Argument Diagramming and Diagnostic Reliability

Collin LYNCH [a,1] Kevin D. ASHLEY [a,b] and Niels PINKWART [c] and Vincent ALEVEN [d]

[a] *LRDC & Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, USA (collinl@cs.pitt.edu)*

[b] *School of Law, University of Pittsburgh (ashley@pitt.edu)*

[c] *Department of Informatics, Clausthal University of Technology, Clausthal, Lower Saxony, Germany (niels.pinkwart@tu-clausthal.de)*

[d] *Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA (aleven@cs.cmu.edu)*

Abstract. Diagrammatic models of argument are increasingly prominent in AI and Law. Unlike everyday language these models formalize many of the the components and relationships present in arguments and permit a more formal analysis of an arguments' structural weaknesses. Formalization, however, can raise problems of agreement. In order for argument diagramming to be widely accepted as a communications tool, individual authors and readers must be able to agree on the quality and meaning of a diagram as well as the role that key components play. This is especially problematic when arguers seek to map their diagrams to or from more conventional prose. In this paper we present results from a grader agreement study that we have conducted using LARGO diagrams. We then describe a detailed example of disagreement and highlight its implications for both our diagram model and modeling argument diagrams in general.

Keywords. Diagrammatic models of argument, ITS, Reasoning with Hypotheticals, AI and the law, Models of Legal Knowledge, Inference and Argumentation, Legal Education

Introduction

Diagrammatic models of argument are a growing area of research in AI and Law [14,7,6]. These models provide a framework for computationally modeling complex legal decision-making by instantiating argument schema and domain-specific critical questions about an argument's potential weaknesses [19]. Diagrams represent arguments as a series of discrete argument moves, an accumulation of moves, or the results of moves made by the participants in the argument scheme. Each diagrammatic model emphasizes different aspects of argument process. Toulmin [17] diagrams, for example, present a static view of an argument representing the inferential support that data confer on claims (i.e., legal conclusions) often mediated through warrants with backing. Easterday dia-

[1] Corresponding Author: LRDC 3939 O'Hara St. Pittsburgh, Pennsylvania 15217

grams [5] by contrast focus explicitly on formal causal inferences suitable for philosophical, empirical, and policy debates.

Legal argument diagrams are meant to facilitate communication by reifying the argument structure thus making the individual moves and structural relationships explicit. They are also intended to make arguments computationally tractable by communicating that structure in a rigorous form [14,7,6]. Developers hope that legal professionals, students, and ordinary citizens will use diagrams to annotate existing arguments and to construct and communicate new claims.

If these goals are to be realized however it is important that humans be able to parse and assess diagrams reliably, an empirical question. It is also important that the models support the flexibility required for open argumentation and do not truncate the argument space in the name of tractability. While many of the existing argument models are subject to strong formalisms, no set of formal rules can easily account for all possible variations. Bench-Capon and Gordon present some examples of disagreement in the relatively constrained domain of civil law [2] where the goal, in both the original text and the diagrams, is to eliminate ambiguity unlike real-world argumentation which often trades on it. Given the range of variation exhibited in professional argumentation and the rigidity of argument models, assessment reliability is neither self-evident, nor assured. Indeed, we expect that domain experts will often disagree about both the appropriate formulation of an argument and the complete meaning of a given structure. This issue is especially acute when dealing with arguments created by students or other non-experts who may disagree as to the appropriate formulation of diagrammatic rules.

Our purpose in this analysis is to focus on the extent to which argument diagrams can be reliably assessed by domain experts. Reliability is important if diagrammatic models are to be employed more widely in both educational and professional settings. Unreliable models are neither usable as a robust communications tool nor as a stable knowledge base. While a great deal of work has been done on written essay grading and tagging, less work has been done on diagram reliability, both for novel argument diagrams and for annotation diagrams of the type we consider here. This assessment therefore is relevant both for ourselves as it supports the usability of our system, and for other authors as a model for their own analyses.

A LARGO-style diagram [13] follows an argument scheme for hypothetical reasoning, that is, the posing of hypothetical cases or artificial patterns of fact often of the form "what if...", to evaluate proposed legal rules for deciding a case. It represents the accumulation of moves proposing a test or rule for deciding a case, challenging the rule as too broad or too narrow or exploring its meaning, and responding to the hypothetical by modifying the rule or by analogizing or distinguishing the hypothetical and the problem scenario. While this form of reasoning occurs in civil law jurisdictions [10] it is common in higher level courts, particularly those in common-law domains, whose decisions set precedents for deciding future cases [1].

Students are rarely taught this form of argument explicitly. Rather, they are expected to pick it up from in-class socratic discussion and written opinions. Such examples, however, are neither very explicit nor very clear. LARGO is designed to support students' understanding of this process by helping them to reify the essential components and relationships in the argument and thus to access complex real-world examples.

In a series of studies, we have been analyzing and comparing LARGO diagrams constructed by Law students at the University of Pittsburgh's School of Law. Our goal in these studies was to examine the use of argument diagrams as an educational tool with a focus on oral argument comprehension. In [1] we presented empirical evidence that features of argument diagrams made with LARGO are correlated with two independent measures related to argumentation ability: standardized test scores that assess the students' ability to evaluate reasoning and arguments and students' number of years in law school. In the present study we turned our attention to the diagnosticity of the argument diagrams. For this study we retained two expert legal instructors who graded a set of LARGO diagrams for factors such as quality, correctness and students' understanding of the model. We presented some preliminary results of this study in [**?**] and in [9]. This work extends those papers with more results and a more detailed analysis of the differences between good graphs.

In the course of their full analyses our graders identified some interesting variations in the argument diagrams. In this paper we present the overall results of our agreement analysis and then examine one point of disagreement in detail. This variation presents an example of complex behavior where different interpretations of the same argument are made explicit in the diagrammatic language. These interpretations in turn affect the structure of the argument as diagrammed. We present this variation and some consequences of it below.

1. Related Work

Prior researchers have developed instructional programs to teach problem-solving, argumentation and reasoning skills through the medium of argument diagrams. Carr [3] developed an instructional program in which law students created novel Toulmin-style argument diagrams in a legal domain. Other instructional systems employing argument diagrams have been developed for teaching reasoning and argumentation skills in philosophy (e.g., critical thinking [18] and causal reasoning [5]) and natural science [16]. The Belvedere system offered students advice, at least in a prototype version, based on the system's analysis of the students' developing natural science argument diagrams [12].

There is much less work on assessment of problem-solving or argumentation skills using the argument diagrams as evidence of students' understanding. In the most relevant related work [8], the researchers manually compared Toulmin-style diagrams of arguments about public health issues (e.g., the desirability of introducing genetically-modified organisms into the food chain). Students constructed the diagrams in two conditions: as a medium for discussion or as a tool for reconstructing a completed debate.

The current research differs from the above work in a variety of respects. First and foremost we are focusing on expert human graders employing agreed-upon criteria rather than a specific scoring algorithm as Lund did. This is similar to the route taken by McClure, Sonak and Suen [11]. However their focus was on the assessment of concept maps rather than functional arguments or procedural annotations. Secondly, most of the above work employs Toulmin-style datum-claim diagrams; LARGO's diagrams correspond to a process model of arguing with hypotheticals. Rather than expressing their own arguments, as in [18], the law students use LARGO to reconstruct experts' arguments which, significantly, are quite naturally represented in terms of the process model [1].

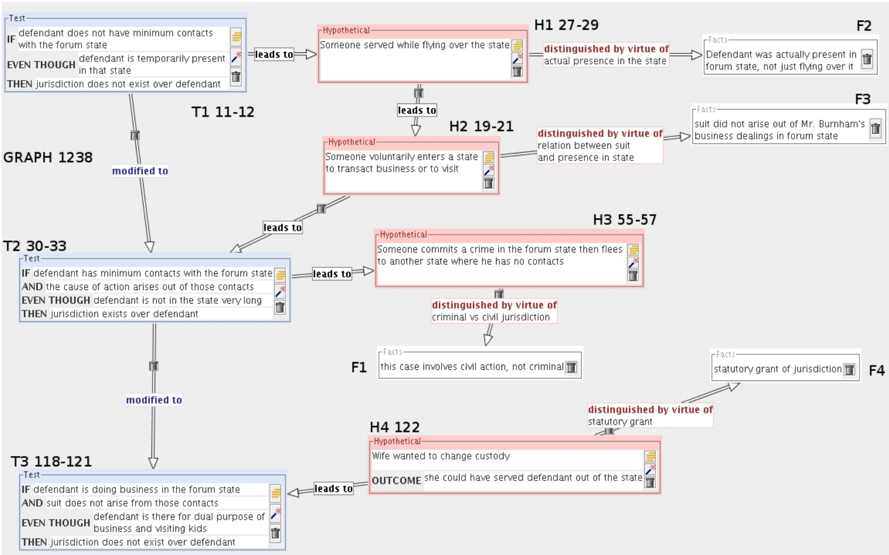

Figure 1. Sample Student Graph

2. Sample Student Diagrams

Figure 1 contains a sample student diagram. This diagram represents a portion of the argument in the case of *Burnham vs. The Superior Court of California* 495 U.S. 604 (1990). This is a standard personal jurisdiction case studied in first-year legal process courses at American law schools. The case concerns a father who was served with divorce proceedings while visiting his children in California. He did not reside in California nor was the marriage formalized there. The question before the court is whether he can be compelled to stand for divorce in California as opposed to his home state where the couple had married. Annotations have been added to the diagram to indicate test *T3* located in the lower-left corner of the diagram, as well as hypothetical *H4* and fact *F4*. As part of these annotations the line numbers to which the test and hypothetical are linked have been added. We will return to this diagram in subsequent sections.

As we noted above, the focus of our diagrams is on the process of argument. In forming arguments before the courts an advocate will routinely pose a *test* or legal rule that, if adopted achieves their desired outcome. The justices (or under other circumstances, the opposing counsel) will respond by posing *hypothetical cases* or what-if scenarios that put pressure on one or another aspect of the test. The advocate will then respond by analogizing or distinguishing the posed hypothetical from the facts of the case at hand and modifying their test as necessary. For more detail on the process model see [1].

As described in [13] students using the LARGO system are presented with an oral argument transcript and a palette of nodes and relations. Of these, the *test* and *hypothetical* nodes may be explicitly linked to the relevant portion of the transcript. Some of these links have been shown in Figure 1. They form the diagram by adding selected nodes to the diagram, drawing relations between them, and annotating the nodes and relations with summaries. Thus enabling them to reify the argument process and facilitating later analysis.

3. Expert Grading Experiment

We have conducted a series of studies with LARGO at the University of Pittsburgh's School of Law including studies with volunteer and non-volunteer first year law students as well as third-year students. Under the U.S. System, law is a graduate degree typically taking three years to complete. First year law students are typically recent graduates of a four-year baccalaureate program while third-year students are soon to graduate and receive legal accreditation.

For the purposes of this study we selected a total of 198 graphs generated by 48 first and 23 third-year law students at the University of Pittsburgh's School of Law, and two law school faculty members. The graphs cover three U.S. Supreme Court cases, *Burnham*, *Asahi Metal Industry vs. Cheng Shin* 480 U.S. 102 (1987) and *Burger King vs. Rudzewicz* 471 U.S. 462 (1985). All three cases center on questions of personal jurisdiction and are standard components of first-year legal process courses.

We engaged a pair of senior law school faculty from the University of Pittsburgh to grade the graphs in a double-blind manner; they were unaware of what class or study group the graphs originated from. Prior to grading, both faculty members trained on the system using the same series of cases as the students. This included producing marked up graphs for all three cases. Each grader's own graphs were made available to him during grading to act as a reference while the other faculty member's graph, minus one dropped due to a clerical error, were inserted into the grading pool resulting in a total of 203 graphs.

The graders were initially provided with a sample of 6 graphs drawn from a different student pool not used in this study and a set of draft grading criteria. They marked up the cases independently and then met to discuss the results and refine their criteria. This process not designed to ensure that the results were "legally sensible" and to avoid any spurious sources of error.

The graphs were provided in anonymized form with each graph being designated by a randomly assigned ID that did not identify the student or study group. While each faculty member saw the same ID for the same graph the graphs were presented in a randomly shuffled order to avoid bias. In addition to the graphs themselves the graders were also provided with a copy of the argument transcript for which the graphs were constructed. Annotations on the graphs of the type shown in Figure 1 indicated the transcript linking. This facilitated lookup when assessing the individual Test and Hypothetical nodes.

Each grader began by partitioning the graphs into one of three equally-sized bins of "poor", "medium" and "good" graphs. They then further divided each bin equally into "better" and "worse" graphs. This binning resulted in a six-point grading scale for the graphs and was defined based upon an initial "gestalt" comparison. The purpose of this initial binning was to avoid the "reassessment" phenomenon whereby graders alter their criteria as they work through a set of materials. Having assigned the "gestalt" grade they then reshuffled the graphs and began assigning detailed grades to the graph reflecting the extent to which the graph covered the essential elements of the argument, its correctness, and the student's understanding of the argument model. They then graded each test and hypothetical node independently in the context of the graph. Finally they assigned an overall grade to each graph reflecting their now more complete judgment of the graph quality. In this respect the gestalt ranking represents their initial visual assessment of the graphs including their essential first impressions while the overall grade reflects the results of their much more detailed judgment.

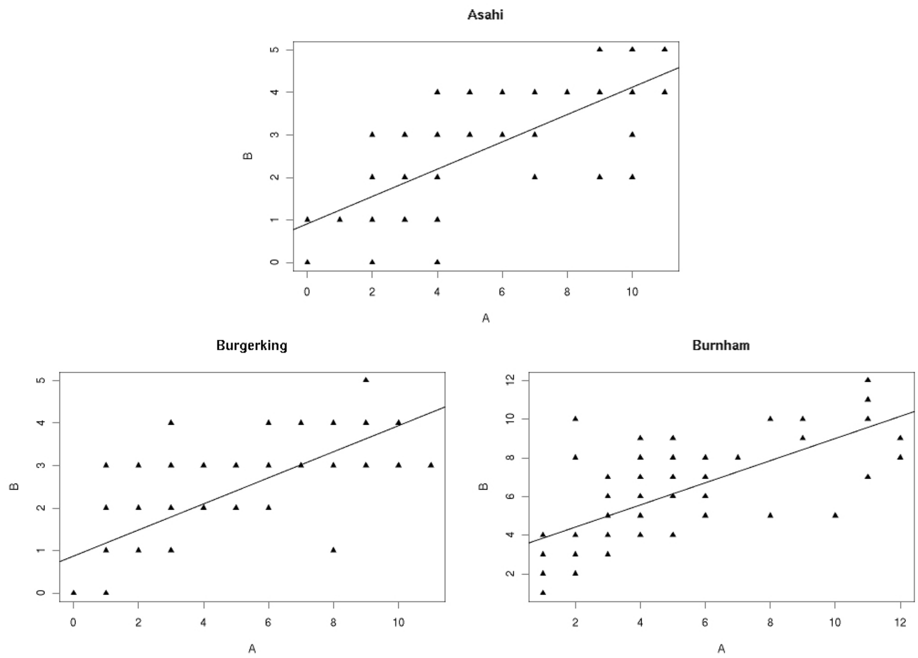

Figure 2. Per-case inter-grader agreement.

	Gestalt Ranking		*Overall Grade*			
			Slope		Intercept	
Case Name	ρ	*p-value*	*est.*	*p-value*	*est.*	*p-value*
Asahi	0.71	$p < 0.001$	0.32	$p < 0.001$	0.92	$p < 0.001$
Burger King	0.73	$p < 0.001$	0.30	$p < 0.001$	0.88	$p < 0.001$
Burnham	0.7	$p < 0.001$	0.57	$p < 0.001$	2.85	$p < 0.001$

Table 1. Inter-grader ranking agreement.

4. Grading Comparisons

Table 1 provides statistics for inter-grader agreement on both the gestalt rankings and overall grades. A number of statistical agreement metrics exist in the literature, notably Cohen's Kappa [4]. Kappa, however, is designed for taxonomic classification where no ordering relation exists between the alternatives. As such it is unsuited to the data here. For that reason we calculated inter-grader agreement on gestalt rankings using Spearman's ρ which is designed to assess agreement between ordered pairs [15]. Spearman's ρ is a rank-based measure of association in the range of $[-1, 1]$ with values near 0 indicating no correlation. Since overall grades are assigned on an absolute scale rather than in a ranking, we used standard linear correlation models to compute inter-grader agreement. Scatterplot representations of the agreement with fitted line models can be seen in figure 2. As the table and diagrams demonstrate, inter-grader agreement was relatively

	A		B	
Case Name	ρ	*p-value*	ρ	*p-value*
Asahi	0.73	$p < 0.001$	0.83	$p < 0.001$
Burger King	0.85	$p < 0.001$	0.85	$p < 0.001$
Burnham	0.88	$p < 0.001$	0.87	$p < 0.001$

Table 2. Intra-grader ranking to score agreement.

high with both graders assigning the same or similar ranks and grades to all graphs. This indicates that, at least with respect to the higher-level grading metrics, the graders are in agreement and can rate the graphs in a consistent manner.

With respect to the intra-grader agreement the results are quite similar. Here we used Spearman's ρ to assess the agreement between a grader's initial gestalt ranking of a graph and the final scores. The results of this analysis are shown in Table 2. As the results indicate, the graders did not change their overall assessments of the graph to a great extent between their initial first-pass rankings and their detailed review.

5. Good Graph Differences

The overall inter-grader agreement indicates that expert legal faculty can reliably differentiate a good diagram from a poor one. This overall agreement, however, does not mean that the diagrams themselves are homogeneous. Even 'good' diagrams exhibit some important structural differences, suggesting that, despite the formal diagram language, there exist multiple acceptable ways to map a given dialogue component in the context of the graph as a whole.

Consider again the graph in Figure 1. As discussed above, the test annotated *T3* is linked to lines 118-121 of the argument. In this extract as shown below, Mr. Sherman, the advocate for Mr. Burnham, is responding to a question from the justice about the lower court's decision. In so doing he describes the lower court's reasoning and advances his own reading of its significance:

116. QUESTION: – maybe we made a mistake granting the case.
117. MR. SHERMAN: It doesn't say that jurisdiction was being upheld on the ground that he has sufficient contacts under the minimum contacts test, and it couldn't have because it was quite obvious that the trial court did not ground jurisdiction on that basis, and it couldn't have because under California law and under this Court's decisions there are insufficient contacts as a matter of law to uphold jurisdiction here for two reasons.
118. The first is that doing business in a state only gives rise to jurisdiction under the traditional test, if that's what the Court thought it was applying, or if that's what this Court wanted to apply, for causes of action arising out of that business. And the cause of action here does not arise out of the business that Mr. Burnham did in California.
119. The wife is seeking to invalidate a marital settlement agreement that she executed in New Jersey. That has nothing to do with the husband's contacts in California.
120. Secondly, California [*17] has held in Modlin v. Superior Court, which is in our briefs, that if a person, a father, combines coming to a state to visit his children with doing some business, the combination of those do not meet the minimum contacts test. That's Modlin v. Superior Court, 176 Cal.App.3d 1176.

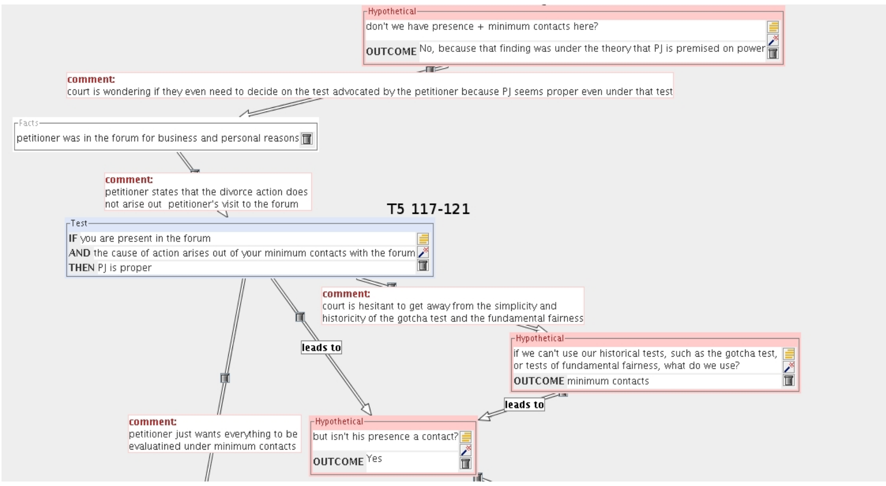

Figure 3. Argument Diagram Segment.

In Figure 1 the user summarized this text with a Test box stating [2]:

IF defendant is doing business in the forum state.
AND suit does not arise from these contacts.
EVEN THOUGH defendant is there for dual purpose of business and visiting kids
THEN jurisdiction does not exist over defendant.

On this reading, in the context of the graph, the test is represented as a modification of the proposed rule from the one articulated in lines 30 – 33 and is related to a subsequent hypothetical *H2* regarding child custody. Grader A ranked the graph in the top bin (*Good better*) and assigned it a score of $\frac{12}{12}$. When grading the graph itself grader A noted that he had not included this portion of text in his own graph but considered this a good alternative interpretation. The test itself was scored highly by grader A. Grader B scored the graph highly but not as highly assigning it the second highest rank and a score of $\frac{9}{12}$. As with grader A, grader B assigned the test full scores.

A different annotation for the same transcript item is shown in Figure 3. In this diagram, the student has omitted the hypothetical located at line 122 and represented lines 117-121 as a test stating:

IF you are present in the forum
AND the cause of action arises out of your minimum contacts with the forum
THEN pj is proper

Unlike the prior student's work, this framing uses a simpler logical format and omits the qualification. Like the previous graph, it was ranked highly by both graders; both assigned it to the top rank and gave a high overall score of $\frac{11}{12}$. However, as an individual test it was scored lower by both graders due to poor connections with the relevant hypotheticals. As this example illustrates, the graders may agree about the overriding quality of an argument diagram without necessarily agreeing on the details of the representations nor does there appear to be a single optimal diagram.

---

[2]Here bold letters represent structural items.

6. Conclusions

Our initial analysis of grader agreement is quite promising. The faculty members were able to define a relatively clear set of grading criteria and to come to agreement about the quality of each graph. This agreement was true both for the overall grades that were assigned after a detailed analysis but also for their first pass gestalt grades. This is important as it indicates that consistent assessments may be made by human experts with reasonable ease. However, the grading process highlighted the fact that often a given argument text could be reasonably represented diagrammatically in a variety of ways. A detailed examination of the graphs indicates a range of variation even between good diagrams both in terms of the structural relationships between nodes and the contents of each element.

Given the open texture of legal rules and concepts and the variability of interpretation endemic in natural language, the variation in argument diagrams is not unexpected, even in an annotation task such as ours. This variation is a key sticking point for argument diagrams or any other formal models of reasoning. The flexibility of natural language facilitates creative argumentation. In order for any model to serve as a robust tool for communication and instruction, it must accommodate this flexibility. Conversely, in order for a model to be computationally tractable it must avoid, minimize, or finesse this variability, a difficult design challenge. While relatively simple logical differences in interpretation of the type we highlighted in the above example may be identified and dealt with, more complex structural differences will be problematic.

We are continuing our analysis of the LARGO diagrams. Having shown that experts' may agree with respect to their general diagnoses, we are now analyzing their detailed rankings. This includes a focus on both the finer-grained analyses provided by the experts, such as individual test and hypothetical grades, as well as other sources of diagnostic information such as written assignments. In the future we plan to compare students' diagrams with their written assignments. Our goal is to identify pedagogically useful diagram characteristics that both predict students' later misconceptions or argument decisions and which may be used to provide guidance in future versions of the system.

Acknowledgments

NSF Grant IIS-0412830, Hypothesis Formation and Testing in an Interpretive Domain, supported this work.

References

[1] K. Ashley, C. Lynch, N. Pinkwart, and V. Aleven. A process model of legal argument with hypotheticals. In E. Francesconi, G. Sartor, and D. Tiscornia, editors, *Legal Knowledge and Information Systems, Proceedings Jurix 2008: 21st Annual Conference*, pages 1–10, 2008.

[2] T. J. M. Bench-Capon and T. F. Gordon. Isomorphism and argumentation. In *International Conference of Artificial Intelligence and Law*, pages 11–20. ACM, 2009.

[3] C. Carr. Using computer supported argument visualization to teach legal argumentation. In *In Visualizing Argumentation*, pages 75–96. London, Springer.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[5] M. Easterday, V. Aleven, and R. Scheines. 'tis better to construct than to receive? the effects of diagramming tools on causal reasoning. In R. Luckin, K. Koedinger, and J. Greer, editors, *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, pages 93–100. Amsterdam, Netherlands: IOS Press, 2007.

[6] T. Gordon. Visualizing carneades argument graphs. *Law, Probability and Risk*, 6(109), 2007.

[7] T. Gordon, H. Prakken, and D. Walton. The carneades model of argument and burden of proof. *Artificial Intelligence*, 171:875–896, 2007.

[8] K. Lund, G. Molinari, A. Sjourn, and M. Baker. How do argumentation diagrams compare when student pairs use them as a means for debate or as a tool for representing debate? *Computer-Supported Collaborative Learning*, 2(273), 2007.

[9] C. Lynch, K. D. Ashley, N. Pinkwart, and V. Aleven. Computational argument as a diagnostic tool: The role of reliability. In *Proceedings of the Fall Symposium on The Uses of Computational Argumentation, Washington D.C., USA*. AAAI, 2009. (in press).

[10] D. N. MacCormick and R. Summers, editors. *Interpreting Precedents: a Comparative Study.* Ashgate/Dartmouth., 1997.

[11] J. McClure, B. Sonak, and H. K. Suen. Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, 36(4):475–492, 1999.

[12] M. Paolucci, D. Suthers, and A. Weiner. Automated advice-giving strategies for scientific inquiry. In *Proceedings 3d Intl Conference Intelligent Tutoring Systems (ITS-1996)*, 1996.

[13] N. Pinkwart, V. Aleven, K. Ashley, and C. Lynch. Evaluating legal argument instruction with graphical representations using largo. In *Proceedings Artificial Intelligence and Education 2007. Marina Del Rey, CA.*, July 2007.

[14] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 13(4):961–980, 2004.

[15] C. Spearman. The proof and measurement of association between two things. *Amer. J. Psychol.*, 15:72–101, 1904.

[16] D. D. Suthers and C. D. Hundhausen. Learning by constructing collaborative representations: An empirical comparison of three alternatives. In P. Dillenbourg, A. Eurelings, and K. Hakkarainen, editors, *European Perspectives on Computer-Supported Collaborative Learning, Proceedings of the First European Conference on CSCL.*, pages 577–584. Maastricht, the Netherlands., 2001.

[17] S. Toulmin. *The Uses of Argument.* Cambridge University Press., Cambridge, 1958.

[18] T. van Gelder. The rationale for rational. *Law, Probability and Risk: Special Issue on Graphic and Visual Representations of Evidence and Inference in Legal Settings.*, 6(1-4):23–42, 2007.

[19] D. Walton. *Argumentation Methods for Artificial Intelligence in Law.* Springer., Berlin, 2005.