Improving CSCL indicators by sharing multimodal teaching and learning Corpora

Christophe Reffay, Marie-Laure Betbeder

Improving CSCL indicators by sharing multimodal teaching and learning Corpora

Christophe Reffay
Computer Science Laboratory,
University of Franche-Comté
Christophe.Reffay@univ-fcomte.fr

Marie-Laure Betbeder
Computer Science Laboratory,
University of Franche-Comté
Marie-Laure.Betbeder@univ-fcomte.fr

Abstract. We point out the need for CSCL community to reach large scale validation for its results by addressing the lack of sharing of interaction indicators and data. The main goal of the Mulce[1] project is a definition for teaching and learning corpora (especially for interaction tracks), a technical format to organize data and a platform for corpus sharing, providing analysis and visualization tools.

Keywords: e-research, sharing corpora, interaction analysis, CSCL indicators.

MOTIVATION

During the last decade a lot of technical propositions for indicators of social or cognitive process monitoring have been made. The very most part of these indicators (including ours) are designed in a given context, where they show some interesting properties and even promise usefulness for the various actors involved in CSCL real situations. Unfortunately, these indicators often stay in the researchers' hands and are rarely used by real actors of the situation. As far as we know, none of them have been validated or at least evaluated by real/concrete actors. The need for validation of these indicators, at least in a given context, becomes crucial if we want this domain to contribute to the distance learning area of the real world. These indicators are also rarely reused in other situations or contexts. With others (Rourke et al., 2001) we think that replicability, reliability and objectivity need to be improved in our work. A path on this direction would be sharing data so that they can be analyzed and compared. Therefore we propose a formalism to describe contextualized interaction data.

PROPOSAL

Our proposal consists in (1) a formalism to describe a learning and teaching corpus and (2) a platform for corpus sharing (Reffay et al., 2008). The formalism defines the information which can be contained in a corpus and the structure of the data. Through the platform, researchers can share their corpora with the community and access the data shared by other members of the community.

Proposal 1: Learning and teaching corpus formalism

Building and recording interaction in an online training

We define a Learning & Teaching Corpus as a structured entity containing all the elements resulting from an on-line learning situation, whose context is described by an educational scenario and a research protocol. The core data collection includes all the interaction data, the training actors' production, and the tracks, resulting from the actors' actions in the learning environment and stored according to the research protocol. In order to be sharable, and to respect actor privacy, these data should be anonymised and a license for its use be provided in the corpus. A derived analysis can be linked to the set of data actually considered, used or computerized for this analysis. The definition of a Learning & Teaching Corpus as a whole entity comes from the need of explicit links, between interaction data, context and analyses. This explicit context is crucial for an external researcher to interpret the data and to perform its own analyses.

Corpus composition and structure

The main components of a learning corpus are:

[1] The Mulce project, leaded by T. Chanier, is supported by the French National Research Agency. http://mulce.univ-fcomte.fr/axescient.htm#eng

- The Instantiation component, the heart of the corpus, which includes all the interaction data, production of the on-line training actors, completed by some system logs as well as information characterizing actors' profile.
- The Context concerns the educational scenario and the research protocol (optional element).
- The License component specifies both corpus publisher's (editor) and users' rights and the ethical elements toward the actors of the training.
- The Analysis component contains global or partial analysis of the corpus as well as possible transcriptions.

The Mulce structure aims at linking the components of the corpus. For example a researcher, while reading a chat session (which belongs to the instantiation component), may have to read the objectives of the activity (which belongs to the pedagogical context).

Instantiation formalism: Actors and environment description

If learning design is a general description of activities involving generic roles and environments, the instantiation phase is concerned by real actors and concrete platforms and tools. It consists in describing (1) the actors (identifier, learning profile, linguistical and cultural aspects, etc.), (2) the technological environments (name, version, URL, etc.), (3) the tools used during the learning activity (technical components actually used) and (4) the groups and their members.

Instantiation formalism: Workspace concept

The hierarchical structure of the learning stage (potentially spread in parallel groups) is captured in the Workspaces element, i.e.: a sequence of "workspace" elements (see figure 1). A workspace is generally linked to a learning activity (of the pedagogical scenario). It encompasses all the events observed during this activity, in the tool spaces provided for this activity, for a given (instantiated) group of actors. A workspace description includes its members (references to the actors registered in the learning activity), starting and ending dates, the provided tools and the traces of interaction that occurred in these tools. In order to fit the hierarchical structure of learning and support activities, a workspace 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 enables to browse the interaction data in many different ways, independent to the concrete storage organization in the XML document.

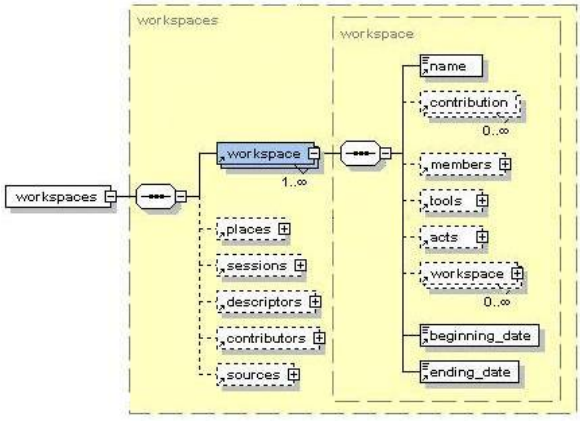

Figure 1. Extract of the XML Schema: workspaces and workspace elements.

Our specification describes communication tools and their features with a great level of precision. The corpus builder can specialize/particularize the schema (i.e., restrict it) to fit the specific tools and features proposed to the learners in a specific learning environment. In the meantime, if a tool cannot be described with the specification, one can augment the schema by adding new elements, in order to take into account the tool's specificities. Both of these mechanisms offer two ways, the specification can be extended to fit the analysis needs. Moreover, recursive workspace description enables the corpus descriptor to choose the grain at which he needs to describe the environment. Thus, a workspace can be used to describe a complete curriculum, a semester, a module, a single activity or a work session (a concept generally related to synchronous learning activities). The workspace concept represents the space and time location where we can find interaction with specific tools.

Interaction traces are stored according to the act's structure. All actions, wherever they come from are described by an act element. Depending on its nature, an act is described by different elements, mandatory or optional like a reference to its author identifier, a beginning and an ending date, an act type and an actual content

(or value). For example, a chat act can have the type in/out (actor entering/leaving), it may contain a message, can be addressed to all the workspace members or to a specific one (particularly if it is a private message). This XML Schema defines the storage structure for many act types, e.g.: forum message, chat act, transcribed voice act, and more. The complete Mulce schema[2] for the instanciation component (structured information data) is available online.

Proposal 2: a Platform for corpus sharing

Sharing corpora

The deposit of a corpus consists in declaring it, describing it by means of general metadata, and uploading its components (described previously). Each component has a specific formalism. These can either be standard formalism such as Learning design (IMS-LD 2003) (used for the context components: educational scenario and research protocol), or the specific formalism described here above for structured interaction data. If these recommended formalisms are used to describe the various components of the uploaded corpus, the researchers will fully benefit from all the tools provided one the Mulce platform to navigate and analyse the entire corpus. Otherwise, the corpus will be downloadable as is by other researchers. Each component is described by its specific metadata. On the Mulce platform, these metadata can be used by a researcher to find corpora that fit particular constraints. For example the researcher can select the corpora concerning its own research interests, either in term of used tools, of targeted audience or learning domain.

Browsing and analyzing corpora

The second part of the platform proposes the visualization, the navigation and the analysis of structured interaction data. We distinguish two parts: the navigation (or visualization) aspect, and the analysis aspects of corpora. The interest of the navigation aspect is twofold. Firstly, the corpus becomes independent to the (evolving) software, where originally interaction took place. This is a major benefit for data longevity and reusability. Secondly, because of the main attention paid on the context of interaction in the Mulce project, the interaction navigator makes explicit links between interactions and their surrounded context. Finally, the researcher can select a part of a corpus by means of requests. He can, for example, select all the interactions of an actor using a specific communication tool. For each of the interactions he can access to the prescribed educational activity. The analysis aspect of corpora concerns the use of tools based on the instantiation component formalism. The XML format being defined, we hope that different analysis tools (including indicator synthesis), coming from various teams, will have a version that can operate on the Mulce structure. The tools proposed on the platform will originate from our research team (Betbeder et al., 2007) or from partnership. For example we have two running collaborations: the Calico[3] project, Tatiana (Dyke et al., 2007). The Calico project aims at proposing different visualization and analysis tools, specialized on discussion forums. Tatiana includes a navigator, a replayer and an annotator. The replayer functionality synchronizes the various data sources. We are currently adapting its XML schema to fit ours and extend its visualization functionalities to other communication tools. We are interested in other collaborations aiming at providing other analysis tools.

To conclude, this work suggests a way to access, share, analyze and visualize *learning and teaching corpora*. To make this technically possible, we propose (1) a formalism which defines, describes and organizes data provided by on-line training and (2) a platform to navigate through, visualize and analyze corpora (currently being integrated) through of a variety of tools.

REFERENCES

Betbeder M.-L., Tissot R., Reffay C. (2007). Recherche de patterns dans un corpus d'actions multimodales. In Nodenot, T., Wallet, J., Fernandes E. (Eds.) *EIAH'2007 Conference:* Switzerland, june, pp. 533-544.

Dyke, G., Girardot, J.-J., Lund, K. & Corbel, A. (2007). Analysing face to face computer-mediated interactions. EARLI,, Eötvös Lorand University, Hungarian Academy of Sciences,.

Reffay, C., Chanier, T., Noras, M. and Betbeder, M.-L. (2008) Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. In *STICEF journal*, Vol. 15, 2008. 25 p.

Rourke L., Anderson T., Garrisson D., Archer W. (2001). Methodological Issues in the Content Analysis of Computer Conference Transcripts. *IJ-AIED*, Vol 12, p. 8-22.

[2] http://mulce.univ-fcomte.fr/metadata/mce-schemas/mce_sid.xsd

[3] Calico is a french research project described here: http://calico.inrp.fr/CALICO