Guiding the Process of Argumentation: The Effects of Ontology and Collaboration

Frank Loll, Niels Pinkwart, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
Email: frank.loll@tu-clausthal.de, niels.pinkwart@tu-clausthal.de

Abstract: Teaching argumentation is challenging, and the factors of how to effectively support the acquisition of argumentation skills through technology are not fully known yet. In this paper, we evaluate the impact of using an argumentation system with different argument ontologies and with collaborative vs. individual use on the outcomes of scientific argumentation. The results of a controlled lab study with 36 participants indicate that simple ontologies may be more appropriate than highly structured ones. In addition, collaborative argumentation lead to more cluttered argumentation maps, including a higher amount of erroneously used and duplicate elements, which indicates that an expected peer-reviewing between group members did not occur. Yet, groups also tended to include more points-of-view in their arguments, leading to more elaborated argument maps.

Introduction

The successful application of argumentation skills is important in many aspects of life. Even though this importance has been widely recognized, many people have problems as they engage in argumentation activities (Kuhn, 1991). In addition, educating students in their argumentation abilities is often not explicitly taught in school (Osborne, 2010) or at least problematic, caused (among other factors) by teacher's time: face-to-face tutoring is still the favored argumentation teaching method, but does not scale up well for larger groups.

One approach to deal with this issue is the use of argumentation systems (cf. Scheuer et al., 2010, for an overview) – tools that engage (groups of) students in argumentation by representing the argument in a graphical fashion (e.g., a graph, table or matrix), some of them providing feedback and intelligent support. While argument modeling has shown to be effective to promote learning in general (Harrell, 2007, Easterday et al., 2007), the specific roles of collaboration and of the argument ontology that is used in an argumentation system (i.e., the given palette of elements which "pre-structure" the input to the system) are still largely unclear and existing studies about their effects (e.g., Janssen et al., 2010; Osborne, 2010; Sampson & Clark, 2008; Schwarz & Glassner, 2007; Schwarz et al., 2000; Suthers, 2003; Toth et al., 2002) are hardly comparable because the specific conditions (population, the tool used, etc.) in which the studies were conducted differ a lot.

With respect to collaboration, it is widely accepted that arguing in groups can be beneficial for learning (e.g., Osborne, 2010; Schwarz et al., 2000). However, even though collaboration has shown to be effective for solving highly intellective problems (e.g., Laughlin et. al., 2006), reasonable collaboration does not occur by nature (Dillenbourg et al., 1995; Rummel & Spada, 2005) and unstructured collaborative argumentation per se will not lead to higher quality arguments (Sampson & Clark, 2008). Instead, it is important to aid the process of collaboration, e.g. though visualizations (Toth et al., 2003; Suthers, 2003, Schwarz & Glassner, 2007), access restrictions (Schwarz & Glassner, 2007) or collaboration scripts (Stegmann et al., 2007).

Ontologies are the elements available for modeling and representing an argument in the computer system. They may be as simple as in Athena (Rolf & Magnusson, 2002), where the elements are just node, pro and con, or as complex as in Rationale (van Gelder, 2007), a system with 6 categories for argument types and a large set of relations. Suthers (2003) highlighted the *representational guidance* of visualizations as well as of the ontologies that these visualizations refer to. However, even though it turned out in a couple of studies (e.g., Suthers (2003), Schwarz & Glassner (2007)) that the presence and the granularity of an ontology did indeed have an effect on the outcomes of argumentation, a concrete answer to the questions which argument elements are important in which domain (or to learn to argue in general) was not systematically investigated so far – Suthers (2003) just noted that a too detailed ontology may confuse students with "*a plethora of choices*" (p. 34).

Schwarz & Glassner (2007) investigated the effects of informal ontologies and floor control on the results of argumentation with respect to the quality of a final argument map (including the number of relevant claims and arguments, references to peers, and chat expressions) as well as to co-elaboration of knowledge. Their findings indicate that a higher structural degree of ontology and the presence of a turn-taking mechanism lead to more relevant claims and arguments and reduced unwanted behavior like off-topic discussions.

Stegmann et al. (2007) evaluated the usefulness of scripts that guide learners in creating single arguments by means of input masks that scaffold the creation of single arguments (consisting of claim, qualifiers and grounds) and a script that scaffolds a typical argumentation process consisting of argument-counterargument chains. The use of these scripts resulted in a gain in formal quality of single arguments as well as in argumentation sequences in online discussions. In addition, the scripts facilitated the acquisition of knowledge about argumentation.

Even though there is some evidence that these structural scaffolds may guide learners to successful interaction and success in learning, it is hard to come to general statements about their usefulness based on the existing studies with their multiple tools in various settings. In addition, possible interaction effects between collaboration and specific argument ontologies have not been investigated systematically: do some argument ontologies have benefits for collaborative usage as compared to others? In this paper, we want to make a next step towards a deeper, more comparable evaluation of the factors that make educational argumentation systems (un-)successful. First, we briefly outline a flexible framework that can be used to easily manipulate important variables. Then, we describe the use of this tool in a controlled lab study that investigates the impact of different ontologies and collaboration on the outcomes of scientific argumentation.

The LASAD System

The web-based LASAD framework for argumentation was designed with a focus on flexibility. Among other features (cf. Loll et al., 2010), the system is able to support different argument ontologies, visualizations, and collaboration settings. On the ontology side, each element, i.e. boxes and relations between them, can be independently defined based on a set of elements (e.g., labeled text fields, references to text passages or to URLs, ratings) including their visualization attributes such as color, size, or border. On the collaboration side, LASAD supports synchronous and asynchronous usage and offers tools like a chat, a list of users currently participating as well as awareness mechanisms such as the tracking of cursor movements of all participants. The use of these configuration options allowed us to set up the conditions for the study presented in this paper. A part of an argument map created via LASAD during the study presented in this paper is shown in Figure 1.

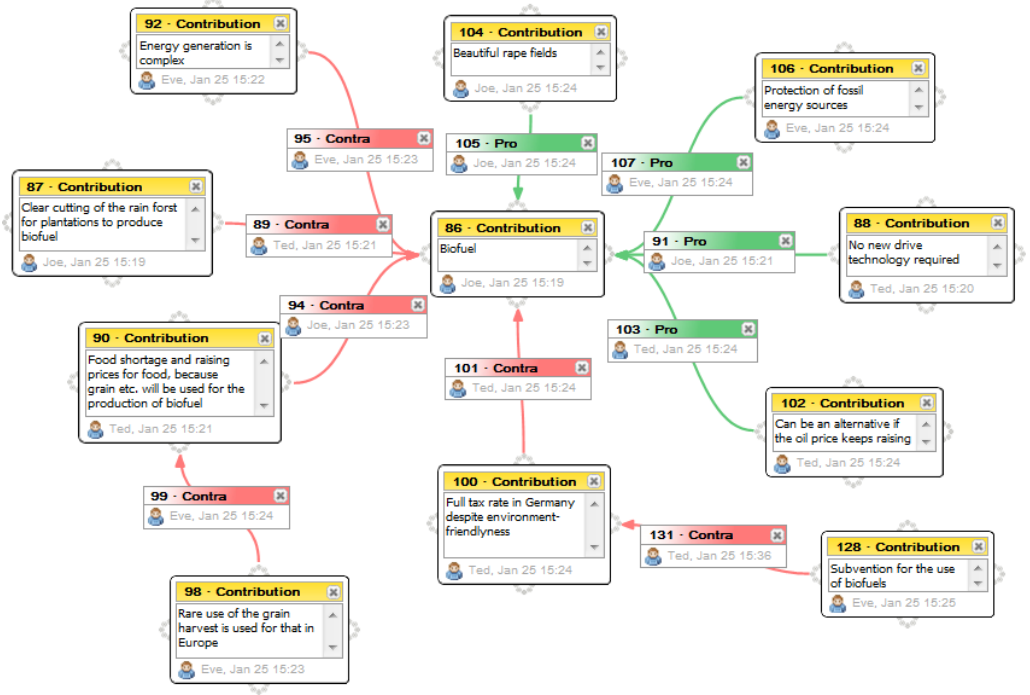

Figure 1. Part of an Argument Map created during the study via LASAD using a domain-independent ontology with general contributions showing arguments for and against biofuel.

Study Description

Inspired by prior research results on educational argumentation technologies (cf. above), our goal was to investigate the effects of using different ontologies in individual as well as collaborative argumentation on the outcomes of scientific argumentation and on the learning effects with respect to argumentation skills and domain-specific knowledge. Whereas prior research results suggest that argument models and ontologies may influence the process of argumentation (e.g., Suthers, 2003), we wanted to systematically evaluate the effects of various argument ontologies (all with graph-based visualizations) used for the same task. A second goal was to investigate how individual and collaborative argumentation differed, also in the context of different argument ontologies, to identify possibilities to aid argumentative processes better. More specifically, we were interested in the following hypotheses derived from prior research (cf. Scheuer et al., 2010):

Hypotheses: Effects of Collaboration

C1. Arguing in groups (as opposed to constructing arguments individually) will lead to a more elaborated argument, i.e. an argument of higher quality, due to different points-of-views of the participants.

C2. In collaborative argumentation activities, students will be more motivated than in individual ones. We hypothesize this based on the fact that discussions with other arguers will lead to a greater variation of the task steps and, hence, to a less monotonous activity. Prior results by Pinkwart et al. (2008) highlighted the importance of motivation to promote good learning results in argumentation activities.
C3. In collaborative sessions, the participation of single users may drop as compared to individual argumentation sessions: shy arguers may stop arguing against a dominant, leading group member.
C4. Collaborative argumentation will lead to more off-topic activities. Prior results by Schwarz & Glassner (2007) showed that groups can tend to get distracted from tasks, which can be detrimental for the overall argumentation process.
C5. Group members will review and respond to each other's arguments, and, hence, the overall number of mistakes will decrease in comparison with individual activities. We hypothesize this because argumentation is not a trivial undertaking: users may oversee their mistakes and, by discussing about parts of the argument, typical mistakes of single users may be revealed and corrected.

Hypotheses: Effects of Argument Ontologyies

O1. The higher the structural degree of an argument ontology is, the higher we expect the overall structure of the argument map to be. This is a direct consequence if the ontology is used correctly and supported by the findings of Schwarz & Glassner (2007).
O2. The more detailed an argument ontology is, the more elaborated the resulting argument will be. The rationale for this hypothesis is that we expect the multiple elements of detailed ontologies to prompt the users to make use of them and, hence, think about how to fill them with appropriate materials. Clark & Brennan (1991) also noted that it is easier to refer to knowledge units which have a visual manifestation, so that the presence of various, different ontology elements may lead to more discussions and, consequently, to a more detailed resulting argument.

Hypotheses: Interaction Effects

I1. For group argumentation, we hypothesize that the used ontology will influence the degree of collaboration: a more complex ontology may increase the need for collaboration (in order to discuss how to use the different elements to build an argument).
I2. In collaborative sessions, highly structured argument ontologies may be detrimental to the quality of the resulting argument (due to the double complexity of keeping track of the group process and using a complicated argument model at the same time), while the scaffolds that more structured ontologies provide may be more helpful in individual usage.

Study Design

To investigate the hypotheses, a mixed 3x2 design was used. The between-subject factor was the argument ontology. Here, the following three different ontologies were used:

1. A simple domain-independent ontology consisting of a general contribution type ("contribution") and three different relation types ("pro", "contra", "undefined").
2. A second, highly structured, domain-independent ontology based on the Toulmin argumentation scheme (Toulmin, 2003). It consists of five contributions ("datum", "conclusion", "warrant", "backing", "rebuttal") and four different relations ("qualifier", "on account of", "unless", "since").
3. A domain-specific ontology which was inspired by the Belvedere (Suthers, 2003) ontology, consisting of three contribution types ("hypothesis", "fact", "undefined") and three relation types ("pro", "contra", "undefined"). This ontology has shown to be effective for scientific argumentation.

The within-subject factor in the study was collaboration. Each participant was required to argue about one topic on his or her own and about another topic in a group of three. To eliminate possible confounds, we used counterbalancing so that half of the participants began with the group phase while the other half began with the single user phase. In the group phase, each participant worked on one machine. The participants were only allowed to communicate via the chat tool integrated in the argument framework. This simulated a remote discussion even though the users were located in the same room (the experimenter was in this room to enforce the rule). Overall, the study took 6 hours per user, including a 1 hour break between two sessions.

Tasks

Each participant worked on two open scientific problems that have no obvious solution. This kind of task choice was motivated by Toth et al. (2002), who used challenging science problems to simulate an authentic argument activity, avoiding a demotivation of students caused by hiding the answer of already solved questions. The Schwarz et al. (2000) results support this decision: their findings include that argumentation is most effective if students are arguing under uncertainty. In our study, the concrete topics for the arguments were:

1. The potential of alternative concepts for automotives (incl. the electronic car, the fuel cell, and biofuel)
2. The German energy mixture in 2030 (incl. nuclear power, fossil fuel, and renewable resources)

For each topic, three different possible positions were prepared. To allow all participants to argue for or against each of these positions, the students were provided with two pages of background information per position. This material, given in form of material chunks (graphs, tables as well as plain text) was typical for scientific argumentation, including facts, examples, statistical data and observations. In addition, there was one page containing material that was common to all positions (e.g., the definition of kilowatt hour for topic 2). The participants were allowed to go beyond the given material in their arguments.

Each session about a topic was split into four slots of 30 minutes each. In each of the first three slots, the participants were given the background material for one of the three positions (e.g., "nuclear power as a future energy") as well as the common materials and were asked to create an argument about this position using the LASAD system. The fourth slot was the used to integrate the three separate positions and to draw a final conclusion to solve the argumentation task. For this last step, the participants were given the materials for all positions again.

Participants & Training

Overall, 36 (under) graduate students (25 male, 11 female) with different majors participated in the study. They were between 19-35 years old (m = 24.64, sd = 3.68) and in semesters 1 to 22 (m = 7.00, sd = 5.62). All participants were either native German speakers or fluent in this language (the complete study was conducted in German). Participation was voluntary and all participants were paid for completing the study. The participants were assigned randomly to all three "ontology" conditions, i.e. in each condition there were four groups consisting of three students each. In all but one group was one female student.

None of the participants had used the argumentation system before. Thus, a short video introduction (15 minutes) to the LASAD system was shown to make sure that all participants had the same basis. All videos consisted of three parts: (1) A general introduction how to interact with the system, (2) an overview of supporting features to work in groups (e.g., chat, cursor tracking), and (3) an ontology dependent part in which the condition dependent features of the system were explained using an example common to all conditions. Finally, the example argument that was presented in the video was distributed among all participants on paper and was available during the complete study.

Tests & Interviews

To test the learning effects caused by the argumentation tool's use, three multiple-choice tests on argumentation abilities as well as two multiple-choice knowledge tests per topic were used. The tasks of the argumentation tests were taken from a list of questions of the Law School Admission Test (LSAT). Each argumentation ability test consisted of four questions, two from the area of logical reasoning and two from the area of analytical reasoning, i.e. we used approved questions that were not law specific. These tests took place before the first session, between the two sessions and after the second session. The order of the tests was counterbalanced. The participants were given 6 minutes (1.5 minutes per question) per argumentation test.

The knowledge tests were centered on the domain of argumentation in the respective study sessions (automotive concepts and energy mix). They were administered immediately before and after the corresponding sessions (in a counterbalanced manner) to measure domain learning. The participants were given 4 minutes (1 minute per multiple-choice question) per knowledge test.

In addition to these two tests, a questionnaire was used to evaluate the usability of the overall LASAD argumentation system. By means of this test, we wanted to check whether certain features of the system might have hindered the students to engage in reasonable argumentation, especially since this was the first larger study with the LASAD system. Here, the standardized *System Usability Scale* (Brooke, 1996) which has shown to be an accepted measure for usability (Bangor et al., 2009), was used.

Finally, we asked the participants in an open interview about their motivation during the study sessions, and about potential problems and ideas for future improvements of the system.

Coding Procedure

The *material* distributed to the participants consisted of unconnected information chunks including relevant as well as non-relevant parts. To be able to check how much of the relevant material was used, three domain experts independently created a list of all the facts that could either be directly taken from the material or directly concluded based on a combination of multiple information chunks. These lists were merged and discussed; the resulting lists (containing 81 entries for topic 1 and 75 for topic 2) were used as a reference for the relevant information that can be extracted from the hand-out material.

To get further insights into the resulting *argument maps*, 6 of 48 maps (one individual map and one collaborative map for each ontology, i.e. 12.5% of all the maps) were coded element-wise independently by two coders with respect to the use of given material. For each element (boxes and relations) in a diagram, the coders

checked if the contained information was based on a fact in the "reference list" or if it was a completely new contribution. The coders also rated the correctness of the used ontology elements (if, for instance, a fact element was actually used to represent a fact). To judge the *structural quality of an argument map*, the coders additionally checked for each of the 6 chosen maps if this map contains (a) a starting hypothesis, (b) a conclusion and (c) a clear grouping of the different positions.

Based on these coding results, the inter-rater reliability was calculated and resulted in a Cohen's κ of 0.60 for the material used and 0.61 for the used elements. Concerning the general structural features (a-c), both coders agreed 100% on each measure. Taking into account the ill-defined nature of argumentation (Lynch et al., 2010), we assumed this level of agreement to be acceptable overall. The remaining elements were then coded by one coder in the same manner as described above. Overall, 5477 elements were manually coded this way.

To measure the *degree of coordination*, also the chat messages were encoded. First, the chats (consisting of 878 messages) were divided independently by two coders into episodes that belong together, e.g. a discussion about where to start with argument modeling. Slight differences were resolved by discussion between the coders. This resulted in an overall number of 196 chat episodes. Based on the chat episodes of three sessions (one per ontology, i.e. 25% of all material), the following four categories were agreed on as a coding scheme for the chat episodes: (1) *content*, (2) *structure*, (3) *coordination*, (4) *off-topic*. Based on this coding scheme, each chat episode within the 12 collaborative sessions was independently coded by two raters. The raters achieved a moderate Cohen's κ of 0.56. However, it turned out that the categories "*structure*" and "*co-ordination*" were often not clearly distinguishable so that these two categories were merged into one, which resulted in a high κ of 0.76. The raters resolved remaining conflicts through discussion.

Results

This study was the first one done with the LASAD framework. As such, we were also interested in the general usability of the system to check for any possible confounds related to weaknesses of the argumentation system we employed. The *System Usability Scale* test resulted in a mean score of 81.46 (which corresponds approximately to a "B" grade). The results thus indicate that the LASAD framework used in this study was perceived as an adequate tool to support argumentation.

Overall Effects on Argumentation Abilities and Domain Knowledge

Based on the scores of the argumentation ability tests ($m(T_1) = 1.611$, $sd = 1.02$; $m(T_2) = 2.056$, $sd = 0.89$; $m(T_3) = 2.083$, $sd = 1.08$; scale ranging from 0 to 4 points), a repeated measures ANOVA was calculated. This showed no statistically significant gains in argumentation skills, but a tendency ($F(2, 66) = 2.907$, $p = 0.062$). The between-subject factor "ontology" did not cause a significant effect ($F(2, 33) = 0.745$, $p = 0.483$).

Regarding the domain knowledge, a significant gain between pre/post-test scores was consistently achieved. In topic 1 (*Potential of Alternative Drive Concepts for Automotives*), the pre-test resulted in $m = 0.92$ ($sd = 0.77$), whereas the post-test resulted in $m = 2.97$ ($sd = 0.88$; based on paired samples t-test: $t(35) = -10.330$, $p < 0.001$; scale ranging from 0 to 4 points). In topic 2 (*The German Energy Mixture in 2030*) the pre-test resulted in $m = 2.31$ ($sd = 1.04$), whereas the post-test resulted in $m = 3.42$ ($sd = 0.84$; based on paired samples t-test: $t(35) = -5.976$, $p < 0.001$). Concerning the gain of domain knowledge, there was neither a significant difference between individual/collaborative use of the system nor between the different ontologies.

The Effects of Collaboration on the Argumentation Outcome

An ANOVA highlighted significant differences between individual and collaborative argument maps as shown in Table 1. In comparison, collaborative argument maps contained a larger amount of elements (i.e. boxes and relations between them) used overall ($F(1, 46) = 18.954$, $p < 0.001$) and a higher percentage of material used twice ($F(1, 46) = 6.983$, $p = 0.011$). Contrary to our expectations, the percentage of given material used did not differ significantly between individual and collaborative argumentation ($F(1, 46) = 0.932$, $p = 0.339$). Instead, group members provided significantly more *own* contributions (not derived from given material) ($F(1, 46) = 13.524$, $p < 0.001$) than individual arguers. Hypothesis C5 (groups will review the work of the members and, hence, will make less mistakes), measured by the percentage of wrongly used elements, has to be rejected ($F(1, 46) = 0.956$, $p = 0.333$). In fact, mistakes made in the group phases were often very similar to those made in the individual phases, e.g. wrong directions of relations. Thus, hypothesis C1 (group work → higher quality) is only partially supported. To measure the motivation of the participants, we analyzed the statements in the personal interviews conducted after the study. Here, all groups agreed (after short discussions) that working in groups was more motivating than working alone (hypothesis C2). This is supported by the observations of the experimenter, who stated that sometimes the participants in the individual sessions made a bored impression, as opposed to the collaborative sessions. Also, the groups always used all the time for their tasks, while some individuals finished early. Among our study participants, the question about the optimal group size for argumentation was discussed controversially. The majority agreed on two to three people arguing together;

larger groups and the resulting growing needs for coordination were seen as potentially detrimental for the overall results.

Table 1: Comparison between individual and collaborative argument maps.

	Individual (n=36)	Collaborative (n=12)
Overall # of used elements in the workspace	m = 104.00 (sd = 29.72)	m = 143.25 (sd = 15.80)
# of own contributions (not derived from given material)	m = 9.97 (sd = 5.43)	m = 18.58 (sd = 10.61)
Percentage of material used twice	m = 5.03% (sd = 4.23)	m = 9.14% (sd = 5.83)
Percentage of erroneous used elements	m = 22.46% (sd = 18.80)	m = 28.83% (sd = 21.73)

Hypothesis C3 (collaboration → participation drop of single users) is not easy to evaluate. We sought to investigate if users, when working together, became less active. To do so, we first computed the proportion of elements of each user in the collaborative sessions (min = 0.16, max = 0.59, m = 0.33, sd = 0.12) – i.e., single users created between 16% and 59% of a collaborative map. Apparently, there were thus no "drop-outs" and no dominating users creating the whole map alone. To represent how active a user is in individual sessions (as compared to his peers), we also computed, for each user, the proportion of his argument elements in his individual session to the sum of elements of all individual maps of his group members (min = 0.19, max = 0.51, m = 0.33, sd = 0.07). These two values resulted in a significant Pearson correlation of $\rho = 0.428$ ($p = 0.009$). Thus, hypothesis C3 can be rejected: users who are generally (in)active in individual sessions exhibit the same attitude also in collaborative sessions.

The hypothesis that working in groups might lead to a large amount of off-topic talk (hypothesis C4) could not be confirmed, as Table 2 shows. In the argument graphs, there were in fact no noteworthy off-topic contributions at all. The chat, embedded in the tool, seems to work quite well to avoid off-topic talk in the map.

Table 2: Overview of average chat episodes per ontology in multi-user maps.

Ontology	# Content Episodes	# Structure & Coordination Episodes	# Off-topic Episodes	Overall
m (Simple)	*6.25 (44.6%)*	*7.25 (51.8%)*	*0.5 (3.6%)*	*14*
m (Toulmin)	*5.00 (30.3%)*	*9.0 (54.5%)*	*2.5 (15.2%)*	*16.5*
m (Specific)	*4.75 (26.0%)*	*11.75 (64.4%)*	*1.75 (9.6%)*	*18.25*
m (Overall)	*5.33 (32.8%)*	*9.33 (57.4%)*	*1.58 (9.7%)*	*16.25*

The Effects of Ontology on the Argumentation Outcome

Based on the structural assessment of the maps (with respect to starting hypothesis, conclusion and clear grouping), no significant difference between different ontology conditions could be identified and, hence, hypothesis O1 (higher structural degree of ontology → improved structure of the argument) has to be rejected. However, users of the Toulmin-based ontology did show a tendency not to use a starting hypothesis ($F(2, 45) = 3.100$, $p = 0.055$), which is not really surprising as this ontology follows a different model of argumentation (beginning with data and then drawing a conclusion) and there is no explicit hypothesis element in the ontology.

Table 3: Overview of wrongly used ontology elements.

Ontology	Average percentage of wrongly used elements
Simple	m = 10.25% (sd = 10.80)
Toulmin	m = 41.29% (sd = 16.48)
Specific	m = 20.63% (sd = 16.57)

A difference between ontologies was found in the percentage of wrongly used elements, e.g., using a hypothesis box to represent a fact or to ignore the direction of a pro relation ($F(2, 45) = 18.082$, $p < 0.001$). A post-hoc Tukey HSD test indicated that there was a significantly higher error rate (shown in Table 3) in the Toulmin condition than in the others ($p < 0.001$ for Toulmin vs. Simple and $p < 0.001$ for Toulmin vs. Specific).

Hypothesis O2 (detailed ontology → elaborated arguments) could be confirmed partly. An ANOVA showed no significant differences ($F(2, 45) = 1.909$, $p = 0.160$) between ontologies with respect to the percentage of given material being used. However, a non-parametric Kruskal-Wallis test indicated that the amount of own contributions (not derived from given material) did differ significantly ($p = 0.034$) as shown in Figure 2.

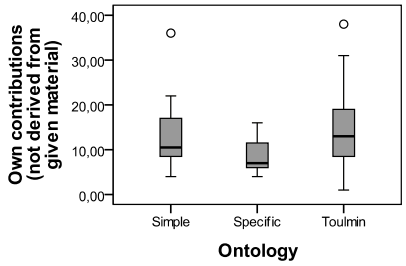

Figure 2. Differences of Own Contributions (Not Derived from Given Material) Used between Ontologies.

Interaction Effects

Regarding hypothesis I1 (ontology will influence the degree of collaboration), we analyzed the number of relations between elements of different authors in relation to the overall number of links as an indicator of the degree of collaboration (since this reflects the inter-relatedness of contributions from different users). An ANOVA did not reveal any significant difference between ontologies ($F(2, 9) = 1.689$, $p = 0.238$). Thus, the hypothesis could not be confirmed. Similarly, the results of a comparison of the number of chat messages used in different ontology conditions did not show any significant differences between ontologies as well (content episodes: $F(2, 9) = 0.212$, $p = 0.813$; structure & coordination episodes: $F(2, 9) = 0.408$, $p = 0.676$; off-topic episodes: $F(2, 9) = 0.568$, $p = 0.586$).

The comparison of the chat messages can be used for the investigation of hypothesis I2 (highly structured ontology will be detrimental to collaborative argumentation) as well, showing that the amount of needed coordination of structure and activities are not dependent on the complexity of the argument ontology. In addition, there was no significant interaction effect between individual / group argumentation and the ontology ($F(2, 42) = 0.605$, $p = 0.551$) in terms of the number of erroneously used elements for argumentation. As such, I2 has to be rejected.

Discussion

Regarding the knowledge and the argumentation tests, the results are not surprising: the increase of domain knowledge was an expected side-effect: if students argue about a topic for a longer time with additional material, the result that they have gained knowledge in this field can be expected. The positive trends shown by the argumentation ability tests is more interesting and needs to be further evaluated in long-term studies – 4 hours use of an argumentation system might not have been enough to come to significant effects at the .05 level.

With respect to collaboration, the results of our study confirm the possible benefit of collaboration for learning argumentation and are in line with prior findings (e.g., Janssen et al., 2010; Osborne, 2010; Sampson & Clark, 2008; Schwarz & Glassner, 2007; Schwarz et al., 2000). Against our hypothesis, groups in our study appeared not to have really checked each other's contributions well, but have argued for or against possible arguments, resulting in more elaborated arguments. This is clearly a point that may be worth future investigations as peer-reviews have shown to be an effective learning strategy (Cho & Schunn, 2007) and their inclusion into argumentation system could be fruitful. Based on a scripted approach, a peer-review process could be enforced in argumentation systems. Contrary to the results of Schwarz & Glassner (2007), the influence of structural aids and collaboration on the amount of off-topic talk could not be confirmed in our study. Possibly, the presence of a separate chat window was sufficient to keep the resulting argument map "clean".

Concerning the guiding function of the ontology, our results support Suthers' (2003) findings. The use of the Toulmin argumentation scheme did lead to a different style of argumentation: While the Toulmin approach is based on data used to draw a conclusion (without any hypotheses), the other ontologies used in our study employ hypotheses that are then backed up with supporting facts. However, we were *not* able to provide evidence that a domain-specific approach is more beneficial for the overall argument quality than a domain-independent one.

In addition, the participants in our study had problems with a highly structured argument ontology, confirming prior findings by Suthers (2003) that a broad range of elements may cause problems for students dealing with it: the Toulmin ontology puts excessive demands onto the students due to its complexity. In fact, there were even students who denied using the ontology correctly at all and only used the colors of the elements as orientation, e.g. using the red "on account of" relation as contra and the green "since" relation as pro. There was no noteworthy difference between the other two ontologies. Limiting, we would like to mention that the students were not familiar with any argument ontology before the study and the theoretical argument model of Toulmin was definitely the most complicated one in our study so that additional training may be required to deal with it. Also, a less elaborated ontology offers simply fewer possibilities to actually use elements incorrectly.

Conclusion & Outlook

In this paper, we contributed to the line of current research that investigates how to effectively support individual and collaborative argumentation. In the study reported in this paper, we systematically varied not only the role of collaboration and argument ontology on their own, but also looked at possible interaction effects between them. Our findings highlight the importance of adequate ontologies: the visual representation (and its complexity) of arguments makes a difference for the resulting overall argument quality. Yet, we were unable to find specific different needs of individuals vs. groups that would correspond to the employed ontologies. This could imply that groups are able to deal with quite complex ontologies even though they also have to manage the complexity of group work at the same time. Further, more detailed investigations are required here: one could argue that the structure provided by an ontology may even support the coordination in groups so that both effects (the detrimental one as well as the supportive one) may have canceled each other out. Finally, our results confirm that groups are able to enrich argumentation with different points-of-views.

In future research, we also plan to carry these results forward into other domains like ethics and legal argumentation and check if the found results are valid across argumentation domains. Additionally, we plan to compare scripted scenarios with unscripted ones to gain further insights how to improve the learning process.

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. International Journal of Usability Studies, 4(3), 114–123.

Brooke, J. (1996). SUS - a quick and dirty usability scale. In P. W. Jordan, B. Thomas, & B. A. Weerdmeester (Eds.), Usability evaluation in industry (pp. 189–194). Taylor & Francis.

Cho, K. & Schunn, C. D. (2007). Scaffolded Writing and Rewriting in the Discipline: A Web-Based Reciprocal Peer Review System. Computers & Education, 48(3): 409-426.

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. L. B. Resnick, J. M. Levine and S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition (pp. 127-148).

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1995). The evolution of research on collaborative learning. In P. Reimann & H. Spada (Eds.), *Learning in humans and machines: Towards an interdisciplinary learning science* (pp. 189-211). Oxford: Elsevier/Pergamon.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). 'Tis better to construct or to receive? Effect of diagrams on analysis of social policy. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), Proceedings of the 13th International Conference on Artificial Intelligence in Education (pp. 93–100). IOS Press

Laughlin, P. R., Hatch, E., C., Silver, J. S., & Boh, L. (2006). Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effects of Group Size. Journal of Personality and Social Psychology, 90(4): 644-651.

Loll, F., Pinkwart, N., Scheuer, O., & McLaren, B. M. (2010). Simplifying the Development of Argumentation Systems using a Configurable Platform. Educational Technologies for Teaching Argumentation Skills.

Lynch, C., Ashley, K. D., Pinkwart, N., Aleven, V. (2010). Concepts, Structures, and Goals: Redefining Ill-Definedness. International Journal of Artificial Intelligence in Education, 19(3): 253-266.

Harrell, M. (2007). Using Argument Diagramming Software to Teach Critical Thinking Skills. Proceedings of the 5th International Conf. on Education and Information Systems, Technologies and Applications.

Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2010). Effects of Representational Guidance during Computer-Supported Collaborative Learning. Instructional Science, 38(1): 59-88.

Kuhn, D. (1991). The Skills of Argument. Cambridge University Press.

Osborne, J. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. Science, 328 (463): 463-466.

Pinkwart, N., Lynch, C., Ashley, K. D., & Aleven, V. (2008). Re-evaluating LARGO in the Classroom: Are Diagrams Better than Text for Teaching Argumentation Skills? In LNCS 5091 (p. 90 - 100).

Rolf, B., & Magnusson, C. (2002). Developing the art of argumentation. A software approach. Proceedings of the 5th International Conference on Argumentation (pp. 919–926).

Rummel, N. & Spada, H. (2005). Can people learn computer-mediated collaboration by following a script? In Fischer, F., Mandl, H., Haake, J. & Kollar, I. (Eds.), Scripting computer-supported communication of knowledge cognitive, computational, and educational perspectives. Dordrecht, NL: Kluwer.

Sampson, V., & Clark, D. (2008). The Impact of Collaboration on the Outcomes of Scientific Argumentation. Science Education, 93(3).

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-Supported Argumentation: A Review of the State-of-the-Art. International Journal of CSCL, 5(1), 43-102.

Schwarz, B. B., & Glassner, A. (2007). The Role of Floor Control and of Ontology in Argumentative Activities with Discussion-based Tools. International Journal of CSCL, 2, 449-478.

Schwarz, B. B., Neuman, Y., & Biezunger, S. (2000). Two Wrongs May Make a Right… If They Argue Together! Cognition and Instruction, 18(4): 461-494.

Stegmann, K., Weinberger, A., & Fischer, F. (2007). Facilitating Argumentative Knowledge Construction with Computer-supported Collaboration Scripts. International Journal of CSCL, 2:421-447.

Suthers, D. D. (2003). Representational guidance for collaborative inquiry. In Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments, pp. 27–46.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to Know": The Effects of Representational Guidance and Reflective Assessment on Scientific Inquiry. Science Education, 86(2).

Toulmin, S. E. (2003). The Uses of Argument. Cambridge University Press, 2nd rev. edition.

van Gelder, T. (2007). The rationale for Rationale. Law, Probability and Risk, 6(1–4), 23–42.

Acknowledgments

This work was supported by the German Research Foundation under the grant "LASAD – Learning to Argue: Generalized Support Across Domains".