Anonymisation de corpus réutilisables

Christophe Reffay, Philippe Teutsch

Anonymisation de corpus réutilisables

Masquer l'identité sans altérer l'analyse des interactions

Christophe Reffay*, Philippe Teutsch**

** LIFC*
16 route de Gray
F-25030 Besançon Cedex
Christophe.Reffay@univ-fcomte.fr
*** LIUM, Université du Maine*
Avenue Olivier Messiaen
F-72085 LE MANS cedex 9
Philippe.Teutsch@lium.univ-lemans.fr

RÉSUMÉ. *La mise à disposition de corpus de traces issues de formations en ligne intéresse la communauté scientifique dans sa démarche d'analyse des interactions humaines produites à travers le dispositif médiatisé. Pour des raisons éthiques, les échanges de corpus doivent garantir l'anonymat des acteurs concernés. Ce travail s'intéresse au processus d'anonymisation d'un corpus pour en permettre un accès plus large. Les principes et l'outil d'anonymisation présentés sont appliqués à un corpus d'interactions en apprentissage des langues. Dans de telles situations, les marques d'identité à masquer sont tantôt des identifiants immuables produits par le système, tantôt des appellations ou interpellations produites par les acteurs eux-mêmes et sujettes à des variations syntaxiques parfois signifiantes et empreintes de culture (surnoms, diminutifs). Enfin, cette contribution pose la question de la frontière à définir entre le contexte socioculturel utile à l'analyse et l'identité de l'acteur à protéger.*

MOTS-CLÉS : *Interactions H-H médiatisées, Corpus de données d'apprentissage, Echange de corpus, Anonymisation.*

1. Introduction

Le travail présenté ici concerne l'étude des interactions humaines à distance, médiatisées par les réseaux dans un contexte d'apprentissage. Dès lors que la situation pédagogique est intégralement médiatisée par un environnement informatique, elle permet un recueil systématique de traces d'activité [CHOQUET et al. 05] et une restitution de ces traces. Cette restitution permet de les rejouer pour améliorer le dispositif (réingénierie), de les caractériser (formation de formateur), ou encore de les interroger (observation de comportements récurrents et analyse didactique) [GLIGOR-CALIN 06]. Les traces qui nous intéressent concernent les interactions produites entre acteurs d'une formation en ligne (FEL) à travers les différents outils de communication utilisés par ces acteurs.

L'intégration d'outils de communication variés dans les plateformes de téléformation modifient toutes les facettes de l'interaction humaine : les conditions, la nature et parfois même les enjeux de ces interactions de FEL où la relation sociale est essentielle [CHANIER 01]. L'étude et la modélisation de ces interactions cherchent à mieux utiliser le potentiel qu'elles représentent pour la conception de dispositifs de FEL. Les interactions étudiées n'ont d'intérêt pour ces recherches que si elles sont issues de situations écologiques, contextualisées, où l'ensemble des dimensions concernées sont concrètement vécues par les acteurs du dispositif.

De nombreuses recherches ont lieu sur les dispositifs de formation en ligne, mais leur scientificité est parfois remise en cause car elles sont rarement réplicables [HENRI & CHARLIER 05]. Pour que le débat scientifique ait lieu, au delà de la publication de recherches sur des corpus « cachés » et inaccessibles, il faut que les données elles-mêmes puissent être consultées et analysées par d'autres chercheurs. De façon duale, pour vérifier sa réplicabilité, une méthode d'analyse doit pouvoir être appliquée à différents corpus pour affiner ses paramètres d'application en fonction du contexte. Cette mise à l'épreuve à des contextes différents permettra à ces analyses d'étendre leur validité et aux outils de démontrer leur robustesse.

Le besoin de réutilisation de corpus a été mis en évidence dans le projet ODIL[1] qui a permis de concevoir un outil de visualisation de contenus de discussions en ligne (ViCoDiLi) indépendamment de leur structure d'origine [TEUTSCH et al. 07]. Le projet Mulce[2] propose une spécification structurant le corpus afin d'en permettre la réutilisation, l'échange et la diffusion [REFFAY et al. 07]. Ces deux actions ont identifié que, parmi les conditions de diffusion et de consultation des corpus, la protection des acteurs est incontournable.

---

[1] ODIL : Outils et Didactique pour l'analyse des Interactions en Ligne : projet (2004-2007) de l'ACI « Éducation et formation », piloté par F. Mangenot et regroupant 7 laboratoires.

[2] Mulce : Multimodal Learning Corpus Exchange : projet (2007-2009) de l'ANR « Corpus et outils de la recherche en SHS » , coordonné par T. Chanier. (http://mulce.univ-fcomte.fr)

Cet article étudie cet aspect critique de la diffusion de corpus d'interactions réelles : la protection des acteurs. Il présente successivement les enjeux et la problématique de l'anonymisation, puis les principes et la méthode d'anonymisation appliqués au corpus de formation en langues Simuligne et un premier bilan de mise en œuvre de l'outil correspondant.

2. Enjeu : protéger l'identité des acteurs lors de l'échange de corpus

Le caractère authentique et incarné des interactions interroge les conditions de diffusion, d'exploitation et d'affichage de ces corpus. Pour des questions de protection des personnes, l'accès aux contributions des acteurs d'une session de formation ne peut se faire qu'à condition de ne pas pouvoir reconnaître les personnes concernées. Les deux extraits ci-dessous illustrent cette nécessité de traitement préalable des corpus avant leur diffusion pour analyse.

> *Nous avons constitué un corpus de textes (courriels) écrits par une personne … âgée d'une quarantaine d'années. … Ce corpus est constitué de 205 courriels issus de sa correspondance personnelle. …* ***Tous les noms propres ont été changés afin de rendre ce corpus anonyme et utilisable par d'autres****. Nous avons ensuite procédé à la suppression des entêtes et des signatures des courriels.*

Figure 1. *Besoin d'anonymisation [BOISSIÈRE & SCHADLE 06, p.4]*

> *Compte tenu des dangers présentés par l'informatisation croissante des données personnelles le souci premier de tout maître de fichier doit être d'anonymiser les applications informatiques dès que cela est possible. Autrement dit, il convient de ne* ***jamais délivrer l'identité des personnes ou des éléments trop identifiants quand cela n'est nullement indispensable à la finalité du traitement****.*

Figure 2. *Nécessité d'anonymisation [MALLET-POUJOL 04, p. 28]*

Pour garantir la protection des acteurs des sessions de formation, les échanges de corpus ne peuvent en conséquence s'envisager que si les données ont été préalablement anonymisées.

3. Problématique : masquer l'identité sans altérer l'analyse des interactions

Bien que l'anonymisation soit imposée par l'éthique et rendue obligatoire par la loi, la dernière phrase (Figure 2) montre qu'elle peut s'appliquer à différents degrés

et qu'il convient de définir les informations « indispensables à la finalité du traitement », à savoir l'analyse du corpus dans notre cas : le tout est donc de déterminer les informations nécessaires à la compréhension et à l'analyse du corpus.

Les plateformes de formation en ligne permettent des échanges de plus en plus riches produisant des corpus d'interactions variées. Le détenteur d'un tel corpus qui souhaite en présenter des extraits à titre d'illustration tout en respectant l'anonymat des acteurs, peut traiter les données manuellement avec des techniques uniformisantes : voiler les visages ou maquiller les voix par exemple. Ce type de traitement peut être généralisé à l'ensemble d'un corpus pour alimenter des analyses qui n'ont pas besoin de distinguer les acteurs (études en linguistique pure ou certaines analyses du discours ou de comportements individuels par exemple). On peut dans ce cas effacer ou cacher toutes les informations personnelles. Pour d'autres recherches, par contre, ces techniques d'uniformisation peuvent perturber, voire empêcher une analyse efficace. L'étude des interactions dans un contexte d'apprentissage des langues (où la langue est objet et moyen d'apprentissage) ou d'apprentissage du travail collaboratif (où les aspects relationnels et conversationnels sont essentiels à la réussite de l'activité), par exemple, nécessite de distinguer chacun des auteurs de contributions pour comprendre la situation. D'un point de vue social, les interactions produites sont influencées par la représentation que chacun a des autres. Cacher ou déformer les informations permettant la perception des identités risque de réduire la portée et la qualité de l'analyse.

Nous cherchons donc à caractériser les constituants de la « personnalité » d'un individu dans un corpus d'apprentissage et d'y situer la ligne de partage entre d'une part l'identification précise de la personne, qu'il est nécessaire de protéger, et d'autre part la connaissance du contexte personnel nécessaire à la tâche d'analyse. L'objectif est de pouvoir tenir compte des besoins des analystes d'un corpus tout en respectant l'anonymat des acteurs de la formation. Notre démarche passe par la définition des données personnelles pour ensuite proposer un processus de transformation de ces données personnelles sans dénaturer le corpus.

Nous proposons de décrire la tâche d'anonymisation selon la terminologie suivante. L'anonymisation consiste à masquer (par effacement ou modification) le nom des intervenants et à empêcher de retrouver leur identité. L'outil effectuant le travail est l'anonymiseur. L'opérateur responsable de la tâche est l'anonymisateur.

4. Principes d'identification et d'anonymisation sur un cas d'étude

Les données liées à la connaissance des personnes impliquées dans un corpus d'interactions produites dans un dispositif de formation en ligne se découpent en deux catégories : d'une part l'identité proprement dite, i.e. les éléments qui permettraient, directement ou indirectement, l'identification des personnes physiques (le patronyme, l'adresse postale ou électronique, les dates et lieu de naissance, la photo, la voix, ou encore l'immatriculation d'un véhicule) et qui doivent donc être

modifiés, d'autre part le profil de la personne, i.e. les éléments qu'elle incarne dans ces interactions et qu'il est nécessaire de préserver pour l'analyse.

Ce traitement particulier d'anonymisation nécessite d'être guidé par l'utilisation à venir du corpus par un tiers et doit pouvoir être réalisé par le détenteur du corpus d'origine qui le « prépare » en vue de le confier à ce tiers. Ce principe et la méthode d'anonymisation sont illustrés par un cas concret de mise à disposition d'un corpus d'interactions en ligne pour une communauté de recherche : le corpus Simuligne.

4.0. Cas d'étude

Simuligne est le nom donné à une formation qui a eu lieu dans le cadre du projet ICOGAD[3] sur une plateforme qui n'est plus disponible aujourd'hui. Il s'agit d'une *simulation globale* [YAICHE 96] proposée en ligne pour la pratique des langues en situation réelle de communication, textuelle et asynchrone. Le scénario invite les apprenants à une production collaborative dans la langue cible. Cette collaboration implique de nombreuses interactions pour organiser, négocier, décider et finalement produire ensemble [REFFAY et al. 02]. Le corpus concerne 40 adultes anglophones en formation continue, 10 natifs francophones étudiants en FLE et 4 tuteurs (1 par groupe). La formation s'est déroulée sur 10 semaines et a produit plus de 12 000 interventions à travers les outils de forum, courriel et clavardage.

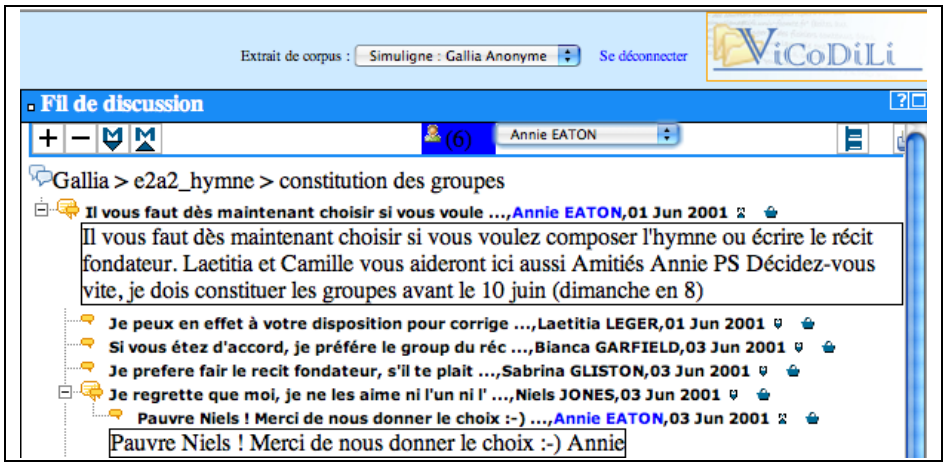

Figure 3. *Visualisation ViCoDiLi du forum "e2a2_hymne" du groupe Gallia*

L'accès au corpus Simuligne a été restauré par l'outil ViCoDiLi (Figure 3) dans le cadre du projet ODIL. ViCoDiLi permet de visualiser l'ensemble des échanges du

[3] ICOGAD : Interaction Cognitives dans les Groupes en formation A Distance : projet (2001-2003) du programme « Cognitique 2000 », regroupant 3 laboratoires, piloté par T. Chanier.

corpus Simuligne à partir d'une structuration XML des contributions de la formation. L'identité des intervenants et le contenu des échanges sont fournis in extenso. L'objectif est de permettre au propriétaire du corpus de le rendre accessible à d'autres équipes sans communiquer l'identité des acteurs de la formation. L'anonymiseur présenté ici se limite au traitement des données textuelles contenues dans des fichiers aux formats non propriétaires. Les données multimédia (photo, audio ou vidéo) posent d'autres difficultés techniques et sémantiques non traitées ici.

4.1. Démarche d'anonymisation

L'anonymisation d'un corpus s'articule autour de trois questionnements : qui effectue la tâche, à quoi correspond-elle, et comment l'effectuer techniquement ?

L'exploitant d'un module d'anonymisation est le détenteur du corpus. Il est garant de l'anonymisation vis à vis des participants de la FEL et prépare le corpus en fonction des attentes et des besoins des destinataires. La modification des informations personnelles est en effet guidée par la connaissance, ou non, des intentions et des besoins des analystes. Si le type d'analyse prévue n'est pas connu, le corpus est mis à disposition de manière relativement aveugle sous la seule responsabilité du propriétaire. Par contre, si l'analyste est assez précis dans sa démarche d'observation et de compréhension des échanges, le corpus peut être anonymisé en tenant compte de ces orientations.

Le processus d'anonymisation gagne donc à être mené à la fois par le propriétaire et par l'analyste utilisateur. Sans chercher à identifier les personnes, ce dernier peut demander des précisions sur les caractéristiques de leur nom (exemple du prénom Bianca qui provoque une discussion très contextualisée Figure 4). On constate alors que chaque anonymisation est particulière et qu'il y a nécessité pour le détenteur à disposer de moyens d'assistance à cette tâche.

Les questions qui se posent à l'anonymisateur sont les suivantes : quels données existent dans le corpus d'origine, quelles données ne peuvent perdurer, quelles données sont attendues par l'analyste dans le corpus produit ? Deux approches complémentaires permettent de répondre. La première s'appuie sur les modèles de corpus disponibles, la seconde sur la réalité des interventions constituant le corpus.

Conceptuellement, les modèles de situations d'apprentissage existants [TEUTSCH et al. 04] distinguent les données liées à l'identité de la personne (nom, prénom, surnom, photo), les données liées à ses caractéristiques sociales (sexe, âge, localisation géographique, langue maternelle) et les données liées à son profil d'apprentissage (niveau et compétences en langue cible, parcours et trajet de formation, situation courante). Parmi ces données, seul le premier lot est sujet à modification, les autres peuvent se révéler indispensables pour certaines analyses.

En ce qui concerne l'identité de la personne, nous distinguons dans le corpus d'une part les composants d'identification manipulés par la plateforme de formation et d'autre part les éléments de désignation personnelle utilisés dans les échanges

eux-mêmes. Les premières données font référence (directement ou indirectement) aux acteurs : nom, prénom, compte d'accès, identifiant, adresse IP, etc. Elles apparaissent sous la forme d'une chaîne de caractères immuable facile à rechercher et à remplacer de façon automatique. C'est le cas du nom de l'auteur d'un message posté dans un forum, de la signature automatique d'un courriel ou des initiales qui précèdent un tour de parole dans un clavardage. Les secondes données se trouvent au milieu des textes produits par les acteurs eux-mêmes : signature, interpellation, réponse ou référence à un ou plusieurs des intervenants (Figure 3). Le traitement est dans ce cas plus délicat puisque les noms des personnes citées à l'intérieur des interactions sont sujets à des variations morphologiques parfois importantes. Ainsi, dans le cas de formations collaboratives à distance et en ligne, les apprenants utilisent souvent des surnoms et diminutifs lorsqu'ils s'interpellent ou qu'ils signent leur intervention. Ces marques de convivialité sont importantes à connaître pour les analystes. En situation d'apprentissage des langues, les noms ou prénoms constituent des repères socio-culturels ou sont teintés de significations abordées dans l'interaction (Figure 4).

Bianca GARFIELD>>en la Colombie mes amies m'appelent contradiction parce que Bianca signifie blanc et je suis un peu marron, tres marron
Camille GUILLON>>Sniff, je n'ai pas de petit nom. Vous pourriez peut-être m'en trouver un ?

Figure 4. *Jeu de mot sur un prénom dans un clavardage de Simuligne*

La recherche et le traitement des interpellations noyées au milieu des interactions, soulignent que l'anonymisation dépasse le cadre purement technique pour aborder un point de vue plus sémantique. L'anonymisation n'est finalement pas simple à modéliser.

Une fois définies ces marques d'identité, il s'agit de définir les techniques permettant de les retrouver dans le corpus et de les y traiter. Différentes stratégies d'anonymisation sont alors envisageables :

- Modifier les patronymes en d'autres noms et prénoms, comme par exemple donner un pseudonyme, conserver le prénom mais effacer le nom, modifier harmonieusement les nom et prénom[4], conserver uniquement les initiales, … Ce type d'anonymisation cherche à rendre le corpus accessible tout en maintenant un rôle spécifique aux données d'identités ;

4 Une équipe anglaise ayant effectué une étude sociologique longitudinale de dix ans sur une population d'adolescents par des entretiens réguliers pour analyser leur rapport à l'adulte (Adulthoods) a constaté des difficultés d'une anonymisation manuelle du corpus en termes de maintien de la cohérence des pseudonymes utilisés.
(http://www.lsbu.ac.uk/inventingadulthoods/archiving/anonymisation/index.shtml)

- Transformer les identités en codes liés aux caractéristiques ou au rôle des acteurs du dispositif (exemple : Tuteur, Apprenant1, Apprenant2, …). Ce type d'anonymisation donne un éclairage particulier au corpus et lui impose une certaine lecture en conséquence ;

- Modifier les patronymes et les compléter par les informations caractérisant le profil du participant (localisation ou langue maternelle par exemple).

L'extrait ci-dessous (Figure 5) donne un exemple de combinaison de ces différentes possibilités.

Dans les interactions rapportées, les participants et les sessions sont anonymés et codés de la manière suivante : une lettre pour le statut (E étudiant, A animateur, C coordinateur local, R responsable de session), un numéro d'ordre dans la session, la désignation de la session (P protosession, C canosession et VR Verba Rebus) et le numéro de la phrase. Par ailleurs les noms des participants ont été modifiés.

Ce qui donne par exemple :
17/03/2004 C1, salon rouge, entre Vasco, Portugais [E229C] et Liliana, Argentine en espagnol [E179C]
[Liliana] Vasco de donde sos ???
[Vasco] eu sou de Portugal, estou em aveiro.
[Liliana] …

Figure 5. *Technique d'anonymisation [DEGACHE 06 p.63]*

4.2. Processus d'anonymisation pour Simuligne dans ViCoDiLi

Cette section présente le processus d'anonymisation utilisé par ViCoDiLi pour le corpus Simuligne. Ce processus s'appuie sur la définition des données d'identité à protéger, sur une table de correspondance attribuant un masque de remplacement à chaque donnée d'identité, et sur un traitement en plusieurs phases du corpus.

Le processus d'anonymisation proposé s'articule en deux phases s'appuyant chacune sur une base d'informations. En aval, le corpus anonymisé est produit à partir d'un ensemble de correspondances entre les formes d'origine des identités et les formes de remplacement de ces identités. En amont, le propriétaire du corpus s'appuie sur l'ensemble des informations individuelles dont il dispose pour préparer la table de correspondance en tenant compte de ce qu'il connaît des acteurs, du contenu des échanges, et des besoins de l'analyse.

Ce processus permet au propriétaire de conserver le profil complet des acteurs afin de toujours pouvoir recréer le lien vers certaines caractéristiques, de définir la logique des équivalences entre les éléments réels d'identité et leurs pseudonymes, et de définir au besoin des équivalences complémentaires à partir d'expressions

repérées dans les échanges. Le principe de conversion entraîne le masquage des noms, prénoms, surnoms et autres diminutifs signalés par l'opérateur en pseudonymes. Le terme « pseudonyme » désigne la forme modifiée de l'identité initiale.

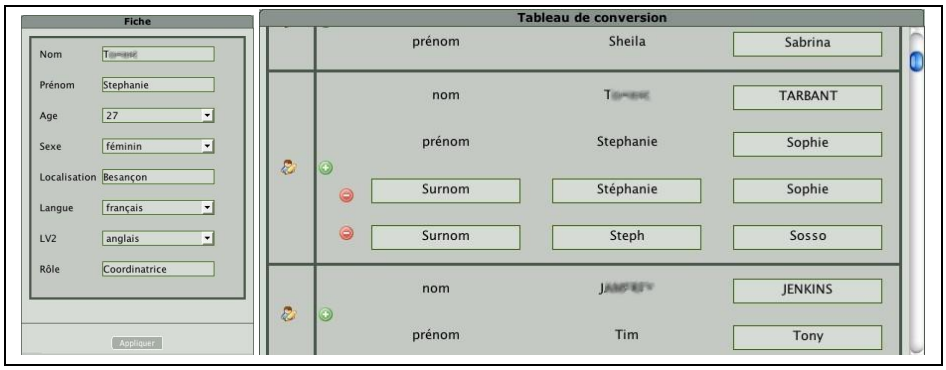

Figure 6. *Interface de l'anonymiseur : Fiches & Table de Conversion* [5]

L'utilisateur anonymisateur dispose d'une interface de description des correspondances entre identité d'origine et identité modifiée (Figure 6). Le système présente dans un premier temps la liste des intervenants connus du corpus (liste issue de la plateforme de formation en ligne via un fichier XML). L'utilisateur peut compléter cette liste avec les surnoms, diminutifs et formes altérées présentes dans le corpus, permettant ainsi de désigner chacune de ces personnes à partir de la déclinaison de leur identité, ce qui impose de bien connaître le corpus ! Le système signale à l'opérateur les doublons repérés dans la table de correspondance. Ces doublons peuvent correspondre à de réels homonymes d'origine, il est alors recommandé de leur attribuer le même masque afin de maintenir l'ambiguïté d'origine. Les doublons peuvent aussi être fortuits (deux masques identiques pour des données différentes dans le corpus d'origine), le système présente alors les différentes formes utilisées pour que l'opérateur vérifie ses déclarations.

La table de correspondance entre identités d'origine et pseudonymes est accompagnée d'un ensemble de fiches (Figure 6). Chaque fiche contient les caractéristiques réelles de l'acteur de la formation : identité complète, âge, localisation, … Ces informations, uniquement connues du propriétaire, ont pour but de l'aider à choisir un pseudonyme en tenant compte, si besoin, de certaines caractéristiques du profil de l'acteur (rôle, sexe, langue, culture, etc.).

Le processus d'anonymisation en lui-même consiste à appliquer les modifications dans le corpus d'origine (fichier XML) en deux phases : modification des identifiants des acteurs dans les en-têtes des interventions puis modification dans le

[5] Les patronymes réels ont été volontairement voilés pour la publication.

corps des interventions. Ce processus transforme le contenu sans altérer la structure XML, ce qui permet à ViCoDiLi de visualiser également le nouveau corpus.

4.3. Évaluation du processus d'anonymisation

Pour estimer la qualité d'une anonymisation, trois points de vue sont nécessaires : celui de l'acteur de l'expérimentation qui est de fait le mieux placé pour juger du caractère anonyme des données ; celui de l'utilisateur anonymisateur (détenteur du corpus) qui pourra juger de l'utilisabilité, de l'utilité et de l'efficacité de l'outil d'anonymisation ; et celui du chercheur étranger à la formation d'origine qui, lui, pourra juger de la lisibilité du corpus ainsi obtenu.

Une partie des forums du corpus Simuligne a été rassemblée pour être anonymisée à titre expérimental par l'outil proposé dans cet article. Un premier bilan de l'outil et de la démarche d'anonymisation est présenté.

Les informations demandées par l'outil sont simples et l'interface est claire. La liste des acteurs, et leur fiche détaillée indiquant leur réelle identité est disponible au moment de construire la table de correspondance ; ce qui facilite grandement le choix des pseudonymes. Le fait de pouvoir sauvegarder et réutiliser la liste de correspondances ainsi que les fiches détaillées est essentiel. Le prototype est fonctionnel et le processus est assez rapide. Les vérifications et traitements effectués par l'outil informent l'utilisateur des risques d'une transformation non réversible.

Les limites du prototype concernent principalement sont architecture logicielle. L'outil ayant été développé comme une application serveur pour faciliter sa conception et son évolution rapide, son exploitation implique la transmission du corpus original et de la table de conversion sur le réseau, ce qui est relativement incompatible avec le caractère confidentiel du processus. La version finalisée de l'outil sera développée comme une application autonome sur le poste de l'utilisateur.

Enfin, la transformation des données est une action jugée délicate par le chercheur détenteur du corpus qui craint d'y perdre des informations utiles à la compréhension du contexte ou à la réalisation des analyses. Ceci renforce l'intérêt pour le détenteur de conserver le corpus d'origine et d'affiner progressivement la table de conversion grâce aux éventuelles incohérences détectées par l'analyste.

Une fois le corpus entièrement anonymisé, les acteurs de la formation seront invités à accéder aux données pour qu'ils puissent vérifier qu'il n'est pas possible de remonter à leur identité dans la nouvelle forme du corpus.

5. Conclusion

L'accès aux corpus d'interactions intéresse la communauté EIAH pour ses études sur la caractérisation de ces échanges et sur l'influence du contexte de formation médiatisée sur l'apprentissage.

D'autres communautés de recherche s'appuient sur des corpus de données normalisés pour croiser leurs recherches, protocoles et résultats. En traitement automatique du langage (TAL), par exemple, l'expérience de la Freebank [SALMON-ALT et al. 04] propose un modèle générique basé sur la TEI[6] et met à disposition des corpus homogènes (e.g. transcription anonymisée de dialogues téléphoniques) pour permettre à différentes équipes de comparer les performances de leurs outils de TAL sur ces corpus. Dans ce domaine, et dans celui des formations en sciences exactes, les informations personnelles et socioculturelles représentent sans doute peu d'intérêt pour les analyses. L'anonymisation peut se faire aisément en affectant un code unique à chaque acteur, sans affecter la qualité des analyses.

Dans des domaines d'apprentissage s'appuyant sur des interactions humaines produites en texte libre, la problématique de l'anonymisation est plus complexe. Ainsi, dans le cas de corpus d'interactions en contexte d'apprentissage linguistique et collaboratif, il est nécessaire de définir des protocoles communs d'anonymisation en vue de leur diffusion. La question se pose dans le cas d'extraction de données pour une visualisation particulière comme pour ViCoDiLi, elle se pose également pour l'accès aux échanges directement sur la plateforme d'origine (exemple des projets Learnet et Galanet où l'accès est limité pour protéger les identités des acteurs).

Le travail sur l'anonymisation a abouti à un questionnement sur le contexte personnel des participants à une FEL et aux rôles que peuvent jouer ces données lors de l'analyse du corpus. On retient d'une part que l'anonymisation de corpus est incontournable, imposée par le respect des personnes, et d'autre part que la compréhension et l'analyse du corpus nécessitent de connaître le contexte de production des échanges en ligne. L'anonymisation d'un corpus est finalement liée à ce qu'y cherche l'analyste : par exemple une anonymisation très codée si on s'intéresse aux rôles de chacun, ou une anonymisation plus contextualisée si on s'intéresse aux inter-relations humaines. La ligne de partage peut alors se réduire au traitement des patronymes et surnoms de la personne. L'anonymisation a donc pour caractéristique de préserver la perception du contexte en protégeant l'identité de la personne physique.

5.0. Bibliographie

[BOISSIÈRE & SCHADLE 06] Boissière, P., Schadle, I., « Proposition d'un cadre méthodologique d'évaluation des systèmes d'assistance à la saisie de textes : Applications aux systèmes Sibylle et VITIPI ». Actes de *Handicap 2006*, Paris, 2006, p. 149-154.

[CHANIER 01] Chanier, T., « Créer des communautés d'apprentissage à distance ». *Les dossiers de l´Ingénierie Educative*, n° 36, Centre National de Documentation Pédagogique, Montrouge, France, 2001, p. 56-59.

6 TEI Text Encoding Initiative. http://www.tei-c.org

[CHOQUET et al. 05] Choquet, C., Luengo, V., Yacef, K., (Eds.), Proceedings of "Usage Analysis in Learning Systems" workshop, held in conjunction with the 12th *Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 2005, 122 p.

[DEGACHE 06] Degache, C., « Aspects du contrat didactique dans une formation plurilingue ouverte et à distance ». Dans Mangenot, F. & Dejean-Thircuir, C. (coord.) Les échanges en ligne dans l'apprentissage et la formation, numéro thématique *Le Français Dans Le Monde*, Recherche & Applications n° 40, juillet 2006, p. 58-74.

[GLIGOR-CALIN 06] Gligor-Calin, L., « Aide à la compréhension du comportement de l'utilisateur par la transformation des traces collectées ». Actes des *Premières Rencontres Jeunes Chercheurs en EIAH*, Evry, France, mai 2006, p. 115-122.

[HENRI & CHARLIER 05] Henri, F., Charlier, B., « L'analyse des forums de discussion pour sortir de l'impasse ». Actes du Symposium, formation et nouveaux instruments de communication, coordonné par Bruillard, Baron et Sidir, Amiens, 20-21-22 janvier 2005. *http://www.dep.u-picardie.fr/sidir/articles/henri_charlier.htm*

[MALLET-POUJOL 04] Mallet-Poujol, N., « Protection de la vie privée et des données personnelles ». *Legamedia*, France, Février 2004. http://www.educnet.education.fr/chrgt/guideViePrivee.pdf

[REFFAY et al. 02] Reffay, C., Chanier, T., Nicolet, J., « Produire ensemble pour apprendre : expérience d'une simulation globale en ligne ». Actes du *Colloque national Apprendre avec l'Ordinateur*, Bordeaux, France, January 2002, p. 24.

[REFFAY et al. 07] Reffay, C., Noras, M., Chanier, T., Betbeder, M.-L., « Contribution à la structuration de corpus de formations en ligne pour un meilleur partage en recherche ». Conférence EPAL : Echanger Pour Apprendre en Ligne, Grenoble, juin 2007.

[SALMON-ALT et al. 04] Salmon-Alt, S., Romary, L., Pierrel, J.-M., « Un modèle générique d'organisation des corpus en ligne ». *Traitement automatique du langage (Tal)*, n° 45/3, 2004, p. 145-169.

[TEUTSCH et al. 04] Teutsch, P., Bourdet, J.-F., Gueye, O., « Perception de la situation d'apprentissage par le tuteur en ligne ». Actes de *TICE'2004*, Compiègne, France, octobre 2004, p. 59-66.

[TEUTSCH et al. 07] Teutsch, P., Dejean-Thicuir, C., Bangou, F., « ViCoDiLi, un outil pour visualiser des contenus de discussion en ligne ». Conférence EPAL : Echanger Pour Apprendre en Ligne, Grenoble, juin 2007.

[YAICHE 96] Yaiche, F. *Les simulations globales mode d'emploi*. Hachette, Paris, 1996, 129 p.