : *P007* identifie de façon unique la *petite ville* dont la *forme normale* du *nom* est *Canet*.

6.2. *Premier marquage : étape 3 (cf. section 3.2.)*

Dans l'outil visé (sur Calico), la phase de marquage pourra s'appuyer sur le concordancier (Concordagora) dont une illustration est donnée en figure 2.

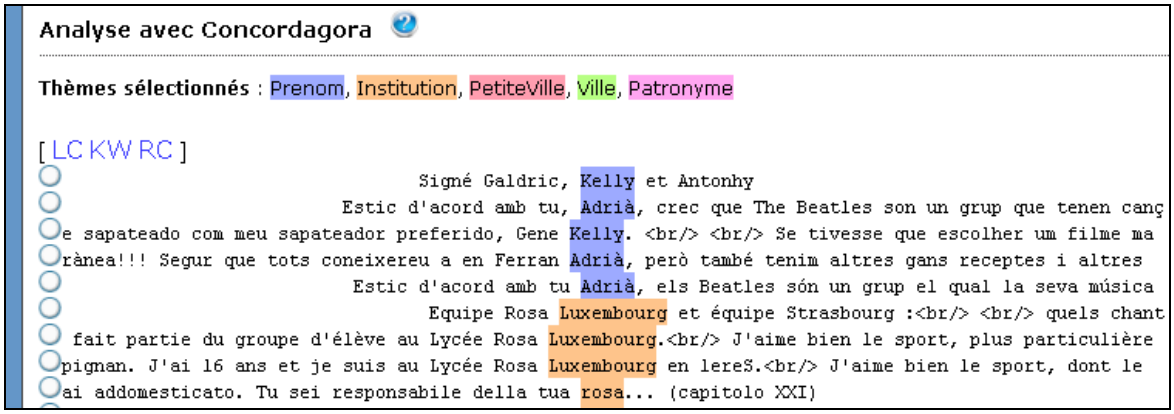

Figure 2. *Exemple d'utilisation du concordancier*

Le concordancier permet d'afficher tous les contextes dans lesquels les différentes graphies reconnues apparaissent. Chaque occurrence reconnue est colorée dans la couleur correspondant à sa catégorie (Prénom, Institution, etc.) Le contexte de l'occurrence surlignée est indispensable à l'utilisateur pour déterminer l'entité qu'elle représente. Par exemple, sur la première ligne, Kelly apparaît au milieu de 2 autres prénoms de participants. On peut en déduire (s'il n'y a qu'une Kelly) que c'est bien le prénom qui désigne la participante (de code F058 du tableau 2). Techniquement (en arrière plan pour le futur utilisateur, mais en parallèle et à la main sur la version locale dans notre test), afin de conserver cette information dans le code XML du corpus, nous balisons cette occurrence de la façon suivante :

```
<calix entity_id="F058" type="Participant" category="firstname"
       case="capitalized" form="normal" modified="wait">Kelly</calix>
```

Mais sur la troisième ligne, une autre occurrence de Kelly apparaît dans un tout autre contexte. Cette occurrence de Kelly est un patronyme qui désigne un personnage public. Il faut donc l'ajouter à la table des entités avant de marquer (par les tags XML) cette nouvelle occurrence de Kelly.

```
<calix entity_id="PP001" type="Public" category="lastname"
       case="capitalized" form="normal" modified="wait">Kelly</calix>
```

Nous voyons sur la figure 2 que la graphie Adrià (enregistrée dans la table 2 comme prénom d'un participant) peut aussi désigner le patronyme d'un personnage public : ici, le cuisinier Ferran Adrià. Le dernier cas visible sur la figure 2 concerne la graphie Rosa (Institution) dont l'occurrence sur la dernière ligne représente un nom commun (la rose en italien). Le cas des noms communs étant fréquent, nous envisageons une option (en un seul click) permettant de catégoriser ces occurrences dispensant l'utilisateur de définir une nouvelle ligne dans la table. Cette occurrence doit malgré tout être marquée dans le corpus :

```
<calix type="Common" case="lower" form="normal" modified="wait">rosa</calix>
```

Cette méthode permet de marquer toutes les occurrences des graphies listées, qu'elles doivent ou non être modifiées par le processus de substitution final. Cette garantie permet au chercheur de différencier (dans le corpus anonymisé), les graphies volontairement inchangées de celles qui auraient été oubliées. L'attribut « modified="wait" » indique que le processus d'anonymisation est en cours et que la décision de transformer ou non cette graphie n'a pas encore été prise. La valeur de cet attribut devra être changée en "yes" ou "no" au cours de la phase finale de substitution.

Dans le corpus « Programme court », à partir des 18 graphies distinctes (9 prénoms et 9 patronymes), le concordancier trouve 110 occurrences. Seules 3 d'entre-elles représentent des entités nouvelles (non définies au départ dans la table).

Partant de 264 graphies (distinctes) dans le corpus « Nomades », le concordancier nous permet de situer 222 occurrences dans leur contexte. 46 d'entre-elles représentent 16 nouvelles entités et 2 noms communs.

6.3. Premières fouilles : étape 4 (cf. section 3.2.)

6.3.1. Fouilles lexicales

Dans le corpus « Programme court », à partir des 18 graphies distinctes, la fouille lexicale génère 25 graphies candidates (présentes dans le lexique) à l'utilisateur. Une seule d'entre-elles a pour initiale une majuscule et représente un nouveau prénom, 2 autres représentent un prénom ou un patronyme de participant écrit tout en minuscules. Deux autres enfin (claire, manuel) auraient pu être des prénoms mais le contexte de leurs occurrences montrent qu'ils sont utilisés comme noms communs. On ne retient donc que 3 graphies sur les 25 proposées.

Partant des 264 graphies pour « Nomades », nous avons du séparer les traitements en 3 passes : 103 prénoms, 108 noms d'utilisateurs et 53 autres (44 patronymes, 4 institutions, 5 villes). Des 103 prénoms initiaux, nous obtenons 95 candidats dont 31 seront retenus. Pour illustrer les variations lexicales, nous présentons dans le tableau 3 les graphies initiales (ayant généré au moins une dérivée retenue) et les graphies dérivées retenues.

Graphies initiales	**Adriana Alèxia Anthony Baptiste Cleissa Eli Elouise Emmanuel Federica Ferran Gabriela Guillem Iñigo Jaqueline Jean José Kelly Léo Mariana Mary Michela Monica Olalla Oleguer**
Dérivées retenues	**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**

Tableau 3. *Variations lexicales des prénoms du corpus Nomades lors de la première fouille lexicale*

Toutes les graphies dérivées retenues (à ce stade) sont ajoutées dans la table graphies-entités.

6.3.2. Fouilles contextuelle

La nouvelle liste des graphies/entités du corpus « Programme court » est passée de 18 à 21 graphies lors de la fouille lexicale (cf. 6.3.1.) Dans le corpus, ces 21 graphies admettent au total 113 occurrences qui sont autant de contextes (gauches et droits) pouvant révéler de nouvelles graphies. En appliquant l'algorithme de l'annexe 1, on distingue 110 contextes gauches. 36 peuvent être automatiquement retenus pour capitalisation puisque toutes les occurrences de ces contextes sont suivies d'une graphie recensées. 46 contextes (18 capitalisables et 28 non capitalisables) sont testés dans le concordancier pour exhiber les 312 occurrences suivies de graphies non encore listées : aucune nouvelle graphie personnelle détectée. Au cours du processus, plusieurs contextes gauches (trop fréquents) ont été rejetés. Deux des 110 contextes contenaient eux-mêmes un prénom recensé et ne pouvaient être capitalisés en l'état. Ils ont permis de généraliser la règle (capitalisable) de composition des noms séparés par un trait d'union : La version contexte gauche (<Nom>-…) renvoie une nouvelle graphie (aussi détectée dans la fouille lexicale). La

version contexte droit (…-<nom>) permet de détecter un nouveau prénom. Parmi les règles généralisées, nous pouvons citer les contextes symétriques (<nom> et …, … et <nom>) en remarquant que (dans ce corpus) le taux de réussite du contexte gauche est de 4/8 tandis que pour le contexte droit il est de 4/4.

Si la fouille contextuelle ne semble pas très payante pour le corpus précédent, elle s'avère tout à fait fructueuse dans sa première application sur le corpus « Nomades ». Le tableau ci-dessous présente les contextes ayant révélé de nouvelles graphies personnelles et les graphies révélées.

Contextes déclencheurs	**Cara ... / appelle… /Merci…/ llamo…/chamam…/accord avec… / concordo com a / meu nom / je m'appel… / diu el… / nombre, … / … , <Nom>/<Nom> i …/ <Nom> et … / … et <Nom>**
Graphies révélées	**adriana, Antonhy, Federiac, fran, iñigo, jean, Kellly, Leo, léo, May, Asenjo, Belle, Bet, Beth, Christine, Fede, Line, Maria, Peimika, Regina**

Tableau 4. *Contextes déclencheurs (fructueux) et nouvelles graphies révélées dans « Nomades » lors de la première fouille contextuelle*

6.4. Fin du processus

Les nouvelles graphies détectées au cours de l'étape précédente de fouille (ex : 4 prénoms pour « Programme Court » et 41 prénoms pour « Nomades »), doivent être ajoutées à la table graphies/entités avant de marquer toutes leurs occurrences (soit : 5 pour le premier corpus et 59 pour le deuxième). Au cours du marquage, on détecte 2 nouvelles entités (une école et un auteur scientifique) dans le premier corpus et 6 (personnages publics) dans le deuxième. Les itérations suivantes (sans l'outil) deviennent difficiles à effectuer. Nous devrions être en mesure de finaliser la détection et la transformation pour permettre le partage des corpus anonymisés au moment de la conférence.

7. Bilan et perspectives

À l'aide de l'outil Colagora de Calico nous avons pu demander à Sandrine Deprez (détentrice du corpus « Nomades ») 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 ne pouvait pas trouver), les 6 autres n'auraient pas échappé à la méthode si elle n'avait pas été appliquée partiellement à la main.

Pour le corpus « Programme court » (entièrement en français), nous avons eu recours à des expressions régulières pour détecter toutes les formes ayant une première lettre majuscule, placées ailleurs qu'en début de phrase. Cette requête (effectuée dans TXM) renvoie 792 occurrences en contexte, desquelles nous avons pu extraire une cinquantaine de graphies faisant référence à 11 lieux (dont 2 trop précis), 6 institutions (dont 2 à masquer), 7 personnes (dont 3 à masquer). Soit au total 8 entités non masquées alors qu'elles auraient du l'être, représentées par 9 graphies et 31 occurrences pour 115 effectivement marquées.

Cette première évaluation est très encourageante pour le corpus « Nomades » tandis qu'elle montre une grande faiblesse de la méthode pour le corpus « Programme Court ». Nous l'expliquons essentiellement par le fait qu'il concerne peu de locuteurs : nous avons donc peu de graphies à entrer dans la phase d'initialisation, il y a moins de contextes récurrents et, puisque le contenu est rédigé dans un bon niveau de langue, nous pourrions avoir recours à un dictionnaire de la langue. Cependant, la piste des expressions régulières utilisée dans la phase d'évaluation peut également être introduite au moment où les 2 autres techniques de fouille ne fournissent plus de nouvelles graphies candidates. Il nous reste à évaluer la réutilisabilité et la rentabilité des règles de contextes capitalisées en les employant sur un corpus semblable (ce qui n'était pas du tout le cas des deux corpus étudiés ici).

Au-delà de la méthode, nous espérons pouvoir développer (dans Calico) un outil qui permette de la mettre en œuvre efficacement pour que les chercheurs puissent enfin partager leurs données, tout en protégeant les acteurs impliqués et ce faisant, améliorer les méthodes de recherche de notre communauté.

8. Références

AFNOR (2000). *Anonymisation - Glossaire et démarche d'analyse et expression de besoins*. Paris : AFNOR. http://www.abs92.com/documents/boite_a_outils /notions_fondamentales/notions_de_stat/2_anomysation.pdf

Allaire, S., & Lusignan, G. (2011). Enseigner et apprendre en réseau : collaborer entre écoles distantes à l'aide des TIC. Anjou : Éditions CEC.

Bereiter, C., & Scardamalia, M. (2003). Learning to work creatively with knowledge. In E. D. Corte, L. Verschaffel, N. Entwistle & J. v. Merrienboer (Eds.), *Unravelling basic components and dimensions of powerful learning environments* (pp. 55-68). Oxford, UK: Elsevier Science.

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. (coord.). *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).

Chanier, T., Lamy M.-N., Reffay, C., Betbeder, M.-L., Ciekanski, M. (2007). 'Corpus global Simuligne'. [Learning and Teaching Corpus]. Mulce.org: Université de Franche-Comté. [En ligne]. Accessible à [oai:mulce.org: mce.simu.all.all, http://repository.mulce.org]

Corti, L., Day, A., & Backhouse, G. (2000). Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. Forum *Qualitative Sozialforschung / Forum: Qualitative Social Research* [Online Journal], 1(3), Art. 7. http://www.qualitative-research.net/index.php/fqs/article/view/1024/2207

Datacite (2012) : http://datacite.org/repolist List of repositories for data access and reuse

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).

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.

Laferrière, T., Hamel, C., Allaire, S., Turcotte, S., Breuleux, A., Beaudoin, J., et al. (2011). L'École éloignée en réseau, un modèle. Rapport-synthèse, octobre 2011. CEFRIO.

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).

Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., et al. (2008). *Automated de-identification of free-text medical records.* BMC Medical Informatics and Decision Making, 8(1), 32. doi:10.1186/1472-6947-8-32

Nelson, B. (2009). Empty Archives. Nature, (461), News feature, pp 160-163, 10 sept. 2009.

Reffay, C., & Teutsch, P. (2007). *Anonymisation de corpus réutilisables : masquer l'identité sans altérer l'analyse des interactions*. Rapport interne, LIFC, 12 pages. http://edutice.archives-ouvertes.fr/edutice-00158877/fr/

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.

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), Art. 29*, http://nbn-resolving.de/urn:nbn:de:0114-fqs0501297

TXM (2012) : http://textometrie.ens-lyon.fr/?lang=fr : Plateforme et outils hors-ligne (version 0.6) de textométrie issus d'un projet ANR 2007-2010.

Annexe 1 : Algorithme de fouille contextuelle (contextes gauches)

```
// Données
G_ent : Liste des graphies répertoriées à repérer
F_Max : Fréquence maximale autorisée pour 1 contexte candidat (défaut : 50)
T_Min : Taux minimum de réussite pour capitaliser une règle(défaut : 20%)

// Initialisation
LC = Ø : Liste (vide) des contextes gauches à traiter
LC_cand  : Ensemble (éventuellement vide) des contextes gauches candidats
LC_good  : Ensemble (éventuellement vide) des contextes gauches retenus
LC_bad  : Ensemble (éventuellement vide) des contextes gauches rejetés

// Déroulement de la méthode
Si LC_good ≠ Ø, Alors {cas ou l'on réutilise des contextes capitalisés}
|   Présenter toutes les occurrences des contextes de LC_good
|   Si l'utilisateur repère de nouvelles graphies G_new Alors
|   |     Les ajouter G_new à la liste G_ent
|   Fin Si
Fin Si

Pour chaque graphie G_i de G_ent, faire :
|   LC = liste des contextes gauches (différents) de taille 1
|   Tant que LC ≠ Ø, Faire
|   |     C_i = first(LC)
|   |     N = Nombre d'occurrences de C_i dans tout le corpus
|   |     Si (N>F_Max) Alors
|   |     |     Ajouter à LC les différents contextes gauches C_i augmentés
|   |     |           d'une graphie (à gauche)
|   |     Sinon
|   |     |     Si (N>1) Alors
|   |     |     |     Ajouter C_i à LC_cand
|   |     |     Sinon
|   |     |     |     Ajouter C_i à LC_good
|   |     |     FinSi
|   |     FinSi
|   |     Retirer C_i de L_C
|   Fin Tant que
Fin Pour

Pour chaque contexte C_i distinct dans {LC_cand \ LC_bad} Faire :
|   Présenter à l'utilisateur toutes les occurrences de C_i suivies de
|         graphies non répertoriées
|   Si il retient un contexte et une nouvelle graphie G_new, Alors
|   |     Ajouter G_new à la liste G_ent
|   Fin Si
      Taux = Nb occurences de Ci devant une graphie référencée / Nb total d'occurrences de Ci
|   Si (Taux >= T_Min) Alors
|   |     Ajouter C_i à LC_good
|   Sinon
|   |     Ajouter C_i à LC_bad
|   Fin Si
Fin Pour

Transformer tous les contextes de LC_good contenant une graphie de G_ent

{Résultat : LC_good = Ensemble des contextes gauches capitalisables}
```