McLaren, B.M., Wegerif, R., Mikšátko, J., Scheuer, O., Chamrada, M., & Mansour, N. (2009). Are Your Students Working Creatively Together? Automatically Recognizing Creative Turns in Student e-Discussions. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-09)*, Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling. (pp. 317-324). IOS Press.

Are Your Students Working Creatively Together? Automatically Recognizing Creative Turns in Student e-Discussions

Bruce M. MCLAREN[a,b1], Rupert WEGERIF[c], Jan MIKŠÁTKO[a], Oliver SCHEUER[a], Marian CHAMRADA[c], Nasser MANSOUR[c]
[a] *Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany*
[b] *Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[c] *University of Exeter, Graduate School of Education, St Luke's, Exeter, U.K.*

Abstract. In this paper, we discuss how Artificial Intelligence (AI) techniques might be brought to bear in automatically recognizing "creative reasoning" in student e-discussions. An AI-based graph-matching algorithm was used to find instances of *deepening* and *widening*, interactional categories that provide evidence of, respectively, explicit argumentation and creative reasoning. A deepening occurs when students provide further argumentation for an on-going perspective. A widening occurs, on the other hand, when a student (or students) attempts to diverge from the current perspective by either questioning it or presenting a new perspective or new idea. Given examples of deepening and widening from real e-discussions, the AI algorithm was able to successfully find similar events within new e-discussions and did so within realistic run-time expectations. Our ultimate aim is to provide a software tool for teachers that will support them in recognizing a range of important dialogic aspects of student e-discussions, such as deepening and widening.

Keywords. Creative Reasoning, Collaborative Learning, Artificial Intelligence Analysis Techniques, Shallow Text Processing

1. Introduction

An emerging trend in education is for students to use collaborative computer-based tools to discuss, debate, and argue with one another about topics presented in a classroom [1], [2]. With these tools students work on separate computers but communicate in a shared workspace, through dragging-and-dropping different types of discussion objects (e.g., "claim", "question"), filling those objects with text (e.g., "I disagree with that claim because …") and linking the objects to other objects using meaningful links (e.g., "supports", "opposes") (cf. [3]). On the ARGUNAUT project [4], we are using the tools Digalo [5] and FreeStyler [6] (http://www.collide.info/index.php/FreeStyler/) to allow students to engage in such graphical e-discussions. In order to allow a teacher to moderate the discussions, we have developed a Moderator's Interface (See Figure 1) that contains alerts (e.g., whether students are swearing) and a variety of awareness indicators (e.g., the relative amount of participation by each student in a discussion).

One type of alert is a so-called "deep" alert, which indicates complex interactions between students in e-discussions (called "clusters"), such as a sequence of back-and-

[1] Corresponding Author; email address: bmclaren@cs.cmu.edu.

forth argumentation. An AI-based graph-matching algorithm, DOCE (Detection of Clusters by Example), has been developed to support such alerts [7]. For instance, in Figure 1, the Moderator's Interface shows the top five "Chain of Opposition" clusters (indicated by the small circles shown in the geometric center of the clusters) in a particular e-discussion. The Chain of Opposition cluster is three or more shapes in length and involves, typically, two students arguing back and forth, each opposing the other's argument. In Figure 1, the teacher has highlighted one of the specific Chain of Opposition clusters for closer inspection.

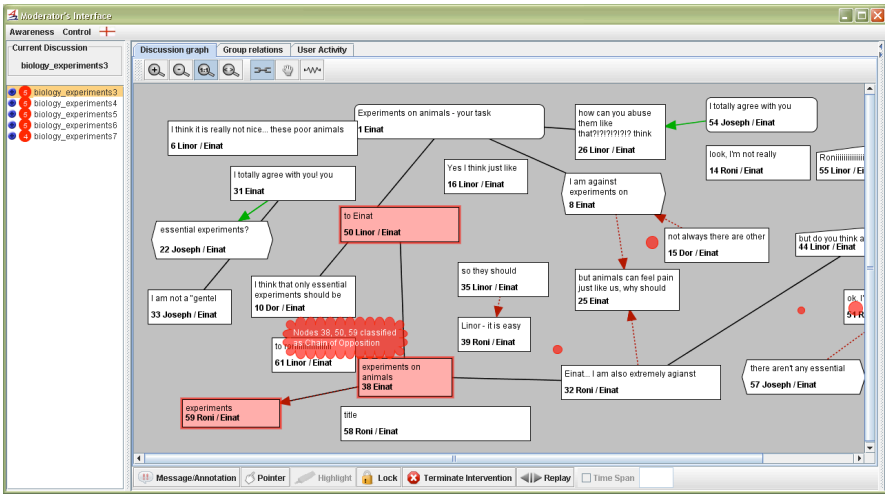

Figure 1: ARGUNAUT's Moderator's Interface showing one of its deep alerts, 'Chain of Opposition', in a selected e-discussion, which is selected from the current active e-discussions shown on the left

Two of the cluster types that are searched for by DOCE, *deepening* and *widening*, are interactional categories that we were particularly interested in identifying in classroom discussions as evidence of, respectively, explicit argumentation and creative reasoning [8]. A *deepening* occurs when students provide further argumentation for an on-going perspective [9], similar in concept to the "Chain of Opposition" clusters. A *widening* occurs, on the other hand, when a student (or students) attempts to diverge from the current perspective by either questioning it or presenting a new perspective [8]. In this paper, we report on some success we had in having our algorithm find such clusters in new e-discussions.

2. A Research Challenge: How Do We Find and Code Creative Reasoning?

In a recent review of the literature, Andriessen [10] presents developments in argumentation theory as moving from abstract and formal studies towards an analysis of actual human dialogues. Most schemes applied to analyzing online argumentation (e.g. those developed originally by Toulmin, Van Eemeren and Walton, see [10] for details) focus on explicit reasoning in the form of claims, challenges to claims and reasons in support of claims. This approach is good at picking up *critical* reasoning but ignores more *creative* reasoning and interaction. In dialogues voices interact in unpredictable ways to produce new perspectives that can enable participants to see the topic of the dialogue in a new way. Such dialogic shared thinking, as a dance of voices and perspectives, is clearly an act of creative reasoning.

So how are we to find such creative reasoning in real discussions? If by widening we mean bringing in new perspectives that enables the participants in a dialogue to see things in a new way and thus expand their understanding, then any widening move in a debate is also a *creative* move[2]. Deepening, on the other hand, is about unpacking assumptions and following chains of entailment and so broadly coincides with the traditional focus on explicit reasoning. With this in mind, we developed a coding scheme both for research purposes and to provide a basis (and data) for the AI techniques discussed later. The scheme included the more traditional focus on explicit reasoning but also looked for the taking of perspectives and the listening to different perspectives in a way that allows for the emergence of creative new perspectives (i.e., insights) that expand the dialogue without necessarily resolving a given problem [11].

To help us code the rather complex online discussions produced in the course of the ARGUNAUT project (called "e-discussions" or "discussion maps" henceforth) we generated *sequence diagrams*, visual representations of e-discussions that serve the purpose of providing an abstracted overview of: (1) the number and length of contribution sequences and (2) the branching of sequences at different points in the discussion. By "zooming in" on the branch points (i.e., by inspecting the cluster of contributions around and including a branching step), we found that the branches often mapped to new perspectives (but not always creative contributions, as per footnote 2) in the e-discussions.

For instance, see Figure 2. In this small cluster of contributions from a discussion map, found using the technique described above, an instance of a creative, new perspective was found. A group of three students is discussing the question: 'Will the Internet bring the world together or deepen its divisions?' The "new perspective" emerges when one student (see the shape in the upper left, the branch point found in a sequence diagram) suggests that the "internet, in a way, is an extension of ourselves" and reflects "our own personalities and self concept." This is a new and unexpected perspective and prompts two other related contributions (upper right and bottom) that continue and expand on this line of discussion.

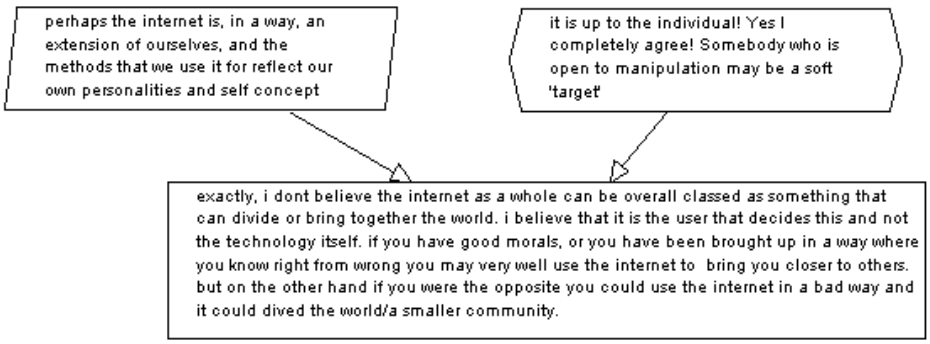

Figure 2: The emergence of a new perspective (upper left) by a student and responses to it by other students

[2] Note, however, that our definition of "widening" excludes new perspectives that lead to non-creative discussion. For example, students might make off-point comments that lead to "new perspectives" but such new lines of discussion are rarely creative and are not labeled as widening in our scheme. Of course, some subjective judgment is involved in defining creativity when a widening occurs in an e-discussion.

Inspection of the abstract sequence diagrams, followed by "zooming in" on branches, such as what is shown in Figure 2, allowed us to somewhat rapidly find and code instances of widening in a variety of maps, more specifically, in a set of 14 e-discussions created by undergraduates and post-graduate students in the UK. In the following sections we describe the DOCE algorithm and the results we obtained in applying the algorithm to these annotated maps.

3. DOCE: Automatically Finding Instances of Deepening and Widening

To search for clusters of interesting student interactions, such as the widening cluster illustrated in Figure 2, we developed an algorithm called ***DOCE*** (***D***etection ***o***f ***C***lusters by ***E***xample) [7]. DOCE is based on the idea of using cluster *examples* to find similar clusters in new discussions, similar to the ideas from case-based reasoning [12], [13]. DOCE operates either by a teacher or researcher selecting an example cluster in a current e-discussion (e.g. connected individual contributions that provide a good example of a particular interaction of interest) or by selecting a pre-defined cluster type (e.g., "widening", "deepening"), which, in turn, loads relevant pre-saved cluster examples from a database. The example cluster(s) (also called a "model graph" in the following text) is then used as a search query for similar clusters across other discussion maps (called "input graphs"). The algorithm uses both structural features (e.g., the types of contributions made by students – for instance, "claim" or "question" – and types of links between contributions – for instance, "supporting" or "opposing") and textual features (i.e., the text provided by the students, unigrams, bigrams, and syntactic structures from that text, extracted by TagHelper [14]) of the discussion map to find similar clusters. The output of the algorithm is a list of matching clusters in the discussion map(s), sorted according to a similarity rating. Due to space limitations, we refer the reader to a more comprehensive description of the DOCE algorithm in [7].

4. Experiment: Finding Deepening and Widening Clusters with DOCE

But how well does DOCE work in finding instances of deepening and widening? To find the answer to this question we conducted an experiment in which we took hand-annotated examples of deepening and widening (annotated by the members of the Exeter team on the co-author list) from actual classroom discussion maps, and tested whether DOCE was able to use those examples to find the *other* examples of deepening and widening in our data set. More specifically, we took 30 annotated examples of deepening and 30 examples of widening from 14 distinct discussion maps, and did the following:

- For each annotated example, we ran DOCE with that annotation as the model graph against all of the other 13 discussion maps with at least 2 annotations
- We considered a *relevant match* to be 70% overlap, e.g., the following annotated example and found cluster would constitute a relevant match, since there is a 75% node overlap (bold-faced nodes overlap):
  - Annotated example: (Node1, **Node3**, **Node4**, **Node5**)
  - Found Cluster: (**Node3, Node4, Node5**, Node6)
- We varied parameters, such as the number (N) of clusters that were returned by DOCE and the relative impact of structural and textual properties on the

similarity score of cluster pairs (e.g., is it more important that texts or shape types are similar?).

- We evaluated recall, precision, and recall+precision on each run of DOCE, metrics typically used in information retrieval:
  - *Recall* represents the number of annotations of type x covered by DOCE within its Top N, divided by the count of annotations of type x in the searched map (value between 0 and 1.0).
  - *Precision* is the number of relevant matches of type x found by DOCE in the Top N divided by N (value between 0 and 1.0).

Unfortunately, since there is no "gold standard" for performing the type of retrieval task done by DOCE, there was no other computational model to compare to DOCE in our experiment. However, in an earlier experiment, reported in [7], we compared DOCE to a simple program that returned random clusters and found that DOCE performed significantly better. While the random algorithm is, admittedly, a low bar to exceed, doing *significantly* better than random demonstrated that DOCE is clearly finding (at least some) clusters of interest. We considered recall to be the most important metric in our experiment, as it was most important to us to maximize the number of interesting returned clusters in a given discussion. The number of relevant matches (i.e. precision) has somewhat lower importance since we as researchers, and humans in general, are typically clever enough to filter out irrelevant matches.

4.1. Results on the Effectiveness of DOCE

The results of the experiment are summarized in figures 3 and 4. Note, first of all, that the *best* results for deepening and widening are quite reasonable (the middle bar for recall, precision, and recall+precision in each of the figures), especially for recall, the metric we consider most important. By "best" result, we mean the human-annotated cluster that led to the highest recall and precision values when used as a model graph to DOCE. For instance, notice that the best deepening model graph (the middle bar in each of the first two sets of three metrics in Figure 3) led to a recall of 0.80 and precision of 0.52. The *average* results, calculated across *all* of the annotated clusters (the leftmost recall, precision, and recall+precision bar in each of the figures), are not good (e.g., the 0.42 recall and 0.27 precision in Figure 3 are very poor). However, focusing on the best results is more important because, by the nature of the DOCE algorithm, only the *best* examples of deepening and widening will subsequently be used as model graphs to DOCE. That is, once one finds the best model for a particular cluster type – or the best set of models – that model (or models) will then be used as a "search probe" for all subsequent searches.

We also tested whether *combining* the results of multiple runs of DOCE might further improve the results. That is, we wanted to answer the question: Can multiple, high-quality clusters lead to even better results than single best clusters in retrieving relevant clusters? (Note that we did not do such an evaluation in our original paper on DOCE, i.e., [7]) We implemented this combination by ranking the results according to the average relevance scores of the three single-best models (We tried combinations greater than three, but this did not improve the results). The third bar in each set of three bars in Figures 3 and 4 depicts these results. Notice that for the deepening cluster results shown in Figure 3 the combination approach did marginally worse (i.e., recall+precision = 1.30 for the combination approach vs. 1.33 for the single best model), but for the widening clusters shown in Figure 4, the combination approach did

a bit better (i.e., recall+precision = 1.49 for the combination approach vs. 1.42 for the single best model).

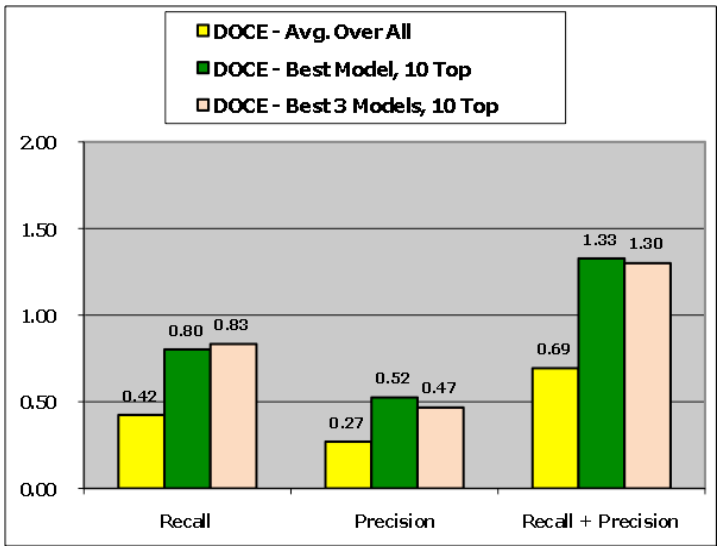

Figure 3: Results of Applying DOCE to the Deepening Clusters

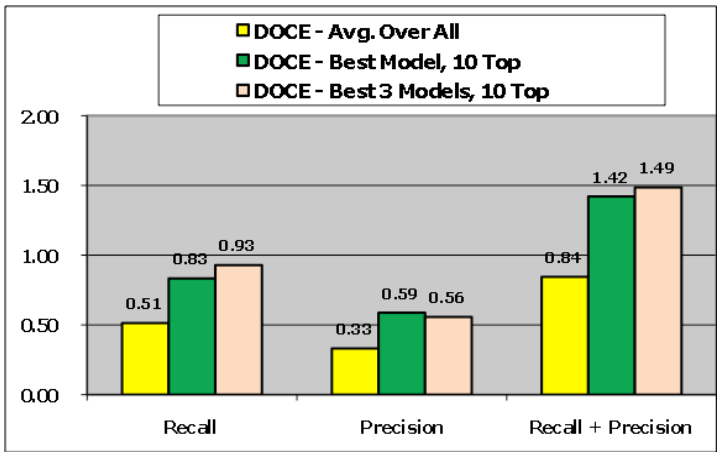

Figure 4: Results of Applying DOCE to the Widening Clusters

It is also important to note that DOCE performed better in identifying widening clusters, the cluster type of most interest to us with respect to this paper, than in identifying deepening clusters. In particular, note that the best widening recall (0.93), precision (0.59), and recall+precision (1.49) in Figure 4 improves upon the best deepening recall (0.83), precision (0.52), and recall+precision (1.33) from Figure 3.

Interestingly, the results we achieved with this data improved upon most of the results reported on a different data set, with different annotated cluster types [7]. More specifically, the DOCE algorithm generally performed better in finding snippets of "creative reasoning" than it did in finding more standard argumentation structures, such as "Chain of Opposition," discussed earlier.

4.2. Results on the Usability of DOCE

The run-time of graph-matching algorithms such as DOCE are known to be NP-Complete [15]. In computational theoretic terms, this means that in the worst case, the run-time of the algorithm grows exponentially with the input size, resulting in non-acceptable run-time behavior even for moderately sized input. However, such algorithms, when applied in practical contexts in limited ways, are often usable without ever bumping into theoretical limits. To explore this, we analyzed the run-time characteristics of DOCE run against the data from our 14 discussion maps.

Not surprisingly, the empirical runtime analysis showed a linear relationship between the text length of the input graph and the pre-processing time (the time needed to extract features from the input and model graph). On average, pre-processing took about 10 seconds per map and peaked at roughly 21 seconds for very large maps (> 60 contributions) with plenty of text. The search time (finding the most similar matches) was highly influenced by the size of the used models: Models of size three were in all cases unproblematic with search times always below five seconds. For models of size four the maximum search times rose to 34 seconds. For models of size five we saw in single cases run-times above one hour, confirming the theoretical assumption of an exponential run-time growth, although in the vast number of cases the run-times were still at an acceptable level.

In summary, it is clear that very large discussion maps can be a problem for DOCE, especially when a teacher uses the algorithm in real-time fashion. On the other hand, the Exeter discussion maps were quite large, created during classroom use over several days, and are almost certainly at (or above) the upper limit of practical map size. Input models (i.e., annotated clusters) did not lead to any excessive search times, as long as the models did not exceed 4 nodes in size. Model graphs that reached five nodes in size led to some extreme cases, but it should be noted that these cases only occurred when a five-node model graph was used to search very large discussion maps. Generally speaking, as long as model graphs do not exceed five nodes they are practically usable, especially when discussion maps are reasonably sized.

5. Discussion and Conclusions

These results, while preliminary, are very encouraging. It appears that the DOCE algorithm is reasonably capable of finding examples of creative reasoning, at least with respect to widening of a discussion, given prior, annotated examples of such reasoning in other discussions. Furthermore, as long as the researcher or teacher is careful not to use too-large model graphs against too-large discussion maps, the DOCE algorithm runs in a practical amount of time. Thus, the DOCE algorithm is a tool that either a researcher or a teacher can use to pinpoint and evaluate creative reasoning in the context of real e-discussions. This technique has the potential to inform moderators when creative and/or critical reasoning are occurring in maps – as well as when it is not occurring, indicating that it might be time to intervene. Of course, we need to do more extensive testing of the algorithm. Thirty example clusters from 14 maps is a small N to experiment with, especially when one considers the size of corpuses typically used in information retrieval experiments.

This work can help researchers explore dialogic theory that views creative reasoning as central to successful student interaction. Just as it is sometimes useful to ‘deepen’ a dialogue by critically testing assumptions and teasing out their implications,

so it is sometimes useful to 'widen' a dialogue by introducing new ideas. Effective thinking for most tasks requires both of these moves; the DOCE algorithm can help explore when and how students engage in these types of interaction in e-discussions.

There is currently a great deal of interest around the world in teaching for creativity, widely seen as one of the core '21st century skills' required for flourishing in the emerging knowledge age [16]. The research described in this paper makes a contribution by associating creativity with the widening that occurs when new voices and perspectives emerge within discussions, and by showing that creativity in this form can be recognized not only be expert humans but also by AI techniques.

Acknowledgements. We thank the entire ARGUNAUT team. The 6th Framework Program of the European Community, Contract No. 027728, sponsored this research.

References

[1] Lingnau, A., Harrer, A., Kuhn, M., & Hoppe, H.U. (2007). Empowering Teachers to Evolve Media Enriched Classroom Scenarios. In: *Research and Practice in Technology Enhanced Learning*, vol. 2 (2) (pp. 105-129).

[2] Schwarz, B. & De Groot, R. (2007). Argumentation in a Changing World. In: *Int'l Journal of Computer-Supported Collaborative Learning*, vol. 2 (pp. 297-313).

[3] Pinkwart, N. (2003). A Plug-in Architecture for Graph Based Collaborative Modelling Systems. In: U. Hoppe, F. Verdejo & J. Kay (eds.) *Proceedings of the 11th Int'l Conference on Artificial Intelligence in Education* (AIED 2003), (p. 535-536).

[4] De Groot, R., Drachman, R., Hever, R., Schwarz, B., Hoppe, U., Harrer, A., De Laat, M., Wegerif, R., McLaren, B.M., & Baurens, B. (2007). Computer Supported Moderation of E-Discussions: the ARGUNAUT Approach. In the *Proceedings of the Conference on Computer Supported Collaborative Learning (CSCL-07)*. (p. 165-167).

[5] Kochan, E.L. (2006). Analysing Graphic-Based Electronic Discussions: Evaluation of Student's Activity in Digalo. In: *Proceedings of the European Conference on Technology Enhanced Learning* (EC-TEL 2006), Crete, Greece, Lecture Notes in Computer Science, Springer, pp 652–659.

[6] Hoppe, H.U. & Gaßner, K. (2002). Integrating Collaborative Concept Mapping Tools with Group Memory and Retrieval Functions. In: G. Stahl (ed.), *Computer Support for Collaborative Learning - Foundations for a CSCL Community (Proceedings of CSCL-2002)*, Boulder (USA), pp. 716-725.

[7] Mikšátko, J. & McLaren, B.M. (2008). What's in a Cluster? Automatically Detecting Interesting Interactions in Student E-Discussions. In B. Woolf, E. Aimeur, R. Nkambou, S. Lajoie (Eds), *Proceedings of the 9th Int'l Conference on Intelligent Tutoring Systems (ITS-08)*, Lecture Notes in Computer Science, 5091 (pp. 333-342). Berlin: Springer.

[8] De Laat, M., Chamrada, M., & Wegerif, R. (2008). Facilitate the Facilitator: Awareness Tools to Support the Moderator to Facilitate Online Discussions for Networked Learning. In: *Proceedings of the 6th Int'l Conference on Networked Learning, Halkidiki, Greece*, 80-86.

[9] Baker, M., Andriessen, J., Lund, K., van Amelsvoort, M., & Quignard, M. (2007). Rainbow: A Framework for Analyzing Computer-Mediated Pedagogical Debates. *Int'l Journal of Computer-Supported Collaborative Learning* (IJCSCL) 2, 315-357.

[10] Andriessen, J. (2008). Arguing to Learn. In: K. Sawyer (Ed.) *Handbook of the Learning Sciences*. Cambridge: Cambridge University press.

[11] Wegerif. R. (2007) *Dialogic, Educational and Technology: Expanding the Space of Learning*. New York: Springer-Verlag.

[12] Kolodner, J. (1993). *Case-based Reasoning*, Morgan Kaufmann Publishers, San Francisco.

[13] McLaren, B.M. (2003). Extensionally defining principles and cases in ethics: An AI model. In: *Artificial Intelligence*, vol. 150, pp. 145-181.

[14] 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. *Int'l Journal of Computer-Supported Collaborative Learning* (IJCSCL) 3 (3).

[15] Garey, M.R. & Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*, WH Freeman & Co. New York, NY, USA.

[16] Wegerif, R., & De Laat, M.F. (in press). Reframing the Teaching of Higher Order Thinking for the Network Society. In S. Ludvigsen, A. Lund & R. Säljö (Eds.), *Learning in Social Practices. ICT and New Artifacts - Transformation of Social and Cultural Practices*: Pergamon.