S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Comprehensive Computational Support for Collaborative Learning from Writing

**Peter REIMANN[a]* , Rafael CALVO[b], Kalina YACEF[c] &
Vilaythong SOUTHAVILAY[c]**
[a]*Faculty of Education and Social Work, The University of Sydney, Australia*
[b]*School of Electrical and Information Engineering, The University of Sydney, Australia*
[c]*School of Information Technologies, The University of Sydney, Australia*
*peter.reimann@sydney.edu.au

Abstract: Learning about subject matter and about writing by collaboratively authoring an electronic document is an important variant of computer-supported collaborative learning. Collaborative writing is particularly often practiced in Higher Education. Our research has the goal to develop comprehensive software support tools for collaborative discipline-based writing, and to study how the team writing process is affected by the use of these tools. This paper describes initial tool developments that integrate computational document analysis methods with process mining methods into a comprehensive writing environment and reports first experiences gained in a undergraduate engineering course.

Keywords: CSCL, writing, text mining, process mining, formative feedback.

Introduction

Writing can be an important form of learning, both of writing itself and of subject matter [1, 2]. We are particularly interested in collaborative forms of writing (for the purpose of learning), which fall into two main categories: *Peer reviewing* where the outcome is an individual document that has been composed by one student and has been reviewed by at least one other student (once or repeatedly), and *collaborative writing* where the outcome is a collaboratively composed and revised document. Collaborative writing (CW), defined by Lowry *et al.* [3] as "..an iterative and social process that involves a team focused on a common objective that negotiates, coordinates, and communicates during the creation of a common document" is a cognitively and organizationally demanding process. As a specialized form of group work it involves a broad range of group activities, multiple roles, and subtasks. When performed by groups that communicate (partially or only) through communication media, the process typically involves, in addition, multiple tools (e.g. phone, mail, instant messaging, document management systems) with different use characteristics.

In CSCL, writing has been seen as a means to deepen students' engagement with ideas and the literature and for knowledge building [4] by jointly developing a text or hypertext. In addition to knowledge building in asynchronous collaboration, synchronous collaborative development of argumentative structures and texts has received much attention [e.g. 5].

The availability of the Internet has made both peer and collaborative writing very easy to implement in schools and universities, and has led to genuinely new forms of writing, such as blogging and wiki writing [6]. In recent years, the rise of so-called 'cloud computing' tools such as Google Documents has led to the availability of almost desktop-quality online writing environments with very little costs to the user. The widespread availability of high-quality technical CS tools does not mean, however, that writing is now performed better. As the use of word processors has not led to better individual writing, so does the availability of wikis engines and Cloud tools not lead by itself to better documents and cooperation, deeper learning, or more satisfaction with the writing process.

Because of the complexity of the CW process, explicit and scaffolded support needs to be provided, in particular for novice writers. Such support generally falls into one of three classes: specialized writing and document management tools, document analysis software, and team process support. Our research focuses on the latter two, the first is provided by commercial vendors (e.g. Google) who provide the writing tools and store the documents written by students. We conjecture that in order to support students in Higher Education effectively in writing together and learning together from the writing process, computational support that has so far been separated should be combined: Namely tools that provide feedback on the *product* of the writing process (drafts) should be combined with tools that can provide visualisations of and feedback on the *team process*.

Accordingly, we combine two computational techniques: semantic analysis, which focuses on extracting knowledge from documents about what the student wrote (or edited) and process mining, which focuses on extracting process-related knowledge from event logs recorded by an information system. In this paper, we describe the architecture of our comprehensive writing environment (CWE) and provide examples for first experiences with an initial implementation.

1. Architecture of the Comprehensive Writing Environment

The CWE framework (see Figure 1) integrates a front-end writing tool that supports collaborative writing activities (manages access writes etc.) and stores all revisions of documents created, shared and edited by groups of writers. with tools for document and process analysis. (Two additional components, a writing assignment management tool for large courses and a automatic question generation tool also exist in first implementations, but are not described here, see [7].) In order to perform analysis of the writing process for particular documents, each revision of a particular document must contain information such as edited text, timestamp of committing change, and identification of the writer. Based on the information such as timestamp and writers' identification of all revisions and event logs of reviewing activities, a process mining tool is used to discover sequence patterns of writing activities (WriteProc). The process analysis provides a way to extract knowledge about writers' interaction and cooperation. The analysis can identify interactions' patterns that lead to a positive outcome and indicate patterns that may lead to problems. In addition, text mining techniques are applied to analyze text-based changes of all revisions of documents (Glosser). The text-based analyses can provide semantic meaning of changes in order to gain insight into how writers develop ideas and concepts during the writing process.

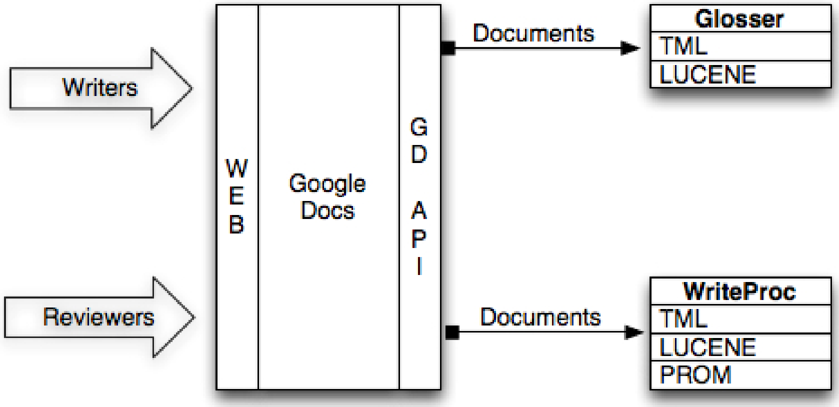
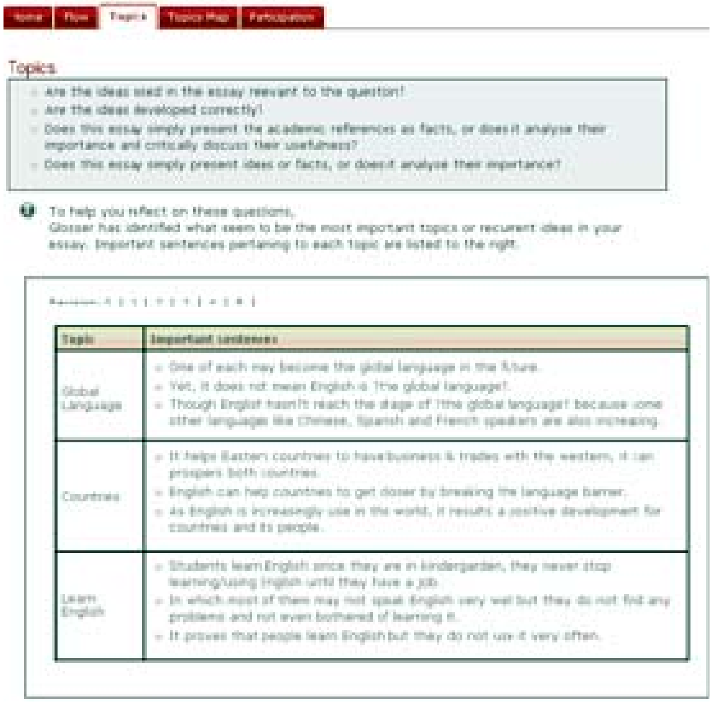


Figure 1: Architecture of WriteProc, a framework for collaborative writing support.

As already mentioned, the front-end writing tool in CWE is Google Docs, a web-based utility with most functionalities for word processing that allows users to share their documents with other team members and to write (almost) synchronously. The Google Document Lists Data API (GDAPI) is used to integrate Google docs into the WriteProc system as shown in Figure 1. The API allows CWE to retrieve and track all versions of documents created, shared and edited among group members. Every time a writer makes changes and edits a particular document, the identification of the writer, the edited content of the document, the timestamp of committing changes and the version number of the edited document are retrieved and stored in CWE's central relational database by using the API. This information extraction is executed seamlessly in the background so that the writers are not interrupted. Using these records, CWE performs document analysis in order to provide feedback on certain aspects of a document (Glosser) and performs process analysis to provide information on the collaboration process (WriteProc).

Glosser is a web-based tools that uses some grammatical but mainly statistical techniques to analyse a document (each version) with respect to parameters such as topics included, relationship between the topics, coherence between paragraphs [8]. The feedback provided by Glosser helps a student to review a document by highlighting the features a document communicates, such as the keywords and topics it includes, and the flow of paragraphs. For the case of collaborative writing, by analysing the content and author of each document revision, it is possible to determine which author contributed which sentence or paragraph and how these contribute to the overall topics of the document. These collaborative features of Glosser can help a team understand how each member is participating in the writing process. The user interface of the Topics feedback tool in Glosser is displayed in Figure 2. The trigger questions at the top of each page are provided to help the reader focus their evaluation on different features of the document. Below the questions is the supportive content called ‘gloss’, to help the reader answer those questions. The ‘gloss’ is the important feature that Glosser has highlighted in the document for reflection. A rollover window on each sentence indicates who wrote it.



Topics

- Are the ideas used in the essay relevant to the question?
- Are the ideas developed correctly?
- Does this essay simply present the academic references as facts, or does it analyse their importance and critically discuss their usefulness?
- Does this essay simply present ideas or facts, or does it analyse their importance?

To help you reflect on these questions,
Glosser has identified what seem to be the most important topics or recurrent ideas in your essay. Important sentences pertaining to each topic are listed to the right.

Revision: [illegible]

Topic	Important sentences
Global Language	One of each may become the global language in the future. Yet, it does not mean English is ?the global language?. Though Englist hasn't reach the stage of ?the global language? because some other languages like Chinese, Spanish and French speakers are also increasing.
Countries	It helps Eastern countries to have business & trades with the western, it can prospers both countries. English can help countries to get closer by breaking the language barrier. As English is increasingly use in the world, it results a positive development for countries and its people.
Learn English	Students learn English since they are in kindergarden, they never stop learning/using English until they have a job. In which most of them may not speak English very well but they do not find any problems and not even bothered of learning it. It proves that people learn English but they do not use it very often.

Figure 2: Glosser screen displaying information on topics found in a document. The upper box shows reflection prompts, the lower table is part of the output from Glosser with main topics found in the text displayed in order of their importance.

WriteProc uses a combination of text statistical techniques and process mining techniques to extract information about the mining process from document changes as well as event logs capturing user behavior. WriteProc is currently under development and will eventually comprise a process mining component plus a module that will provide process visualizations for students. The (web-based) visualization module is not developed yet. For the analytical part of WriteProc we currently use PROM [9]; in its final form, WriteProc will use algorithms as contained in PROM, for instance, but made available as web services, independent of the PROM user interface. We describe WriteProc in the context of two case studies in more detail below.

Both Glosser and WriteProc use TML, a multipurpose text mining library (http://sourceforge.net/projects/tml-java/) that implements the natural language processing (NLP) and machine learning techniques that analyse the actual content of the document revisions. TML provides a comprehensive set of text mining algorithms and scaffolds every stage of the text mining process. TML integrates the open source Apache Lucene search engine, the Stanford NLP parser and the Weka machine learning libraries, and is itself open source. TML provides functionalities for the pre-processing of documents, tokenising, stemming and stop-word removal.

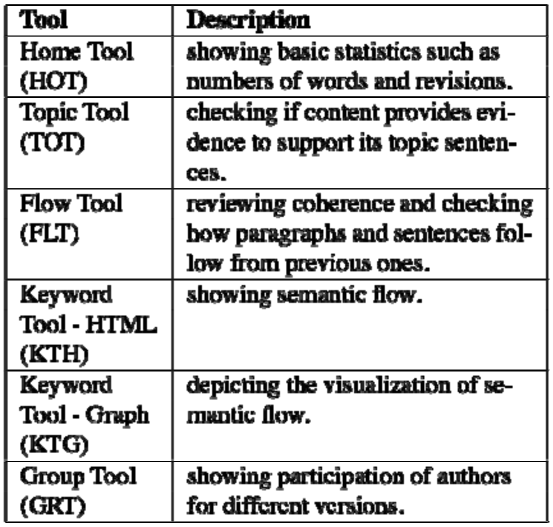
2. Sequence and Process Analysis of Collaborative Writing

With sequence and process analysis methods one can uncover regularities ("patterns") contained in information pertaining to temporal order and duration of events. From a learning perspective, it is particularly interesting to find out if there are sequence-dependent regularities in data that correlate with measures such as learning gains or motivational aspects, such satisfaction with group work. In order to demonstrate how such forms of analysis can be integrated into an online writing environment, we illustrate the use with two case studies.

2.1 Analysis of Glosser Logs

The data come from 58 engineering students enrolled in a course on e-business. In pairs of two they had to write a Project Specification Documents (PSD) for their proposed e-business projects. Each pair had to submit one PSD of between 1,500 and 2,000 words. Students were required to write their PSD on Google Docs and share the documents with the course instructor. They were asked to submit their PSD using Glosser. Two other students who were members of different groups reviewed the submitted PSD. Students had one week to review each other's documents and submit their feedback. After getting feedback on their documents from their peers, students could revise and improve their writing if necessary before submitting the final version one week later. Before the submission of the final version they also used Glosser. The total event log file of the system consisted of usage data of Google Docs and Glosser for three weeks. In addition to this log file, the marks of the final submissions of the PSD together with a very good understanding of the quality of each pair through the semester was used to correlate behaviour patterns to outcomes.

The event data type analysed by us were the choice of review tools in Glosser, each of which corresponds to a tab in the interface that can be opened--and the underlying analysis activated--by clicking on the respective tab (see Figure 3 for descriptions). The question was if there is any systematic relation between patterns of use of these review tools and the resulting quality of the document (expressed by grading).



Tool	Description
Home Tool (HOT)	showing basic statistics such as numbers of words and revisions.
Topic Tool (TOT)	checking if content provides evidence to support its topic sentences.
Flow Tool (FLT)	reviewing coherence and checking how paragraphs and sentences follow from previous ones.
Keyword Tool - HTML (KTH)	showing semantic flow.
Keyword Tool - Graph (KTG)	depicting the visualization of semantic flow.
Group Tool (GRT)	showing participation of authors for different versions.

Figure 3: Review tools available in Glosser

The event corpus analysed comprises 4,677 events logged on students' work on 29 documents. The development of each of these documents was treated as a process case, and we distinguished eight event types: use of each the six review tools displayed in Figure 3, opening the Google document (ROD), and accessing the review tool (TOR).

From the event log of our case study data, we extracted the process model shown in Figure 4, which represents the process common to all the groups; we used the PROM Fuzzy Miner (see [10] for more details on this algorithm and an application to group decision making). Groups began with events of opening a particular document (ROD). Then, the reviewing tool was requested (TOR). After that, different reviewing activities were performed, in no particular order. The process reiterated until users logged off or closed their browsers.

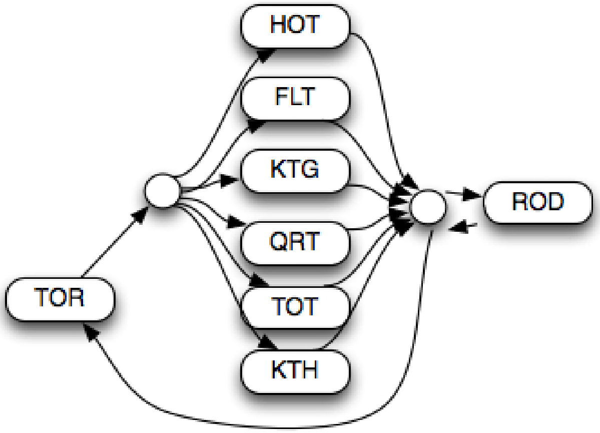


Figure 4: Process model for the use of reviewing tools across all pairs.

We were interested in finding out more about differences between groups and relations to success criteria. ProM provides a Performance Sequence Analysis plug-in to find the most frequent paths in an event log. (This algorithm also uses performance data, in particular duration of events, but we do not elaborate on this due to lack of space.) Figure 5 shows the most frequent pattern in terms of transitions between event types for Group 1 (received a mark of 8/10) at the top and Group 29 (10/10) at the bottom. As one sees, there are no dramatic differences between these two groups. In general, in these case study data there were no substantial differences between groups' behavior that were correlated with success criteria. Such differences were also not found when looking at frequency data. For more details, see [11].

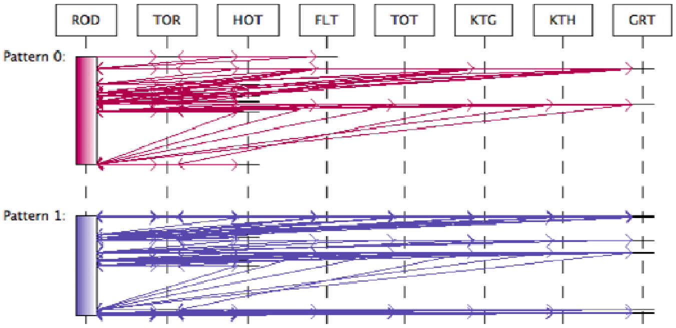
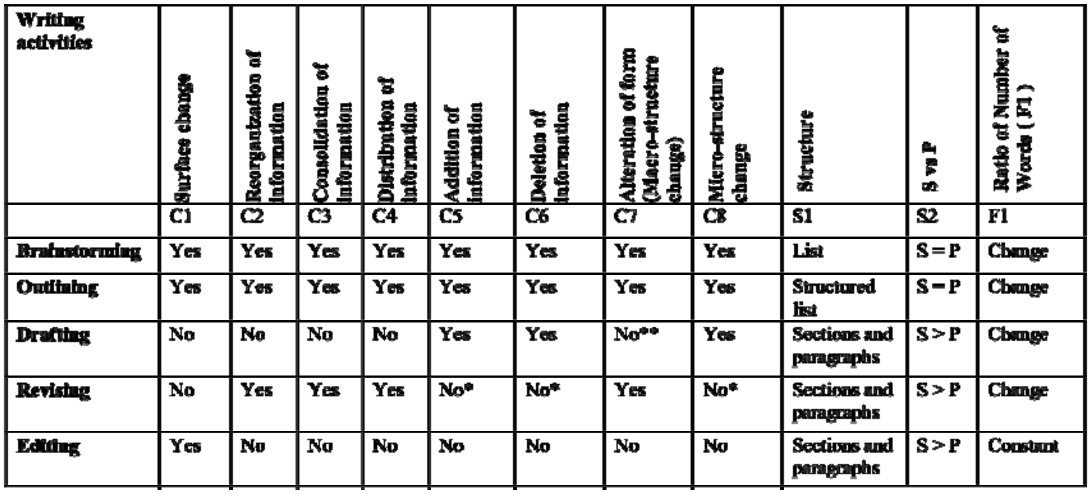


Figure 5: Performance Sequence diagram for Pair 1 (top) and Pair 29 (below)

2.2 Analysis of Google Logs in Terms of Team Writing Activities

While mining for sequence patterns in log file data may be informative for the researcher, the question remains to what extent this kind of information is also informative for the student (and teachers/lecturers, for that matter). We would contend that event-based log file visualizations are by and large not very useful as *feedback* to learners unless the visualized information pertains directly to pedagogically relevant processes. For the visualization described above, no pedagogical (prescriptive) model exists that would suggest any specific sequencing of reviewing activities, and hence the information as visualised has limited feedback value. We do, however, have prescriptive models for the larger process of group writing, for instance based on the taxonomy suggested in [3]: (Writing) teachers often formulate at least partial orders on sequences of Brainstorming, Outlining, Drafting, Revising, and Editing. Hence, process visualizations on this level would constitute potentially valuable feedback as students can compare their group's sequence with, for instance, an ideal writing sequence. In any case, describing behavior in such event categories that encompass the *semantics* of collaborative writing activities is inevitably more informative than a description of behavior sequences in terms of activities in the software interface (Glosser for instance).

Here we illustrate this point and demonstrate how the records stemming from writers' activities in Google Docs can be semantically interpreted. Data from the same students and course as in Section 2.1 were used, but this time we build on the database of records of activities in Google Docs. The analysis proceeded in multiple steps: First, after initial data cleaning the Google Docs data log (time-stamped versions of a document along with information which user performed the changes) was interpreted in terms of *individual* Writing Activities and their effects on a document (e.g. topic shifts, change of coherence, see the top row in Figure 6). For this, text-statistical methods were employed, using the TML library (see section 1) and methods of latent semantic indexing (we build in particular on work described in [12]). In a second step, these individual Writing Activities and document changes were related to *collaborative* writing activities (the Lowry taxonomy, see first column in Figure 6) by means of heuristics. The heuristics were implemented computationally, so that the heuristic mapping could be performed automatically [see 13 for details].



Writing activities	Surface change	Reorganization of information	Consolidation of information	Distribution of information	Addition of information	Deletion of information	Alternation of form (Macro-structure change)	Micro-structure change	Structure	S vs P	Ratio of Number of Words (F1)
	C1	C2	C3	C4	C5	C6	C7	C8	S1	S2	F1
Brainstorming	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	List	S = P	Change
Outlining	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Structured list	S = P	Change
Drafting	No	No	No	No	Yes	Yes	No**	Yes	Sections and paragraphs	S > P	Change
Revising	No	Yes	Yes	Yes	No*	No*	Yes	No*	Sections and paragraphs	S > P	Change
Editing	Yes	No	No	No	No	No	No	No	Sections and paragraphs	S > P	Constant

Figure 6: Heuristics that lead from changes in documents (columns) to identification of group writing phases

After this extensive pre-processing, we are left with an event sequence that we can interpret pedagogically, in the context of collaborative writing: for each document produced (by a pair of students in our case) we have a sequence of events in terms of the Lowry taxonomy, which we then can subject to sequence and process analysis, as sketched in Section 2.1.

3. Conclusions

Collaborative online writing provides CSCL researchers with rich data and at the same times comprises an important element of academic work and 'knowledge work' in general. We have demonstrated how building on a globally available 'cloud' writing tool (Google Docs) one can add analytical services that have the potential to support learning from writing. We have also demonstrated how one can move from log file visualizations to visualizations of process that capture the semantics of writing. On this level the notion of a holistic 'process' as different from a 'sequence of steps' becomes meaningful: teachers typically provide students with a sense of how the overall team writing process should proceed, linking all the elements of the taxonomy into a coherent whole, a students will, if things go well, strive to realize this process in their team work. (For more on the distinction between sequence and process in temporal data see [14]). How students react to these new affordances for learning will be the focus of our future studies.

Acknowledgements

This research was in parts supported by a Discovery Grant from the Australian Research Council and a TIES grant from the University of Sydney.

References

[1] Bereiter, C. and M. Scardamalia, *The psychology of written composition*. 1987, Mahwah, NJ: Lawrence Erlbaum Publishers.

[2] Galbraith, D., *Writing as a knowledge-constituting process*, in *Knowing what to write: conceptual processes in text production*, T. Torrance and D. Galbraith, Editors. 1999, Amsterdam University Press: Amsterdam. p. 139-150.

[3] Lowry, P.B., A. Curtis, and M.R. Lowry, *Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice.* Journal of Business Communication, 2004. **41**(1): p. 66-99.

[4] Scardamalia, M. and C. Bereiter, *Knowledge building: Theory, pedagogy, and technology*, in *The Cambridge Handbook of the Learning Sciences*, R.K. Sawyer, Editor. 2006, Cambride University Press: New York. p. 97-115.

[5] Amalesvoort van, M., J. Andriessen, and G. Kanselaar, *Representational tools in computer-supported colloborative argumentation-based learning: How dyads work with constructed and inspected argumentative diagrams.* Journal of the Learning Sciences, 2007. **16**(4): p. 485-521.

[6] Stern, S., *Producing Sites, Exploring Identities: Youth Online Authorship*, in *Youth, Identity, and Digital Media*, D. Buckingham, Editor. 2008, The MIT Press: Cambridge, MA. p. 95-117.

[7] Calvo, R.A., et al., *Collaborative writing support tools on the cloud*. manuscript under review.

[8] Villalon, J., et al., *Glosser: Enhanced feedback for student writing tasks*, in *8th IEEE International Conference on Advance Learning Technologies (ICALT) (Santadar, Spain, July 1-5, 2008; accepted March 2008)*. 2008: Santadar, Spain.

[9] Aalst, W.M.P.v.d., et al., *PROM4.0: Comprehensive support for real process analysis*, in *Petri nets and other models of concurrency (Proceedings 28th International Conference on Applications and Theory of Petri Nets and Other Models of Concurrency, ICATPN 2007, Siedcle, Poland, June 25-29, 2007)*, J. Kleijn and A. Yakovlev, Editors. 2007, Springer: Berlin. p. 484-494.

[10] Reimann, P., J. Frerejean, and K. Thompson, *Using process mining to identify models of group decision making processes in chat data*, in *Computer-supported collaborative learning practives: CSCL2009 conference proceedings*, C. O'Malley, et al., Editors. 2009, International Society for the Learning Sciences: Rhodes, Greece. p. 98-107.

[11] Southavilay, V., K. Yacef, and R.A. Calvo, *WriteProc: A framework for exploring collabortive writing processes*, in *Australian Document Computing Symposium*. 2009: Sydney, Australia.

[12] Villalon, J. and R.A. Calvo, *Single document semantic spaces*, in *The Australian Data Mining Conference*. 2009.

[13] Southavilay, V., K. Yacef, and R.A. Calvo, *Process mining to support students' collaborative writing*, in *Educational Data Mining Conference*. 2010: Pittsburgh, PA.

[14] Reimann, P., *Time is precious: Variable- and event-centred approaches to process analysis in CSCL research.* International Journal of Computer-supported Collaborative Learning, 2009. **4**: p. 239-257.