Multimodal learning and teaching corpora exchange

Christophe Reffay, Marie-Laure Betbeder, Thierry Chanier

Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project

Christophe Reffay*

Science Technology Education & Training (STEF)
École Normale Supérieure de Cachan
61 av. président Wilson,
94235 Cachan Cedex, France
E-mail: christophe.reffay@ens-cachan.fr
*Corresponding author

Marie-Laure Betbeder

FEMTO-ST Institute,
Université de Franche-Comté,
16 route de Gray, 25030 Besançon cedex, France
E-mail: marie-laure.betbeder@univ-fcomte.fr

Thierry Chanier

Language Research Laboratory (MSH-LRL),
Université Blaise Pascal,
4 rue Ledru, 63057 Clermont-Ferrand Cedex, France
E-mail: thierry.chanier@univ-bpclermont.fr

Abstract: In order to make replication possible for interaction analysis in online learning, the French project named Mulce (2007-2010) and its team worked on requirements for research data to be shareable. We defined a *learning and teaching corpus* (LETEC) as a package containing the data issued from an online course, the contextual information and metadata, necessary to make these data visible, shareable and reusable. These human, technical and ethical requirements are presented in this paper. We briefly present the structure of a corpus and the repository we developed to share these corpora. Related works are also described and we show how conditions evolved between 2006 and 2011. This leads us to report on how the Mulce project was faced with four particular challenges and to suggest acceptable solutions for computer scientists and researchers in the humanities: both concerned by data sharing in the Technology Enhanced Learning community.

Keywords: E-science; Data sharing; Learning and Teaching Corpus; LETEC; research data repository; publication and replication datasets.

Reference to this paper should be made as follows: Reffay, C., Betberder, M.-L., Chanier, T. (2012) '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"…

Biographical notes: Christophe Reffay is researcher in Computer Science at the Educational Science laboratory STEF of the École Normale Supérieure of Cachan, France. He is involved in the domains of Technology Enhanced Learning and Computer Supported Collaborative Learning. His work is dedicated to collaboration analysis and tool conception to capture, analyse and represent social activity in collaborative learning environments. He conducted Social Network Analysis on communication data issued from the "Simuligne" experiment. Since 2007, he has been involved in the Mulce project (presented in this paper). His current interests are Social Network Analysis, indicators for collaboration in Technology Enhanced Learning environments and research data sharing.

Marie-Laure Betbeder is an Associate Professor of computer science at the FEMTO-ST institute in Besançon, France. For the last ten years, she's been involved in the domains of technology enhanced learning and computer supported collaborative learning. Her early works were focused on the taylorability of computer artifacts as a support for learning environments. Since 2007 she has worked on the Mulce project (presented in this paper). She is now involved in another French national project aiming at proposing an environment for controlled language writing support.

Thierry Chanier is professor at the Université Blaise Pascal, France. His main domain of research over the 20 past years has been in technology enhanced language learning or Computer Assisted Language Learning (CALL). Since 1999, he has been mainly concerned with online distance learning. He studies interactions in multimodal environments, where groups of learners of different countries and languages learn collaboratively. His has been the leader of the Mulce project. He is a member of the editorial board of several CALL journals, and co-creators of open access journals. He has been mandated in 2003 by the French Minister of Education and Research to develop the first open archive in Humanities in France.

1 Introduction

Jim Gray, a renowned computer scientist, presented a new paradigm of scientific research, labelled "E-science" or "data-intensive science". It is characterised as the fourth paradigm in the research cycle after experiment, theory, and simulation.

> "We have to do better at producing tools to support the whole research cycle […]. Today, the tools for capturing data […] are just dreadful. After you have captured the data, you need to curate it before you can start doing any kind of data analysis, and we lack good tools for both data curation and data analysis. Then comes the publication of the results of your research, and the published literature is just the tip of the data iceberg." (Gray, 2007: xvii)

Although Gray took his examples from the fields of biology and environmental sciences, in the TEL community, we are also directly concerned with this paradigm. Data have to be shared. Tools for organising and analysing these data need to be much improved as well as the links between data and the most visible part of our work, namely publications.

First and foremost we must be concerned by the extensiveness of data collection and by the description of the context. Studying collaborative online learning, in order to understand this specific type of situated human learning, to evaluate scenario, or to improve technological environments, requires accessibility to interaction data collected

from various participants in the learning situations. However, interdisciplinary communities involved in this research have not been able to characterise a shareable scientific object according to a comprehensible methodology. On one hand, one finds subsets of data, not contextualised with respect to the pedagogical and technological learning situations. On the other hand, raw data are inextricably tangled in specific software using proprietary formats. A simple collection of students' online interaction data does not represent a scientific object, as Kern and Warshauer (2004) emphasised in the language learning field:

> "Researchers must carefully document the relationships among media choice, language usage, and communicative purpose, but they must also attend to the increasingly blurry line separating linguistic interaction and extra linguistic variables. […] Studies of linguistic interaction will likely need to account for a host of independent variable: the instructor's role as mediator, facilitator, or teacher; cross-cultural differences in communicative purpose and rhetorical structure; institutional convergence or divergence on defining course goals; and the affective responses of students involved in online language learning projects." (Kern, Ware and Warshauer, 2004, p.248).

In this article, we develop a new scientific object created in 2006 by the Mulce project, namely the *LEarning & TEaching Corpus* (LETEC). The LETEC structure has been used to organise data compiled from a variety of collaborative online learning and teaching situations into various sorts of corpora. We also present the Mulce repository which is the location where data can be shared in order to facilitate the comparison of analyses.

Collaborative online learning situations have a number of variables which are difficult to control. These variables make the comparison of scientific results difficult and the replication of a given learning and teaching experience near impossible. For example, applying the same learning design to a new cohort of learners does not imply that phenomena observed within the first cohort will occur in the latter. Replication in ecological contexts being impossible to obtain, we worked to make interaction and production traces issued from Learning Management Systems (LMS), available to the whole research community. This was the Mulce strategy to make these situations comparable and re-analysable.

Since the beginning of the Mulce project, the international situation has changed in various scientific fields with the apparition of mandates for open access to research results, including access to data. Other TEL repositories have appeared, alongside discussions around tools for data analysis, or around data structure appropriate for the encoding of interactions. The question of finding a common framework is set. Challenges faced in the Mulce project, whose audience includes TEL researchers, differ from questions raised among the ITS and CSCL communities. This persuaded us to propose a flexible framework which can improve the quality of TEL research and, at the same time, the creditability (Rourke et al., 2001; King, 2007) of researchers belonging to our heterogeneous community.

After this introduction, the paper is organised into three main sections before a short conclusion. Section 2 presents the achievements of the Mulce project. Section 3 gives an overview of the changes and progress made on the data sharing topic since 2006. The experience of the Mulce project with respect to four important challenges is reported upon in section 4 with suggestions to improve corpora building, deposit and reuse for the TEL community.

2 Achievements of the Mulce project

This section is organised into two subparts. We first present the main questions which, since its outset, the Mulce project was confronted with. The second subpart describes the answers given through the corpus definition and structure and the Mulce repository.

2.1 Main objectives and questions faced by the Mulce project

At the launch of our project in 2006, with sponsorship from the French national research council (ANR), our main objective was to propose an open space for sharing corpora and research data concerning online interactions for the researchers (Chanier and Ciekanski, 2010; Reffay and Betbeder, 2009). This objective forced us to think about the requirements of sharing corpora. They concern human, technical, and ethical aspects, which we detail below.

2.1.1 Structure of a coherent dataset

In order to share data of online training interactions, the dataset should be structured and coherent. It should include all the information needed in order that a researcher who did not participate in the course can understand the situation, i.e.: (1) the context of the educational scenario, (2) the original research questions, (3) the educational and technical context and (4) the interactions which occurred. It appeared essential to begin by proposing a definition and a corpus structure which would be human readable, machine readable and reusable.

2.1.2 Data longevity

In order to share data, collected online interactions should be independent from any platform and stored in an independent formalism. Although it may be easier to replay interactions in their original platform, access to the platform in its specific version is not necessarily long-lasting. It is, thus, important to collect tracks stemming from training courses in a tool agnostic form.

2.1.3 Human readability

The training context information is a crucial source of data. It is essential for the understanding of the situation. The educational context can be described from two viewpoints: through metadata, such as LOM ones, or through the scripting of educational instructions. The scripting formalism is a theme studied in the community, for example in IMS-LD (IMS), or LDL (Ferraris, Martel and Vignollet, 2007). But further information needs to be added such as the description of the technological context, including the platform, the tools and their features.

2.1.4 Machine readability

For the interaction data to be handled, reused and analysed, it must be recorded in an interpretable and interoperable formalism. We needed a format that enabled the description of interactions but that was independent of the medium in which they were produced. This formalism has to be able to describe, for example, chat interaction or

interaction using a conceptual map tool and to connect it to the session place, the actors and the activity described in the educational instructions. In order to infer some semantically valid information from raw data, we need a structured and constrained computable description.

2.1.5 Open data access

To allow research data to be shared, the solution lies in a repository where corpora can be visualised and shared. Currently there exist few studies relative to such repositories. Concerning the exchange protocol, the Open Archive Initiative (OAI-PMH, 2002) is often used in open archives. It allows the metadata descriptions of the records in an archive to be collected and ensures the interoperability between repositories and harvesters. Our repository, however, needed to be part of a larger databank which could ensure permalinks through stable OAI identifiers.

2.1.6 Personal data conformance and access to data

The problem is twofold: 1) rights and ethics for collecting, managing and distributing the data ; 2) license for accessing the data.

Applications collecting information that describe: users, their activities and their interactions, must take care on "sensitive" data. They have to conform to the laws which regulate the collection, treatment and transfer of personal data. According to the OCDE, seven principles relative to the protection of personal data must be tackled by computer systems before being displayed by institutions. These principles are labelled: opinion, purpose, consent, disclosure, access and responsibility.

2.2 Main contribution and results

In this subpart we explain how we dealt with the questions outlined in Section 2.1 in order to build the open-access repository (Mulce-repository, 2012). Further information may be found on our web site which documents the project (Mulce-documentation, 2012).

2.2.1 Learning & Teaching Corpus definition: structure of a coherent dataset.

We define a Learning & Teaching Corpus (LETEC) as a structured entity containing all the elements resulting from an online learning situation, whose context is described by an educational scenario and a research protocol. The core data collection includes all the interaction data, the course participants' productions, and the tracks, resulting from the participants' actions in the learning environment and collected according to the research protocol. In order to be shareable, and to respect participant privacy, these data should be anonymised and a license for their use be provided in the corpus (see next paragraph). A derived analysis can be linked to the set of data actually considered, used or computed for this analysis. An analysis consists of a data annotation/ transcription/ transformation, properly connected to its original data. It can be merged into the corpus itself, in order for other researchers to compare their own results with concurrent analyses or to build their complementary analyses upon these previously shared results.

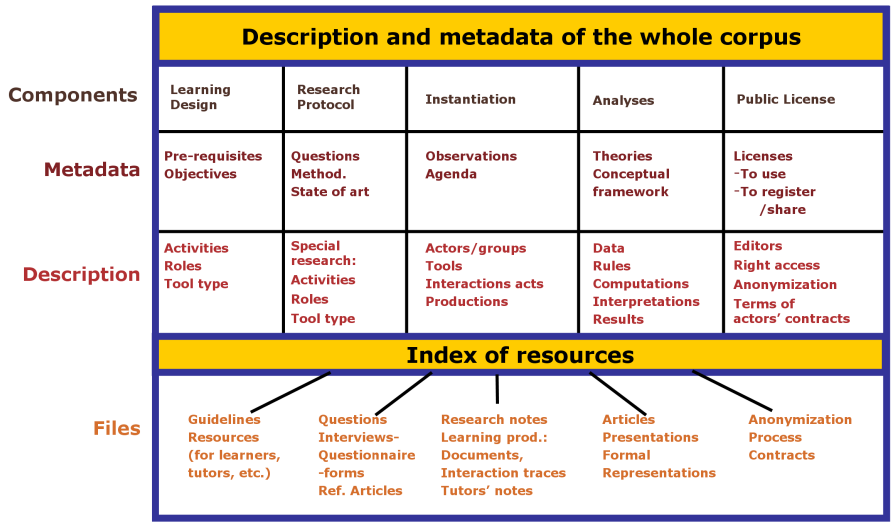
The definition of a LETEC corpus as a whole entity comes from the need for explicit links, between interaction data, context and analyses. This explicit context is crucial for external researchers to interpret the data and to perform their own analyses.

2.2.2 Corpus composition and structure: Data longevity, human readability and personal data conformance

The main components of a LETEC (see Figure 1) are as follows:

- The Instantiation component is the heart of the corpus. It includes all the interaction data, the online course participants' productions, complemented by some system logs, as well as information characterising participants' profiles (see details hereafter);
- Both of the Learning Design and Research Protocol components define the context and answers the question of human readability;
- The License component relates to ethics and access rights (see hereafter);
- The Analysis component contains global or partial analyses of the corpus, as well as possible transcriptions or analytics.

Figure 1 Teaching and learning corpus: the main components in a Content Package (see online version for colours)



The Mulce structure aims to organise the corpus components in a way that enables subparts of components to be linked. For example, a researcher, while reading a chat session in the instantiation component, will be able to read the objectives of the activity in which this session took place. The activity is part of the pedagogical context described in the learning design component.

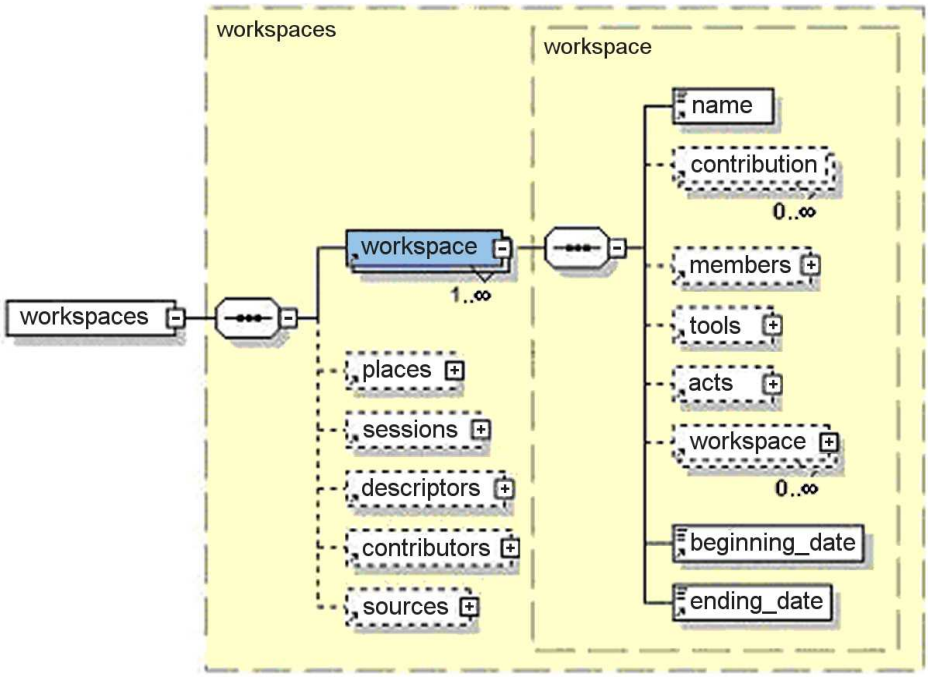
A standard exchange format is also required to download the whole corpus, which may include hundreds of files. We chose the IMS-CP formalism (IMS) as the global container. Its XML formalism allows the description of metadata at different levels, and includes an index pointing to the set of heterogeneous resources. Learning design and research protocol components can use IMS-LD structure as their organisation model. If

they are only described by a simple text, this text can be defined as an *item* element of the IMS-CP *organization* element. However, the Instantiation component is more specific and has to capture and organise the collected tracks of the situation, played out by the participants. To respect participant privacy, these data should be anonymised. We decided to define a specific XML schema for this organisation: the Structured Interaction Data (SID) model (Mce_sid, 2012).

2.2.3 The core component Instantiation: machine readability

The hierarchical structure of the learning stage is captured in the *workspaces* element that contains a sequence of *workspace* elements (see Figure 2).

Figure 2 Extract of the XML Schema: The Workspaces element (see online version for colours)



The *workspaces* element defines lists of:

- *places* that organises the space. Each *place* element defines a reference and description of a virtual or physical place;
- *sessions* that splits the time into meaningful periods. Each *session* element defines a reference and description for a dedicated period of time like a chat session or any other mainly synchronous activity;
- *descriptors* or tags that may be used by researchers in their analyses by associating *interaction* acts to a set of these descriptors in order to categorise or count units for each category;

- *contributors* for the corpus like: researchers, developers, compilers, recorders, inputters, etc.;
- *sources* are identified records. A source element is generally a reference to an audio or video track.

A *workspace* is generally linked to a learning activity defined in the learning design component. It encompasses all the events observed during this activity, in the tool spaces provided for this activity, for a given instantiated group of participants. As shown in Figure 2, a *workspace* description includes: its *members* as references to the participants registered in the learning activity, *starting* and *ending dates*, the *tools* and the interaction tracks or *acts* that occurred using these tools. In order to fit the hierarchical structure of learning and support activities, a *workspace* element can recursively contain one or more *workspace* elements.

The lists of *places*, *sessions*, *descriptors*, *contributors* and *sources* defined in the *workspaces* element can be referenced by *workspace*, *contribution*, or *act* elements. For example, descriptors may list identified categories so that each *act* of the *acts* element list could refer to one or more of these categories. This principle allows the interaction data to be searched or visualised in many different ways independent of the concrete storage organisation in the XML document.

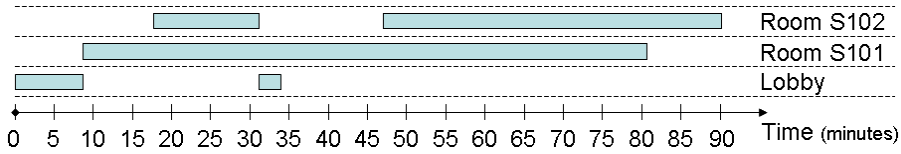
Our specification can describe communication tools and their features quite detailed. For example, some forum tools track when a reader opens a message. Such data can be used for a social networks analysis, considering only opened messages as valid relationships between writers and readers. Such detailed features need to find a place in the specification because they are needed by some subsequent analysis. The corpus builder can specialise or particularise the schema by restricting or augmenting it. These mechanisms are necessary to fit some specific features of tools in a given learning environment and some corresponding specific analysis needs.

Moreover, recursive workspace descriptions enable the corpus compiler to choose the level of details at which he needs to describe the environment. A workspace can be used to describe a complete curriculum, a semester, a module or a single activity. The workspace concept represents the space and time location where we can find interaction with identified tools. This concept, implemented in (Mce_sid, 2012) has the same modularity as the EML learning units (Koper and Tattersall, 2005). These workspaces can be directly explored on the Mulce repository. Once a LETEC is selected, the user can "visit the hierarchical structure of workspaces". Interested readers might like to explore one or both of the following LETEC structures of workspaces that illustrate how this concept can be used in different ways for two typical situations:

- Complex activities structure in a mainly asynchronous platform: *Simuligne* LETEC (Chanier et al., 2007). For sake of server size and workload restrictions, only one group among four is visible through the Mulce repository interface. The workspace of the group named Aquitania is organised as a tree of encapsulated workspaces with six levels (deep) and 19 leaves.
- Sessions in an audio-synchronous platform for two groups, split in parallel subgroups simultaneously in separated virtual rooms: *Copéas* LETEC (Chanier et al., 2008). The corpus represents a course given in 8 sessions for each of the two groups. At the course level, the workspace contains a list of 16 workspaces: one for each

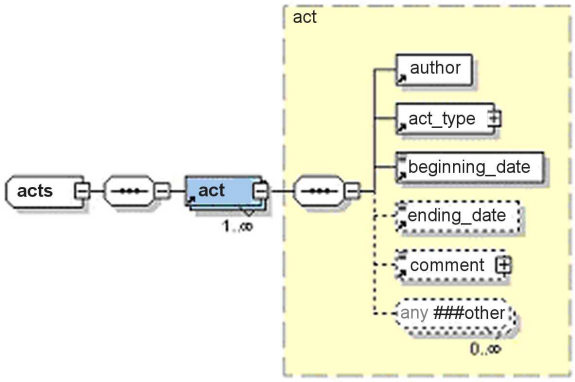
group and session. At the session level, the workspace may contain different workspaces representing various spaces and time slots where participants where interacting in a virtual room during the session. For example, the structure of session R8 is a list of five workspaces collecting interactions of the group R that split in different ways over three different spaces during 90 minutes. Figure 3 presents the way the subgroups are using the rooms during the session R8. On this figure, each rectangle shows a time slot where a given room has been used by some participants to interact via voice, chat or collaborative tools.

Figure 3 Copéas: subgroups moving through virtual spaces during the session R8 (see online version for colours)



Each workspace may be concerned by different interaction tools. Interaction may occur in devices and tools which can be as different as: forum, blog, text-chat tools or collaborative production tools. As examples of such tools, we can cite a collaborative conceptual map editor, a collaborative word processor and a collaborative drawing tool.

Figure 4 Extract from the XML Schema – the act concept (see online version for colours)

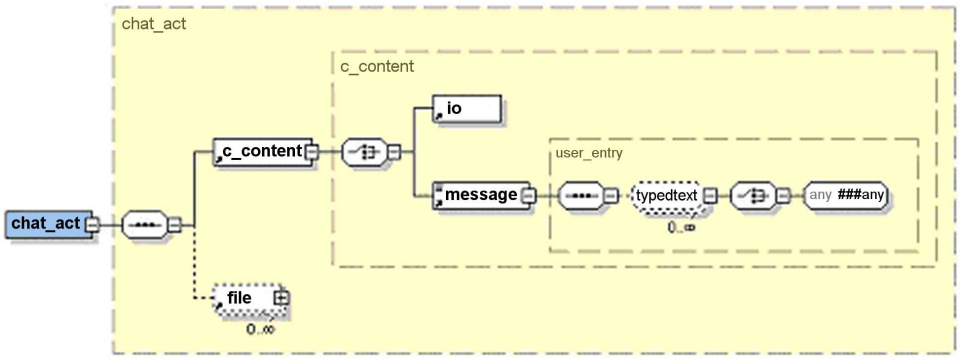


Interaction tracks are stored according to the *act*'s structure as presented in Figure 4. All actions are described by *act* elements. An *act* necessarily refers to its *author* identifier (defined in the members list – Figure 2), and a *beginning_date*. Depending on the nature of the act (*act_type*), an optional *ending_date* can be specified. The *act_type* element is a selector. The actual content and structure of an act depend on its type. Once selected this type among *mail_act* or *chat_act* for example, the data of the specific act type can be stored in the corresponding structure.

For example, a *chat_act* (see Figure 5) is either an enter/leave act (*c_content*: io), or the content of the message that can be addressed to all the workspace members or to a

specific member (e.g. if it is a private message). A chat act can contain an attached document (*file*) which in turn is described by a name, a type and a date.

Figure 5 Extract from the XML Schema – the chat act concept (see online version for colours)



2.2.4 The License component: collecting, distributing and accessing the data

The License component specifies the corpus publisher's and the users' access rights, as well as ethical elements concerning the course participants. Prior to data collection, researchers inform participants and ask them to sign an informed consent form. It details the questions of private protection, the intellectual rights of the author(s) and the consent they have obtained to release data online. In order to deal with these issues, we had to develop methods to anonymise data included in the Instantiation component. The license agreed by users' when accessing the repository is a Creative Commons license. However, users need to be minimally identified beforehand, in order to moderate the use for research or educational purposes.

2.2.5 A repository to share and visualise corpora: open access data

Once data have been collected, structured and described by metadata, they are deposited on the Mulce repository.

Being connected to other Open Archive Initiative repositories (Nelson and Warner, 2002), Mulce repository allows metadata to be shared and our corpus objects to become visible to the whole community. In order to ensure permalinks, Mulce appears in the general directory *Open Language Archives Community* (Olac-Mulce, 2012; Simons and Bird, 2007), not only because some of its corpora belong to the field of language learning, but also because the instantiation component of Mulce corpora include large language data, stemming from online learning situations and the corresponding human-human interactions. Thanks to this directory, each corpus has a permalink. This guarantees accessibility for the community of researchers. We also have formally defined an extension of the OLAC linguistic type entitled "language and teaching corpus".

In every corpus, each component is described by its specific metadata. In the Mulce repository, these metadata can be used by a researcher to find corpora that fit particular constraints. A form provides a selection of corpora according to the following criteria:

participants (students, tutors, native speakers), technologies (asynchronous LMS, audio-graphic conference, discussion forum, chat, …), pedagogical dimensions (global simulation, intercultural scenario, English and ICT, …), learning fields (French as foreign language, English for ICT, …), analysis tools (forum analysis, synchronised multimodal layouts, social networks analysis, …), language used, interaction and modalities (spatial-, spoken-, textual-, iconic-, multimodal scaffolding language, …). The result of this request is a list of corpora matching the criteria. Each corpus in this list is presented with synthetic information. At this point, the user is provided with three links that lead to: (1) the detailed description of the corpus based on its metadata, (2) a specific interface to browse its hierarchical workspace structure, or (3) a zip file containing the whole corpus to be downloaded.

2.2.6 Linking analyses to the main corpus: notion of distinguished corpus

In the field of natural language processing, the Freebank project (Salmon-Alt, Romary and Pierrel, 2004) created French corpora annotated on several linguistic levels like structure, morphosyntax, syntax and co-reference. The main objective was to make them available online for research purposes. The seminal idea suggested in the *Freebank* project was the progressive enrichment of corpora by mean of external annotations made afterwards by other researchers on the corpus. This idea has been implemented in the Mulce project by the concept of *distinguished corpus* that includes a particular analysis of a selected part of the main (LETEC) corpus. The LETEC includes all original structured interaction data, while the distinguished corpus only contains transformed data, useful for the specific analysis. It refers to a selection of the original data, in the main corpus. In a distinguished corpus, every tool used for the analysis is described and referenced, in order to allow other researchers to redo or extend the analysis (see section 4 for examples).

3 What is new in this arena since 2006?

In 2003 the Berlin declaration defined the notion of "open-access contribution". The Berlin process (2012) maintains the list of its signatories (currently 372) and invites the scientific community to its annual "Berlin conference".

As a result of this general worldwide evolution, new policies have been implemented in various national research structures. For example, following the National Science Foundation (NSF, USA), the JISC (UK) has opened a new research program (2011-2013) for the management of research data that:

> "[…] is recognised as one of the most pressing challenges facing the higher education and research sectors. Research data generated by publicly-funded research is seen as a public good and should be available for verification and re-use. In recognition of this principle, all UK Research Councils require their grant holders to manage and retain their research data for re-use, unless there are specific and valid reasons not to do so." (JISC, 2011).

The emergence of such policies also gave support to the development of concept, framework and systems to achieve data share in various research communities.

In the TEL community Markauskaite and Reimann (2008) have emphasised the potential benefits of sharing research data. Teams of researchers all over the world have

developed repository tools (JISC RPS, 2009). A new portal (Datacite, 2011) provides a list of repositories available for the sharing of research data in every scientific field and develops proposals for citing data sets. At the time of writing, this list contained around 150 sites. Only three of them concern education: PSLC Datashop, Mulce and JEDI (JEDI, 2011).

For the Intelligent Tutoring System (ITS) field, the PSLC DataShop presented in (Stamper et al., 2010) provides a data repository including datasets and a set of associated visualisation and analysis tools. These data can be uploaded as well-formed XML documents that conform to the Tutor_message schema. The datasets are fine-grained, principally automatically generated by ITS and focus on action/feedback interaction between learners and virtual tutor tools. The PSLC Datashop is probably the most successful story if we consider its impact on the ITS research community: it became a standard for an Intelligent Tutoring System to be evaluated with the associated tools. The learning curve obtained by confronting learners' traces with the measurement tool has become a kind of standard for quality measurement (Martin et al., 2011). We may attribute this success to the very constrained model of interaction supported by PSLC Datashop: the model itself directly inherits its specificities from the behaviourist theory of learning. For systems based on other theories like interactionist or constructivism, enabling more complex human interaction, such models may not apply.

JEDI (2011) is a repository whose aim is to preserve and distribute data collected in education and other related fields for research, educational and governmental administrative purposes. 140 data sets and reports (in Japanese) are currently available on the JEDI repository. Data are supposed to be processed with the R software.

In the sphere of recommendation tools, the theme team "dataTEL" (dataTEL STT, 2010) has studied several repositories and data sets to see how suitable they are for use as a test bed for the benchmarking of recommenders algorithms and tools (Drachsler et al., 2010).

All aforementioned repositories mainly deal with homogeneous data collected in a specific context. For example, PSLC Data Shop only manages data issued from ITS according to a specific format. The data processed in this case are Human Computer Interaction. In social sciences, Dataverse (King, 2007) and JEDI only consider flat representations like surveys.

Repository framework and system for digital object of many kinds flourished since the 2000's. The Open Archive Initiative (MPH) became a standard in 2002. In 2006, the vast majority of objects shared among scholars are written documents. Since 1998, many projects like the Public Knowledge Project founded by John Willinsky, have produced software and frameworks to improve sharing among scholars of different communities. In the library domain, we find many databases for written documents or their metadata descriptors. Such general purpose descriptors are defined in the Dublin Core specification since 2000. But specific communities took another way by creating databases of specific objects like in biotechnology information. E.g.: the genome-specific resources. We can now find many systems (JISC RPS, 2009) that make digital object collections visible and accessible. Digital objects may be of many concrete file formats.

In the Mulce repository, we deal with the hierarchical complexity of learning activities: from an entire experiment involving the whole set of participants in many different spaces over several months, to a single activity for a subgroup, in a single interaction tool. For each intermediary level of this complexity, there is a corresponding description of the context. The complexity is generally due to the learning organisation.

The heterogeneity of data is a consequence of the research protocol and data collection. Both learning design and research protocol descriptions are needed by future analysts to correctly interpret Computer Mediated Communication that is the central material for a typical LETEC in the Mulce repository.

This raised the question of tools interoperability and common technical structures to hold interaction data.

In the past five years, tools processing structured data have been developed in the ITS field. We have already cited those attached to PSLC Datashop. .

Concerning the structure for interaction data, in the CSCL community, an interesting framework: DELFOS (Osuna, Dimitriadis and Martínez, 2001) defines an XML based data structure (Martinez, De la Fuente and Dimitriadis, 2003) for collaborative actions in order to promote interoperability between analysis tools, readability for automated tools and adaptability to different analytic perspectives. Some of these authors joined the European research project on Interaction Analysis (JEIRP–IA) and reported in (Martinez, Harrer and Barros, 2005) a template describing Interaction Analysis tools and a common format.

Beyond the infrastructures, formats and tools, we want to emphasise two major authentic experiences of data sharing among researchers in the CSCL community.

The diversity of epistemological beliefs and research methodologies has led several researchers to question how such a plurality could become a source of productive scientific discourse rather than disagreement or balkanisation. Between 2008 and 2011, a series of five workshops were conducted to address this issue. Some preliminary results of this effort were reported in (Suthers et al., 2011). Sharing of datasets and multiple analyses of these datasets were investigated. However, it was difficult to determine the cause of differences in interpretation without being able to return to the primary data. In the final two workshops, to as great an extent as possible, primary data and analytic representations were shared and used to combine multiple viewpoints. The tool Tatiana has been used by this group to collect, synchronise and compare a variety of annotations, analytic representations and to support the exploration of data and their context (Dyke, Lund and Girardot, 2009; Reffay, Dyke and Betbeder, 2012).

Another important and visible experiment in research data sharing in the CSCL field has been led by G. Stahl on the spring fest 2006 session resulting from interaction in the Virtual Math Team environment (VMT). Multimodal chat sessions were collected and delivered to 28 external researchers from 11 countries, 18 institutions and eight different research fields. Each researcher applied her/his own analysis methods and tools to process these interaction data. The result is reported in (Stahl, 2009). One of the two VMT data sets used in this series of analysis has been edited in a Mulce structure (oai:mulce.org:mce-vmt-letec-teamc) and is now available in our repository.

4 Challenges and solutions according to the Mulce experience

Considering the impulse of some researchers, the support of institutions and the current development of research data repositories, one might believe that sharing research data has become easy. However, are individual scientists ready to spend time making their data shareable and accessible through data repositories? Nelson (2009), for example, asserts that open archives stay empty in spite of the intellectual agreement from all disciplines as regards to the benefits of data sharing. In some disciplines, including

physics, mathematics and computer science, communities are populating data repositories. This is also the case for some specific data banks in fields such as geophysics, biodiversity, ecology, Protein Data Bank, GenBank, …

> "But those discipline-specific successes are the exception rather than the rule in science. All too many observations lie isolated and forgotten on personal hard drives and CDs, trapped by technical, legal and cultural barriers — a problem that open-data advocates are only just beginning to solve." (Nelson, 2009, p.160)

In this section, we detail four challenges a data repository may face. For each challenge, we report the Mulce project experience and draw some perspectives to face them. As already mentioned, the motivation for sharing research data varies from one discipline to another. LETEC corpora correspond to a multidisciplinary field. It not only concerns computer science, but also educational science, domain specific learning sciences, like applied linguistics when, for example, language learning is at stake. Every discipline has its own methodology, units of analysis (Fjuk and Ludvigsen, 2001) and viewpoints concerning the observed and recorded data. They also have long traditions of well-established content analysis methods where a systematic organisation of data and processing on a large scale is not the rule. This is the reason why, in the description of the following challenges, we will sometimes make the distinction between researchers either in the humanities or in computer science.

4.1 Challenge 1: exchanging corpus and reusing data

Do researchers in TEL actually reuse data provided by others? Do they make their own data available to others? Does the Mulce repository actually ease data reuse?

4.1.1 Experience

Chris Teplovs who was not part of the Mulce project, reused the Simuligne corpus and extracted all the discussion forum messages from their Mulce structure without our help. The content is mainly in French whereas he is an English speaker. He then processed the data in his own tool *Knowledge Space Visualizer*. This was a confirmation that Mulce structure is coherent and quite easy to handle for a computer scientist (Reffay et al., 2011).

In Computer-Assisted Language Learning it is a common habit for a researcher to analyse data collected from a learning situation in which s/he designed or helped design. Consequently it is worth mentioning that Ciekanski, an applied linguist involved in the Mulce project, reused data from the LETEC Copéas, an experiment with which she had no connection (Ciekanski and Chanier, 2008). Furthermore, Lamy: another applied linguist co-designer of the Copéas experiment but who did not participate in the compilation nor organisation of the LETEC, has reused part of its data and recently published on this data seven years after the course ran (Lamy, 2012). These reuses and publications have only been possible because of the permanent repository access offered to researchers from different countries and institutions.

As already mentioned in section 3, the VMT corpus had been shared by many researchers. Before we put its data in a Mulce structure, it has been shared as a simple spreadsheet listing interaction acts and a sophisticated replay of the short session. In the

same field, we can also mention the Calico platform to share discussion forum and use analysis tools specialised in such Computer Mediated Communication.

4.1.2 Perspectives

It is too early to provide any statistics about Mulce repository usage: observing data reuse, new analyses and publications is a long-term project. It requires considering research differently, not always processing a one pass analysis on his/her own data, and avoiding constantly switching from one experiment to another without reconsidering it and without making cross-references.

Besides reusing data for producing new analyses from various perspectives, some researchers in humanities look forward to extracting data in order to develop resources for teachers training. In applied linguistics for example, such resources are analysed by students to detect efficient ways of using text-chat or various modalities in multimodal platforms. Opening up research data repositories based on learning situations to teachers is worthy of new research projects.

4.2 Challenge 2: Building an exhaustive, well structured and contextualised corpus: A hard work for scientists from different disciplines

What kind of learning session and data can be stored in the Mulce structure? How external researchers can reuse this structure to build their own corpora?

4.2.1 Experience

The Mulce repository is currently fully operational for publishing / deliver corpora which have been carefully structured. At present, 34 corpora have been released; six of them being LETEC / global corpora, the remaining distinguished corpora. The Mulce structure has proved to be flexible enough to gather and organise data from a variety of TEL situations which differ culturally or technologically. They deal with groups of learners resident in different countries and cultures like: Germany, the United Kingdom, Colombia, the USA or France. In the central instantiation component of our oldest LETEC Simuligne, we mainly assembled textual data coming from emails, forum discussions, text-chat interactions. Progressively we have extended our XML schemas to encode multimodal actions occurring in audio-graphic environments including audio, text, collaborative tools (e.g.: text editor, concepts maps, slides presentations), blogs, and recently 3D environments (e.g. *Second Life*) where verbal and non-verbal acts are transcribed (Wigham and Chanier, in press). Hence the Mulce project offers an original contribution to other data compiled by the TEL community, only made possible thanks to multidisciplinary collaborations. Let us now report the type of difficulties that new researchers face when considering building a LETEC corpus with their own data.

As regards computer scientists, even if it is often time consuming, it seems quite easy to collect, transform and store large collections of various data coming from different tools, but more difficult to document the pedagogical context and the research protocol.

For researchers in humanities, besides cultural differences in ways of compiling and analysing data, literacy in organising data with XML and schemas is scarce. In order to bypass this situation and publish new LETEC corpora, Mulce members brought support to researchers in order to let them compile their data and crosslink them between

interactions, learning design and research protocol. XML manifests were built by Mulce members.

4.2.2 Perspectives

The documentation of the XML Mulce schema should be improved. Where computer scientists are concerned, we have to better explain the necessity of a well described context for data interpretation by researchers from other disciplines. We are developing LETEC and distinguished corpora templates which can be more easily filled in for researchers in the humanities. We set specific trainings (Eurocall, 2010), and designed a website which we continuously update and which documents our process and methodology (Mulce-doc, 2012). Other developments of software connectors from common LMS and communication tools to the Mulce structure may reduce the gap for them.

4.3 Challenge 3: Making the deposit of a new corpus, available for a non Mulce member

Why a corpus deposit still needs a contribution from a Mulce member? How can we avoid this problem? Is it a priority?

4.3.1 Experience

Corpora deposited in the Mulce repository have only been edited by members of the team project and all except the aforementioned VMT have been compiled by the same team. New corpora are currently processed by young researchers who edit data from learning situations where Mulce members were not involved. The reader may wonder whether some part of the TEL community will contribute to the Mulce repository.

Today, the deposit of a new corpus on the Mulce repository is a tricky process that involves various interventions on different resources: 1 – specify criteria for the search process within Mulce database, 2 – check the integrity of the XML manifest of the corpus and the presence of all files on the server, 3 – update the OAI-PMH with the various metadata to be broadcasted to the OAI harvesters.

4.3.2 Perspectives

Obviously, developing an end user interface which could support this three-step process would solve the problem. This will be necessary when (and if) we get too many requests from external researchers to deposit ready made corpora. At present, a few researchers have shown initial interest but few have completed the process of building sharable corpora. Consequently, the development of this type of functionality is not a priority.

4.4 Challenge 4: Making connections with analysis tools

What are the benefits of tools connected Mulce corpora? How many tools are currently connected?

4.4.1 Experience

The current situation is quite uncomfortable because of the unbalanced effort for researchers to organise data compared to the immediate benefit they can get out of it. When the project started, we expected that when it ended other researchers would be able to connect their own tools to the Mulce structure in order to analyse interaction data in the Mulce collection of corpora. This has been the case for the Tatiana tool (Dyke *et al.*, 2009) with various corpora which contain multimodal interactions (cf. section 2.2). We also made a connexion to the set of Calico tools (Giguet *et al.*, 2009; Calico, 2011) dedicated to discussion forum analysis with data extracted from the LETEC Simuligne (oai:mulce.org:mce-simu-forum-all). This has been technically possible thanks to our common choice of XML format family and the transformation facilities in the XML world. It has been eased by the presence in these projects of computer scientists working collaboratively with researchers in the humanities. The connexion between Mulce format and both of Tatiana and Calico tools directly offers new perspectives for the whole community to analyse or re-analyse Muce corpora with the facilities and representations that these tools enable.

Experience gained from Calico partners helped us realise that even transforming a forum into a simple flat XML structure required support for researchers in educational science, for instance.

4.4.2 Perspectives

Considering the latter difficulties, the first two connexions with Tatiana and Calico tools can be considered as an encouraging starting point. Greater corpora reuse is necessary to multiply viewpoints on data and expand the range of analysis tools that could be easily connected in order to motivate scientists from the TEL community.

5 Conclusion

The TEL community is concerned with what Gray called the fourth paradigm for scientific research, i.e. “data-intensive science”. For the improvement of our research methodology and tools, we claim that data collected in learning situations need to be shared. The first question is how extensive such compilation must be and what the context description attached to the learning situation needs to be in order to obtain a true corpus which is an object worthwhile of scientific consideration.

We described here how the Mulce project responded appropriately to online collaborative learning situations where human to human interactions are prevalent and occur simultaneously in heterogeneous environments. The Learning & Teaching Corpus (LETEC) is a structured entity containing all the elements resulting from an online learning situation. Achievements of the Mulce project, which began in 2006, were presented, including its open access repository.

During the last five years the international context in scientific research has moved forward with a clear emphasis on the need to share and cite research data. We gave an overview of work undertaken in TEL around repositories, tools, and ways of structuring data. This certainly shows the vividness of our field but eludes the project of a unified approach.

We finally pointed out four of the challenges faced by data repositories that aim at sharing data between scientists from different disciplines. For these, in the particular case of LETEC corpora, we reported on the Mulce experience and its current situation and suggested some perspectives.

6 References

Berlin Declaration (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

Berlin Process (2012). http://oa.mpg.de/lang/en-uk/berlin-prozess/

Calico (2011). *Presentation of the project, list of tools and publication*. http://www.stef.ens-cachan.fr/calico/en/calico.htm

Chanier, T. and Ciekanski, M. (2010) 'Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage' (Benefits of Sharing Corpora when Analyzing Online Interactions: an Example of Methodology Related to a Databank of Learning and Teaching Corpora.), *Apprentissage des Langues et Systèmes d'Information et de Communication (ALSIC)*, 13. DOI : 10.4000/alsic.1666

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

Chanier, T., Reffay, C., Betbeder, M.-L., Ciekanski, M. (2008). 'Copéas'. [Learning and Teaching Corpus]. *Mulce.org*: Université de Franche Comté. [online]. Available at [oai:mulce.org: mce-copeas-letec-all, http://repository.mulce.org]

Ciekanski, M., Chanier, T (2008). 'Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment'. *ReCALL*, vol. 20 (2), Cambridge University Press. 162-182. [online]. Available at [doi:10.1017/S0958344008000426 , http://edutice.archives-ouvertes.fr/edutice-00200851]

DataCite [online] *Project to help researchers to find, access, and reuse data*. http://datacite.org

DataTEL STT (2010). *STELLAR Theme Team dataTEL: A Data Set Framework for Recommender Systems in Technology Enhanced Learning*. Lead by H. Drachsler. http://www.stellarnet.eu/instruments/theme_teams/#DATATEL

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H., Friedrich, M. and Wolpers, M. (2010) 'Issues and considerations regarding shareable data sets for recommender systems in technology enhanced learning', *Procedia Computer Science*, 1(2), pp.2849-2858.

Dyke, G., Lund, K. and Girardot, J.-J. (2009) 'Tatiana: an environment to support the CSCL analysis process', in *CSCL'2009: Proceedings of the International Conference on Computer Supported Collaborative Learning*, Rhodes, Greece, pp. 58-67.

EUROCALL (2010). *Workshop on "Dissemination and comparison of research findings: developing Contextualized Learning and Teaching Corpora (LETEC)*", Bordeaux. http://ubpweb.univ-bpclermont.fr/HEBERGES/mulce/spip.php?article29

Ferraris, C., Martel, C. and Vignollet, L. (2007) 'Helping Teachers in Designing CSCL Scenarios: a Methodology Based on the LDL Language', in *Proceedings of Computer Supported Collaborative Learning Conference*, pp.193-195.

Fjük, A. and Ludvigsen, S. (2001) 'The Complexity of Distributed Collaborative Learning: Unit of Analysis', in *Proceedings of the European Conference on Computer Supported Collaborative Learning*, Maastricht.

Giguet, E., Lucas, N., Blondel, F.-M. and Bruillard, E. (2009) 'Share and explore discussion forum objects on the Calico website', in *CSCL'2009: Proceedings of the International Conference on Computer Supported Collaborative Learning*, Rhodes, Greece. pp. 174-176.

Gray, J. (2007/2009) 'Jim Gray on eScience: A Transformed Scientific Method in Hey, T., Tansley, S. and Tolle, K. (Eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. WA : Microsoft Research. xxvii-xxxi. [online]. Available at http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

IMS [online] *IMS global learning consortium* http://www.imsglobal.org/

JEDI [online] *Educational Data Archive*. http://essrc.hyogo-u.ac.jp/jedi/

JISC (2011). *Digital infrastructure: Research management programme - Managing research data*. Joint Information Systems Committee (UK). http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx

JISC RPS (2009). *Repository Project* Survey by JISC. http://www.dspace.org/images/publications/repository-software-survey-2009-03.pdf

Kern, R., Ware, P. and Warshauer, M. (2004). Crossing frontiers: new directions in online pedagogy and research', *Annual Review of Applied Linguistics*, 24, pp. 243-260.

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

Koper, R. and Tattersall, C. (2005) *A Handbook on Modelling and Delivering Networked Education and Training*. ISBN 978-3-540-22814-1. Springer Verlag.

Lamy, M.-N. (2012) 'Personal Learning Environments: Concept or Technology? Click If You Want to Speak: Reframing CA for Research into Multimodal Conversations in Online Learning', *International Journal of Virtual and Personal Learning Environments*, 3(1), pp. 1-18.

Markauskaite, L. and Reimann, P. (2008) 'Enhancing and Scaling-up Design-based Research: The potential of E-Research', in *the proceedings of the International Conference for the Learning Sciences*, Utrecht, The Netherlands.

Martin, B., Mitrovic, A., Koedinger, K., and Mathan, S. (2011) 'Evaluating and Improving Adaptive Educational Systems with Learning Curves', *Journal of User Modeling and User Adapted Interaction*, 21(3), Springer.

Martinez, A., De la Fuente, P. and Dimitriadis, Y. (2003) 'Towards an xml-based representation of collaborative action', in the *Proceedings of International Conference on Computer Supported Collaborative Learning*, Bergen, Norway, pp. 14-18.

Martinez, A., Harrer, A. and Barros, B. (2005) *Library of Interaction Analysis Tools*, Deliverable D.31.2 of the JEIRP IA (Jointly Executed Integrated Research Project on Interaction Analysis Supporting Teachers & Students' Self-regulation). KaleidoScope.

Mce_sid (2012). *Schema for the Structured Information Data (instantiation component) of a Mulce corpus.* http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/mce_sid.xsd

Mulce-repository (2012). *Open access repository where LETEC corpora may be downloaded.* Mulce.org. http://repository.mulce.org

Mulce-doc (2012) *Web site explaining the Mulce methodology and commenting scientific events around the project*. http://mulce.org

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

Nelson, M. and Warner, S. (2002) *The Open Archives Initiative Protocol for Metadata Harvesting*, Lagoze, C., Van de Sompel, H. (eds). Version 2.0.

OAI-PMH (2002). *The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0*. Document Version 2002/07/05. Open Archive Initiative. http://www.openarchives.org/OAI/openarchivesprotocol.html

Olac-Mulce (2012) *Description of Mulce corpora within the OLAC directory* http://www.language-archives.org/archive/mulce.org

Osuna, C., Dimitriadis, Y. and Martínez, A. (2001) 'Using a Theoretical Framework for the Evaluation of Sequentiability, Reusability and Complexity of Development in CSCL Applications', in the *Proceedings of the European Computer Supported Collaborative Learning Conference*, Maastricht, The Netherlands.

Reffay, C. and Betbeder, M.-L. (2009) 'Sharing corpora and tools to improve interaction analysis', *Learning in the Synergy of Multiple Disciplines - Fourth European Conference on Technology Enhanced Learning*, France.

Reffay, C., Teplovs, T. and Blondel F. M. (2011). Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. *In Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings. Volume II – short papers*, Hans Spada, Gerry Stahl, Naomi Miyake, Nancy Law (Eds.), ISLS, pp. 846-850, 4-8 july, 2011, Hong Kong, China.

Reffay, C., Dyke, G. and Betbeder, M.-L. (2012) 'Data sharing in CSCR: towards in-depth long term collaboration', in Juan, A., Daradoumis, T., Roca, M., Grasman, S., Faulin, J. (Eds.), *Collaborative and Distributed E-Research: Innovations in Technologies, Strategies and Applications*, IGI Global http://www.igi-global.com ,pp. 111-134. ISBN 978-1-4666-0125-3

Rourke, L., Anderson, T., Garrison, D. R.and Archer, W. (2001) 'Methodological Issues in the Content Analysis of Computer Conference Transcripts', *International Journal of Artificial Intelligence in Education*, 12.

Salmon-Alt, S., Romary, L. and Pierrel, J.-M. (2004) 'Un modèle générique d'organisation des corpus en ligne', *Traitement automatique du langage* (Tal), 45(3), pp. 145-169.

Simons, G., Bird, S. (2007) OLAC: Open Language Archives Community. http://www.language-archives.org/ http://www.language-archives.org/OLAC/metadata.html

Stahl, G. (2009) *Studying Virtual Math Teams*, New York, Springer. [online] Available at http://gerrystahl.net/vmt/book/

Stamper, J.C., Koedinger, K.R., Baker, R.S.J.D., Skogsholm, A., Leber, B., Rankin, J., and Demi, S. (2010) 'PSLC DataShop: A Data Analysis Service for the Learning Science Community', *in the Proceedings of Intelligent Tutoring Systems*, LNCS, 6095, pp. 455-456.

Suthers, D.D., Lund, K., Rosé, C., Dyke, G., Law, N., and Teplovs, C. (2011) 'Towards productive multivocality in the analysis of collaborative learning', in the *Symposium at CSCL 2011*, Hong Kong, China, pp. 1015-1022.

Wigham, C.R. and T. Chanier (in press). "A study of verbal and nonverbal communication in Second Life. the ARCHI21 experience". ReCALL 25(1), Cambridge Journals. [online] Available at http://edutice.archives-ouvertes.fr/edutice-00674138