Extending validation of tools and analyses in CSCL situations: How to collaborate on interaction analysis?

Christophe Reffay, Marie-Laure Betbeder

Extending validation of tools and analyses in CSCL situations: How to collaborate on interaction analysis?

Christophe Reffay, Marie-Laure Betbeder, *Computer Science Laboratory of the Franche-Comté University, France.*

Introduction

The Mulce Project aims at developing a server that would allow researchers to share Learning and Teaching Corpora (Letec). Our main goals are: first, to facilitate the work of researchers in the CSCL field by reusing existing corpora instead of creating new experiments, collecting and organizing data, and secondly, to connect these corpora to shared analyses and visualization tools. We also hope that this process of sharing would deepen and widen the validity of research tools and analyses. The limited space of our publications generally does not allow the authors to include their data nor the detailed context of their source experiments, which means that these results cannot be reused nor reproduced.

In the first part of this paper we present some of our work trends related to the workshop topics. In particular, visualization and analysis tools for synchronous and asynchronous interactions. In the second part we discuss the lack of validity for indicators and analysis tools and results in this particular domain. The last part presents the Mulce project which is a tentative to face this stated lack of validity.

Our work on Analysis and visualization tools

In previous research, presented in CSCL2003, we built an automatable computing process based on Social Network Analysis to evaluate the cohesion of a learning group using e-mail and forum messages (Reffay & Chanier, 2003). Another work, mainly conducted during the phd thesis of A. Mbala (Mbala, 2003) and presented in ITS2002 (Intelligent Tutoring System) (Mbala et al., 2002), is a multi-agent system that tries to predict if some learners are less and less participating, in order to prevent their abandon. We have also worked on the analysis of data resulting from a tailorable framework to support collective activities in a learning context (Betbeder & Tchounikine, 2003). We analysed act's contents and worked out the proportion of acts related to the following four categories: activity achievement, group organization, environment tailoring and socialization.

In collaboration with a research team in linguistics, we are also involved in the field of multimodal interaction analysis (Betbeder et al., 2008), combining quantitative (macroscopic level) and content analysis (microscopic level) on micro actions that each participant can realize in an audio-graphic synchronous collaborative environment (audio talking turns, chat acts, votes, paragraph production in a shared text document, objects in a shared concept map or whiteboard). Our contributions in this field deal with (1) pattern discovering in sequences of such actions (Betbeder et al., 2007), (2) a visualization tool (Betbeder et al., 2008) which emphasizes the intertwinement of the actors' synchronous acts. In such environment, actors can interact by using different modalities at the same time. Due to the importance of time and sequence of acts, such phenomena are hardly visible through a database representation.

We are currently working on navigation through corpora. This includes the selection of corpora and visualization of archived interaction acts. Once the corpus has been chosen, a system of requests provides selection of parts of the corpus by considering for example: time restriction, selected communication tools, author or group or string search in interaction contents. The resulting set of acts

can be visualized in the appropriate form, according to the communication tool they are recorded from. This heterogeneous set of acts is organised by our XSD schema, specified by the Mulce project, presented in (Reffay & al., 2008) and available on the Mulce project site[1].

The need of validation

In CSCL domain, it seams to us that the lack of validity of results and tools becomes more and important. We can precise our view on this problem in the following terms:

- Lack of transparency and availability (accessibility for other researchers) of interaction data resulting from an online learning situation. These data are unusable for others. Furthermore, the pedagogical and scientific contexts of these learning situations are rarely published.

- Too many analysis tools are developed by a given team and only tested in a (unspecified) given context. Interaction and research data being unreachable for the rest of the community, the scientific results cannot be replicated or reproduced (Rourke et al., 2001; Garrison et al., 2006). As a consequence, there is no possible comparison between methods or results concerning one set of interaction data and another.

Studying collaborative online learning, in order to either understand this specific type of situated human learning, or evaluate scenario and associated devices, or improve technological environments, requires accessibility to interaction data collected from various actors (e.g. learners, teachers, tutors, etc.) that participate to learning situations. Recent international research publications and scientific events are related to this topic. But interdisciplinary communities involved in this research have not been able to characterise a sharable scientific object according to a comprehensible methodology.

On the one hand, we have partial data, not contextualised with pedagogical and technological learning situation elements, or else raw data that are inextricably tangled in specific software using proprietary formats. A simple collection of student's online interaction data is not a scientific object for the research community focused on online learning. In the language learning field, this idea is emphasised by Kern, Ware and Warshauer as follows:

> *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 et al., 2004)

Our research domain is not only concerned by learning, but more widely by all pedagogical aspects and viewpoints and particularly by teaching. Then, some studies aim at "[…] gather evidence about the effects of instructional conditions of instruction" (Chapelle, 2004: 594). Success in such studies requires gathering the context elements, and in particular those characterising the pedagogical situation. From a methodological point of view, this leads to link up the various data sources in order to create a scientific object, worthy of analysis. This idea is emphasised by the following excerpt concerning interaction in discussion board:

---

[1] Mulce data structure available here: http://mulce.univ-fcomte.fr/metadata/mce-schemas/mce_sid.xsd

Research in computer mediated communication in education aims at describing complex phenomena by using content analysis methods that flavour partially only some aspects of the communication. The method should be able to consider the discourse as a situated verbal interaction, in its various dimensions: linguistic, situational (in the universe of reference and interaction situation) and hierarchical constraints of the discourse.

(Henri & Charlier, 2005)

On the other hand, inaccessibility to research data, that is in fact the reality in the quite whole international community, hinders online learning situations from being considered as a scientific object: it impedes verifications, controversy, replication, refinement, multiple analyses, etc. Even if multiple analyses are strongly encouraged by research heads, they are still exceptions that we can illustrate by the studies of Kramsch and Thorne (Kramsch & Thorne, 2001) and (Thorne, 2003). Analysing some interaction pieces, extracted from a learning situation informally described in (Kern, 2000), they gave a different interpretation to explain the failure in this interaction between learners of different mother tongues (Kern et al., 2004:p251). Reanalysis can be motivated by various factors like verification (in the previous example), use of alternative content analysis methods (i.e. special issue of Computers & Education (Valke & Martens, 2006)), comparison of results provided by distinct disciplinary approaches (Corbel et al., 2006), etc.

However, except this contrastive or alternative view, one can consider that analysis approach is by nature a cumulative process, built by distinct research teams, supported by previous analysis, each of these giving its set of annotations. In this sense, the transcription process of audio-oral or multimodal interaction (Vetterr & Chanier, 2006; Ciekanski & Chanier, 2008) is the necessary first step before any other analyses. In the same way, we can consider that chat sessions or forum sets have to be prepared in a target format from row data in the proprietary structures tangled in the various platforms in order to explicit speech turns or messages, speaker/scripter, etc.). This first step, giving the content in a textual and structured form, can be followed by a first level of annotation coming from a conversational analysis, and by a second one concerning a discourse analysis. Such research practices are already well rooted in the NLP (Natural Language Processing) research domain, where, from corpus extracts, distinct researchers can cumulate different level of description/annotation: morphologic, syntactic, anaphoric, etc. (Salmon-Alt et al., 2004).

A research community becomes mature by sharing (contextualised) resources, tools and practices. Resources and tools are components, aside our publications, of the scientific contribution that our research directions ask us to put in open access (Berlin, 2003; Chanier, 2004:121). A necessary condition to share such resources is open access that should be supported by a set of techniques and protocols (standardisation, interoperability, metadata, etc.). We also need to solve the problems of rights in our research domain and, concerning learning and teaching studies, involving human beings in social experiments; the major ethic question has been quoted by Chapelle in the following terms:

Any discussion of technology in second language research would not be complete without raising the ethical challenges that researchers face in SLA [Second Language Acquisition] research in general and particularly in research involving the collection and archiving of personal performance data that reveal personal attributes (Chapelle, 2004: 599).

Mulce project

In order to widen and strengthen the scientific approach in online learning domain, and specifically analysis of online interaction in learning situations, we launched the MULCE (*MULtimodal Corpus Exchange*) project (Mulce, 2007). The first step of this project has been to define (and exemplified) the notion of "Learning and Teaching Corpus". Its main objectives are:

- ➢ To define the data structure of a "Learning and Teaching Corpus",
- ➢ To specify and develop a technical support to effectively share such corpora, integrating OLAC specifications (Open Language Archives Community) and OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting),
- ➢ To rebuild 2 of our global corpora (Simuligne and Copeas) to be accessible through out this platform, and documented according specifications of a "Learning and Teaching Corpus".

We propose (1) a formalism to describe learning and teaching corpora and (2) a platform to share them among the research community. 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.To share a corpus, a researcher has to provide metadata describing the corpus' components and upload a file describing each component. While accessing a corpus, an identified researcher is provided with a variety of tools to browse the corpus components, to navigate through the contextualized interaction data, to visualize and to analyze them.

Our major goal is to develop efficient tools and technical environments to help the wide variety of actors involved in online teaching and learning.

REFERENCES

Berlin (2003). Berlin declaration on *" Open Access to Knowledge in the Sciences and Humanities "*. Munich, Institut Max Planck. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

Betbeder, M.-L., Ciekanski, M., Greffier, F., Reffay, C., and Chanier, T. (2008). Interactions multimodales synchrones issues de formations en ligne : problématiques, méthodologie et analyses, *In STICEF journal*, 15, (to appear).
http://sticef.univ-lemans.fr/num/vol2008/06-betbeder/sticef_2008_betbeder_06.htm

Betbeder, M.-L., & Tchounikine, P. (2003). Symba: a tailorable framework to support collective activities in an learning context. In J. Favela and D. Decouchant, (Eds.), *Procs of the 9th Int. Workshop on Groupware (CRIWG 2003)*, Springer-Verlag, Autrans, France, 90-98.

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: Environnements Informatiques pour l'Apprentissage Humain*, Switzerland, 533-544.
http://edutice.archives-ouvertes.fr/edutice-00158881/

Chanier, T. (2004) *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche* ? Paris, L'Harmattan. http://archivesic.ccsd.cnrs.fr/sic_00001103/fr/

Chapelle, C.A. (2004). Technology and second language learning: expanding methods and agendas. System, 593-601.

Ciekanski, M. & Chanier, T (2008). Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. *The journal of Computer Assisted Language Learning*. Recall, 20 (2), Cambridge University Press, 162-182.
http://edutice.archives-ouvertes.fr/edutice-00200851/

Corbel, A., Girardot, J.-J., and Lund, K. (2006). A method for capitalizing upon and synthesizing analyses of human interactions. In W. van Diggelen & V. Scarano (Eds.), *Workshop proceedings Exploring the potentials of networked-computing support for face-to-face collaborative learning. EC-TEL 2006 First European Conference on technology Enhanced Learning*, October 1, Crete, 38-47.

Garrison, D.R., Cleveland-Innes, M., Koole, M., Kappelman, J. (2006). Revisting methodological issues in the analysis of transcripts: Negotiated coding and reliability. *The Internet and Higher Education*, 9(1), 1-8.

Henri, F., & Charlier, B. (2005). L'analyse des forums de discussion Pour sortir de l'impasse. In Baron G-L., Bruillard E., Sidir M. (Dir.), *Symposium Symfonic. Formation et nouveaux instruments de communication*. Amiens, Université de Picardie, janvier. http://archive-edutice.ccsd.cnrs.fr/edutice-00000897

Kern, R. (2000). *Literacy and Language Teaching*. Oxford: Oxford University Press.

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

Kramsch, C., & Thorne, S. L. (2001). Foreign language learning as global communicative practice. In D. Block & D. Cameron (dir.), *Globalization and language teaching*. Londres, Routledge, 83-100.

Mbala, A.,. Reffay, C., and Chanier, T. (2002). Integration of automatic tools for displaying interaction data in computer environments for distance learning. In S.A. Cerri, G. Guardères, and F. Paraguaçu (Eds.), *ITS'02, Intelligent Tutoring System conference*, volume 2363 of *LNCS*, Springer-Verlag, France, 841-850.

Mbala, A. (2003). Analyse, Conception, spécification et développement d'un système multi-agents pour le soutien des activités en formation à distance. Phd thesis in computer science of University of Franche-Comté.

Mulce (2007): English version of the MULCE Project homepage. *Multimodal Learning Corpus Exchange* (2007-2009). http://mulce.univ-fcomte.fr/axescient.htm#eng

Reffay, C. & Chanier, T. (2003). How social network analysis can help to measure cohesion in collaborative distance-learning? *In proceeding of Computer Supported Collaborative Learning conference (CSCL'2003)*, Bergen - http://edutice.archives-ouvertes.fr/edutice-00000422

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 (Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation)*, 15, (to appear). http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01.htm

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*. http://aied.inf.ed.ac.uk/members01/archive/vol_12/rourke/full.html

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

Thorne, S. L. (2003). Artifacts and cultures-of-use in intercultural communication. *Language Learning & Technology*, 7(2), 38-67.

Valcke, M., & Martens, R. (2006). Methodological Issues in Researching CSCL. *Special issue of Computers & Education*, 46(1), 1-104.

Vetter, A. & Chanier, T. (2006). Supporting oral production for professional purpose, in synchronous communication with heterogeneous learners. *The journal of Computer Assisted Language Learning*. Recall, 18(1): Cambridge University Press, 5-23.
http://edutice.archives-ouvertes.fr/edutice-00080316/