

# Anonymizing and sharing corpora of online training courses

## Masking the identity while preserving the context for written exchange analysis

Philippe Teutsch<sup>1</sup>, Frédéric Piat<sup>1</sup>, Christophe Reffay<sup>2</sup>

<sup>1</sup>LIUM, Université du Maine, Avenue Messiaen, F-72085 le Mans cedex 9, France

<sup>2</sup>LIFC, Université de Franche Comté, 16 route de Gray, F-25030 Besançon cedex, France

Email : {Philippe.Teutsch, Frederic.Piat}@univ-lemans.fr

Christophe.Reffay@univ-fcomte.fr

**Abstract:** The scientific community working on Technology Enhanced Learning and more specifically on Computer Supported Collaboration Learning (CSCL) is interested in making available some well-described corpora of online courses. This will permit the analysis of the same set of human interaction from different points of view and with different methodologies coming from various disciplines. This work addresses the issue of the necessary (for ethical and legal concerns) anonymization process to be applied on a corpus to share it with a larger research community. This contribution looks for the right borderline permitting to save the social and cultural context and hide efficiently the identity of the actor in order to protect his privacy. The principles and tools presented in this article are applied to a corpus of textual interaction in language learning.

## 1. Introduction

The work presented here deals with the study of computer supported human interaction at a distance, in a learning context. The learning situation being mediated by a Learning Management System (LMS), we know we are able to systematically collect the logs of activities for all participants (Choquet *et al.*, 2005) and then compute a representation of these logs. The data we are dealing with are textual messages produced and exchanged by actors of an online learning situation (students, teachers and tutors), by means of computer mediated communication tools like mail, forum and chat.

The integration of different communication tools in a LMS can modify all parts of human interaction: conditions, nature and even the issue of interaction in learning at a distance where social interaction is essential (Chanier, 2001). Modelization and analysis of such interactions should help us in the design or the evolution of online learning environments and scenarios. Such analyses are valuable only when interaction is produced in an authentic learning situation, where real actors are involved in a concrete, cultural and social interaction.

Much research is being conducted on online distance learning situations, but their scientific rigor should be improved by replicability that is rarely possible (Henri & Charlier, 2005). In order for the scientific debate to take place, we need more than the results published in scientific articles: data and hidden corpora should be available to other researchers to make possible concurrent analysis on the same data. Conversely, a proposed analysis process or method should be applied to different corpora in order to widen its validity and robustness.

The ODIL<sup>1</sup> french project established the need for corpus reuse and designed a visualization tool for online discussion contents (ViCoDiLi, Teutsch *et al.*, 2008), independent from the LMS used by the actors. The Mulce<sup>2</sup> project aims at specifying a structure for a corpus in order to make it reusable and sharable (Reffay *et al.*, 2008). Both research actions have shown that, among the various conditions that allow sharing and publication of corpora, protection of actors is an essential first step.

This paper focuses on the critical aspect in sharing real-life online learning corpora: i.e. the protection of actors. We first present the issue of anonymization and the stakes involved. The principles and methods of the anonymization process are then presented and applied to the “Simuligne” corpus, a data collection from an online language learning situation. Finally, we present the corresponding tool designed and developed for this project.

## 2. Protecting actors' identities: an essential stake

The recorded interactions being authentic, our scientific community has to take care when opening them to a wider audience like other researchers or people that may process the data for uncontrolled purposes. For ethical reasons, access to the actor's contributions is available only if the actor's identity cannot be recognised. In order to illustrate the preprocessing needed before sharing such a corpus, we translated both of the following French citations:

---

<sup>1</sup> ODIL : Outils et Didactique pour l'analyse des Interactions en Ligne, ACI project (2004-2007)

<sup>2</sup> Mulce : Multimodal Learning Corpus Exchange, ANR project (2007-2009), <http://mulce.univ-fcomte.fr>

*We have collected texts (e-mails) written by a forty year-old person... This corpus contains 205 messages of his personal mailbox. ... All names have been changed in order to make this corpus anonymous and usable by others. We then removed all headers and signatures from these e-mails.*

**Figure 1.** Anonymization need (Boissière & Schadle, 2006, p.4)

*Taking into account the danger of increasing digitization of personal data, the first concern of any file administrator should be to anonymize the data as soon as possible. In other words, identity or identifying elements should never be delivered when they are not necessary for the ultimate goal of the process.*

**Figure 2.** Anonymization necessity (Mallet-Poujol, 2004, p. 28)

Consequently, in order to guarantee the protection of the actors of learning sessions, data anonymization is required before any corpus exchange.

### **3. Masking identity while preserving the interaction context**

Even if anonymization is mandatory (by law) and imposed by ethical concerns, the last sentence of the second citation shows that its application can be considered at different levels depending on the information necessary for the aim of the process, that is in our case the corpus analysis. In the end, the question boils down to: what is the information needed to interpret and analyse the corpus?

Communication tools in LMS allow increasingly rich exchanges, producing a wide variety of interaction data. The corpus' holder who wants to share some pieces as illustration, keeping actors' anonymity, can process those pieces manually by using techniques like face blurring or voice distortion. Such a process can be applied to the whole corpus when downstream analysis needs no distinction between actors (e.g.: pure linguistics, some behaviour or discourse analysis). In this case, all identity elements can be simply hidden or removed. However, for a lot of other research, such drastic techniques may impede or altogether prevent an efficient analysis. To understand the situation and carry out their analyses, other studies on interaction need to distinguish the various actors (i.e. message authors). This is the case for example in the following contexts: language learning (where language is both considered as learning object and means of expression), collaborative learning (where human relationship and conversation are essential to succeed). Consequently, to distort or hide identification elements may reduce the quality or impact of the analysis.

We have to characterize the components defining the "persona" of an actor in an online learning situation. Among these components, we should define the borderline between: the identity of the individual that must be protected on one hand, and the personal information context needed by the analysis on the other hand. We have to take into account the needs expressed by analysts and the necessity of keeping the actors' anonymity. To reach this goal, we first give the definition of personal data and then propose a transformation process that will not distort the corpus.

### **4. Case study of identification and anonymization principles**

The data concerning actors' personality in an online learning situation can be separated in two categories. In the first one, we have the identity itself: i.e. the information enabling, directly or indirectly, the identification of the physical individual (lastname, surface mail or e-mail address, date and place of birth, photo, voice, or registry number of a vehicle). In the second one, we find the various information elements that describe the actor's skills (mother tongue, personal interests, computer or communication skills, experience, cultural environment, etc.). If elements of the first category must be hidden to protect the actor's anonymity, those of the second one should be kept for the analysis needs.

We suggest that this particular anonymization process should be driven by the needs of subsequent analyses (that will be made by a third party), but it must be performed by the holder of the corpus. The holder "prepares" the data that will be analysed by a third party. This principle and the resulting method are illustrated on an ecological corpus that has to be shared by a research community: the "Simuligne" corpus.

#### **4.1 Case study**

Simuligne is the name of a training course given within the ICOGAD<sup>3</sup> project on a platform that has become today obsolete and unavailable. It consists of an on-line global simulation to practice languages in real-life communication settings, textual and asynchronous. The scenario prompts the learners to collaborate to produce

<sup>3</sup> ICOGAD : Interaction Cognitives dans les Groupes en formation A Distance, research program "Cognitique" (2001-2003)

written messages in the targeted language. Numerous interactions to organise, negotiate, decide and finally produce together are necessary to produce a genuine collective work.

The corpus contains data from 40 English-speaking adults following evening classes, 10 French native speakers studying "French as a foreign language" and 4 tutors (1 per group). The training course lasted for over 10 weeks and produced over 12,000 messages through forums, emails and chats.

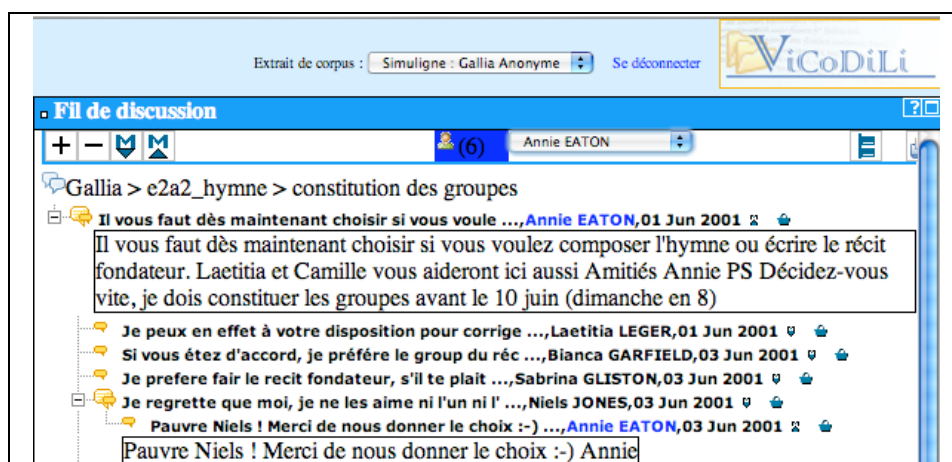


Figure 3. Display by ViCoDiLi of the forum "e2a2\_hymne"

Access to the Simuligne corpus was restored with the "ViCoDili" tool developed in the context of the "Odil" project (see Figure.3). This tool allows the user to visualise all the communications contained in the Simuligne corpus, using an XML structure of all contributions in the training course. The participants' identities and the messages' contents are fully disclosed. The goal is to allow the corpus owner to share it with other research teams without disclosing the participants' identities. The anonymizing tool presented here (called "anonymizer" in what follows) is limited to processing text data. Multimedia data such as photos, audio and video require different processing techniques constituting whole research areas in themselves, and will not be treated here.

#### 4.2 The anonymization approach

Anonymizing a corpus requires answers to three general questions: who does it, what to do exactly, and how to do it?

The user of the anonymization software is the corpus owner. He is responsible for guaranteeing the anonymity of the participants to the on-line training course, and prepares the corpus according to the needs and expectations of end-users of the anonymized corpus. Modification of personal information is guided by the knowledge, or lack thereof, of the intentions and needs of downstream analysts. Thus the anonymization process can be improved both by the owner and the analyst of the corpus. Without trying to identify individuals, the latter can ask for precisions regarding name characteristics (e.g., the name "Bianca" giving rise to a very specific exchange, Figure 4). We then realize that each anonymization has its own constraints and that the owner needs tools to optimize the process within these constraints. He has to answer the following questions: which personal and identifying data are present in the original corpus, which ones cannot be kept, and which ones are needed by the analyst in the final corpus? Two complementary approaches can yield answers. The first one relies on the corpus structure, the second one on the actual original content of the interventions making up the corpus.

*Bianca GARFIELD>> In Columbia my friends call me "contradiction" because my name means "white" but I'm rather dark-skinned; actually, really dark*

Figure 4. Explanations regarding a first name in a Simuligne chat

On a theoretical basis, the current models of learning contexts distinguish three kinds of data (Teutsch *et al.*, 2004): those relating to the individual's identity (first/last name, picture,...), those relating to social characteristics (gender, age, mother tongue,...) and those relating to his/her learner profile (target-language proficiency level and skills, academic profile or history, current situation, ...). Among these data, only the first kind has to be systematically modified, while both of the others may need to remain unchanged for subsequent analyses.

Regarding the individual's identity, we distinguish the identifying data handled by the training platform on one hand, and on the other hand those used in the messages themselves.

The former refer directly or indirectly to the actors: first/last name, login, id, IP address and so on... All appear as a uniquely-defined character string, easy to automatically search for and replace. This is the case of the name

of the author of a message posted in a forum, of the automatic signature of an email, or of the initials preceding the message in a chat.

The latter can be found in the midst of texts produced by the actors themselves: signature, calling, answer or reference to one or several other actors (for example Laetitia, Camille, Niels and Annie in Figure 3). Processing of this information is in this case much more complex, given that the names cited inside the messages can be subject to many, and sometimes very different, morphological variations. Indeed, in the case of collaborative, remote on-line training courses, learners usually use nicknames when signing or calling each other, and it is important to analysts to recognise these. In a language learning context, first and last names can be socially and culturally marked or they can carry a meaning discussed about in the interaction.

The search for and the processing of the callings of other people, spread in the midst of all messages, show that anonymization goes way beyond a purely information-processing technical issue, getting to more semantic issues. Modeling anonymization does not appear to be so straightforward.

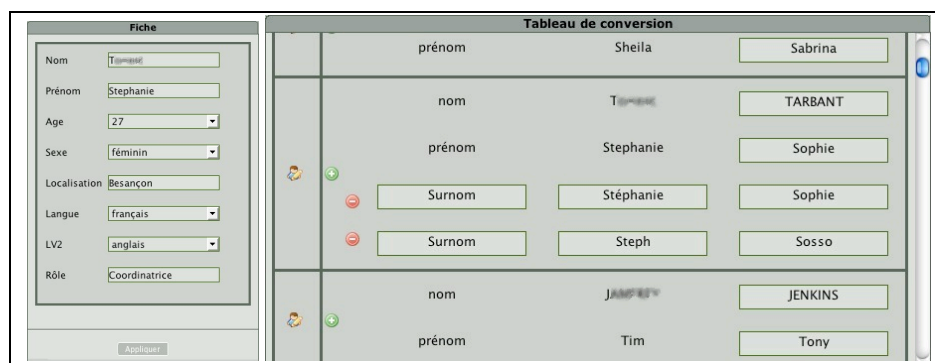
After all identity markers have been defined, we have to choose which techniques to use to find and process them in the corpus. We can then imagine several anonymization strategies:

- Change names into other first and last names, for instance by attributing a masking name, by keeping first while deleting the last names, by harmoniously modifying them, by keeping only initials,...
- Transform the identities into codes directly linked to the characteristics or to the role of the actor (e.g., Tutor, Learner#1, Learner#2,...). This kind of anonymization focuses on a particular aspect of the corpus and pushes the reader towards a particular interpretation.
- Modify the names and complete them with profile information (mother tongue for instance)

### 4.3 Anonymization process for Simuligne in ViCoDiLi

This section presents the anonymization process used by ViCoDiLi for the Simuligne Corpus. This processing of the corpus is multi-phased and relies on the definition of the identifying data to protect, and on a conversion table associating a mask substituted to each identifying data. Downstream, the anonymized corpus is produced from a list of associations between the original character string of the identity and their replacing forms. Upstream, to prepare the conversion table, the corpus owner relies on all the individual data available while taking into account his knowledge of the actors, of the content of interactions, and of the analysis requirements.

This process allows the owner to keep the complete profile of the actors which can be useful for three tasks: restoring at any time the link to some of the characteristics, defining the logic behind the equivalence between the real-life identifying data and the masking names, and if needed defining further equivalence for expressions spotted in the exchange. The conversion principle implies the replacement of first and last names, pseudos or other nicknames spotted by the operator with new appropriate masking names.



**Figure 5.** Screenshot of the anonymizer interface: Forms and conversion table.

Figure 5 shows the interface available to the operator in charge of the anonymization, which describes the association between original and modified identity. At first the system displays the list of the actors known from the corpus (the list is extracted from the on-line training platform through an XML file). The user can complete this list, adding nicknames and altered forms found in the corpus.

The system warns the user when doubles appear in the conversion table. These doubles can refer to actual original homonyms, it is then recommended to substitute to their name the same masking name so as to maintain the original ambiguity. The doubles can also appear by accident (two identical masking names associated to different original data in the original corpus), in which case the system displays the different forms used so that the operator can check his choices of masking names.

A set of forms comes along with the conversion table between original identities and masking names. Each form contains the real characteristics of the actor of the training course: complete identity, age, location... This information, only known from the owner, can be useful to help him choose for the actor a masking name that

will take into account some of the characteristics of his profile such as his role, gender, language, culture and so on...

The anonymization process in itself consists of modifying the original corpus (XML file) in two phases: modifying the actors' identifiers first in the prompts before their messages, then inside the body of all messages. This process alters the XML file's content while preserving its structure, so that ViCoDiLi can also display the new corpus.

#### **4.4 Evaluation of the anonymization process**

To evaluate the quality of an anonymization, three points of view have to be taken into account: that of the actor of the training course who is the best judge of the anonymous nature of the transformed data; that of the user (the corpus owner) who can judge the usefulness, the usability and the efficiency of the anonymization tool; and that of the researcher foreign to the original on-line training course who will be able to judge the readability of the final corpus.

To test our tool, we gathered part of the forums of the Simuligne corpus. We present here a first assessment of the tool's performance and of the anonymization process.

The information required by the tool is simple and the interface is clear. The list of actors and their detailed forms indicating their real identities are available to the user when he builds the conversion table, to make easier the choice of a masking name. It is essential to be able to save and reuse the table as well as the detailed individual forms. The prototype functions well and the process is carried out rather quickly. Checks and processes carried out by the software warn the user regarding any risk of an irreversible transtraining course.

The main shortcoming of the tool is mostly its software architecture: the tool has been developed as a server application to simplify its design and allow frequent upgrades, so its use implies transmitting the original corpus and conversion table over the network, which is not compatible with the confidentiality of the process.

Finally, transforming the data is a rather critical action for the corpus owner who could be concerned about losing any information, which could impede context understanding or further analysis. Hence the owner is advised to save the original corpus and refine step by step the conversion table to correct a possible incoherence (e.g. between masking names and cultural profile) detected by the end-user.

### **5. Conclusion**

The research community involved in the development of computer environments for learning is interested in accessing interaction corpora, to be able to characterise these exchanges and to understand how the context of on-line training course influences human learning.

Other research communities rely on standard format corpora of data to cross-check their research, protocols and results. In the field of language processing for instance, the Freebank experiment proposes a Text Encoding Initiative based generic model to grant free access to homogenous corpora (e.g. anonymized transcriptions of telephone dialogs) to allow different teams to compare the performances of their tools. In this domain, and that of hard-science learning, personal and socio-cultural information is probably of little interest to analysts. Anonymization can easily be done by attributing a unique code to each actor, while preserving the quality of analyses.

The anonymization problem gets more complex when applied to learning domains relying on free text-based human interaction. Therefore in the context of collaborative language learning, we need to define standard anonymization protocols to share the corpora. The question remains open in the case of data extraction for a particular display method as for ViCoDiLi, or in the case of direct access to exchanges on the original platform.

Working on anonymization has led to questions regarding the personal profile of the participants to an on-line training course and regarding the role these personal data can play in corpus analysis. To analyse and fully understand a corpus, we need to know the context in which on-line exchanges are produced; but corpus anonymization is necessary to respect privacy. Thus conducting this process is essentially about finding the borderline between protecting the privacy of physical persons and preserving the context.

### **6. Bibliographie**

- Boissière, P., Schadle, I. (2006). Proposition d'un cadre méthodologique d'évaluation des systèmes d'assistance à la saisie de textes : Applications aux systèmes Sibylle et VITIPI. *Handicap 2006*, Paris, p. 149-154.
- Chanier, T. (2001). Créer des communautés d'apprentissage à distance. *Les dossiers de l'Ingénierie Éducative*, n° 36, Centre National de Documentation Pédagogique, Montrouge, France, p. 56-59.
- Choquet, C., Luengo, V., Yacef, K., (Eds., 2005). Proceedings of "Usage Analysis in Learning Systems" workshop, held in conjunction with the 12th Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, 122 p.

- Henri, F., Charlier, B. (2005). L'analyse des forums de discussion pour sortir de l'impasse. Symposium Formation et nouveaux instruments de communication, coordonné par Bruillard, Baron et Sidir, Amiens. [http://www.dep.u-picardie.fr/sidir/articles/henri\\_charlier.htm](http://www.dep.u-picardie.fr/sidir/articles/henri_charlier.htm)
- Mallet-Poujol, N. (2004). *Protection de la vie privée et des données personnelles*. Legamedia, France. <http://www.educnet.education.fr/chrgt/guideViePrivee.pdf>
- Reffay, C., Chanier Th., Noras, M., Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche, Revue *STICEF*, Volume 15, <http://sticef.org>
- Teutsch Ph., Bourdet JF., Gueye O. (2004). Perception de la situation d'apprentissage par le tuteur en ligne, TICE'2004, Compiègne (France), p. 59-66.
- Teutsch Ph., Bangou F., Dejean-Thircuir Ch. (2008). Faciliter l'accès aux échanges en ligne et leur analyse, le cas de ViCoDiLi, Revue *STICEF*, Volume 15, 2008, <http://sticef.org>