The Calico Platform: Multilingual Monitoring of Online Discussions

Emmanuel GIGUET, Nadine LUCAS, GREYC, CNRS – Université de Caen Basse-Normandie – ENSICAEN
Email: Emmanuel.Giguet@info.unicaen.fr, Nadine.Lucas@info.unicaen.fr

François-Marie BLONDEL, Eric BRUILLARD, STEF, ENS Cachan – INRP, UniverSud
Email: francois-marie.blondel@stef.ens-cachan.fr, eric.bruillard@creteil.iufm.fr

Abstract: In this article, we present the Calico website, a shared space where researchers and practitioners in education share and explore discussion forum objects coming from different e-learning platforms. The platform is briefly described. The focus is set on the different kinds of representation provided by the Calico toolkit.

Introduction

Asynchronous online discussion forums are used in a wide range of contexts in education. Analysing and building representations for the large amount of data underlying these forums often requires sophisticated methods and tools (Hrastinski & Keller, 2007). Several techniques like social network analysis (de Laat *et al.*, 2007), text mining (Fujitani *et al.*, 2003) or data mining (Romero & Ventura, 2007) have been used to extract indicators and visualize results that are significant for participants. Rosé *et al.* (2008) provide a review on automatic collaborative learning processing.

Bratitsis & Dimitracopoulou (2007) developed the DIAS discussion forum system with several integrated interaction tools that offer a wide range of indicators to all discussion users: students, tutors, teachers and researchers alike. Li *et al.* (2007) propose a multidimensional analysis framework that supports interaction analysis, text analysis and social network analysis. But in those systems like DIAS or Knowledge Forum for instance, the analysis tools are only available through the platform that supports the discussion forum itself.

Comparing analyses of forums coming from various contexts remains a difficult task. One challenge lies in the fact that applying various tools to the same forums requires a lot of transformations. To facilitate comparisons in content analysis, Law *et al.* (2007) argue for a unified toolset for analysing CSCL stressing on the fact that "the tools for different analysis are not integrated so that lots of time are wasted in transforming data into different formats for the different analysis". Offering new open services and tools for the sharing, the exploration and the comparison of fora is at the core of the Calico initiative presented in this paper.

One major issue of the Calico project is to make forums easier to read, explore and analyse. In this perspective, a shared workspace has been developed with novel tools that propose several ways to display the contents of a forum, to compute quantitative and qualitative indicators about authors, interactions, topics and to offer new ways to display global or local information about a forum.

The Calico shared space

The Calico research network associates 4 research laboratories and 6 colleges of education with the goal of developing a better understanding of distant collaborative learning and providing tools for researchers and practitioners for better management study and analysis of discussion fora. The main purpose of this network is to share data, methodology, needs, tools, and analyses between researchers and teachers, allowing different views on content and interaction analysis.

The Calico website offers a shared space dedicated to researchers and practitioners for analyzing "Computer Mediated Communication" (CMC) objects. It was originally created for sharing discussion forum objects from e-learning students' platforms but it now handles other communication objects such as mailing lists.

The Calico website (http://www.crashdump.net/calico/) is a CMC object sharing website where users can upload, view, study and share CMC objects. Unregistered users can watch and study public anonymous CMC objects on the site, while registered users are permitted to upload an unlimited number of objects. Some objects are available only for the Calico special interest group, while private CMC objects are strictly available for their owner.

Sharing and exchanging CMC objects raises various difficulties including legal and technical aspects of sharing such documents. The privacy aspects have also to be considered. The Calico website provides light anonymisation features for discussion fora. Full automated anonymisation is not provided since it may transform significant parts of messages (Reffay & Teutsch, 2007).

The technical issues of sharing and exchanging CMC objects stand in three points: the specification of exchange formats, the management of dynamic sources, the management of large sources.

Sharing and exchanging CMC objects is a new need. Standard exchange format did not exist when we started this work in early 2000's. Most platforms handle their own data format and few ones include export

techniques. The Calico and Mulce (Multimodal Learning Corpus Exchange, http://mulce.univ-fcomte.fr/axescient.htm) initiatives started the design of an exchange format for such data quite simultaneously. While the Calico XML exchange format, named XmlForum, allows the representation of discussion forum objects, the Mulce XML exchange format allows the representation of general CMC objects and includes detailed meta-information (Reffay *et al.*, 2008).

The XmlForum exchange format proposal has been designed by B. Huyn Kim Bang and E. Giguet in 2005. The format is quite simple and figure 1 illustrates its structure.

```
− <forum name="OS Lounge">
 − <message id="146">
   − <header>
       <author>Mike C.</author>
       <datetime>21/08/2007 11:22</datetime>
       <subject>The picture thread</subject>
     </header>
   − <body>
     − <span class="postbody">
         Please post your picture here so that I can match your name and face.
         <br/>
         [...]
       </span>
     </body>
   </message>
 − <message id="147">
   − <header>
       <msgref id="146"/>
       <author>Alex L.</author>
       <datetime>21/08/2007 15:04</datetime>
       <subject>Alex L.</subject>
     </header>
   − <body>
     − <span class="postbody">
         This is me and Johnny C. at the Rio in Vegas before WSOP.
         <br/>
```

Figure 1. This anonymous forum excerpt, written in XmlForum format, shows two timestamped posts. The second message, posted by Alex L. on 2007/08/21 at 15h04. refers to the first one, posted by Mike C. at 11h22, initiating the thread.

Sharing and Exchanging CMC Objects with the Calico website

The design of an exchange format is mandatory but still does not solve the whole exchanging and sharing problem. Exchanging and sharing discussion objects requires the conversion of discussion forum to XmlForum. This is a key point. The conversion is achieved by connectors or spiders that translate e-learning platform proprietary format to XmlForum open format. We already built up two *spiders*, one for discussion objects coming from *BSCW* (*Basic Support for Cooperative Work*) platforms (http://www.bscw.de/) and another one for discussion objects powered by *phpBB*, one of the most commonly used open solution (http://www.phpbb.com/), and one *connector* for ASPFRM discussion objects coming from the *DIAS* system (Bratitsis & Dimitracopoulou, 2007). C. Reffay designed a connector that converts discussion coming from *Mulce* environment. A converter for the *Moodle* forum module (http://moodle.org), and a spider for *Sympa* mailing list (http://www.sympa.org/) should be released in a near future.

Exploring CMC Objects with the Calico toolkit

The Calico toolkit is made of multilingual, user-centered, exploration tools dedicated to CMC Objects. Prior to the exploration, the CMC Objects must have been uploaded on the Calico website. In the following sections, we will focus on five components of this toolkit.

Reading and Filtering CMC Objects with ShowForum

The toolkit includes a basic tool named ShowForum to display and read all available information related to a forum. The messages can be displayed with two layouts : the list layout, and the thread layout (see figure 2). The list layout simply displays the message in chronological order. The thread layout displays the message by threads, in chronological order of the initial message of the thread. Each thread is shown as a “tree”, focusing on the relation between posts and replies. A feature allows dynamic anonymization of the author name mentioned in the header of the messages. This feature is shared with the other exploration tools.

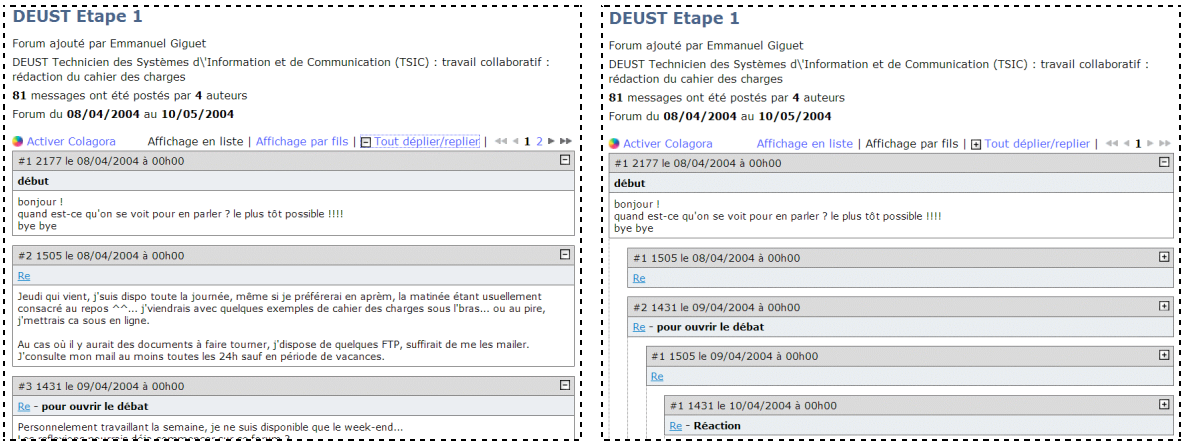

Figure 2. ShowForum displays discussion forum with two layouts: the list layout and the thread layout

Preferences allows the user to filter the messages according to several combined criteria: authors, date range and thread subjects. These preferences will be considered by all the other exploration tools so that analysis can be performed on different views of interest.

Creating chronological thumbnails of CMC Objects with Anagora

Anagora provides a graph representation to visualize overlapping discussion threads over time on a single screen. Its special feature is to calculate the best resolution for a forum to fit on a screen by choosing the most appropriate time scale according to data (Giguet & Lucas, 2009).

Anagora highlights high activity in a forum, through discussion overlap (see figure 3). A discussion thread (on the same topic) is shown as a red block (here in darkgray), horizontally spreading according to its duration, and vertically spreading according to its number of messages. When clicking on a block, the title of the thread and the number of contributors appear along with dates. Discussion threads are displayed on rows called chronograms. There are as many chronograms as overlapping discussions. In figure 1 there are at most 4 ongoing discussion threads at the same time, during the first decade. Chronograms are stacked, with the first chronogram placed at the bottom of the screen, simultaneous overlapping discussion threads are placed above.

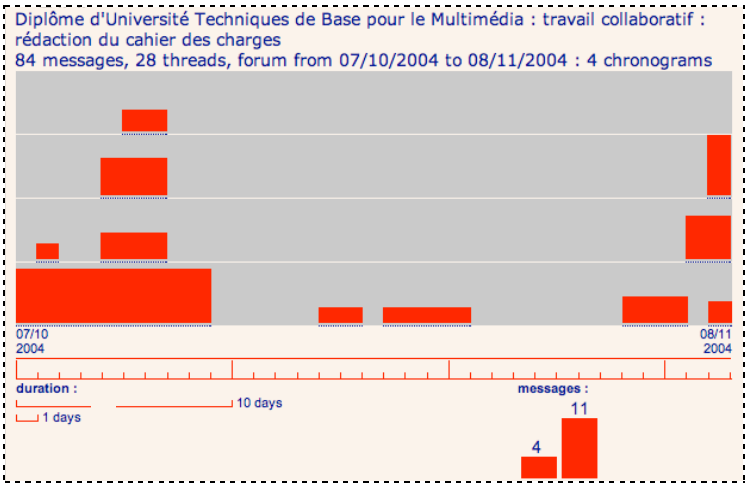

Figure 3. Anagora chronograms for a small group (DUTBM task 1) with typical peaks of simultaneous discussions at start and before the end of the forum

Focusing on the CMC Objects content structure with Themagora

Themagora parses discussion forum objects as collective discourse directed either by a specific task achievement goal or by a communication goal. It is a multilingual robust tool, providing "forum tiling" but also "discussion nesting". It uses the differences in the structure of messages along with stylistic statistical data to segment and organize the content of discourse. No external resources are needed. The output of this adaptive parser is a scalable view on collective discussion. Such views are used to browse and navigate in large forums, or to compare discussion progress between small learners groups (Lucas & Giguet, 2008).

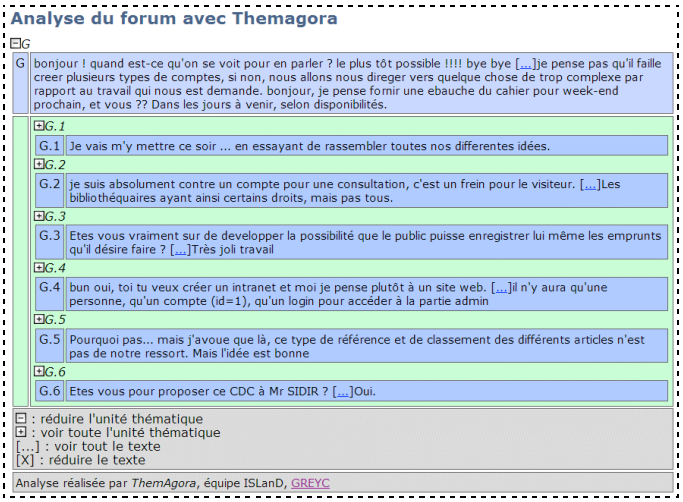

Figure 3. Themagora provides a scalable view of the forum, based on "forum tiling" and "discussion nesting"

Building and Locating topics in CMC Objects with Colagora

Colagora is a tool based on ThemeEditor (Beust, 2002). It allows the user to build up lexical topics, according to his interest. Then it allows to explore the forum through these highlighting filters.

Topics of interest are defined with simple word lists. These word lists are either uploaded in the Calico website, or defined interactively from the forum lexicon. Colagora is directed by the user's needs as far as lexicon is concerned: the tool counts word occurrences and displays the whole word list sorted by frequency or by alphabetic order. The user can defined word topics with the words that truly appear in the forum, and with their existing variation due to derivation or misspelling. Colagora then highlights every matching word of the forum with the color linked to the topic (see figure 4a below).

Like discussion forum objects, topic objects can be shared with other members. When topic objects are selected, they are used by other tools to colour their own representation (see below Bobinette).

Time, threads and topics: Bobinette

Bobinette was first developed by Huynh Kim Bang and Bruillard (2005) to solve reading problems occurring with classical forum interfaces. Bobinette offers both global and local views of a forum. It uses the chronological axis to display beads (representing posts) on a thread of discussion drawn as an horizontal line. Simultaneous threads are represented as additional lines below the first one. The main topic can be visualised by colour, and special messages, typically starters containing many questions, are highlighted by a question mark. Any post can be clicked open for closer reading.

Bobinette has the capability to compute statistics about word topics for a forum and for each post. The content of selected posts can be displayed and the topic words highlighted. (see figure 4b).

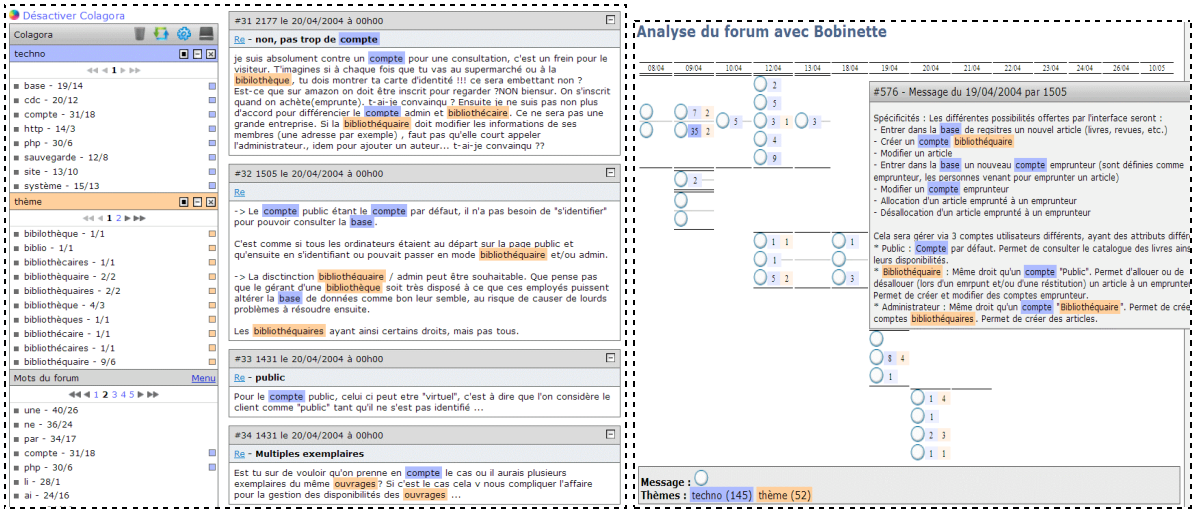

Figure 4. a) Colagora helps defining topics interactively, facilitating the handling of misspelled words. b) Bobinette offers global and local views and uses active Colagora topics to highlight words of interest

Discussion

One major result of the Calico website is to make forum objects easier to share and explore in education. Thus, the website contributes to the dialogue between researchers and practitioners. The Calico shared workspace includes novel tools that propose several ways to display the contents of a forum, calculate quantitative and qualitative indicators about authors, interactions, topics and offer new ways to display global or local information about a forum.

The Calico website hosts about 50 CMC Objects, including discussion forum objects and mailing list objects in four languages (French, English, Greek and Vietnamese). These objects were uploaded by members of the Calico network and by other researchers, using different e-learning platforms. About 45 different forums have already been uploaded, the smaller counts 12 posts (6 authors) and the larger 545 posts (248 threads, 74 authors). Mailing lists have been uploaded and reach 926 posts, 365 threads, 37 authors.

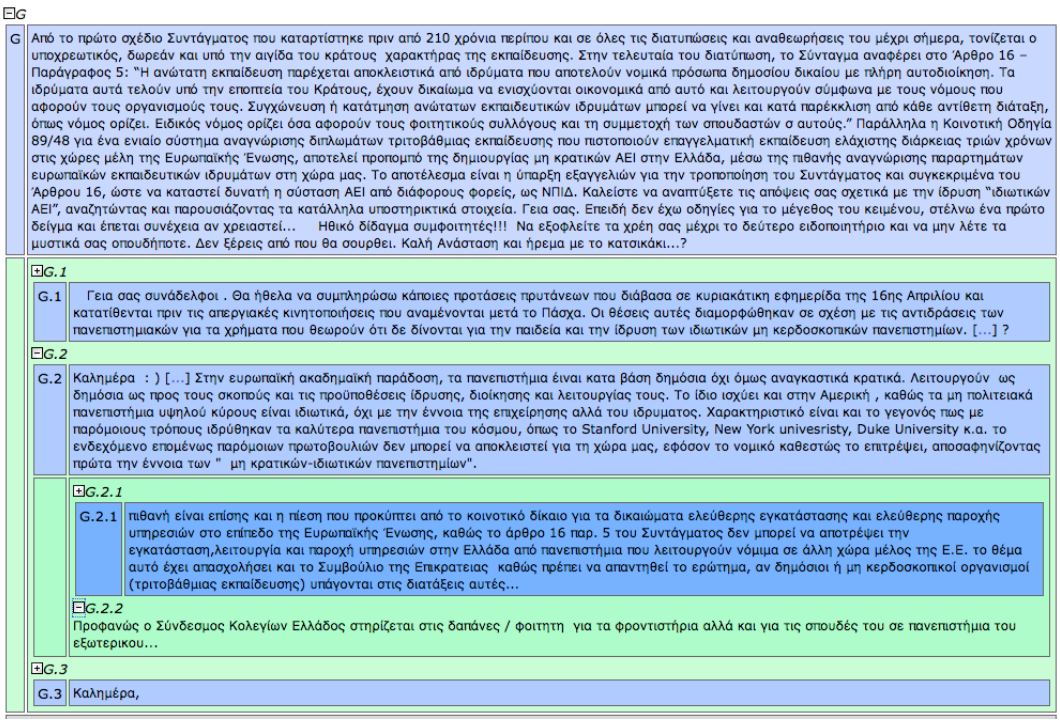

Figure 5. The Calico platform handles Greek here with Themagora

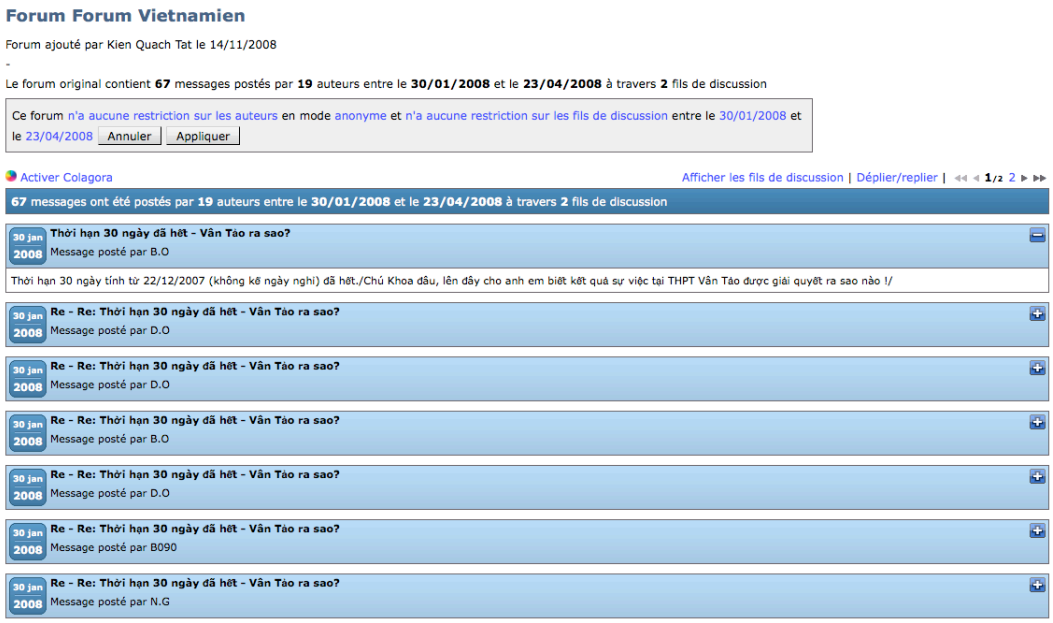

Figure 6. The Calico platform handles Vietnamese here with Showforum

These promising results are due to a pragmatic XML exchange format, named XmlFormat, combined with the availability of connectors and spiders to convert from online discussion software (e.g., BSCW, phpBB).

Perspective

We now consider the extension to other CMC objects, for instance, chat rooms, instant messaging, blogs. Other languages are also considered: exploring forum objects in Turk and Arabic is under way.

The platform should now handle larger objects, scaling to 100.000 posts since we start a new collaboration that implies the management of discussions within a community of several hundred of maths teachers over 3 years, representing approximately 30.000 messages per year on 30 discussion forum objects.

The improvement of the existing tools will focus on higher interactivity of the displays, better management of scalability, multilinguism, while allowing comparison between forum objects or forum excerpts.

We look forward to exchange more discussion objects to improve language coverage and also to discuss ways of improving analysis, interface by relying on colleagues expertise.

References

Beust P. (2002). Un outil de coloriage de corpus pour la représentation de thèmes. In JADT 2002 : 6èmes Journées internationales d'Analyse statistique des Données Textuelles. Saint Malo, France.

Bratitsis, T., & Dimitracopoulou A. (2007). Interaction Analysis in Asynchronous Discussions: Lessons learned on the learners' perspective, using the DIAS system. In C. A. Chinn, G. Erkens & S. Puntambekar (Eds.), Proceedings of the Computer Supported Collaborative Learning (CSCL) 2007, 87-89.

Fujitani, S., Mochizuki, T., Kato, H., Isshiki, Y., Yamauchi, Y. (2003). Development of Collaborative Learning Assessment Tool with Multivariate Analysis Applied to Electronic Discussion Forums. In G. Richards (Ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003*, 200-203.

Giguet, E. & Lucas, N. (2009). Creating discussion threads graphs with Anagora, *CSCL 2009*, Rhodes, ICLS.

Hrastinski, S., & Keller, C. (2007). Computer-mediated communication in education: A review of recent research. Educational Media International, 44(1), 61-77.

Huynh Kim Bang, B., & Bruillard, E. (2005). Vers une nouvelle interface de lecture pour des forums de discussion dédiés à des élaborations collectives. In Saleh I. & Clement T. (eds.), Créer, jouer, échanger, Actes e H2PTM'05, Paris: Lavoisier, p. 43-56.

de Laat, M., Lally, V., Lipponen, L. & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. In International Journal of Computer-Supported Collaborative Learning, Vol. 2, No. 1. March, 87-103.

Law, N., Yuen, J., Huang, R., Li, Y. & Pan, N. (2007). A Learnable Content & Participation Analysis Toolkit for Assessing CSCL Learning Outcomes and Processes. In C. A. Chinn, G. Erkens & S. Puntambekar (Eds.), Proceedings of the Computer Supported Collaborative Learning (CSCL) 2007, 408-417.

Li, Y., Liao, J., Wang, J., & Huang, R. (2007). CSCL Interaction Analysis for Assessing Knowledge Building Outcomes: Method and Tool, In C. A. Chinn, G. Erkens & S. Puntambekar (Eds.), *Proceedings of the Computer Supported Collaborative Learning (CSCL) 2007: Mice, Minds and Society*, 428-437.

Lucas, N. & Giguet, E. 2008. Robust adaptive discourse parsing for e-learning fora. In "The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008)", 730-732, Spain.

Reffay, C., Chanier, T., Noras, M., & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. STICEF, 15. www.sticef.org

Reffay, C., Teutsch, Ph. (2007). Anonymisation de corpus réutilisables. http://edutice.archives-ouvertes.fr/edutice-00158877

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A. & Fischer, F. (2008) Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in CSCL. International Journal of Computer-Supported Collaborative Learning 3 (3), 237-271.

Acknowledgments

This work is funded by the French ministry of education, research and technology. The design of the Calico user interface and tools has benefited from helpful discussion with all participants of the Calico network.