AIED Applications in Ill-Defined Domains

Proceedings of a workshop held during AIED 2007,
the 13th International Conference on Artificial Intelligence in Education

Marina del Rey, CA
July 10, 2007

Vincent Aleven
Kevin Ashley
Collin Lynch
Niels Pinkwart

Workshop co-chairs

Preface

This volume constitutes the proceedings of the workshop on AIED Applications in Ill-Defined Domains, held in conjunction with the Thirteenth International Conference on Artificial Intelligence in Education in Marina Del Rey, California (AIED 2007). This workshop is the second in what we hope will become a series devoted to the special challenges and opportunities of developing intelligent tutoring systems and other AIED applications for Ill-Defined Domains. The first took place at the Eighth International Conference on Intelligent Tutoring Systems (ITS 2006) in Jhongli, Taiwan.

AI-supported educational systems have made great strides in recent years both as research tools and teaching applications. Most of the AIED research and development to this point have been in well-defined domains such as physics, mathematics, or chemistry. Such domains are characterized by a well-accepted theory or model that makes it possible unambiguously to classify problems as correct or incorrect. Typically quantitative, well-defined domains are often taught by human tutors using such unambiguous problems as training examples. Such domains are particularly amenable to model-tracing tutoring systems. Operationalizing the domain theory makes it possible to identify problems for study, provide a clear problem solving strategy, and assess results definitively based on the existence of unambiguous answers. Help is readily provided by comparing the students' problem-solving steps to the existing domain models.

Not all domains of teaching and inquiry are well-defined; indeed most are not. Domains such as law, argumentation, history, art, medicine, and design are ill-defined. Ill-defined domains lack well-defined models and formal theories that can be operationalized; typically problems do not have clear and unambiguous solutions. Often even well-defined domains are increasingly ill-defined at the edges where new knowledge is being discovered. For these reasons, ill-defined domains are often taught by human tutors using exploratory, collaborative, or Socratic instructional techniques.

Ill-defined domains present a number of unique challenges for researchers in Artificial Intelligence in Education, but also exciting opportunities. The challenges include: 1) Defining a viable computational model for aspects of underspecified or open-ended domains; 2) Development of feasible strategies for search and inference in such domains; 3) Provision of feedback when the problem-solving model is not definitive; 4) Structuring of learning experiences in the absence of a clear problem, strategy, and answer; 5) User models that accommodate the uncertainty of ill-defined domains; and 6) User interface design for AI-supported educational systems in ill-defined domains where usually the learner needs to be creative in his actions, but the system still has to be able to analyze them. These challenges also present opportunities; if the AIED community learns how to address them systematically, it can finally branch out from the traditional domains into newer arenas of higher and professional education where complex problem-solving under conditions of uncertainty represent the norm.

The papers contained in this volume demonstrate promising approaches toward developing, applying, and evaluating AI-supported educational systems for ill-defined domains, addressing the challenges mentioned above.

This volume contains four long and two short research papers presenting work in a variety of domains. Some of these, like legal argumentation and psychology demonstrate the potential for applying AIED approaches to some new and largely unexplored fields of pedagogy. Other papers deal with more seemingly well-defined domains involving medicine, causal reasoning, and language learning, but illustrate how even these involve the need for interpretation, trial-and-error under conditions of uncertainty, and dealing with ambiguity.

Apart from the different domains described in the papers of this volume, the methods and tutoring approaches also vary. Some papers show attempts to adapt to more ill-defined fields paradigms that have been successful in well-defined domains such as a constraint-based approach. Others show new methods for dealing with the problems of building AI-supported educational systems in ill-defined fields, including diagrammatic representations of arguments and complex solutions, improved conceptual scaffolding, temporal Bayesian networks and computer-based simulations, or specialized tutoring agents.

Jerry Andriessen, Paul Brna, Jill Burstein, Rebecca Crowley, Andreas Harrer, H. Chad Lane, Susanne Lajoie, Liz Masterman, Bruce McLaren, Antoinette Muntjewerff, Katsumi Nitta and Beverly Woolf have reviewed the paper submissions for this workshop. Thank you for helping us to organize this interesting event. Last but not least, we thank the participants of the workshop for contributing their ideas and research results.

Vincent Aleven, Kevin Ashley, Collin Lynch, and Niels Pinkwart
June 2007

Organizing committee

Vincent Aleven, Carnegie Mellon University, USA
Jerry Andriessen, University of Utrecht, The Netherlands
Kevin Ashley, University of Pittsburgh, USA
Paul Brna, University of Glasgow, UK
Jill Burstein, Educational Testing Service, USA
Rebecca Crowley, University of Pittsburgh, USA
Andreas Harrer, University of Duisburg-Essen, Germany
H. Chad Lane, Institute For Creative Technologies, USC
Susanne Lajoie, McGill University, Canada
Collin Lynch, University of Pittsburgh, USA
Liz Masterman, Oxford University, UK
Bruce McLaren, German Research Center for Artificial Intelligence, Germany
Antoinette Muntjewerff, University of Amsterdam, The Netherlands
Katsumi Nitta, Tokyo Institute of Technology, Japan
Niels Pinkwart, Clausthal University of Technology, Germany
Beverly Woolf, University of Massachusetts, USA

Table of contents

1 Towards the Design of A Representational Tool To Scaffold Students' Epistemic Understanding of Psychology in Higher Education
Katerina Avramides and Rose Luckin

11 Themis, a Legal Agent-based ITS
Ig Bittencourt, Evandro Costa, Baldoino Fonseca, Guilherme Maia, and Ivo Calado

21 A Framework for Building Intelligent Learning Environments in Ill-defined Domains
Vu Minh Chieu, Vanda Luengo, Lucile Vadcard, and Dima Mufti-Alchawafa

31 The logic of Babel: Causal reasoning from conflicting sources
Matthew W. Easterday, Vincent Aleven, and Richard Scheines

41 Mapping and Validating Case Specific Cognitive Models
Geneviève Gauthier, Susanne P. Lajoie, and Solange Richard

51 Argument diagramming as focusing device: does it scaffold reading?
Collin Lynch, Kevin Ashley, Niels Pinkwart, and Vincent Aleven

61 Resolving Ambiguity in German Adjectives
Amanda Nicholas and Brent Martin

Towards the Design of A Representational Tool To Scaffold Students' Epistemic Understanding of Psychology in Higher Education

Katerina AVRAMIDES[a] and Rose LUCKIN[b]
[a]*IDEAS lab, Dept. of Informatics, University of Sussex, Brighton, UK*
[b]*London Knowledge Lab, Institute of Education, University of London, London, UK*

Abstract. The paper calls for a consideration of students' understanding of the nature of knowledge and knowing (termed their *epistemic cognition*) in the design of AIED applications in ill-defined domains. The importance of students' epistemic cognition is discussed with reference to both the characteristics of ill-defined subject matter in general, and to specific domains. It is suggested that, although common characteristics can be identified across many domains, the nature of knowledge, such as how knowledge is justified, is different in each area of study. Hence a domain-specific consideration is also necessary in designing effective applications. The paper discusses an interview-based study of psychology students' epistemic cognition in the context of writing a formally assessed essay. The findings inform the preliminary phase in the design of a representational tool to scaffold student learning. Although current representational tools are powerful, they do not scaffold students' epistemic understanding of the subject matter. The present design aims to address these issues.

Introduction

The distinction between ill-defined (or ill-structured) and well-defined subject matter has, typically, been made in the context of problem-solving [1]. Well-defined problems consist of a well-defined statement that presents all elements of the problem to the solver, and a finite number of operations that can be applied to reach a solution. A solution is unambiguously either correct or incorrect. In contrast, the initial and goal states of ill-defined problems are subject to interpretation, there is no formal structure to the problem-solving process, and the adequacy of solutions is judged against ill-defined criteria. This distinction has been taken to the level of domain [2, 3]. Well-defined domains are defined as those in which many phenomena are described consistently across cases by scientific principles and formal models. Examples of such domains are chemistry and physics. Ill-defined domains are those in which the phenomena under investigation cannot be conceived of within a well-defined framework, as the concepts involved cannot be ascribed a well-defined meaning. Moreover, the methods of investigation and analysis are subject to the same ill-definedness. Examples of such domains are psychology, philosophy, art and history.

Corresponding Author: Katerina Avramides. IDEAS Lab, Dept. of Informatics, University of Sussex, Falmer, Brighton BN1 9QU, UK. E-mail: K.Avramides@sussex.ac.uk

Kitchener [4] discusses the different processes involved in solving well-defined and ill-defined problems and argues that solving ill-defined problems is different in that it engages a level of processing above cognition and metacognition, which she terms *epistemic cognition*. This level of cognitive processing "is characterised as the processes an individual invokes to monitor the epistemic nature of problems and the truth value of alternative solutions" (p.225). There is a growing body of research on people's, particularly students', epistemic cognition under a variety of terms, such as epistemological beliefs, personal epistemology and epistemic resources [5-8]. At the risk of simplifying a complex concept, it can be said that a sophisticated understanding of knowledge and knowing entails an understanding of the complex, socially constructed nature of knowledge. Findings strongly suggest that epistemic cognition is linked to academic learning, and that the majority of students lack such an understanding of knowledge [9].

The paper first considers the characteristics of ill-defined domains and the difficulties they pose to learning. The concept of epistemic cognition and the importance of considering it in the design of AIED applications are then discussed within the context of ill-defined domains. More specifically, epistemic cognition is discussed in relation to both the general characteristics of ill-defined domains and the nature of knowledge and knowing within specific domains. It is argued that there are important domain differences in how knowledge is developed and how it is justified. Hence, it is suggested that, although ill-defined domains share common characteristics, it is necessary to also consider the nature of knowledge in specific areas of study when designing AIED applications. The paper then focuses on the domain of psychology and discusses preliminary design considerations of a representational tool to scaffold student learning. The design is grounded in an interview-based study of psychology students in higher education.

1. Ill-defined domains

1.1. Characteristics of ill-defined domains

The notion of ill-definedness has been considered, predominantly, within the context of problem-solving, but also at the level of domain [2, 10]. Similar characteristics have been identified at both levels of analysis. This section considers two analyses that come from an educational perspective. Jonassen [11] defines ill-defined problems by the following criteria: (a) they involve unknown elements, (b) there exists no unambiguously correct solution, there may be multiple solutions or no solution, (c) many paths exist to solving the problem, the validity of which cannot be judged by absolute criteria, and (d) solvers are often required to make personal judgements. Lynch et al. [10] consider ill-definedness at the level of domain and identify the following five characteristics from a review of the literature: (a) the lack of unambiguous criteria by which to verify the validity of solutions to problems, (b) that the development of formal theories is not compatible with the nature of ill-defined domains, (c) that even at a novice level solving problems in ill-defined domains involves a process of design and not application of formal theories, (d) the ubiquity of concepts that cannot be ascribed an absolute definition, and (e) that problems cannot be decomposed into independent subproblems. Both the above analyses consider similar

issues. Ill-defined subject matter is characterised by concepts that cannot be ascribed a precise meaning. Hence the issues we are dealing with, how we reason with them, and how we evaluate our reasoning, also cannot be defined precisely.

1.2. Learning in ill-defined domains

Even within a constructivist framework, students must be given some form of 'building blocks' with which to construct. These 'building blocks' must be accepted by the learner at face value, because they form the domain itself, the questions that are asked and the ways in which these are addressed. Learners must comprehend what the knowledge constructing enterprise within the given domain is about and what tools are used to develop knowledge, before they can go on to develop their personal understanding of it. Given the above characteristics of ill-defined domains, these 'building blocks' cannot be defined precisely. The complexity and ambiguity are present at the novice level. Understanding knowledge in ill-defined domains and how to construct it requires an understanding of how to work with this complexity and ambiguity to reach conclusions. In other words, understanding the processes involved in justifying knowledge. This relates to one's understanding of the nature of knowledge and knowing (*epistemic cognition*), which is discussed in the following section.

2. Epistemic cognition

2.1. Defining epistemic cognition and its importance in learning

Epistemic cognition is a slippery concept that is difficult to discuss in concrete terms. It is, broadly, defined as people's ideas about the nature of knowledge and knowing [9]. It is, typically, conceived of as deriving from philosophical epistemology, which is concerned with the nature of knowledge, its sources and limits [12]. Beginning with William Perry's [8] empirical work on students' intellectual development, educational psychologists became interested in individuals' understanding of the nature of knowledge and knowing. However, the link between philosophical epistemology and epistemic cognition is not direct, although this is not clear in the literature. Epistemic understanding is conceived of in an educational sense. For example, when considering people's ideas about the source of knowledge, researchers are not referring to their beliefs about the fallibility of the senses, but to their understanding of knowledge as something that is constructed by the learner, rather than coming from authority. For the purpose of illustrating what epistemic cognition refers to (though an oversimplification), it can be said that conceptions range from viewing knowledge as a direct representation of an objective truth, to understanding that knowledge is relative to methods of observation and conceptual analysis with some positions better supported than others.

Several theoretical frameworks have been developed to describe this aspect of people's thinking (e.g. personal epistemology [9], epistemological resources, [5], epistemological beliefs [6]). Although there are similarities, in essence, each approach reflects a different conceptualisation and each dictates a different methodological approach. These differences are fundamental. Is epistemic cognition a trait-like characteristic that develops in 'stages', a multidimensional system of independent

beliefs, or a set of context-dependent resources? Can it be 'measured' independently of context by interviews and questionnaires, or only deduced from specific contexts? This lack of a coherent theory and methodology has impeded research into this important aspect of learning going beyond a narrow audience within educational psychology. However, despite this theoretical abstruseness, research to date strongly suggests that it plays a significant role in learning. Empirical findings have related epistemic cognition to various aspects of learning, such as cognitive processing strategies, conceptual change learning and academic achievement (see [9] and [13] for an overview of the main frameworks and empirical work).

The approach adopted in the present research will be described at a general level, as it is beyond the scope of this paper to consider theoretical issues in detail. Epistemic cognition is defined here as people's understanding of how knowledge is justified (what makes a belief knowledge). In other words, not only a high level idea that some positions are better supported than others, but an understanding of why. It is conceived of as a context-dependent conception and not a stable belief that guides behaviour. In accordance with a sociocultural approach, it is postulated that epistemic cognition is formed through the way that knowledge is communicated to learners from experts, written materials, and their discussions with peers. The implications of this are that the way that we design teaching applications will impact on how learners understand the nature of the subject matter. And that, in order to understand how students' conceptions are formed and how we might support a more 'sophisticated' understanding, we need to study how they engage with subject matter in specific educational contexts.

2.2. Epistemic cognition in relation to ill-defined domain characteristics

A theoretical analysis places epistemic cognition at the centre of learning in ill-defined domains. Students' understanding of how knowledge is justified will determine their conception of what domain knowledge is and how to go about understanding it. For example, students that conceive of knowledge as a direct representation of reality that is justified by direct observation will likely seek the 'truth' amongst conflicting interpretations and alternative solutions. For example, a history student that views knowledge as accurate accounts of past events will likely conceive of knowledge constructing as identifying the truth in historians' interpretations. Even if students understand that knowledge is a socially constructed representation that emerges from considering empirical evidence and reasoned argument, they may not fully understand how empirical evidence is based on theoretical frameworks. For example, a psychology student that understands that 'intelligence' can be defined in many ways, but lacks an understanding of the dependency of empirical evidence on the method by which it was collected, is likely to draw unjustified conclusions from it.

Therefore, in communicating the nature of ill-defined subject matter, we need to also communicate to students that concepts can be defined, valid solutions can be reached in spite of uncertainty, and that personal judgements need to be justified. There is strong evidence that many students view knowledge as 'discovery' of reality or an utterly subjective enterprise [9]. Thus we need to consider what conception students are forming from the educational experiences we design and how we can support them in understanding how knowledge is constructed. This should be considered at the level of domain. Each domain deals with different areas of experience and so the nature of the

subject matter and, consequently, how knowledge of it is justified are different, as discussed in the following section.

2.3. Epistemic cognition in relation to specific domains

Although domains can be classified as well-defined or ill-defined, there are important differences between them. For example, the design of a usable interface to an airport control system, answering a question on the impact of colonisation on the culture of aborigines, critiquing the literary status of a particular novel, diagnosing a medical condition, considering the legal justification of a war, and considering the influence of the home environment on children's self-esteem, are all ill-defined subjects. However, the nature of knowledge in each is quite different, as is the process of answering them and justifying the validity of that answer. For example, what counts as justification in philosophy is reasoned argument, what counts as justification in cognitive science is empirical evidence derived from a reasoned theoretical framework, and what counts as justification in a court of law is reasoned argument based around empirical evidence that is of a different nature.

Knowledge domains have been categorised along other dimensions, such as Biglan's [14] classification along the hard/soft, pure/applied and life/non-life dimensions. Each categorisation will necessarily generalise on domain differences. The well-/ill-defined dimension is not criticised as irrelevant. Rather, it is argued that if the aim of AIED applications is to teach domain knowledge, this distinction is not adequate on its own, particularly if we consider what the nature of knowing is in each domain.

It is a difficult task to consider the nature of knowledge and knowing. This is not least because there is no correct way of conceiving of it. Scientists, science educators and philosophers of science disagree between and amongst themselves about the nature of human knowledge [15]. Moreover, ill-defined domains in particular are characterised by the lack of a single paradigmatic approach to knowledge development. However, this does not mean we can ignore the issue. In designing educational technology we need to consider what conception of knowledge and knowing we are communicating to learners.

2.4. The design of AIED applications to support epistemic understanding

Research within the AIED community has considered the impact on learning of many learner characteristics, such as motivation and metacognition [16]. Particularly as there is an increased interest in designing applications for ill-defined domains, it is timely to consider the impact of epistemic cognition on learning. This is not meant to imply that epistemic cognition is not relevant to learning within well-defined domains. For example, there is substantial research that investigates how to convey the notion of "science-in-the-making" as opposed to "science-as-discovery", particularly within the context of collaborative systems [17]. However, it is especially relevant to teaching ill-defined subject matter, as working with knowledge from a novice level requires an understanding of the relative validity of different theoretical ideas and methodological tools.

Computer technology is potentially well-suited to teaching the epistemic nature of ill-defined subject matter [2, 18, 19]. For example, non-linearity allows the context-

sensitivity of concept definitions to be illustrated by linking multiple definitions of concepts to different contexts. The structure of arguments and counterarguments can be represented visually. Case studies and evidence for and against a claim can be represented in various mediums. The strength of links between claims and arguments or evidence can be represented diagrammatically.

Some research into collaborative systems has considered learners' epistemic understanding [17]. This has focused on issues of representing knowledge to learners. For example, Belvedere [20] allows students to link hypotheses to data that support or falsify it. SenseMaker [17] uses argument maps to scaffold an understanding of the relationship between theory and evidence. However, such systems to date have focused on well-defined domains. Mapping representations have been used to represent ill-defined domains (for example [21, 22]), and although they are complex and powerful, their design does not relate to an epistemic understanding of the subject matter. Some systems scaffold an understanding of the complexity and structure of the subject matter [23], but not an understanding of the validity of claims and how this is assessed. The aim of the research described in the following sections is to explore how students' epistemic cognition shapes their approach to constructing knowledge of ill-defined subject matter in the domain of psychology. The further aim is from this research to develop a representational tool to explore how engaging them in representing material in a particular way might challenge their conceptions of the nature of knowledge.

3. A study of students' epistemic cognition in psychology in higher education

3.1. Study description

3.1.1. Study design

The aim of the study was to explore the impact students' epistemic cognition may have on the way they approach learning in an ill-defined domain. The domain was psychology, and the learning context was a formally assessed essay. Essay writing is not simply a process of utilising knowledge that has already been constructed (as in an exam), but involves a process of research and learning. Thus it allowed the opportunity to explore how students' epistemic cognition might shape the learning process within this context.

3.1.2. Participants

Eight participants were recruited from psychology courses at the University of Sussex that required them to write an essay as part of their formal assessment. Half were undergraduate students in their second year of study (all female with an age range of 20 to 41) and the other half were taught postgraduate students (3 female and 1 male with an age range from 25 to 37). They were paid for their participation in this study.

3.1.3. Data collection

Participants were interviewed twice, once before and once after they had completed their essay. Interviews were semi-structured and lasted approximately 45 minutes. They were asked to keep any handwritten notes, keep track of literature searches and

also include diary-like comments on anything that stood out during the writing process. They were also required to write their essay on a Microsoft Word document that was set-up with a macro to save a version of the document every 15mins.

The first interview focused on the essay writing process and included questions on their view of essay writing as a form of formal assessment, their perception of what structure an essay should have, commonalities in their writing process from past experience, and the tools they use (e.g. paper, mind-maps, word-processing). The second interview focused on the way they organise the material and prompted them to explain the specific subject of their essay, whether they have formed an opinion on the subject, whether they believe it is possible to form an opinion, their assessment of their knowledge and how they justify this assessment.

3.1.4. Missing data

Participant G dropped out after the first interview. The data from the Microsoft Word logs was incomplete for participants E, F and H, and participant B did not use the correct file.

3.2. Study findings and design implications

The data was analysed as separate case studies, as the small number of participants did not allow for any aggregation or statistical analyses. The aim was an in-depth analysis of individual students' approach. The analysis is not yet complete, so this section does not present the final findings from the study, only a subset. It is also beyond the scope of this paper to consider the theoretical aspect of the analysis that relates to the definition and study of epistemic cognition. Space limitations do not allow a detailed consideration of each case study or even the report of detailed quotes from the interviews. However, meaningful themes can be drawn out from the data. A few of these are presented together with design implications for a representational tool.

3.2.1. Knowing does not involve having a personal perspective

Participants B, C, D, E, F and H emphasise their lack of expertise that prevents them from being able to critique research findings. Participant E says she may sometimes disagree with experimental design, but would not have a better idea of how to design the study. Participant D reports she has great difficulty in forming her own opinion, as there is always conflicting evidence. Moreover, in the way they approach researching and writing their essays they do not appear to be trying to form their own opinion of the material. The issue is not that they report they cannot claim an opinion. Rather it is that, despite this, they mostly rate their knowledge as high. This indicates, that knowing for them is equated with knowing of 'stuff', of experts' opinions within the field, not of forming their own conception of it. It can of course be argued that, although the interview questions emphasised a broader context of knowing outside of the given assignment, they were possibly considering knowledge within the limits of their student status. However, although they cannot be expected to have such depth of understanding, as learners it should be something they strive for and consider part of knowing. A representational tool that highlights agency in knowing as a part of a knowledge representation may at least probe them to consider a more active role of themselves as learners in the knowledge constructing process. One way of achieving

this might be to require students to evaluate their confidence in the certainty of claims and the degree to which they understand the links between claims and empirical evidence (possibly using a colour coding scheme). For one participant at least, the interviews suggest that when asked to elaborate on their knowledge they began to question their knowledge ("yeah, quite a lot I don't know actually").

3.2.2. Considering the context of knowledge

Some participants indicated they had difficulty understanding the context-dependency of ill-defined concepts. For example, participant C does not consider the specific definitions of educational practice and culture that are adopted in the research she discusses and that this is only a subset of possible definitions. She does discuss the relative validity of different studies. However, only in terms of problems with their design, not the theoretical framework they are based upon. Participant E discuses that there are many factors that could be impacting on the social problem of bullying, but does not consider the contextual issues in the study of these factors. A representational tool could highlight this context-dependency of concept definitions by prompting students to specify how they were defined in a particular study. This could also highlight issues in comparing findings across studies and how, in the domain of psychology, the conceptual analysis of an issue impacts on what is considered important in empirical investigations.

3.2.3. Dealing with an unlimited conceptual space

Unsurprisingly, participants found it hard to narrow down the essay question to a scope they could manage. Participant C deals with this difficulty by first deciding what she wants to say, according to what argument she thinks is easiest to support, and then fitting the evidence to support her argument. She says "when writing your paper you can tweak it to make it look more valid, use stuff that supports your point and leave stuff that doesn't". The scope of her paper is quite limited, though she does not consider this an issue, and her self-assessment of her knowledge on the question is very high. Participant D talks in some detail about the difficulties she encountered when beginning a university course in history, where there was far less structure in the material that was given to her. She spent hours reading information, unable to filter through what was relevant. She has now consciously adopted the strategy of first reading textbooks and deciding what she wants to say and then focussing on specific research articles. She gives quite a broad overview of the topic, and indicates she has difficulty conceiving of it in a narrow sense. She says that the more she thinks about it the more things there were to include. Both participants face difficulty with the lack of boundaries in the conceptual analysis of the essay question. One could legitimately question how much understanding participant C has gained from the experience, as she deals with the difficulty by treating the exercise as a 'game'. Participant D attempts to form a broader picture of the issues, but is overwhelmed and is unable to integrate all the information into an understanding of her specific topic. She still, also rates her knowledge as quite high.

Both participants may be assisted by representing the broader picture of the issues, not only of the areas they have read about, but also of those that they are aware exist, but they have not had the time to study. Also indicating their perceived level of understanding in each area might also help to guide their learning. In the case of

participant C, it may help her understand the limited scope of her conceptual analysis and possibly prompt her to re-evaluate her assessment of her knowledge. Participant D may be aided in visualising what she does know and the adequacy of this for the purposes of the given essay, as opposed to fearing that she has not covered enough material. She may then be able to integrate and make sense of the material that she has covered.

The issue is that they are novices and, obviously, cannot cope with the full scope and complexity of the material. But they also have no conception of what their scope of understanding is or any guidance as to what level of consideration is appropriate. Knowledge of the specific issue cannot be isolated, it is embedded in the larger picture, which D is aware of, but does not know how to cope with, and C does not appear to be aware of. There is a need to represent that knowledge cannot emerge from a narrow consideration of a topic, but, equally, does not require a consideration of every conceivable issue.

3.3. Study limitations and future steps

The small number of participants allowed an in-depth analysis of their experience in this specific learning context. It also meant that the extent to which the findings can be generalised is limited. However, the results suggest that these students at least were facing difficulties in understanding the nature of their subject and that this had an effect on the particular learning experience. Given the ubiquity of essay writing as a form of formal assessment in higher education, this will not be a problem confined to a handful of learning experiences. We need to consider how the way we are representing knowledge to students, and the way we are asking them to represent their own knowledge, impacts on their thinking and learning.

The design process of the current tool is still at the early stages. Further research will explore what representations may be more effective in communicating a desired conception of knowledge and knowing. At the moment it is conceived of as a concept-mapping tool that students will use during the researching and writing of an essay. It will require them to identify aspects of their thinking such as their conceptual analysis of the relevant issues, the sources on which they based this and the adequacy of it. It will also include a representation of the empirical evidence that they use to support claims and how it addresses the issues identified in the conceptual analysis. It will also prompt them to rate (and possibly justify) their understanding. The results of an evaluation of such a tool will inform whether requiring students to represent their knowledge within a particular framework can influence the way they conceive of it.

4. Concluding comments

The aim of this paper was to highlight the importance of epistemic cognition in learning in ill-defined domains. Domains differ in the way they justify knowledge claims and this is an integral part of developing domain knowledge. There are, of course, many factors that affect learning in any given context. Research within the AIED community has evolved from its early roots in intelligent tutoring systems to designing technology-enhanced learning contexts of increasing complexity, such as collaborative learning and augmented reality. It has also expanded into a theoretical

framework of learning that goes beyond the cognitive to include the metacognitive and motivational aspects of learning. The present research takes this a step further and calls for an exploration of how students' epistemic cognition impacts on learning and how the design of AIED applications can scaffold students understanding of the nature of knowledge and knowing.

References

[1] Jonassen, D., *Instructional Design Models for Well-Structured and Ill-Structured Problem-Solving Learning Outcomes.* Educational Technology Research and Development, 1997. **45**(1): p. 65-94.

[2] Spiro, R.J., et al., *Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains.* Educational Technology, 1991. **31**: p. 24-33.

[3] Aleven, V., et al. *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*. in *8th International Conference on Intelligent Tutoring Systems*. 2006. Jhongli (Taiwan): National Central University.

[4] Kitchener, K.S., *Cognition, metacognition, and epistemic cognition.* Human Development, 1983. **26**: p. 222-232.

[5] Hammer, D. and A. Elby, *On the Form of a Personal Epistemology*, in *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*, B.K. Hofer and P.R. Pintrich, Editors. 2002, Erlbaum: Mahwah, NJ. p. 169-190.

[6] Schommer-Aikins, M., *Explaining the Epistemological Belief System: Introducing the Embedded Systemic Model and Coordinated Research Approach.* Educational Psychologist, 2004. **39**(1): p. 19-29.

[7] Baxter Magolda, M.B., *Evolution of a Constructivist Conceptualization of Epistemological Reflection.* Educational Psychologist, 2004. **39**(1): p. 31-42.

[8] Perry, W.G., *Forms of intellectual and ethical development in the college years: A scheme*. 1970, New York: Holt, Rinehart & Winston.

[9] Hofer, B.K. and P.R. Pintrich, *The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning.* Review of Educational Research, 1997. **67**(1): p. 88-140.

[10] Lynch, C., et al., eds. *Defining Ill-Defined Domains: A Literature Survey*. Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems, ed. V. Aleven, et al. 2006, National Central University: Jhongli (Taiwan).

[11] Jonassen, D., *Towards a Design Theory of Problem Solving.* Educational Technology Research and Development, 2000. **48**(4): p. 63-85.

[12] Dancy, J., *An Introduction to Contemporary Epistemology*. 1985, Oxford: Blackwell.

[13] Hofer, B.K. and P.R. Pintrich, eds. *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*. 2002, Lawrence Erlbaum Associates, Inc.: Mahwah, NJ.

[14] Biglan, A., *Relationships between subject matter characteristics and the structure and output of university departments.* Journal of Applied Psychology, 1973. **57**(3): p. 204-213.

[15] Alters, B.J., *Whose Nature of Science?* Journal of Research in Science Teaching, 1997. **34**(1): p. 39-55.

[16] Looi, C.-K., et al., eds. *Artificial Intelligence in Education: Supporting Learning Through Intelligence and Socially Informed Technology*. 2005, IOS Press: Amsterdam.

[17] Sandoval, W.A., et al., *Designing Knowledge Representations for Learning Epistemic Practices of Science*, in *The Annual Meeting of the American Educational Research Association*. 2000: New Orleans.

[18] Jacobson, M.J. and R.J. Spiro, *Hypertext Learning Environments, Cognitive Flexibility, and the Transfer of Complex Knowledge: An Empirical Investigation.* Journal of Educational Computing Research, 1995. **12**(4): p. 301-333.

[19] Jonassen, D., *Using Cognitive Tools to Represent Problems.* Journal of Research in Technology and Education, 2003. **35**(3): p. 362-381.

[20] Suthers, D.D., *Towards a Systematic Study of Representational Guidance for Collaborative Learning Discourse.* Journal of Universal Computer Science, 2001. **7**(3).

[21] Nicholson, P. and R. Johnson, *MetaMaps: Assessing, understanding of large, complex or distributed knowledge domains.* Education and Information Technologies, 1999. **4**(3): p. 297-312.

[22] McAleese, R., *The Knowledge Arena as an Extension to the Concept Map: Reflection in Action.* Interactive Learning Environments, 1998. **6**(3): p. 251-272.

[23] Pinkwart, N., et al. *Towards legal argument instruction with graph grammars and collaborative filtering techniques*. In M. Ikeda, K.D. Ashley, & T.W. Chan (Eds.) Proceedings of the *8th International Conference on Intelligent Tutoring Systems*. 2006. Berlin: Springer Verlag.

Themis, a Legal Agent-based ITS

Ig BITTENCOURT[ab], Evandro COSTA[a], Baldoino FONSECA[a], Guilherme MAIA[a], and Ivo CALADO[a]

[a]*Federal University of Alagoas -Computer Science Institute Tabuleiro dos Martins, Postal Code 57.072-970, Brazil, Maceio -AL GrOW - Group of Optimization of the Web*
[b]*Federal University of Campina Grande, Brazil, Paraíba*
e-mail: ibert@dsc.ufcg.edu.br

Abstract. As an interesting example of ill-defined domain, Law domain has been challenged AI-ED system researchers. In this context, Law students have little chance to deal with realistic situations, requiring to apply real cases, rules, and different viewpoints. To address these issues, we introduce an agent-based Intelligent Tutoring System (ITS) applied to the mentioned domain. Then, we defined an agent-based architecture to support multiple views of domain knowledge, improving the quality of student-ITS interactions and the learning success of the students. Each tutoring agent from the system contains a hybrid knowledge-based system that combines Case-Based Reasoning (CBR) and Rule-Based Reasoning (RBR). In addition, each agent adopts the Reinforcement Learning Algorithm aiming at identifying the best pedagogical strategy by considering the student profile. This paper focuses on both architecture and the mentioned Artificial Intelligence techniques into a Legal System. A case study to demonstrate the feasibility of the system is presented.

Keywords. Artificial intelligence and law, intelligent tutoring systems, case-based reasoning, rule-based systems, reinforcement learning.

Introduction

AIED system researchers have been challenged to approach ill-defined domains, such as example Law domain. Particularly in legal domain, several researches provides evidences that involve Law students with real cases, rules, and different viewpoints of knowledge is often recognized as important to their successful learning, such as [2,11,18,17]. Furthermore, the use of a hybrid solution to the problem solving is also motivated due to the structure of the juridical system[1]. For instance, legislation is the main Legal research, where magistrates make their decision based on the code and laws, originating case solutions. In addition, Legal Intelligent Tutoring Systems (ITS) are a kind of complex, domain-oriented software systems which are frequently pointed out by researchers as suitable applications for the multi-agent approach [8].

To address these issues, we introduce the so-called Themis, an ITS applied to Legal domain, according to the multi-agent architecture derived from Mathema model [7]. The main goal of this model is to increase the opportunities for students to construct their own knowledge through a problem-based learning approach. Moreover, Themis may also solve problems by using CBR or RBR or a combination of them. CBR has been used to check the similarity between old cases to justify new problems and RBR to evaluate the rules of Normative Knowledge. In addition, to improve the pedagogical interaction, the system adopts the Reinforcement Learning Algorithm aiming at identifying the best pedagogical strategy by considering the student profile.

In the presented approach, the idea is to engage Law students into interactions with ITS based on the resolution of Legal problems and their consequences on other tutoring activities, concerning the Penal Law. The starting point of these interactions occurs when ITS submits a penal situation to Law students. Then, they

[1] Civil Law, also known as Continental Law or Roman Law has been used in the system

will learn two fundamental but different skills of Legal problems. First, know how to identify relevant cases and Legal concepts (Normative Knowledge, for instance) of the cases. Second, know how to use them effectively as examples justifying position in a Legal argument.

Altogether this paper focuses on both architecture and the mentioned Artificial Intelligence techniques into a Legal System. A case study to demonstrate the feasibility of the system is presented.

1. Related Work

Some related works were developed taking into account legal tutoring or hybrid reasoning involving CBR and RBR.

In [1], Aleven proposes an intelligent learning environment designed to help beginning law students learn basic skills making use of arguments with cases.

An ITS for Legal domain, using Rule-Based System and approaching problem-based learning as pedagogical strategy is presented in [21]. This proposal refers to a novel ITS approach applied to Legal domain, using hybrid reasoning (CBR and RBR). It also describes the modeling of multiple views of domain knowledge, providing two-way interaction in a problem-solving process.

[5] combines both the blackboard architecture and distributed AI methods for creating hybrid systems. This means that both RBR and CBR run concurrently giving as output the best result produced by one of the inference mechanisms.

[19] describes a Dutch expert legal system, focused on the domain of landlord-tenant law. It combines knowledge groups like legislation, legal doctrine, expert knowledge and case law.

In [25], the project uses a distributed artificial intelligence approach, operating in the area of credit law that combines CBR and RBR independently. First, the system infers using RBR, thereafter CBR, if RBR does not succeed.

Although [1] and [21] propose an educational system, an intelligent mechanism (Reinforcement Learning) to improve the pedagogical activities were not found. In addition, [25], [19], and [5] do not propose an education system.

This paper proposes a novel ITS approach applied to Legal domain, using hybrid reasoning (CBR, RBR and Q-Learning). Moreover, Themis has an ontology-based approach in order to model domain knowledge, student interaction and pedagogical activities. Equally important, Q-Learning has been used to improve the quality of ITS-Student interaction.

2. Agent Architecture

Figure 1 shows the *Themis* architecture. The system is composed by mediator agent, persistence agent, and an agent society. In the infra-structure layer was used the Jade Framework, because it implements the interoperability standards for agent communication (FIPA [10]).

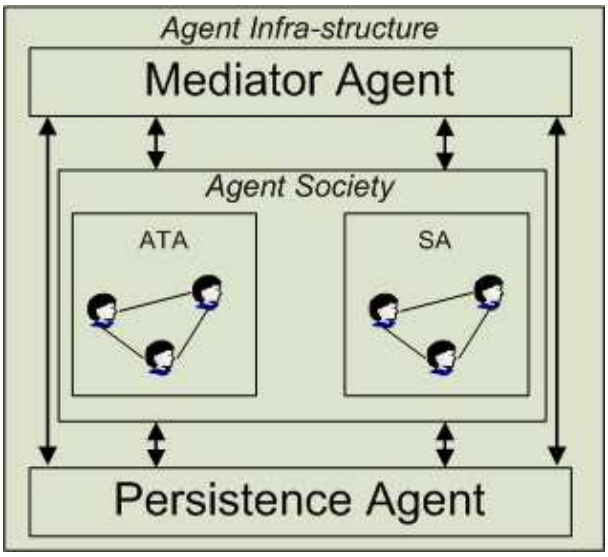


Figure 1. Themis Architecture.

The agent society is composed by artificial tutoring agents (ATA) and support agents (SA). While **ATA** represents an agent-based ITS acting into a specific domain. These agents are responsible for problem solving and providing information to students and each ATA has a domain model ontology (it has the features of a legal domain and sub-domains associated), student model ontology (it is composed by interaction information and the knowledge that a student already learned), and pedagogical model ontology (which is divided in i) strategies, which are defined as an elaborated plan of action built by instructors based on the educational theory and ii) tactics, which are atomic actions that can be used into a strategy.). The **SA** provides assistance to ATA agents through inference engines. The support agents are: 1) CBR Agent which is responsible for evaluating the similarity between the jurisprudence and a penal situation, 2) RBR Agent which is responsible to infer by using normative knowledge and 3) Q-Learning Agent: it is used in student-ITS interaction in order to choose the best tactic in a specific situation.

Finally, mediator agent assures the communication between graphical interface and agents, while persistence agents assure the communication with the knowledge bases.

3. Agents Implementation

This section describes Artificial Tutoring Agents and Support Agents.

3.1. Artificial Tutoring Agents

The Autonomous Tutoring Agents were modeled based on the Mathema Model [7] through the reuse [20,6] and development of top ontologies.

3.1.1. Domain Ontology

The characteristics of the domain is overcame through a multi-dimensional view of the knowledge (external view), which helps a partitioned view (conducting an internal view) of the domain. The external view represents a domain interpretation of a body of knowledge, while the internal view represents a partition of the domain D. Moreover, each partition of D leads to a sub-domain that are mapped into curriculum

structures. The curriculum is composed of pedagogical units (pu), as follows:

$$Curric = \{pu_1, pu_2, ..., pu_n\}, \quad (1)$$

Curric represents a curriculum and its associated pu_i. Also, each pu_i corresponds to a set of problems and each problem contains concept and results that assist the resolution process. Finally, each problem is associated with conceptual content to support the student, as shown in Figure 2.

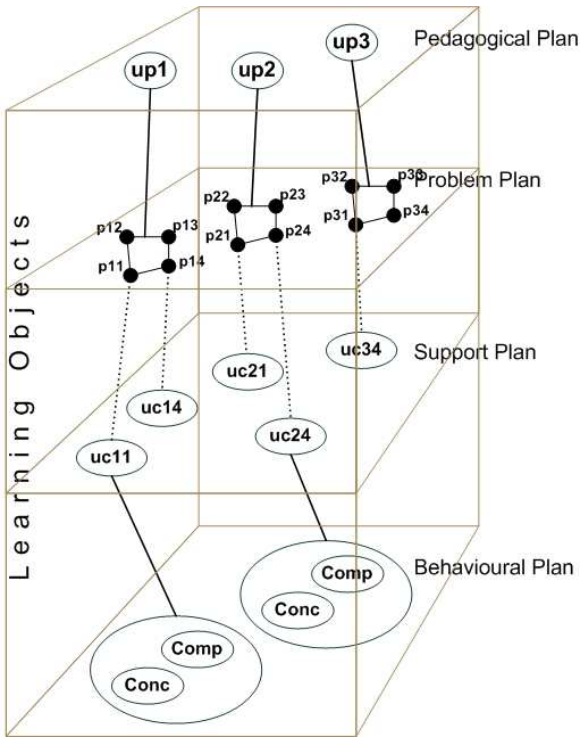


Figure 2. Pedagogical Structure of the Domain Ontology.

3.1.2. Student Ontology

The information necessary to this ontology are i) *Static Information*: the student information that do not change during the student-system interaction like name, telephone, address, so on and ii) *Dynamic Information*: the student information that change during the student-system interaction. Figure 3 approaches interaction features between the student and the system. Another important point is that the ontology keeps interaction information such as evaluation of problems, student activities, student knowledge state, learning goals, and so on.

3.1.3. Pedagogical Ontology

The pedagogical model construction was based on the works [9,14]. The Strategy used was problem-based learning and the tactics are: increase the problem difficulty degree; decrease the problem difficulty degree; same difficulty degree; change the sub-domain; change the issue and change the problem issue to past issue.

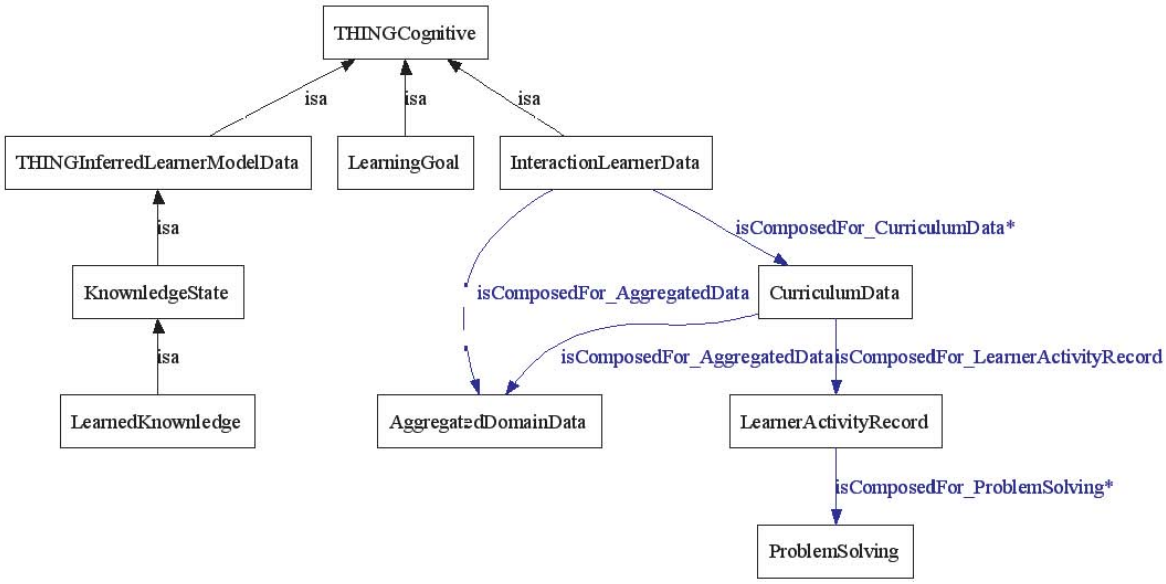


Figure 3. Students Ontology.

3.2. Support Agents

3.2.1. CBR (Case-Based Reasoning) Reasoning

The knowledge was represented by n attributes $A= \{a_1,a_2,...,a_n\}$where each attribute has a weight $W= \{w_1,w_2,...,w_n\}$, for more details on knowledge representation and similarity functions, see [15]. The similarity function between two cases is defined in Equation 2:

$$SIM(C_1,C_2)= \sum_{i=1}^{n} (w_i * sim(a_{c1}, a_{c2})) \quad (2)$$

While the retrieval process was done in a sequential way, the reuse and revision phases were not used, because jurisprudence can not be adapted.

The case attributes used are: Co-authorship (participation of other person at the crime), crime qualification, kind of action, crime modality, attempt (if have or not the attempt), result (if the result was favorable to the lawyer or to the prosecutor) and CBR is used according to the algorithm below.

```
Initialize Evaluate(studentSolution);
Initialize CBRCycle();
casesBase ← select casesSolution from Ontology;
Execute Retrieve from CBRCycle;
Select similarCase;
Select similarityValue;
```

Algorithm 1: The student solution evaluation algorithm.

3.2.2. RBR (Rule-Based Reasoning) Agent

It is responsible for the rules evaluation in the Legal ontology where the rules were modeled considering the Normative Knowledge which enables the whole validation of a penal situation. In addition, were modeled 49 rules to infer about the domain. Follow an example of a rule developed using the Jess [12] environment and integrated within Protege [22]:

```
(bind ?article new Article) (defrule concept ("corporalLesion") ?article getInstance() )
```

The interactions between the Law students and the ITS in the problem solving can happen in two ways: (i) when the student submits a penal situation to tutoring system; (ii) when the tutoring system submits a penal situation to the student. The hybrid reasoning mechanism, CBR and RBR, can work together with the legal ontology to solve problems submitted by the student or by the tutoring system.

When the *student submits a penal situation to the tutoring system* it tries to solve the penal situation using both CBR and RBR, and the interaction algorithm was implemented as follows:

```
Initialize Evaluate(studentProblem);
Initialize RBRInfer();
rbrSolution ← try infer from NormativeKnowledge;
Initialize CBRCycle();
casesBase ← select jurisprudence from Ontology;
Execute Retrieve from CBRCycle;
Select similarCase;
Select similarityValue;
BuildSolution(rbrSolution, similarCase);
```

Algorithm 2: The evaluation student problem algorithm.

On the other hand, when the *tutoring system submits a penal situation to the student*, the student describes the solution according to her/his knowledge and only then, the ITS evaluates the student solution according to the algorithm below.

```
Initialize Evaluate(studentSolution);
Initialize CBRCycle();
casesBase ← select casesSolution from Ontology;
Execute Retrieve from CBRCycle;
Select similarCase;
Select similarityValue;
```

Algorithm 3: The student solution evaluation algorithm.

3.2.3. Q-Learning Agent

Some researches pointed out Reinforcement Learning in pedagogical activities approaching the feasibility of the algorithm [3,23,16]. These researches stated that students have different learning style, and these styles can be acquired through the analysis of the interaction. That‘s why a Reinforcement Learning Algorithm was used in order to improve the teaching ITS skills.

This agent aims to learn an action policy that maximizes the expected long-term sum of values of the

reinforcement signal, from any starting state [4]. In the present work, the problem is defined as a Markov Decision Process (MDP) solution.

The chosen of better strategies has been modeled as a 4-tuple (S, A, T, R), where:

- S: set of strategy and MATHEMA Context pairs.
- A: finite set of strategies.
- T: S × A→ Π (s): state transition function represented for the probability value, signalizing the betters strategy to be chosen.
- R: S × A: it is described as a utility value, defined for the similarity of the attributes, mapped as a reward function.

It was used in the e-learning environment a proposal approached in [4] that implements an algorithm which is used in the action choice rule which defines what action must be performed when the agent is in state s_t. The heuristic function (Equation 3) included was:

$$\pi(s_t) = \begin{cases} argmax_{at} \left[\hat{Q}(s_t, a_t) + \xi H_t(s_t, a_t)\right] & if q \leq p, \\ a_{random} & otherwise. \end{cases} \quad (3)$$

.

- H: S× A→ Ris the heuristic function.
- ε : it is a real variable used to weight the influence of the heuristic function.
- q: it is a random uniform probability density mapped in [0, 1] and p(0 ≤ p≤ 1) is the parameter which defines the exploration divided for exploitation balance.

• a_{randon} is a random action selected among the possible actions in state s_t. Then, the heuristic value $H_t(s_t,a_t)$can be defined as shown in Equation 4:

$$H(s_t, a_t) = \begin{cases} max_a \, \hat{Q}(s_t, a) - \hat{Q}(s_t, a_t) + \eta \, if a_t = \pi^H(s_t), \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise. \end{cases} \quad (4)$$

Initialize Q(s, a)
Repeat:
Visit the s state
Select a strategy using the choice rule
Receive the reinforcement r(s, a) and observe next state s ‘.
Update the values of H_t (s, a).
Update the values of Q_t (s, a) according to:

$$\hat{Q}(s,a) = \hat{Q}(s,a) + \alpha \left[r + \gamma max \hat{Q}(s', a') - \hat{Q}(s,a)\right]$$

Update the s ← s’ state
Until some stop criteria is reached,
where $s = s_t$, $s' = s_{t+1}$, $a = a_t$ e $a' = a_{t+1}$

Algorithm 4: The Heuristics Algorithm.

3.3. Graphical Interface

The Interface is responsible for showing information about the system. Moreover, the Interface Agent works as an assistant, in other words, it is made a step-by-step to the student describes the problem.

- problem specification: the steps for the problem specification are:
  1. personages specification (name, age, deficiencies, ...).
  2. relation among the personages (father, mother, son, brother, brother in law, friend, ...).
  3. fact specification:
     (i) Did the murder happen ?
     (ii) What are the personages positions (victim, killer, accomplice, witness, ...)?
     (iii) Which was the gun (slashing/piercing object, poison, revolver, ...) used ?
     (iv) What was the crime reason (revenge, ordered)?
     (v) What are the personages conditions (drunken, strong emotion, sleeping, ...)?
- problem solution.

4. An Illustrative Example

This section presents a student-ITS interaction in order to illustrate the functionalities of the system.

Suppose that the student is working for the first time with the ITS, so the student answers a set of question about Legal issues and then, the knowledge level of the student is defined. Below, it is exploited an example where the student submits a problem to the system.

4.1. Case

Problem: John arrives in his home and see Maria and Joseph (John's brother), sleeping in the bed, naked. Then John overdrew his gun and shot against Maria, which dies.

When the student specifies a problem, the system considers the rules and the cases, evaluating the attributes[2]:

- Personages = John, Maria, and Joseph.
- Relationship among the personages = Brother(John, Joseph), Married(John, Maria).
- Personages positions: Victim(Maria), Accused(John), and Witness(Joseph).
- Personage's deficiency: it specifies if the patient has some physical deficiency that can be considered, for example, a case in which the victim can not protect itself = Maria sleeping in the bed.
- Fact (attempt well successful or not): if the crime was materialized = yes.
- Gun used: the gun is very important, because it can characterize how serious was the crime = gun.
- Reason of the crime: it specifies if the crime was perpetrated for revenge, ordered, among others = adultery.

Solution: The solution is divided into two views: The Prosecutor view who tries to increase the punishment and the Lawyer view that tries to decrease the punishment.

Prosecutor View:

- Normative Knowledge -Qualified Homicide: Art. 121, ï£¡2ï£¡, IV; Doctrine -Qualified Homicide can be used when happens a crime through research that makes difficult or impossible the defense or the offended person, by the fact the victim was sleeping.
- Jurisprudence -Summary: JURI. Qualified Homicide. Research that turn defense of the offended person impossible. Victim Sleeping. [...] Below follows the rule used to prosecutor view solution.

1. Rule
   If victim = 'impossible defense' or Fact = 'concretized', then Article = 121 and Paragraph = 4 and item = IV

[2] The context of the attributes is considered relevant in Brazilian Code

When the student describes a rule, the system attempts to infer about the characteristics and mapping them in the doctrinaire concepts.

Lawyer View 1:

- Normative Knowledge -Self-Defense: Art. 23. Doctrine -Self-Defense can be used when the author has his honor stained for the victim.
- Jurisprudence -Summary: Homicide -Self-Defense of the honor -Accused that, [...].

- Site: `http://jus2.uol.com.br/doutrina/texto.asp?id=980`;

1. Rule

If AccusedCondition = 'self-defense', then Article = 23

Lawyer View 2:

- Normative Knowledge -Privileged Homicide: Art. 121, ï£¡1ï£¡; Doctrine -Privileged Homicide can be used when the author acts through strong emotion.
- Jurisprudence -Summary: JURI. Qualified Homicide. Cohabitation. Condemnation for Privileged Homicide.

1. Rule *If CrimeReason = 'adultery' then AccusedCondition = 'strong emotion' or AccusedCondition = 'depression'*
2. Rule *If AccusedCondition = 'strong emotion' and Fact = 'concretized', then Article = 121 and Paragraph = 1*

In the case, three solutions were returned to the ITS. The ATA Agent 121_2, ATA Agent $121_1$and ATA Agent 23 were used to solve the case, where each solution represents one agent. In addition, Both RBR and CBR agents were used.

5. Final Remarks and Future Work

To sum up, this paper proposed the so-called Themis, a hybrid ITS which provides students with problems and appropriate tutorial feedbacks. The prototype has been used with three types of knowledge domain (Jurisprudence, Normative Knowledge, and doctrines). At the moment, the Case-Based Reasoning model and Rule-Based System that integrate Jurisprudence, Normative Knowledge, doctrines, and the application of the corresponding Legal concepts in the problem solving process were developed. Technologies such as JADE [24], JESS [12], Protégé [22] were used on the development of the prototype.

It is planned a new version of the Themis that includes: (i) to create the strategy structure to the pedagogical model in others parts of the tutor; (ii) to create the student modeling structure to the student model, enabling the holistic view of each individual student to be stored, allowing the tutor to be highly personalized [13]. Finally, it is planned to evaluate the current system with undergraduate students to improve the system's robustness and learning evaluations.

References

[1] V Aleven, Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment, Artif. Intell., Vol. 150, 2003, pp. 183–237.

[2] V Aleven and K D Ashley, What law students need to know to win, in Proceedings of the 4th international conference on Artificial intelligence and law in (ICAIL '93), New York, NY, USA, 1993, ACM Press, pp. 152–161.

[3] A L Baylor and Y Kim, Simulating Instructional Roles through Pedagogical Agents, International Journal of Artificial Intelligence in Education, Vol. 15, 2005, pp. 95–115.

[4] R A C Bianchi, Uso de heurísticas para a aceleração do aprendizado por reforço, PhD thesis, Escola Politécnica, Universidade de São Paulo, 2004.

[5] L K Branting, Reasoning with portions of precedents, in Proceedings of the 3rd international conference on Artificial intelligence and law in (ICAIL '91), New York, NY, USA, 1991, ACM Press, pp. 145–154.

[6] W Chen and R Mizoguchi, Leaner model ontology and leaner model agent, Cognitive Support for Learning -Imagining the Unknown, , 2004, pp. 189–200.

[7] E Costa, A Perkusich, and E Ferneda, From a tridimensional view of domain knowledge to multi-agents tutoring systems, 1998, pp. 61–72.

[8] E B Costa, Um Modelo de Ambiente Interativo de Aprendizagem Baseado numa Arquitetura Multi-Agentes, PhD thesis, Universidade Federal da Paraíba, Campina Grande, 1997.

[9] B d Boulay and R Luckin, Modelling human teaching tactics and strategies for tutoring systems, International Journal of Artificial Intelligence in Education, Vol. 12, 2001, pp. 235–256.

[10] FIPA, The Foundation for Intelligent Physical Agents. http://www.fipa.org, 2005.

[11] M V C Guelpeli, C H C Ribeiro, and N Omar, Utilização de Aprendizagem por Reforço para Modelagem Autônoma do Aprendiz em um Tutor Inteligente, Simpósio Brasileiro de Informática na Educação in (SBIE '03), , 2003, pp. 493–502.

[12] JESS the Rule Engine for the Java Platform. http://herzberg.ca.sandia.gov/jess/, 2003.

[13] V Kumar and D Brokenshire, An ontological framework to collect and disseminate user model data,

in I2LOR: ELearning for the Future: from Content to Services, 2nd Annual Scientific Conference of LORNET research network, Vancouver, Canada, accepted for publication, 2005.

[14] V Kumar, J Shakya, C Groeneboer, and S Chu, Toward an ontology of teaching strategies, in Proceedings of the ITS'04 Workshop on Modelling Human Teaching Tactics and Strategies, Maceió, 2004, 2004, pp. 71–80.

[15] R Lee, Pesquisa Jurisprudencial Inteligente, PhD thesis, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, 1998.

[16] K N Martim and I Arroyo, Agentx: Using reinforcement to improving the effectiveness of intelligent tutoring system, in Intelligent Tutoring System in (ITS '04), Maceió, Brazil, 2004, pp. 564–572.

[17] A Muntjewerff and J A Breuker, Evaluating PROSA, a system to train legal cases, in Artificial Intelligence in Education, J Moore, C Redfield, and L Johnson, eds., FAIA-Series, Amsterdam, 2001, IOS-Press, pp. 278–290. ISSN:0922-6389, ISBN:1 58603 173 2.

[18] T A O'Callaghan, J Popple, and E McCreath, Shyster-mycin: a hybrid legal expert system, in Proceedings of the 9th international conference on Artificial intelligence and law in (ICAIL '03), New York, NY, USA, 2003, ACM Press, pp. 103–104.

[19] A Oskamp, R F Walker, J A Schrickx, and P H v. d Berg, Prolexs divide and rule: a legal application, in Proceedings of the 2nd international conference on Artificial intelligence and law in (ICAIL '89), New York, NY, USA, 1989, ACM Press, pp. 54–62.

[20] J S P. Dillenbourg, A framework for learner modelling, Interactive Learning Environments, 2, 1992, pp. 111–137.

[21] G Span, Lites, an intelligent tutoring system for legal problem solving in the domain of dutch civil law, in Proceedings of the 4th international conference on Artificial intelligence and law in (ICAIL '93), New York, NY, USA, 1993, ACM Press, pp. 76–81.

[22] Stanford, Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu, 2000.

[23] J Tetreault and D Litman, Using reinforcement learning to build a better model of dialogue state, in 11th Conference of the European Chapter of the Association for Computational Linguistics in (EACL '06), Trento, Italy, 2006.

[24] Tilab, Java Agent Development Framework. http://jade.tilab.com/, 2005.

[25] G Vossos, J Zeleznikow, T Dillon, and V Vossos, An example of Integrating Legal Case Based Reasoning with Object-Oriented Rule-Based Systems: IKBALS II, in Third International Conference on Artificial Intelligence and Law in (ICAIL '91), New York, NY, USA, 1991, ACM Press, pp. 31–41.

A Framework for Building Intelligent Learning Environments in Ill-defined Domains

Vu Minh Chieu [a], Vanda Luengo [b], Lucile Vadcard [b], and Dima Mufti-Alchawafa [b]
vmchieu@umich.edu, {Vanda.Luengo, Lucile.Vadcard, Dima.Mufti}@imag.fr
[a] *School of Education – University of Michigan, United States*
[b] *Laboratoire CLIPS – UMR CNRS/UJF/INPG 5524, France*

Abstract. A critical problem of instruction in ill-structured and complex domains has been how to help students attain a deep understanding of a complex concept. Solutions for this problem are usually very costly. For example, the practical course in medical education often requires one-to-one assistance of the expert-teacher for the student to be able to account for a great diversity of complex and real clinical situations. Intelligent learning environments could provide significant help for instruction in ill-defined domains. In this paper, we show how to exploit advanced technologies such as temporal Bayesian networks and computer-based simulations to help the student in advanced learning of complex concepts such as the sacro-iliac screw fixation in orthopedic surgery.

Keywords. Intelligent tutoring systems, student modeling, didactics modeling, medical education, 3D computer-based simulations.

Introduction

An ill-structured and complex domain is a domain in which cases or examples are diverse, irregular, and complex [18]. Advanced learning in ill-structured and complex domains such as medicine and literature gives rise to a difficult problem: What one has to do to attain a *deep understanding* of a complex concept [18]. Deep understanding means that students are prepared to be ready to apply conceptual knowledge in a domain where the phenomena occur in irregular patterns, and to use knowledge in a great variety of ways that may be required in a rich domain. In France, for example, the training of the sacro-iliac screw fixation in orthopedic surgery (Figure 1) is usually organized into two separate courses: (1) a theoretical course in which students are engaged in the acquisition of declarative knowledge (e.g., definitions and examples of key concepts), (2) an important practical course in which students are engaged in the acquisition of procedural and *pragmatic* knowledge (obtained by experience) with costly one-to-one assistance of an experienced surgeon-teacher (e.g., the application of concepts in a diversity of real clinical cases). We consider the sacro-iliac screw fixation as a concept in an ill-defined and complex domain because there are many different solutions and different ways to arrive at the same solution to a given problem, some or which might not be in the "mind" of the expert surgeon-teacher, that is, some solutions and/or ways to arrive at solutions are not predictable.

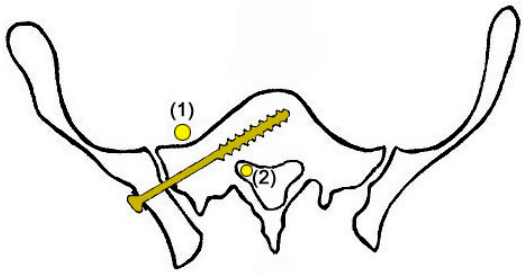


Figure 1. Sacro-iliac screw fixation allows posterior lesions of the pelvic ring of the hip bone to be fixed to the body of S1. This may be performed percutaneously. The danger is a screw course outside of the bone with risk of injury to the lumbo-sacral trunk (1) and the roots of the cauda equina (2).

In earlier work [19], we have shown the importance of a bridge between the previous two courses: an intelligent learning environment as an intermediate phase of learning, which provides an operative dimension of knowledge before the real situation. Several authors have also claimed that the introduction of computers could provide significant help in medical education [6], but on the condition that real underlying educational principles are integrated [8], and particularly that individual feedback is stressed [14]. The critical question in this paper is how the intelligent tutoring system can provide learners with appropriate feedback on their solutions, especially on the ones that are not predictable.

In this paper, we present *a technological framework that could be used to build intelligent learning environments in ill-defined domains*. It is mainly based on an appropriate use of computer-based simulations, temporal Bayesian networks, Web semantic, and fine-grained analysis of didactics[1]. To show the usefulness of the framework, we illustrate the development of TELEOS (Technology Enhanced Learning Environment for Orthopedic Surgery).

In the following sections, we first introduce theoretical framework for the design of our learning environment. Secondly, we present the main results of our didactical analysis. Thirdly, we show the development of a multi-agent platform including a simulation agent, a diagnosis agent, a didactical decision agent, a Web course agent, and a clinical cases agent, which are the core of the learning environment. Finally, we discuss our framework regarding related work and we show promising directions for future research.

1. Theoretical Framework

We take as a fundamental hypothesis for our research that "Errors are not only the effect of ignorance, of uncertainty, of chance […], but the effect of a previous piece of knowledge which was interesting and successful, but which now is revealed as false or simply not adapted" (Brousseau's theory of didactical situations, [3][2], p. 82). In other words, a misconception has a domain of validity, otherwise it would not exist as such. Therefore, the key difference between a misconception and a knowing is that for the former there exists a refutation that is known at least to an observer.

Brousseau's theory goes even beyond the fact of recognizing mental constructs source of errors as knowings. It states that some of these knowings likely to be falsified are necessary to learning: the student's trajectory may have to pass by the (provisional) construction of erroneous knowings because the awareness of the reasons why a knowing is erroneous is necessary to the construction and understanding of a new knowing.

For us, according to the previous theory, the "milieu" for the apprenticeship must be organized to foster learning by producing relevant feedback to the learner's actions. We assume that the system can produce relevant feedback for the apprenticeship if it reacts regarding an internal validation of the learner's solution process. In other words, the system feedback is based on local consistency checks of the learner's actions rather than on an expert's *a priori* solution [11].

For didactical analysis and knowledge representation, we base our work on the cK¢ (conception, knowledge, and concept) model that provides a computational framework for didactical research [2]. We choose this model because of two main reasons: (1) it is adapted to our working hypothesis concerning the essential role of the controls in the action; and (2) it facilitates the analysis of the knowledge to be formalized and implemented in the system.

The aspect of this model that concerns our work is the *conception* conceptualization (conception is the instantiation of the knowing ascribed to a subject by a situation). It conceptualizes a conception as: a set of problems (P); a set of operators (R) involved in solutions of problems from P; a representation system (L) allowing the representation of P and R; a control structure (Σ). The first three components are the key features identified by Vergnaud ([21], p. 145) in order to characterize a concept. The fourth one is introduced for the following reason: Validation is a key aspect of problem solving, so the presence of the control structure (Σ) in the previous conceptualization aims at making a meta-level explicit, with respect to action. The crucial role of control in problem solving has already been pointed out (e.g., [16]): the control elements allow the subject to decide whether an action is relevant or not, or to decide that a problem or sub-problem is solved. Thus, a problem-solving process can be described as a succession of solving steps: $\sigma(r(p(l)))$ with $\sigma \in \Sigma$, $r \in R$, $p \in P$, and $l \in L$. We illustrate more about our formalization in the following section about didactical analysis.

2. Didactical Analysis

The aim of the didactical analysis is to identify the quadruple (P, R, L, Σ) and relationships among its components. In the teaching of complex and ill-defined domains such as medical education, didactical analysis must be hard because of the complexity of knowledge in those subject domains [19]. For instance, it is not easy for the expert surgeon to describe the correct process of sacro-iliac screw fixation *completely*. To validate a solution in a particular case (e.g., the patient with a very hard bone), the expert surgeon sometimes

[1] Didactic mentioned in this paper is an originally francophone term, which designates the study of teaching and knowledge acquisition in different subjects. Didactic is thus different from pedagogy by the central role of the subject contents and by its epistemological dimension (i.e., the nature of knowledge to be taught).

[2] This reference is the most important one in English, which was translated from a French version. The French versions were published earlier (1978, 1982, 1988, etc.).

uses *pragmatic* knowledge, which has not been described in any theoretical courses. For example, here is the expert's explication to the learner about one of the X-rays, taken at the mid-course of the problem-solving process:

> So, here you've got the two dense lines, you see, there, but on the other hand it [the pin] is a bit too much behind... you see, it is a bit too much behind, it should have been much more by here, but the entry point is ok, **we won't modify it, but do not pass over the midline, furthermore he [the patient] has got a very hard bone, so you don't need to have a very well anchored threading**.

The bold-font text in the previous extract shows a part of the expert's professional experience (pragmatic knowledge). *It should be noted that the previous solution has not been in the "mind" of the expert surgeon-teacher until he or she encounters this situation in practice*. This kind of decision is more or less common in complex and ill-defined domains [19], meaning that pragmatic or tacit knowledge plays a key role in those domains [24]. Thus, on the one hand it is important to help the student mastering pragmatic knowledge. On the other hand, because the student may use this kind of knowledge (correctly or incorrectly) during his or her solution validation, it is also critical to take it into account in student diagnosis in order to give relevant feedback to the student.

The approach we used for didactical analysis is based on the framework proposed by Pastré [12] in which the author has advocated that, to improve the professional ability of the learners, it is necessary to be able to analyze how their action is organized, which knowledge and which strategies they apply, which obstacles they encounter. In other words, it is necessary to be able to make a cognitive analysis of applied competences and their development. According to the approach, the analysis process is composed of two main consecutive phases: the preparation phase and the observation and interview phase. The aim of the preparation phase is to master background knowledge of the subject domain in order to best prepare for the observation and interview phase, which in turn helps to collect subject domain knowledge as maximally as possible, including declarative, procedural, and especially pragmatic knowledge. At the moment we were writing this text, 6 sacro-iliac screw fixation interventions had been videotaped and analyzed. Because procedural knowledge in sacro-iliac screw fixation is quite simple, we take into account only declarative and pragmatic knowledge in this work. Table 1 shows the main results of the analysis. The controls have been systematized by our didactical researcher, and almost all of them, especially pragmatic ones, have been validated with an expert surgeon-teacher.

Table 1. Problems, operators, controls, and representation system in sacro-iliac screw fixation

Problems	In sacro-iliac screw fixation, the validation of a solution depends on the characteristics of the problem, for instance, the type of the pelvis fracture, the bone quality of the patient. We name those characteristics "didactical variables" [3]. In our point of view, taking into account the didactical variables is useful for creating the set of problems, which can be used to devise learning situations for the student. Examples of problems include "**validate a predefined trajectory for a sacrum fracture with normal-density bone**", "**define a trajectory for a pure disjunction with normal-density bone**".
Operators	The didactical analysis allows the detailed description of the process of sacro-iliac screw fixation, which can be summarized as follows. The surgeon first inserts a guide pin in the bone through the skin (percutaneously, i.e., without incision). He makes the pin progressing in the bone, taking several X-rays to validate the pin course at different steps of the progression. During this phase (pin insertion), several attempts can be made by the surgeon. Once the pin's trajectory gives satisfaction, the screw fixation phase will be performed: a screw is inserted along the pin, which will make the right bones' compression for the treatment of the fracture. Last, the pin is retired and one suture point is made to close the pin's entry point. Concerning the teaching and learning of this kind of intervention, we determine that the most crucial phase is the pin insertion. So, the operators we identified are: **introduce an entry point**, **orientate the pin**, **advance the pin**, **take an inlet view**, **take an outlet view**, **take a lateral view**, **take a face view**, **restore** (i.e., put the pin back to the previous chosen entry point and direction), and **validate the pin course**. Note that the combination of the four views helps the surgeon examine the surgical situation as completely as possible.
Controls	Presently, we have identified about 100 controls organized into two groups, according to their epistemological dimension, for example, "**if the pin is well positioned, then it is up the anterior cortical bone of the iliac wing on the inlet view**" (Σ8, declarative control), "**if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient**" (Σ14, pragmatic control).
Representations	These are also important, but we have not studied them yet for the time being.

Note that the main difference between the surgeon-teacher (an expert) and the surgeon-learner (a novice) is that the teacher has almost all of the controls in her "mind", and for each control she knows how and when to use it correctly for her decision, whereas the learner may not have some controls in his "mind", and sometimes he uses a control incorrectly for his decision: out of its validity domain. For instance, after perceiving that the pin comes too near from one precise anatomic part of the pelvis bone (e.g., its anterior cortex) on the inlet view, meaning that the pin is "too low" on the inlet view, the expert with Σ14 (if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient) in the mind decides to restart and corrects the entry point downwards (which is valid in the context of sacro-iliac screw fixation), whereas the novice may not have that control in the mind (and thus not correct the pin course) or may have correct understanding of that control (and act as does the expert) or may have incorrect understanding of that control in the mentioned context (and therefore correct the entry point up-

wards). Also note that to decide which control(s) to be used in a given situation, the surgeon (learner or teacher) needs to examine the situation (principally the X-rays) to determine its characteristics regarding the domain constraints. We use the term "situations variables" (SV) to describe the characteristics of a given problem-solving situation. The value of a SV can only be observed by the surgeon when he or she does a relevant action, for instance, "take an inlet view" to know about the value of SV1 “the pin touches the anterior cortex of the pars lateralis sacri on the inlet view”.

3. TELEOS as an Intelligent Learning Environment

Figure 2 shows a simulation-based, multi-agent architecture of our learning environment. A new characteristic of this architecture in comparison with traditional ones has been that we separate the diagnosis from the didactical decision to be able to study and validate them separately. The condition for them to work together is in the core of the model: the diagnosis must be able to identify the controls that intervened during the problem-solving activity; the didactical decision has to be made according to the diagnosed controls. In other words, besides the “standard” feedback provided by the simulation component (e.g., the pin’s trajectory: intra-osseous or extra-osseous) to the learner, the didactical decision component gives him or her pedagogical feedback: another problem to solve, a redirection to a precise part of the online associated course, or a clinical case to consult. Those kinds of feedback must be produced in such a way that fosters learning, that is, helping the learner understand when and why a control is used correctly.

Thus, the main feature of this environment is that we do not evaluate the learner’s behavior by comparing his or her actions with an *a priori* expert solution. We interact with the student according to the relevance of his or her actions in relation with the problem-solving situation. This model allows us to proceed to an internal validation of the learner’s activity, taking into account his or her problem-solving process. To illustrate how TELEOS might help the student in the problem-solving process, in the following sub-sections we go into details the functionality of the main agents of TELEOS by considering a scenario presented in Table 2 as an example.

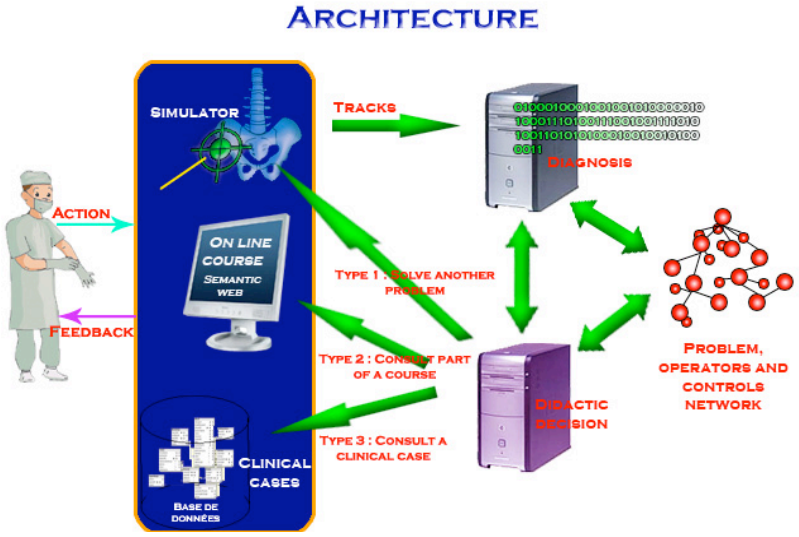


Figure 2. Global architecture of TELEOS

Table 2. A scenario in sacro-iliac screw fixation

Action ID	Description
1	Introduce an entry point for the pin course
2	Orientate the pin
3	Advance the pin
4	Take an inlet view: the pin comes too near from the anterior cortex of the pars lateralis sacri on the inlet view, meaning that the pin is too low on the inlet view
5	Take an outlet view: the position of the pin is correct on the outlet view
6	Restore the pin
7	Introduce another entry point for the pin course
8	Orientate the pin
9	Advance the pin
10	Take an inlet view: although the pin comes a little far from the anterior cortex of the pars lateralis sacri on the inlet view, the position of the pin is still incorrect
11	Take an outlet view: the position of the pin is correct on the outlet view
12	Validate the pin course

3.1. Simulation Agent

On the basis of the results of our didactical analysis presented in Table 1, especially the set of problems and the set of operators, we have been able to build a 3D simulation (Figure 3). The student is shown a 3D pelvis representation, with skin and landmarks (see the object at the top of Figure 3), he or she can turn this 3D object to see the bone structure (see the object at the bottom of Figure 3). The student has to position a pin (see the stick at the top of Figure 3) and to advance it in the simulated body. His or her actions are free, that is, the allowed movements are continuous and he or she can restart the activity at any time (remove the pin, change its entry point, etc.). The user, as the expert in real situation, can ask for X-ray controls during the activity. Note that the 3D pelvis representation and X-rays have been constructed from bones of real patients.

Once the validation is done, by clicking a “Confirm” button (Figure 3), the environment provides the student with various feedback: a "transparency" view that makes the skin disappear, and thus allows the visualization of the validated pin course; the number of attempts, the number of extra-osseous trajectories validated, the number of X-rays taken; and pedagogical feedback.

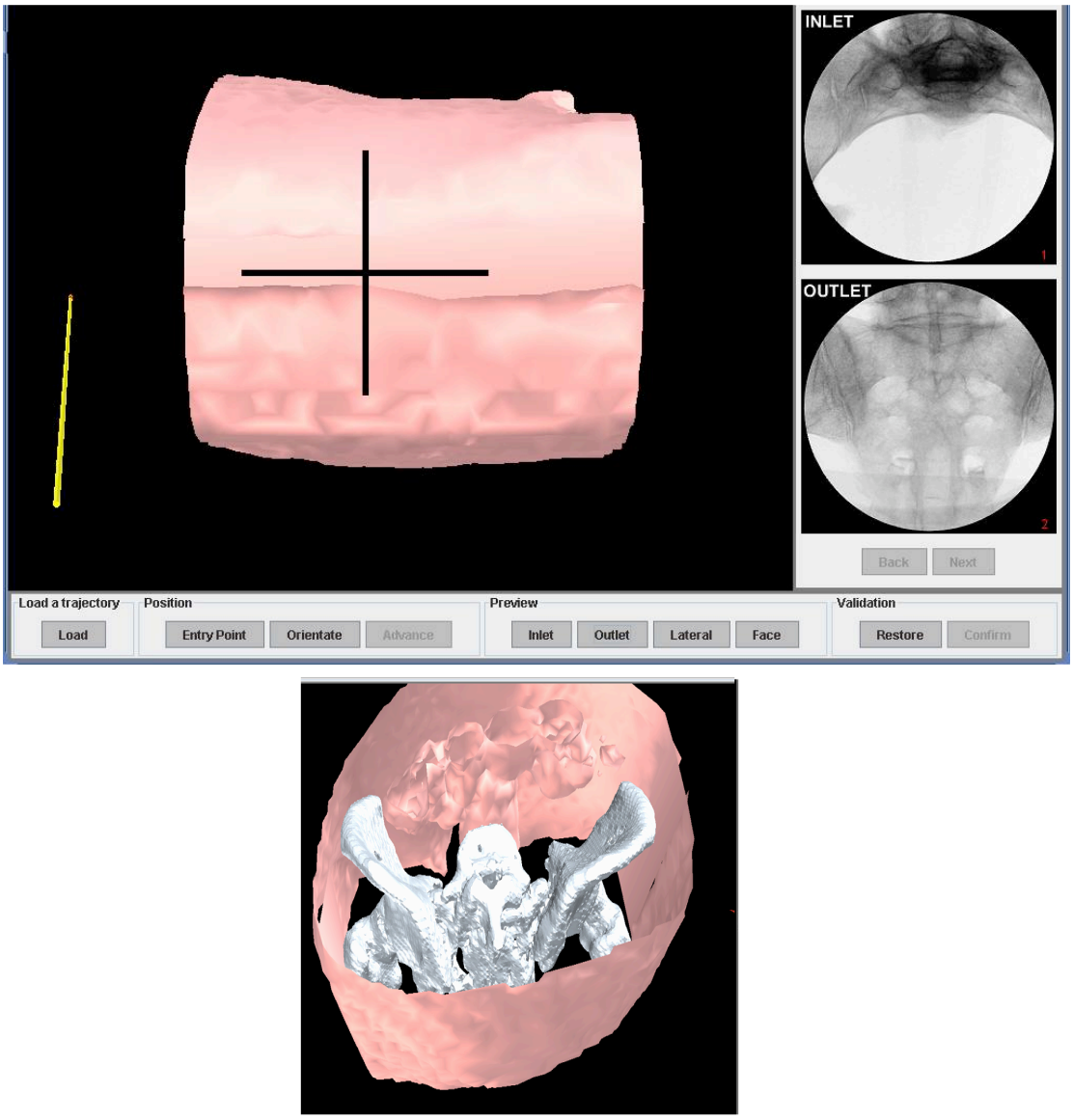


Figure 3. Java-3D simulation interface for sacro-iliac screw fixation

Regarding pedagogical feedback, the aim of the simulation agent is to provide the diagnosis agent (see Figure 2) with the learning traces produced by the student, that is, the course of actions (e.g., take an inlet view) and the evolution of the pin course (e.g., regarding Σ14: if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient, the simulator needs to identify the distance between the pin and the anterior cortex of the pars lateralis sacri on the inlet view). On the basis of those learning traces, the diagnosis agent will be able to diagnose the cognitive state (i.e., the use of controls) of the learner. The next sub-section explains more about this diagnosis.

3.2. Student Diagnosis Agent

Diagnosing the learner's understanding about a certain control exactly could be hard. For example, in the scenario shown in Table 2, after the learner does Action 4 (take an inlet view: the pin comes too near from the anterior cortex of the pars lateralis sacri on the inlet view, meaning that the pin is too low on the inlet view), it could be difficult to diagnose his or her understanding about the related controls (e.g., Σ14: if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient): the student may not have these controls in the mind and make such a pin course randomly, or the student may have correct understanding of these controls but make an incorrect pin course because of the lack of experience (even an expert sometimes must make several tries to arrive at a correct pin course), or the student may have incorrect understanding of these controls and therefore make such an incorrect pin course. That is why researchers in the field often use an intuitive approach (e.g., Bayesian networks) for student diagnosis (i.e., cognitive diagnosis) in this case [7, 9].

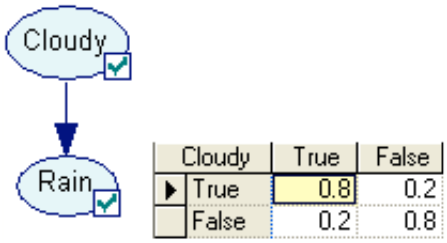
In automated cognitive diagnosis, the temporal dimension has been taken into account [9] to make diagnosis result more accurately. For instance, in the previous scenario, after the student makes an incorrect solution in the first course of actions (Action1 - Action 5) and then he or she makes a good correction in the second course of actions (Action 6 – Action 11), we may confirm that the student has a correct understanding about Σ14, with high probability. Temporal Bayesian networks [15] have been exploited to model such temporal dimension in student diagnosis. In the next sub-sections, we show how to exploit temporal Bayesian networks to implement the diagnosis agent.

3.2.1. Temporal Bayesian Networks

A Bayesian network is a directed, acyclic graph with the following properties:

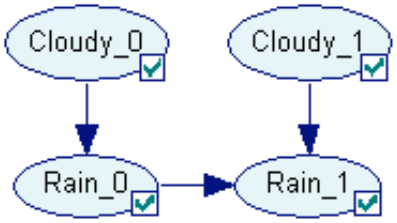
- Each vertex in the graph represents a random variable.
- There is an edge from **X** to **Y** ≠ **X**, whenever **Y** is dependent on **X**.
- Each vertex is labeled with a conditional probability table (CPT) that quantifies the effect of its parents. The out-neighbors of a vertex are called children, the in-neighbors are called parents. A vertex without out-neighbors is called root.

Figure 4 shows an example of Bayesian networks, considering the dependence of “Rain” on “Cloudy”. Figure 5 shows an instance of temporal Bayesian networks (in which stochastic processes are modeled), taking into account the fact that if it rains today it will probably rain tomorrow.



Cloudy	True	False
True	0.8	0.2
False	0.2	0.8

Figure 4. A simple Bayesian network



Cloudy_1	True		False	
Rain_0	True	False	True	False
True	0.9	0.8	0.3	0.2
False	0.1	0.2	0.7	0.8

Figure 5. A simple temporal Bayesian network

3.2.2. Student Modeling

The diagnosis component aims at detecting the student's use of controls during his or her problem-solving process. For each control we consider the following three states:

- **BPV**: This state stands for "brought into play in a valid manner". It means that the student has the control in his or her "mind" and his or her understanding about the control is correct, so he or she may know when and how to use it correctly.

- **BPI**: This state stands for "brought into play in an invalid manner". It means that the student has a misunderstanding about the control in a specific context (he or she may have correct understanding about the control in another context).
- **NBP**: This state stands for "not brought into play". It means that the student does not have the control in his or her "mind".

Because there is no evidence about student's knowledge at the beginning of the learning session, the probability is equally initialized for the three states BPV, BPI, and NBP (see also Table 3 in Section 3.2.3). We consider three cognitive states because we believe that they are helpful enough for our didactical decision component. One can surely add more states to obtain more fine-grained diagnosis results.

3.2.3. Student Diagnosis

The main task for the development of the diagnosis agent is to build a temporal Bayesian network. Figure 6 shows a model consistent with the cK¢ framework described in Section 1. The sub-model "operators" contains nodes representing operators, "evolution_variables" contains nodes representing the evolution of the student's pin course, "correction_variables" is used to model the student's correction of the pin course, and "controls" is used to model control nodes. Figure 7 illustrates the modeling of Σ14 (if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient), as an example. The approach for modeling every control is the same, and can be summarized, as follows:

- *Identify the situation variable(s) related to the control*. For instance for Σ14, the situation variable is "the distance between the student's pin and the anterior cortex of the pars lateralis sacri on the inlet view". The value of this variable is calculated by the simulation agent (see Section 3.1).
- *Create the intermediate variables that model the temporal dimension*. The scenario shown in Table 2 indicates that it is useful to consider the values of the same situation variables at different time in order to model the temporal dimension. Taking into account this, Figure 7 shows three intermediate variables for Σ14. The two "evolution" variables (deterministic nodes) are used to partly describe the learner's solution at present and that at the most previous point in time (each point in time corresponds to an action performed by the learner). The "correction" variable (also a deterministic node) is used to describe the student's correction behavior. It has three parents: two "evolution" variables and one operator variable related to the control being modeled (e.g., "take_inlet" in Figure 7 for Σ14). We consider that the learner can only be aware of a situation (e.g., an error) when he or she does an appropriate operation (e.g., taking an inlet view to know the distance between the pin and the anterior cortex of the pars lateralis sacri). The main point here is the "correction" node, which can take the following values: (1) `correct` (e.g., the distance is correct, according to domain constrains identified in didactical analysis); (2) `no_correction` (e.g., the distance is the same at present and at the previous time); (3) `good_way` (e.g., although the distance is incorrect, the correction is in a good direction, see Actions 4 and 10 in Table 2); (4) `bad_way` (similar to `good_way` but the correction is in a bad direction); and (5) `no_information` (a value by default, which is useful at the beginning of the problem-solving process). The calculation of the "correction" variable value is mainly based on the values of the "evolution" variables, which in turn are computed from a set of IF-THEN rules identified in didactical analysis. The didactical variables (e.g., the bone quality, see Table 1), that is, the contexts of problems, are also modeled in those IF-THEN rules. Because of limited space, we could not go into details those rules.
- *Create the control nodes*. Figure 7 shows two chance nodes (`sigma14_0`, `sigma14_1`) in the case of Σ14. `sigma14_1` represents the diagnosis result at the present and has three parents: `sigma14_0`, `validate_pin_course`, and the "correction" node. `sigma14_0` is used to model the learner's cognitive behavior before and at the most previous point in time. `validate_pin_course` is considered because in our point of view validating an incorrect solution is different from making an incorrect pin course and restarting and correcting the error(s). We have subjectively filled the CPT for `sigma14_1`; the CPT is the same for every control in the same group (e.g., declarative or pragmatic, see Table 1). In the future we shall apply machine-learning techniques [17] to fill those CPTs.

Table 3 shows a part of the diagnosis result for the scenario presented in Table 2. It can be interpreted, as follows: After Action 4, because the learner makes an incorrect distance between the pin and the anterior cortex, the outcome NBP (not brought into play) of Σ14 (if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient) is increased. After Action 10, because it seems that the learner corrects the pin course in a good way, the outcome BPV (brought into play in a valid manner) of Σ14 is increased. After Action 12, however, because the learner validates an incorrect solution, the outcome BPI (brought into play in an invalid manner) of Σ14 is increased. A preliminary subjective evaluation of our researchers toward the diagnosis results of a number of scenarios is positive.

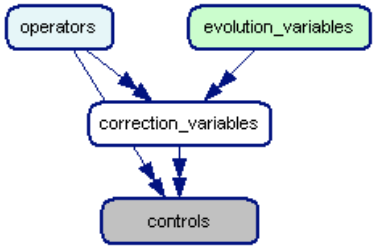


Figure 6. A model of temporal Bayesian network for student diagnosis

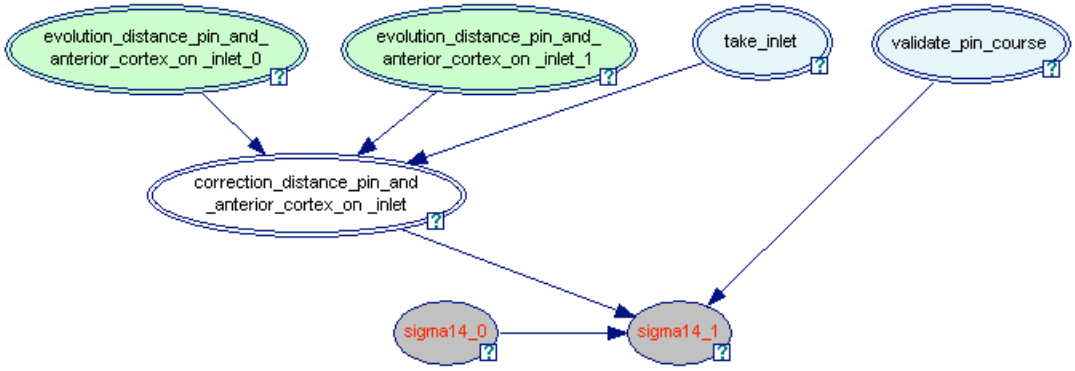


Figure 7. A part of the temporal Bayesian network for modeling the diagnosis of Σ14 (sigma14)

Table 3. Diagnosis result for Σ14

(BPV = brought into play in a valid manner, BPI = brought into play in an invalid manner, NBP = not brought into play)

Action ID	Action	Learning Traces	Diagnosis Result
1	entry point	none	BPV=0.33, BPI=0.33, NBP= 0.34
2	orientate	none	BPV=0.33, BPI=0.33, NBP= 0.34
3	advance	none	BPV=0.33, BPI=0.33, NBP= 0.34
4	take inlet	distance_pin_and_anterior_cortex_on_inlet=1	BPV=0.20, BPI=0.20, NBP= 0.60
5	take outlet	distance_pin_and_sacral_foramen_on_outlet=6	BPV=0.20, BPI=0.20, NBP= 0.60
6	restore	none	BPV=0.20, BPI=0.20, NBP= 0.60
7	entry point	none	BPV=0.20, BPI=0.20, NBP= 0.60
8	orientate	none	BPV=0.20, BPI=0.20, NBP= 0.60
9	advance	none	BPV=0.20, BPI=0.20, NBP= 0.60
10	take inlet	distance_pin_and_anterior_cortex_on_inlet=3	BPV=0.44, BPI=0.24, NBP= 0.32
11	take outlet	distance_pin_and_sacral_foramen_on_outlet=5	BPV=0.44, BPI=0.24, NBP= 0.32
12	validate	distance_pin_and_anterior_cortex_on_inlet=3 distance_pin_and_sacral_foramen_on_outlet=5	BPV=0.23, BPI=0.57, NBP= 0.20

3.3. Didactical Decision Agent

On the basis of the result calculated by the diagnosis agent, the decision agent will be able to provide relevant feedback to the student, as follows: Firstly, it determines the control(s) as the target of feedback (e.g., Σ14: if the pin touches the anterior cortex of the pars lateralis sacri on the inlet view then it is too ventral on the body of the patient). Secondly, it identifies the apprenticeship objective of feedback for the chosen target (e.g., help the student explore a concept further or help the student understand a misconception). Thirdly, according to the target and the objective, the agent chooses the most relevant form of feedback from a number of predefined forms in the learning environment (another problem to solve, a Web content to read, or a clinical case to examine). Finally, according to the form, the agent formulates the content of the form.

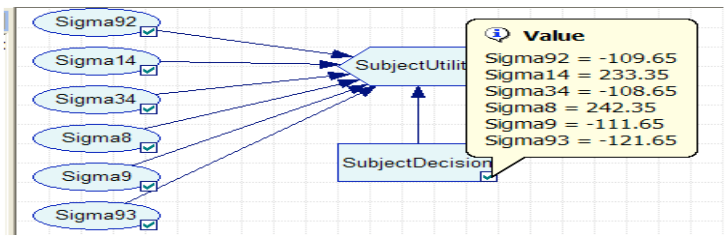


Figure 8. An influence diagram for the target decision of feedback

We use influence diagrams [15] to determine the target control(s) for the feedback. In Figure 8, there are control nodes from the Bayesian network described previously, an apprenticeship utility node (the hexagonal one), and a target decision node (the rectangular one). In order to apply the inference in the diagram, we defined the apprenticeship utility function (see more details in [10]).

3.4. Web Course Agent and Clinical Cases Agent

Presently, the didactical decision agent provides three types of feedback (see also Figure 2): (1) asking the learner to solve another problem with the simulation agent in order to refine the diagnostic result and to help him or her examine a diversity of real situations, (2) insisting the learner to revise an appropriate part of the theoretical course provided by the Web course agent (by appropriate we mean, e.g., an explanation of a concept related to the target controls identified by the didactical decision agent), and (3) exhorting the learner to examine a clinical case (provided by the clinical cases agent), which illustrates the post-consequences of a given pin course more or less similar to his or hers.

The Web course and clinical cases agents are based on the Virtual Observatory described in the VOEU project [23]. The development of the clinical cases agent is mainly grounded on a patient cases database provided by a supporting clinic. The development of the Web course agent is mainly grounded in a Web-based theoretical course and Web semantic techniques [20]. For example, the Web semantic component receives, from the didactical decision agent, the error(s) that must be considered. The error(s) is (are) analyzed in order to produce a webpage (see the left illustration of Figure 9) containing a set of hyperlinks to particular contents of the online course, which are closely related to the error(s) (see the right illustration of Figure 9).

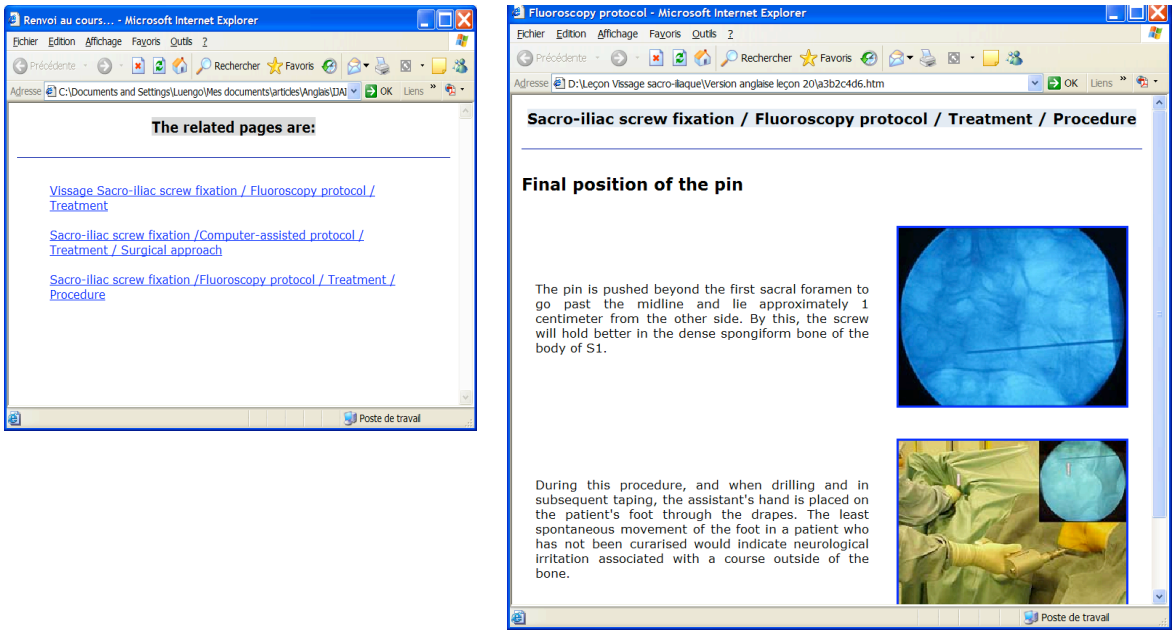


Figure 9. Production of dynamic feedback in relation to the error "pin progression"

4. Discussion and Conclusion

A number of cognitive approaches, for example, overlay [5], buggy [4], model tracing [1], conception [22], have been used to build intelligent learning environments. Very few researches, to the best of our knowledge however, have concentrated on complex and ill-structured domains in which particular kinds of knowledge such as pragmatic one play a key role. In many researches, the authors are versed in the subject knowledge and this knowledge is well documented in many textbooks. In complex and ill-defined subjects such as medical education, however, knowledge used in problem solving is very complex and not completely reported in standard materials such as textbooks. That is why we have argued for a didactical analysis to understand the nature of knowledge being used in problem solving as completely as possible. Indeed, this fine-grained analysis of didactics could provide significant help in implementing a robust domain component for student diagnosis. Of course, this work is domain-dependent and time-consuming.

Regarding the use of diagnosis result for supporting students, existing approaches and ours have offered more or less similar features such as suggesting students to explore appropriate learning contents or case studies in order to "fill the gap" or to solve problems that are adapted to the student's current knowledge state. A key difference of our approach from others has been that because the diagnosis result is "fine-

grained" (i.e., at level of controls), the feedback to the student is "fine-grained". For example, the system can suggest the student to explore a particular content page (units of the learning content) instead of a chapter or a section, or at a given time during the student's problem-solving process the system can select a problem that could be used to refine the diagnosis result as rapidly as possible. Another difference is that we give feedback to the student only at the moment after he or she validates his or her solution to a given problem. We believe that this strategy is helpful (even necessary) in the education of complex domains because even experts in such a domain often need to make several tries before arriving at a correct solution.

Our main affirmation in this paper is that an appropriate use of computer-based simulations, Web semantic, temporal Bayesian networks, fine-grained analysis of didactics based on a robust theoretical framework, that is, the theory of didactical situations [3] and the cK¢ model [2], could be an effective way to build intelligent learning environments, especially in ill-defined domains. Indeed, in this paper we have presented a technological framework (i.e., a multi-agent architecture, an operational approach for student diagnosis and didactical decision, a Web-semantic-based platform for theoretical course) that could be reused for building intelligent learning environments for complex concepts. Constraint-based modeling [11] could also be another technique for developing similar intelligent tutoring systems.

For the time being, we have developed all of the agents separately and we have integrated most of them together, except for the didactical decision one. In the future, after the whole learning environment is available, we shall carry out empirical studies by using both quantitative and qualitative methods [13] in order to know the impact of our intelligent feedback on learning. We shall also look at the usefulness and the effectiveness of the didactical results (the assumption is that the didactical analysis may need to be reinvestigated in order to improve, e.g., the effectiveness of student diagnosis and didactical decision).

References

[1] Anderson, J.R., Farrell, R., & Sauers, R. (1984). Learning to Program in LISP. *Cognitive Science*, 8, 87–129.

[2] Balacheff, N. (2003). Conceptual Framework. In S. Soury-Lavergne (Eds) *Baghera Assessment Project, Designing an Hybrid and Emergent Educational Society* (pp. 3–22). Grenoble : Les cahiers du laboratoire Leibniz (http://www-leibniz.imag.fr/LesCahiers).

[3] Brousseau G. (1997). *Theory of Didactical Situations*. Dordrecht: Kluwer Academic Publishers edition and translation by N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield.

[4] Brown, J.S., & Burton, R. (1978). Diagnostic Models for Procedural Bugs in Basic Mathematical Skill. *Cognitive Science*, 2, 155–192.

[5] Burton, R., & Brown, J.S. (1982). An Investigation of Computer Coaching for Informal Learning Activities. In D. Sleeman, & J. Brown (Eds) *Intelligent Tutoring System*s. Orlando: Academic Press.

[6] Eraut, M., & du Boulay, B. (2000). Developing the Attributes of Medical Professional Judgement and Competence. *Cognitive Sciences Research* (paper 518). Retrieved March 25, 2007 from http://www.cogs.susx.ac.uk/users/bend/doh.

[7] Henze, N., & Nejdl, W. (2001). Adaptation in Open Corpus Hypermedia. *International Journal of Artificial Intelligence in Education*, 12, 325–350.

[8] Lillehaug, S. I., & Lajoie, S. P. (1998). AI in Medical Education: Another Grand Challenge for Medical Informatics. *Journal of Artificial Intelligence in Medicine*, 12(3), 1–29.

[9] Mayor, M., & Matrovic A. (2001). Optimising ITS Behaviour with Bayesian Networks and Decision Theory. *International Journal of Artificial Intelligence in Education,* 12, 124–153.

[10] Mufti-Alchawafa, D. (2006). La Représentation Informatique de la Prise de Décision Didactique. *Premières Rencontres Jeunes-Chercheurs sur les EIAH*. Evry, France.

[11] Ohlsson, S. (1992). Constraint-based Student Modeling. *International Journal of Artificial Intelligence in Education*, 3(4), 429–447.

[12] Pastré, P. (1997). Didactique Professionnelle et Développement. *Psychologie Française*, 42(1), 89–100.

[13] Rieber, L.P. (2005). Multimedia Learning in Games, Simulations, and Microworlds. In R.E. Mayer (Eds.), *The Cambridge Handbook of Multimedia Learning* (pp. 549–567). NY: Cambridge University Press.

[14] Rogers, D.A., Regehr, G., Yeh, K.A., & Howdieshell, T.R. (1998). Computer-assisted Learning versus a Lecture and Feedback Seminar for Teaching a Basic Surgical Technical Skill. *American Journal of Surgery*, 175(6), 508–510.

[15] Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice-Hall.

[16] Schoenfeld, A. (1985). *Mathematical Problem Solving*. New York: Academic Press.

[17] Sison, R., & Shimura, M. (1998). Student Modeling and Machine Learning. *International Journal of Artificial Intelligence in Education*, 9, 128–158.

[18] Spiro, R.J., Feltovich, P.J., Jacobson, M.J., & Coulson, R.L. (1991, May). Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-structured Domains. *Educational Technology*, 31, 24–33.

[19] Vadcard, L., & Luengo, V. (2005). Réduire l'Ecart entre Formations Théorique et Pratique en Chirurgie : Conception d'un EIAH. In P. Tchounikine, M. Joab, & L. Trouche (Eds) *Environnements Informatiques pour l'Apprentissage Humain* (pp. 129–139). Paris : Institut National de Recherches Pédagogiques.

[20] Luengo, V., & Vadcard, L. (2005). Design of Adaptive Feedback in a Web Educational System. In P. Brusilovsky, R. Conejo, E. Millán (Eds) *Adaptive Systems for Web-Based Education: Tools and Reusability* (pp. 9–17), Workshop at Artificial Intelligence in Education.

[21] Vergnaud, G. (1991). La Théorie des Champs Conceptuels. *Recherches en Didactique des Mathématiques*, 10(2/3), 133–170.

[22] Webber, C. (2004). From Errors to Conceptions – An Approach to Student Diagnosis. In J.C. Lester, R.M. Vicari, & F. Paraguacu (Eds) *Intelligent Tutoring Systems* (pp. 710–719). Berlin: Springer (Lecture Notes in Computer Science, Vol. 3220).

[23] Wu, T., Zimolong, A., Müller, G., Vadcard, L., Huberson, C., & Langlotz, F. (2002). *The Virtual Observatory*. Final Deliverable (n°23.03). European Union Project IST-1999-13079 (http://vou-caos.vitamib.com).

[24] Herbig, B., Büssing, A., & Ewert, T. (2001). The Role of Tacit Knowledge in the Work Context of Nursing. *Journal of Advanced Nursing*, 34(5), 687–695.

The logic of Babel: Causal reasoning from conflicting sources

Matthew W. Easterday[1], Vincent Aleven, Richard Scheines
Human-Computer Interaction Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA, 15213

Abstract Ill defined problems lack structure partially because there is no agreed upon way of representing the problem. In this follow-up study, we examine how diagrams help students learn to analyze policy arguments. Our previous work asked students to predict the effect of a policy intervention based on testimonies from conflicting sources, and showed that teaching students a formal, diagrammatic procedure improved students' predictions. In this study we looked at *how* students and experts use diagrams so that we could a) identify errors in student reasoning and b) start to develop a cognitive model of construction and interpretation of causal diagrams. We thus conducted an informal protocol analysis on how 4 novices and 3 experts solved causal reasoning problems using 1) text, 2) text and a correct diagrammatic representation, and 3) text with a diagramming tool. We found that many of the errors in causal reasoning stemmed not from the difficulty of using diagrams per se, but from conflicts of background knowledge with the provided testimonies. Some participants demonstrated a diagram confirmation bias, i.e. they reinterpreted the diagram syntax to reach a conclusion more consistent with their beliefs. Other participants made arguably normative "errors," i.e. they correctly interpreted the sources' claims in the testimony, but judged their own knowledge to be more credible. Allowing students to apply arbitrary background knowledge poses a problem for intelligent tutors that require a fully specified problem space. We conclude that tutors may be able to distinguish between confirmation bias and normative uses of background knowledge by asking students to explicitly add their background knowledge to the diagram.

Introduction

A rational, participatory democracy depends on an informed citizenry, one that can reason about the conflicting policy claims of multiple sources (Gore 2007). For example, if the U.S. Secretary of Defense states that "preemption will decrease a weapons of mass destruction (WMD) threat," while a policy analyst questions the claim, asserting instead that "international sanctions and foreign aid are a better way to reduce anti-American sentiment," and weapons experts argue that "rogue regimes with nuclear material are likely to increase proliferation," citizens must be able to weigh the various claims and judge the likelihood that a suggested policy (preemption) will lead to the desired outcome (a decreased WMD threat)–they must make logical judgments from "babel."

Voss (2005) describes seven features of policy problems that make them ill structured including their: lack of a clear goal state, having no objectively correct answer, etc. We are especially interested in Voss's third feature: the lack of an agreed upon strategies for representing policy problems. Unlike algebraic word problems that have

1 The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

well-defined procedures for converting words into equations (the formal representation), and for solving the equations, there is no agreed upon representation of a policy problem. In this paper we consider the question: how *should* solvers represent ill defined policy problems in order to make inferences about the probable effect of a policy intervention. For example, if a citizen decides that the policy analyst and the weapons experts are credible, how should he then judge the likelihood that preemption will lead to a decreased WMD threat? If, as Simon (1981) conjectured, "…solving a problem simply means representing it so as to make the solution transparent," then the representation issue is not only central for policy, but for ill defined problems as well.

Work on external representations suggests that diagrams might improve reasoning (Larkin, & Simon 1987; Ainsworth, 2006; Harrell, 2004; Scaife & Rogers, 1996, Cox, 1999, Novick & Hurley, 2001, Mayer & Moreno, 2002, Bauer & Johnson-Laird 1993, Pinkwart, Aleven, Ashley & Lynch, 2006), but not which specific diagrams should be used for policy problems, or how to design them. Since arguments about policy often hinge on causal assumptions or inferences, we are focused on the benefits to learners of causal diagrams, a type of diagram that maps out causal relations claimed to exist within a topic area (e.g., the effects of preemption or international sanctions). Such diagrams and the associated causal theory (Spirtes, Glymour, Scheines 2000) lend a significant amount of structure to the domain, but do not render it well-defined. Even after a solver commits to a causal representation, there is no single correct way to represent the causal factors or to choose the grain size at which to represent those factors.

In prior work, we showed that causal diagrams can be helpful to students as they learn to interpret brief (experimenter-written) policy texts (Easterday, Aleven & Scheines, 2007). We found that providing students with a causal diagram that summarizes a particular policy text helps them do better in interpreting that text: students take advantage of the diagram to make better predictions about the effects of policy interventions described in text. We also found that having students practice constructing diagrams for policy texts supports learning how to interpret new texts, even when (as would be the case for a newspaper article) the new texts are not accompanied by a causal diagram. Thus, the previous study implies that an intelligent tutoring system that helps students in constructing causal diagrams will contribute to their skill in reasoning about policy. In order to develop intelligent tutors it is important to understand students' strategies for analyzing a policy texts and constructing diagrammatic representations of the texts as well as the difficulties that students experience in that process. Addressing these questions will yield a greater understanding of argumentation in ill-defined domains and the use of diagrammatic representations in that context.

Previous research has demonstrated a variety of errors (Kuhn 1991) and expert-novice differences (Voss 1983) in policy reasoning. These studies however were not focused on the use of diagrams. Following in their footsteps, the goals of the follow-up study were to gain insight into how students and experts used diagrams to make inferences about the effects of a policy in order to a) identify errors in reasoning, and b) to inform the design of a cognitive tutor.

1. Method

Task and Intervention

As in the previous study, we gave participants short, fictional, policy texts (Figure 1).

Childhood obesity is now a major national health epidemic. A number of facts are widely agreed upon by the public and scientific community: exercise decreases obesity, and eating junk food increases obesity. It's also clear that people who watch more TV are exposed to more junk food commercials.

Parents for Healthy Schools (PHS), an advocacy group which fought successfully to remove vending machines from Northern Californian schools, claims that junk-food commercials on children's television programming have a definite effect on the amount of junk food children eat. In a recent press conference, Susan Watters, the president of PHS stated that "...if the food companies aren't willing to act responsibly, then the parents need to fight to get junk food advertising off the air."

A prominent Washington lobbyist Samuel Berman, who runs the Center for Consumer Choice (CCC), a nonprofit advocacy group financed by the food and restaurant industries, argues that junk food commercials only "influence the brand of food consumers choose and do not not affect the amount of food consumed." While Mr. Berman acknowledges that watching more TV may cause people to see more junk food commercials, he remains strongly opposed to any governmental regulation of food product advertising.

Recent studies by scientists at the National Health Institute have shown that watching more TV does cause people to exercise less.

Figure 1. Policy text on obesity.

We then asked participants to answer questions like: "According to the combination of claims made by the CCC and NHI, will making kids watch less *TV*, decrease childhood *obesity*?" According to the procedures taught in the experiment, students should notice that the CCC denies the effect of TV on obesity *through commercials*, and the NHI claims an effect of TV on obesity *through lack of exercise*, so the correct answer is *yes*, TV will affect obesity (according to the claims of the CCC and NHI). These questions simulate the difficult task of assembling the claims of multiple sources to predict the likely affect of a policy intervention. Note that real policy problems like terrorism or the environment are far more complex, however students find even these simple texts to be quite challenging.

Participants received policy information in one of the following forms:

1. **Text (only)** in which the case studies was presented in text only (Figure 1).
2. **Diagram (+ text)** in which the case study was presented as text accompanied by a correct, diagrammatic representation of the case study (Figure 2).
3. **Tool (+ text)** in which the case study was accompanied by a computer tool with which participants could construct their own diagrams.

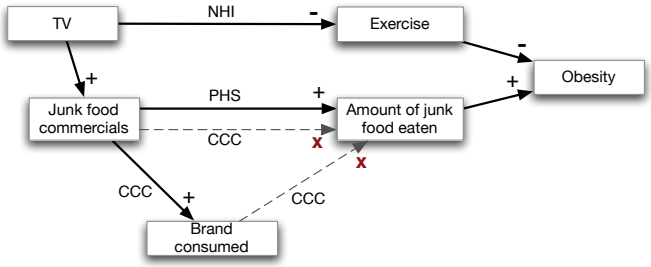


Figure 2. A causal diagram representing the case study on obesity. Boxes represent causal variables, and arrows represent either positive (+), negative (-), or no (x) influence of one variable on another. An annotation on the arrow (e.g. PHS) identifies the source making the causal claim.

Note that to solve the question about TV increasing obesity (according to the CCC and NHI) using the diagram in Figure 2, participants should notice the arrow from TV to exercise labeled "NHI", and the unlabeled arrow from exercise to obesity (representing common knowledge), as they are taught during the experiment.

Participants and setting

In this protocol analysis study, we examined 3 "experts" from the Carnegie Mellon University (CMU) Philosophy Department, all of whom have a PhD in philosophy and who have conducted original research on causal reasoning, and 4 CMU student "novices." All participants were offered $20, however all experts declined payment.

Research design
In this protocol analysis, participants were asked to "talk aloud" as they completed an on-line lesson on causal reasoning. We assigned one novice and one expert to each of the text, diagram and tool conditions, and a second novice to the diagram condition to collect more data. After a pretest involving a case study on the environment similar to Figure 1, each participant received a 4 page, interactive, on-line tutorial on causal reasoning that also included diagrams like Figure 2 for the diagram and tool groups. To make the training for the text as close to identical as possible, every diagrammatic explanation in the diagram/tool training was matched by an equivalent prose explanation in the text training. Following training, we tested all participants on the case study in Figure 1, presented as text only to the text participants, as text with a correct diagram to the diagram students, and as text with a tool to the tool participants.

Data collection and analysis
To measure performance, participants were tested on 10 multiple choice, causal questions (e.g. "According to the PHS, will making kids exercise more reduce the number of junk food commercials they watch?"). Participants could answer either: a) *yes* there would be a causal effect (e.g. making kids exercise more would reduce the number of junk food commercials they watch), b) *no* there would be no causal effect, or c) *inconclusive* the sources explicitly disagree about the causal effect.

To capture process information, participants' speech and on-screen behavior were recorded with a screen capture program. Because we intend to build tutors for automated knowledge tracing and instruction, we did not create a coding manual or use multiple coders to analyze the video; we instead used the screen recordings to identify processes and errors to inform the design of a cognitive model.

2. Results

Text (Novice 1, Expert 1)
Both the novice and expert in the text condition performed quite poorly. Novice 1 scored 20% while Expert 1 scored 0% on the first half of the questions, after which Expert 1 ended the experiment stating: "My brain is fried." While Expert 1's performance seems abysmal, recall that in this condition, Expert 1 did use his standard tool (a causal diagram) and, unlike Novice 1, Expert 1 realized the difficulty of completing the task without a diagram. These scores compare with a chance score of 33% (guessing randomly between the 3 options of yes/no/inconclusive), and the average score of 41% on the same condition in the previous study. This performance underscores the difficulty of reasoning about even simple causal systems using text alone. We take the ubiquity of causal diagrams in causal reasoning research, the poor results of the text group in the previous study, and the difficulty of using text by both the novice and expert in this study, as an indication to focus our future efforts on diagram use.

Tool + Text (Novice 2, Expert 2)
In the previous study, students who were given case studies as text accompanied by a diagramming tool scored an average of 40%, performing no better than the text group. Given the poor performance of this diagram construction group in the previous study, we expected Novice 2 to have difficulty with diagram construction. In fact, both Novice 2 and Expert 2 made better diagrams than most observed in the previous study.

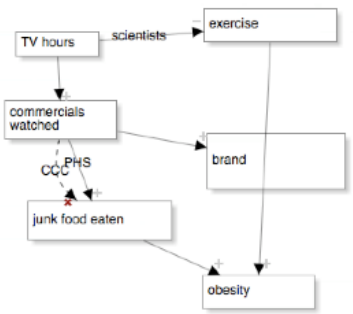


Figure 3. Expert 2's diagram. The experts diagram is virtually correct.

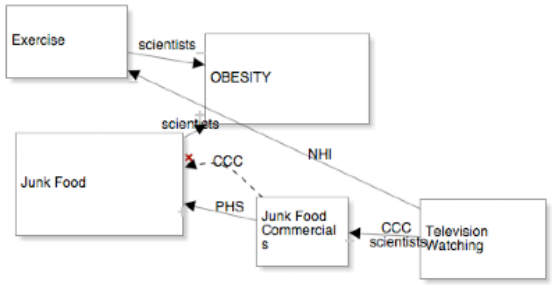


Figure 4. The Novice 2's diagram. Note omitted "brand" variable and the mislabeled links from junk food to exercise.

Despite making relatively good diagrams, small errors in diagram construction sometimes lead to relatively large errors in interpretation. For example, by mislabeling the two arrows pointing to obesity, Novice 2 might answer every question about obesity, (the focus of the case study,) incorrectly. While both their diagrams contained errors, Expert 2's diagram (assuming it was used correctly to answer the test questions) would have lead to the correct answer on 100% of the questions, whereas Novice 2's diagram would have lead to the correct answer on 20% of the questions.

However, both participants often used their diagrams incorrectly. If we grade them on their ability to use the diagram they constructed, (in an algebra word problem where students must translate the problem into an equation and solve the equation, this would be like grading students based only on whether or not they solved the equation correctly, even if the equation they started with was incorrect), Expert 2 used the diagram correctly on only 60% of questions, Novice 2 on 20%. If we grade them based on whether they got the right answer, we find that (coincidentally,) Expert 2 answered only 60% of the questions correctly, Novice 2, 20%.

Diagram + Text (Novice 3 & 4, Expert 3)

During training, participants in the diagram and tool conditions were taught a procedure for interpreting diagrams. We thought that giving students a correct diagram in addition to the text would improve their reasoning. Although the previous study showed that providing a correct diagram did indeed improve performance, this study shows that participants' background knowledge sometimes overrules the conclusions implied by the diagram, and sometimes leads to a reinterpretation of the diagram syntax to reach a conclusion more consistent with background beliefs.

Participants overrule the diagram with background knowledge when they make:

- *Override errors*, where the reasoner correctly reads the graph, but decides that their background knowledge is more credible. This can be normative if the reasoner: a) makes separate and correct predictions about the both the evidence provided and their beliefs, and b) explicitly claim their beliefs to be more credible than the evidence provided.
- *Speculation errors*, such as adding information to the diagram about what a source *would say*, given what that source has already said.

Participants selectively reinterpret the diagram (confirmation bias) when they make:

- *Diagram interpretation errors*, such as confusing observation and intervention, i.e. believing that an arrow showing that A causes B, can also mean that B causes A.
- *False uncertainty errors*, such as interpreting a lack of an arrow by a source as indicating that "we don't know what the source thinks" instead of that "the source makes no claim."

And sometimes participants simply interpret the diagram incorrectly when they make:

- *Combination errors,* where the relevant paths are noticed, but not combined correctly to make the proper inference.
- *Impasse errors,* such as giving up on the diagram (and text) altogether.

Question	Type of error		
chain (correct answer: "yes")	Novice 3	Novice 4	Expert 3
a According to the NHI, will making kids exercise more reduce childhood obesity?	+	+	+
b According to the NHI & CCC, will making kids watch less TV decrease childhood obesity?	combination?	+	combination?
none (correct answer: "no")			
e According to the PHS will watching TV cause children to exercise less?	+	+	uncertainty & speculation
f According to common knowledge, will making children watch less TV decrease childhood obesity?	+	uncertainty	+
common cause (correct answer: "no")			
g According to the NHI, will making kids exercise more reduce the number of junk food commercials they watch?	interpretation	interpretation	+
h According to the NHI, will reducing the number of junk food commercials children watch reduce childhood obesity?	+	uncertainty	uncertainty
common effect (correct answer: "no")			
i According to common knowledge, will making kids exercise more reduce the amount of junk food they eat?	+	impasse	+
j According to the PHS, will making kids exercise more reduce the number of junk food commercials they watch?	+	impasse	override

Table 1. Errors made by diagram participants. The first column shows each of questions asked on the test, grouped by the underlying causal structure of the answer.. Cells indicate that the question was answered correctly ("+"), or the type of error made. No errors were made on questions *c* & *d* (not shown).

Diagram + Text: Novice 3

Novice 3 had the best performance, answering 80% of questions correctly, (as compared with an average score of 49% for diagram students in the previous experiment, putting Novice 3 in the top 15 percentile of that group). For the most part, Novice 3 did not seem to reference background knowledge at all, but seemed to consistently apply the diagram interpretation procedure to each question. To the extent that Novice 3 ignored his background knowledge, he fit the pattern of a ***diligent novice,*** executing the procedure as instructed.

Diagram + Text: Novice 4

Novice 4 answered 50% of the questions correctly, much closer to the average score of 49% observed in the previous study. We could characterize Novice 4's behavior as including far greater interference from background knowledge:

1. On questions *f* and *h*, Novice 4 made *uncertainty errors*, concluding that if there are no arrows (path) between the relevant variables, then it is inconclusive whether the source would say one variable would affect the other. Note that in the diagram interpretation procedure, participants were instructed that if there are no arrows, the source would *not* claim that one variable would affect the other, i.e. the answer is *no*. On question *f*, Novice 4 explains:

   "it doesn't say anything on here... I can't tell from there, so from looking at that, that would be inconclusive..."

   ...and on question *h* Novice 4 explains:

   "it doesn't say anything about junk food commercials, so that would be inconclusive."

Novice 4 did not consistently apply this "no path means inconclusive" reasoning however. In fact, Novice 4 inferred that the answer was *inconclusive* when her background knowledge (that TV does affect obesity) contradicted the correct answer of *no*:

"I would assume that if you're watching TV you're not playing…that would lead to less children being obese."

The quotes suggest that Novice 4 wanted to answer *yes* according to her background knowledge, and selectively reinterpreted the meaning of an absence of an arrow when the correct interpretation contradicted her belief. After answering question *f,* we asked why she chose *inconclusive* rather than *no*, to which she responded:

"…my feeling is to go for *yes*, so I kind of compromised and went for inconclusive."

2. Novice 4 (and Novice 3) also confused observation with intervention on question *g*, incorrectly interpreting an arrow from TV to exercise as also meaning that exercise decreases TV, a *diagram interpretation error*:

   "Well without looking at that I would say 'yes', but looking at this…so kids are exercising more, then they watch less TV, which means they have, watch less junk food commercials. But the question is…'will making children exercise more, reduce the number of commercials they watch'. I don't know about reading the graph backwards, its confusing. Well I'm going to say 'yes'."

   Note that the novice did not regularly infer that an arrow denoting that A causes B also means that B causes A. Novice 4 only made that error when it was consistent with her background knowledge, as can be seen in the above quote.

Unlike Novice 3, Novice 4 used a far greater amount of background knowledge, which unfortunately seemed to hurt performance. Although Novice 4 performed better than participants in the text condition, Novice 4 was not able to reconcile the diagram with her background knowledge, thus undermining the usefulness of the diagram. The reinterpretation of the diagram syntax to reach conclusions consistent with one's belief might be thought of as a kind of *diagrammatic confirmation bias.* Novice 4's behavior seems closer to that of a **confused novice** which is more representative of students in the previous study.

Diagram + Text: Expert 3

Expert 3 made more errors than expected, with an overall performance of 60%. While Expert 3, like Novice 4 made heavy use of background knowledge, Expert 3 seemed better able to separate his background knowledge from the predictions of the diagram, treating the two as separate entities not necessarily requiring reconciliation.

1. Expert 3's first error, on question *j*, was an *override error.* Expert 3 explicitly recognized a difference between his background knowledge and the claims of the diagram, and then explicitly chose to go with his background knowledge:

   "Naturally I would assume that the PHS people would say "yeah it will reduce the number of junk food commercials they watch" because in fact, this guy up here, I think most people would think is actually a, uh, uh, goes both ways.... However, I'm supposed to answer the question based on what's been given to me so far... So I'm going to say the answer I'm supposed to give is 'no', but quite frankly, well you know what, I'm going to give the answer I think is right given the sorts of things I've got here, which is that its actually inconclusive. "

2. Expert 3 also made a *speculation error* on question *e,* where, given the fact that the PHS wants to decrease junk food advertising, Expert 3 inferred that the PHS would also agree with the NHI that TV would decrease exercise (note this would

be a reason to make kids watch less TV and therefore less junk food commercials). This speculation error was combined with an uncertainty error:

> "Well I'm willing to bet the PHS would absorb… well it's inconclusive, we don't know what the PHS thinks, we aren't given any context. ...So I'm going to say inconclusive, because I was not given that piece of information. Moreover, I think the PHS would presumably accept those kind of studies."

Like Novice 4, Expert 3 often applied his background knowledge to the problem. Unlike Novice 4, Expert 3 did not seem to be as confused by the diagram, as shown by his lack of diagram interpretation and impasse errors. When the diagram did not match his beliefs, he explicitly stated that the diagram is wrong and that he chose to rely on his background knowledge. For this reason, we characterize Expert 3 as a **truculent expert**. It may be that Expert 3 has more robust diagram reading skills than Novice 4, which would prevent him from questioning the implications of the diagram, and move on to the question of whether or not the diagram is correct.

Diagram + Text: summary

To clarify the differences between the diligent novice, truculent expert, and confused novice, we characterize the three patterns in Figure 6.

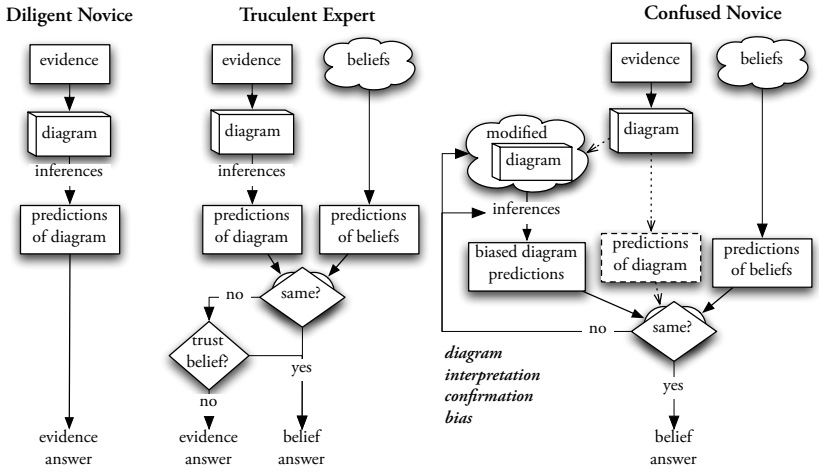


Figure 6. Three different patterns of reasoning.

3. Implications: Tracing background knowledge with tutors

The protocol data makes it clear that the task and measures developed in the previous study cannot detect whether a participant is making an incorrect statement as a truculent expert behaving normatively, or as a confused novice exhibiting confirmation bias. Redesigning the task and measures so that students must make their background knowledge explicit will allow us to monitor how they are using background knowledge. Once we can detect how they are using background knowledge, we can both evaluate their performance, and provide better tutoring. To detect whether students are applying their background knowledge normatively, we could do the following:

1. Ask students to make three inferences, one based on the diagram, one based on their background knowledge, and one based on both.

2. Use an editable (rather than fixed) diagram that allows students to add arrows (but not variables) representing their background knowledge to the diagram.
3. If measures 1 and 2 show that students have not made the correct inference based on the diagram or their background knowledge, we can ask them to highlight the arrows on which they based their decision. Then, for each highlighted, (or relevant, unhighlighted) arrow, we can ask whether the student: a) wants to add/reject a causal relation based on their own knowledge, b) speculate that a source would add/reject a causal arrow, or c) did not notice the arrow.

This procedure would allow us to retrospectively detect each of the errors observed in the protocol analysis. Figure 7 shows where the various interpretation errors arise during diagram interpretation, and how the proposed measures should be able to distinguish between normative uses of background knowledge, and confirmation bias.

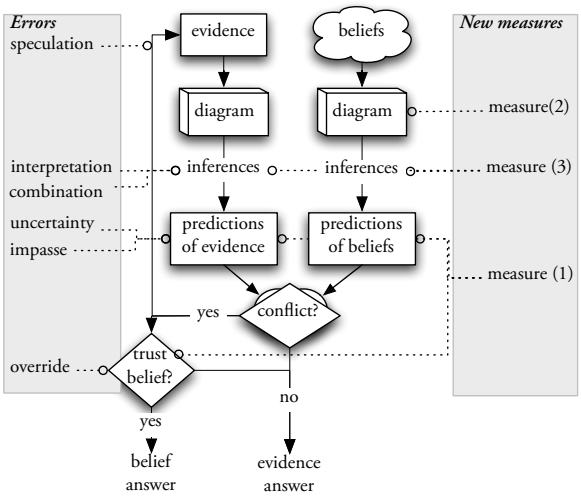


Figure 7. Errors of diagram interpretation and measures to detect them.

4. Discussion

This follow up study to Easterday, Aleven and Scheines (2007) showed that students' errors with diagrammatic representations stem not so much from the difficulty of the diagram construction or interpretation procedures per se, but rather the way in which the procedures conflict with students' background knowledge and informal reasoning.

Educational research tells us that in all domains, teachers (and tutors) "...must draw out and work with the preexisting understandings that their students bring with them." (NAS, 2000, p. 19). While background knowledge affects how students solve problems in general, it plays a more complicated role in ill defined problems. Simon (1973) describes a system for solving ill structured problems as:

> ...a combination of a GPS, which at any given moment finds itself working on some well structured subproblem, with a retrieval system, which continually modifies the problem space by evoking from long-term memory new constraints, new subgoals, and new generators for design alternatives. (p. 192).

...as opposed to "...bringing all of the potentially relevant information in long-term memory together once and for all at the outset, to provide a well structured problem space that does not change..." (p. 192). The fact that people use background knowledge

in ill defined problems to continuously modify the problem state undermines cognitive tutors that rely on a fully specified problem state. How can a tutor trace knowledge and provide feedback in problems where the student is allowed, even encouraged, to apply background information not known to the tutor?

The study showed that background knowledge not only plays a role, but often manifests itself as a kind of *diagrammatic confirmation bias*. Kuhn (2005) describes instances in which students fail to *coordinate theory and evidence*, i.e. misinterpret a given set of facts (evidence) in order to support preconceived beliefs (theory). Kuhn rightly suggests not that students should ignore background knowledge, but rather that they should make judgments about evidence and background knowledge separately, so that they may compare the two (p. 72). Kuhn's admonition suggests that tutors for ill defined problems should not ask students to "[leave] their common sense at the door," but rather trace how students correctly (and incorrectly) apply their background knowledge. Tutors for ill defined problems do not have the luxury of assuming that, as in an algebra problem, the initial given facts determine a unique solution because in ill defined problems, students' background knowledge can sometimes trump the given facts.

Looking toward the future challenges in this line of research, this study suggests that tutors for policy argument must be able to monitor when and how students apply their background knowledge.

References

Ainsworth, S. E. (2006). DeFT: A conceptual framework for learning with multiple representations. *Learning and Instruction*, 16(3), 183-198.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, 4(6), 372-378.

Cox, R. (1999) Representation construction, externalised cognition and individual differences. *Learning and Instruction*, 9(4), 343-363.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). 'Tis better to construct or to receive? The effects of diagram tools on causal reasoning. *Proceedings of the 13th International Conference on Artificial Intelligence in Education.*

Gore, A. (2007). *The assault on reason*. New York: Penguin Press.

Harrell, M. (2004). *The improvement of critical thinking skills in What Philosophy Is* (Tech. Rep. No. CMU-PHIL-158). Pittsburgh, PA: Carnegie Mellon University, Department of Philosophy.

Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge University Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.

Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, 12(1), 107-119.

National Research Council. (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.

Novick, L. R., & Hurley, S. M. (2001). To matrix, network, or hierarchy: That is the question. *Cognitive Psychology*, 42(2), 158-216.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006). Toward legal argument instruction with graph grammars and collaborative filtering techniques. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 227-236.

Scaife, M. & Rogers, Y. (1996) External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, 45(2), 185-213.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, 4(3), 181-201.

Simon, H. A. (1981). The sciences of the artificial (2nd ed.). Cambridge, MA: MIT Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Voss, J. F. (2005). Toulmin's Model and the solving of ill-structured problems. *Argumentation*, 19(3), 321-9.

Voss, J. F., Tyler, S. W., & Yengo, L. A. (1983). Individual differences in the solving of social science problems. In R. F. Dillion, & R. R. Schmeck (Eds.), Individual differences in cognition (pp. 205-32). New York: Academic Press.

Mapping and Validating Case Specific Cognitive Models

Geneviève Gauthier, Susanne P. Lajoie and Solange Richard
McGill University, Canada

Abstract: This study proposes a visual methodology to validate complex solution processes in the context of ill-structured problem solving. This experiment is anchored in the context of medical case-based teaching. The case validation activity proposed is modeled on the authentic case presentation practice performed by physicians. We are using a computer-based learning environment (BioWorld) to present a standardized set of cases to expert teachers who are asked to solve the case and do a think-aloud protocol while solving the cases. We are developing a methodology that addresses both knowledge elicitation as well as knowledge validation for solving and reflecting on ill-structured medical problems. More precisely, this study examines the effectiveness of visual support tools to help physicians verify their diagnostic thinking. In so doing our goal is to build and validate case specific cognitive models.

Introduction

Problem based learning (PBL) is not a new approach in medical education, it is used to teach clinical reasoning and problem solving skills in a number of medical schools (Barrow, 1994; Barrows & Tamblyn, 1980; Koschman, Kelson, Feltovich, & Barrow, 1996). The use of cases for teaching is as old as storytelling and Cox (2001) argues that this way of transmitting knowledge provides a meaningful framework to embed all the objectives and sub-objectives related to a complex patient case. A case presentation in medicine generally consists of a detailed analysis of a patient case but depending on the instructor's prior experience and the facilities in which the patient is seen the solution to these cases varies substantially. Case development work for BioWorld (a computer-based learning environment (Lajoie, Lavigne, Guerrera, & Munsie, 2001)) led us to note significant differences in the thinking and decision making processes involved in complex case solution. Data on the case creation phase demonstrated both validity and reliability issues when working with medical staff and students. This lack of consistency forced us to address the issue of validity and reliability of ill-structured solutions in a more systematic manner.

Case presentation activities are used to teach diagnostic reasoning. Diagnostic reasoning about patient cases share the same components of ill-structured problem solving as defined by Jonassen (1997) in that solving patient cases involve a) plenty of unknown elements, b) there is not one correct unambiguous solution, c) there is more than one way to reach a diagnosis and there are usually multiple ways to reach an acceptable answer (often referred as differential diagnosis) c) there is no absolute criteria or way to validate the answer, and d) case resolution often involves ethical and personal judgments.

This research aims at modeling and building on the case presentation activity that occurs in medical education. We do not intend to replace or compete with the face-to-face case presentations but it aims at documenting and building on key elements related to this practice. In this paper we first describe the instructional context and the computer-based learning environments we use to support and study diagnostic reasoning. We then explain how the validation activity became a key element of the case creation process. We also explain why and how the sampling of detailed solution processes and explanations of expert teachers can lead to the construction of cognitive models.

Cognitive tools to support and study diagnostic reasoning

BioWorld, is a computer-based learning environment that was first designed to promote scientific reasoning in high school students. It provides a realistic environment for students to learn about diseases through solving specific patient cases (Lajoie et al., 2001). Solving a patient case in BioWorld not only consists of submitting a good diagnostic but it also requires students to select and organize evidence that supports and justifies decisions through the case resolution process.

Pilot work with medical students, residents and staff physicians was conducted using BioWorld cases and conclusions recommended the use of this learning environment for medical education (Faremo, 2004). One key aspect of adapting BioWorld to a medical audience is to revise and construct cases at an appropriate level of difficulty. In our attempts to create and develop valid cases in medical education, we have experimented with different methodologies and scenarios to structure case creation. The companion authoring tool, CaseBuilder (Lajoie et al., 2001) which was designed to allow both instructors and researchers to modify cases easily also enables us to explore instructional activities for content creation and revision. Creating cases for an interactive computer environment implies documenting not only the questions and information related to the acceptable answer but it also requires the inclusion of plausible distracters or possible questions learners might have while trying to solve the case.

Creating a case can also be referred to as a problem generation activity, which is an instructional technique that requires the learner to assemble or construct a problem. In a problem generating task the learner needs to choose a specific case to construct or modify the problem they choose to explore and analyze if all the elements are defensible and could make sense for potential problem solvers. Silver (1994) includes both problem modification and problem construction in his working definition of the technique. The problem creator has to make sense of a situation and determine which elements can contribute to the solution and which other elements need to be present as distracters to increase the level of difficulty of the problem. In this context the case builder becomes a cognitive tool that supports our exploration of this learner-centered knowledge building activity.

Our data on the case creation phase demonstrated both validity and reliability issues when working with medical staff and students. As we raised the level of complexity of the cases in BioWorld we encountered challenges in the design and validation of solutions for these complex cases. As mentioned above, the solution to a case in BioWorld does not only consist of the final answer but it also requires a list of prioritized supporting evidences related to this answer. Consequently in the case creation activity, when medical students were constructing cases they had to provide and list the evidence supporting a good diagnosis. We found discrepancies between their hypothetical answer (answers they had planned) and their actual solution (the one recorded when they ran through the case in BioWorld). We first hypothesized that students were maybe not qualified enough to provide a clear answers so we asked a medical expert to do the same case twice. The expert was not aware of having to solve the same case twice; patient names were changed, a 10 day delay between testing occurred and we presented the expert with other similar cases in between the two cases in question. Again we obtained non-identical answer for a relatively simple case of diabetes. Our last attempt to address the inconsistent supporting evidence was to ask a second medial expert to do the exact same diabetes' case. Results were consistent with our two previous experiences and this lack of consistency forced us to address the issue of validity and reliability of solutions in a more systematic manner.

Tracking expert solution processes and explanations

To address the variability of the case's solutions we decided to construct a case validation activity. The simple list of evidence for justifying and explaining the answer was not sufficient to show where and how expert differed in their problem solving processes. Therefore the validation activity was designed to scrutinize every step along the way by including think-aloud protocols of individual participant. We justify this investment of time and resources by using expert teachers that have the ability and experience to solve and explain theirs though processes to others. Additionally, the validation activity has revealed itself to be a key component of the case creation process. The activity provides motivation and feedback to the medical student who acts as case creator. On the other hand when teachers solve cases created by students it gives them clear examples of students' misunderstanding of content and interrelationships of the different components involved in the diagnostic of cases. The validation activity consists of a simulation of a case presentation for medical teachers. Participants are asked to think aloud (do a think-aloud protocol) and provide explanations as they solve a case in BioWorld. The level of the cases and their explanation is at the undergraduate level. From a research perspective, we want to capture and record the strategies experts use to synthesize the information about a disease as well as how they structure and communicate this information in both oral and written forms. Expert teacher can provide us with relatively clear "path" of the decision process as well as explanations and verbalization about the metacognitive strategies they use while solving the cases. This validation activity will be used as a blueprint to build a cognitive model for each of our cases.

Sampling individual and collective problem representation

Protocol analyses are used to explore domain knowledge, to describe what are key elements and how knowledge is structured and used during a problem-solving task (Ericsson & Simon, 1993). Whereas protocol analyses of well-defined problems can result in clear problem solving sequences the analysis of ill-structured problems can be more complex given there is more then one way to reach a solution. We do not aim at conducting an exhaustive task analysis of all the possible solutions path or options but to sample and represent two to five solution paths for each case. The goal is not simply to build an expert path and use it to compare to novices' performance but to build a partial problem space representation that can evolve as more people do theses cases. The visual representations are built to offer a short summary of the though processes with the relative importance of specific steps to the resolution of the case. When interacting with theses representations participants can "zoom in" and open sub-layers to access details, related explanation and exact verbal transcript from the verbal protocol.

The problem space is constructed in three phases. The initial representation built by the researcher summarizes the decision-making process. It is used with experts to have them validate and reflect on the resolution path of the problem. We ask expert to first validate the summary of their case resolution and then select and categorize section of their decision path. Experts are asked to select which elements are absolutely

necessary to the case resolution, which ones are necessary and which one adds useful information. The second version incorporates the changes and categorization done by the experts. This categorization of decision path reveals the relative weight of specific steps and facilitates a comparison between experts. In the third version of the representation we merge experts path to show similarities and differences in the sequence of decisions leading to acceptable answer(s) for a specific case.

Goal and purposes of the visual representation

This validation activity serves multiple purposes. The visual representation is a tool to build and to communicate partial data to participants. As our initial participants are expert medical teachers we hope to extract pedagogical models for teaching specific cases and not only experts' case solution processes. Their experience teaching concepts related to each cases and their ability to predict what learners at the undergraduate level can understand will help us validate and improve content. The actual visual representation could also be incorporated into the computer-based learning environment to teach students. However we hope to use these qualitative blueprints to implement better scaffolding and feedback mechanism into our computer-based learning environment.

Situating the methodology

The use of a diagram, for the partial analysis of data is not a common procedure but Henderson, Yerushalmi, Heller & Kuo (2003) have found that visual maps are useful to analyze complex interview data. They found that concept maps reflected participants' conceptual understanding of the topic, clearly showed relationships between concepts and were useful to show similarities and differences between participants. The use of the term 'concept map' by these authors can be misleading as they do not use it in the way Novak and Cañas (2006) describe in their work. Henderson, Yerushalmi, Heller & Kuo used the Cmap software as a knowledge visualization tool to provide a visual overview of their protocol analysis. The diagram is not constructed by the participant but by the researcher from the verbal protocol. We chose Henderson et al.'s technique as a starting point to develop our own methodology to show a clear link between raw data and the participants' conceptual and procedural knowledge while solving a case.

Pilot Study

Participants

Our subjects were two medical teachers from the medical school at a Canadian university. One subject was an internist and the one was a gastroenterologist.

Materials

Questionnaires and case index

Participants were administered a questionnaire to control for their general practice, recent clinical experience and overall teaching experience. In a post-questionnaire we asked about their experience with cases related to the one they had solved in this study. Participants rated the cases for the level of difficulty and complexity after solving each of them.

Cases

The three cases presented to participants were diabetes mellitus type 1, hyperthyroid and pheochromocytoma. Theses cases were developed around similar set of symptoms and patient characteristics. Patient cases have similar symptoms to force participants to compare and contrast competing differential diagnosis and allow researcher to compare the solutions.

Software

Cases were composed using CaseBuilder but BioWorld software was used to present the cases to participants. Transcription and coding was done using Transana (Woods & Fassnacht, 2007) and the visual representation was built using Cmap software (Novak & Cañas, 2006).

Procedure

Phase 1

Two participants solved three fictitious patient cases in BioWorld. This computer-based learning environment presents patient information interactively. The environment allows the participant to navigate the problem space yet it is structured enough to allow for sequential presentation of case information. The task begins with a problem statement that presents a patient case. The participant selects relevant information from the case information and selects an initial hypothesis and confidence level in their hypothesis. As participants go through the different phases of the case resolution by selecting evidence and ordering tests they are asked to think-aloud and explain their reasoning as if they were doing a case presentation to undergraduate students. After participants have completed each case the verbal transcript and computer log are chronologically combined into one protocol (see example in table 1).

Line	Transcript of Verbal Data		BioWorld Log Data		
…	Transcript	Time verbal	Time BioWorld	Evidence	Action
25	So ah then I guess I should underline the evidence that I have here?	:07:42:			
26	R: yep	07:47			
27	E: So age is important. And then she is on medication, it's very important to know what the medication is.	07:55	07:55	37 year old	add evidence
28	And then just the high blood pressure in a 37 year old is a-, makes you think commonest thing is still essential hypertension but you have to start thinking possibly of secondary causes especially if her blood pressure is really high.	08:07	08:08	medication	add evidence
29	OK uh, frequent headaches is a very important thing if you combine the frequent headaches with the episodes of flushing that really makes you think a lot of a pheochromocytoma.	08:24			

Table 1: Merging verbal transcript and computer log

The researcher uses this protocol to constructs a visual representation of the solution process. Each node or item on the diagram is linked to original statements or action from the protocol. Items can be regrouped or nested into main nodes if participants explicitly combine them or if they represent pre-identified elements/actions from the reasoning process. This visual representation is a summary of their transcript, yet the link to the original data is easily accessible by mouse-over as shown in this screen capture of figure 1.

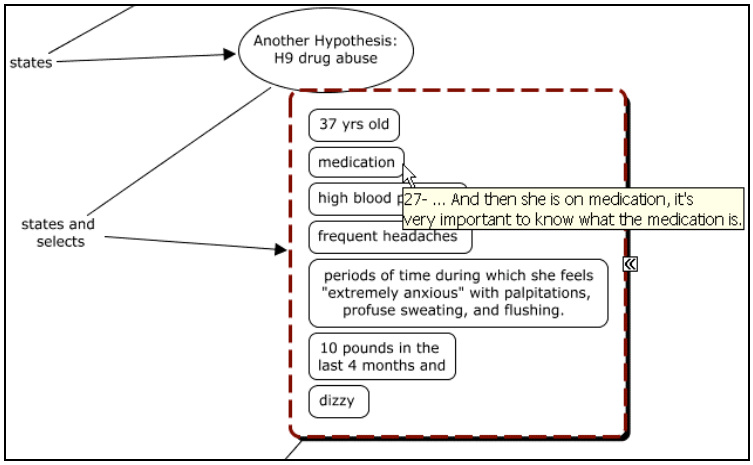


Figure 1: Extract of visual representation of the path with mouse-over

Phase 2

In phase two, a paper version of the diagram is presented to participants for validation. Participants can refer to the protocol and decide to modify or elaborate the initial diagram of their case resolution. Once participants have validated our summary they are asked to select which elements are absolutely necessary to the case resolution, which ones are necessary and which one adds useful information. In the activity participants are asked to color in red elements of the path that are absolutely necessary to solve the case; in yellow elements that are necessary to solve the case; and in blue extra information that is useful but not crucial for solving the case. This leads to a decision path to which we can assign weighting to selected elements. We chose to assign weights of 5, 3 and 1 to better differentiate the importance of each element and be able to use numerical values to assess solutions later on. The three points scale was used for practical reasons to differentiate items of relative importance. We may need to modify the categorization of this rubric if we find it difficult to apply to future sets

of cases. Table 2 below summarizes the resulting grid. Figure 2 shows a section of the path that was categorized with colors in parenthesis.

	Red (+5)	Yellow (+3)	Blue (+1)
Key elements	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)

Table 2: Grid of weights for categorization of elements

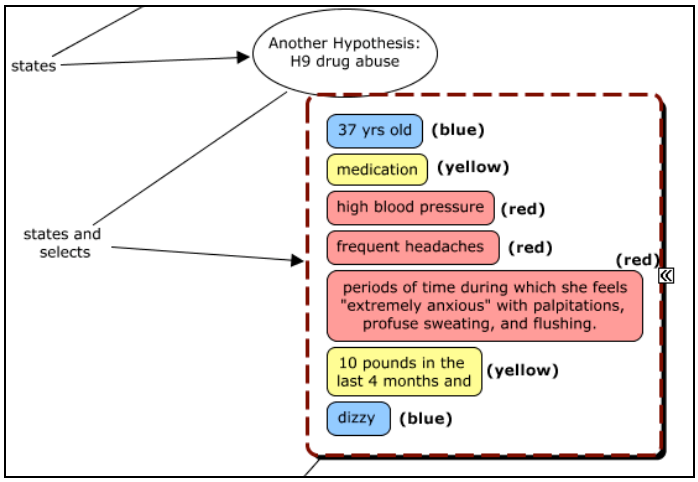


Figure 2: Categorized section of the path

Phase 3

In phase three the researcher combines the visual representation of both expert for each case. As seen in Figure 3 the representation shows where experts' decision process is similar and where it differs. Unfortunately the static figure does not allow the reader to explore details in each layer of the decision path but it gives a good overview of the representation.

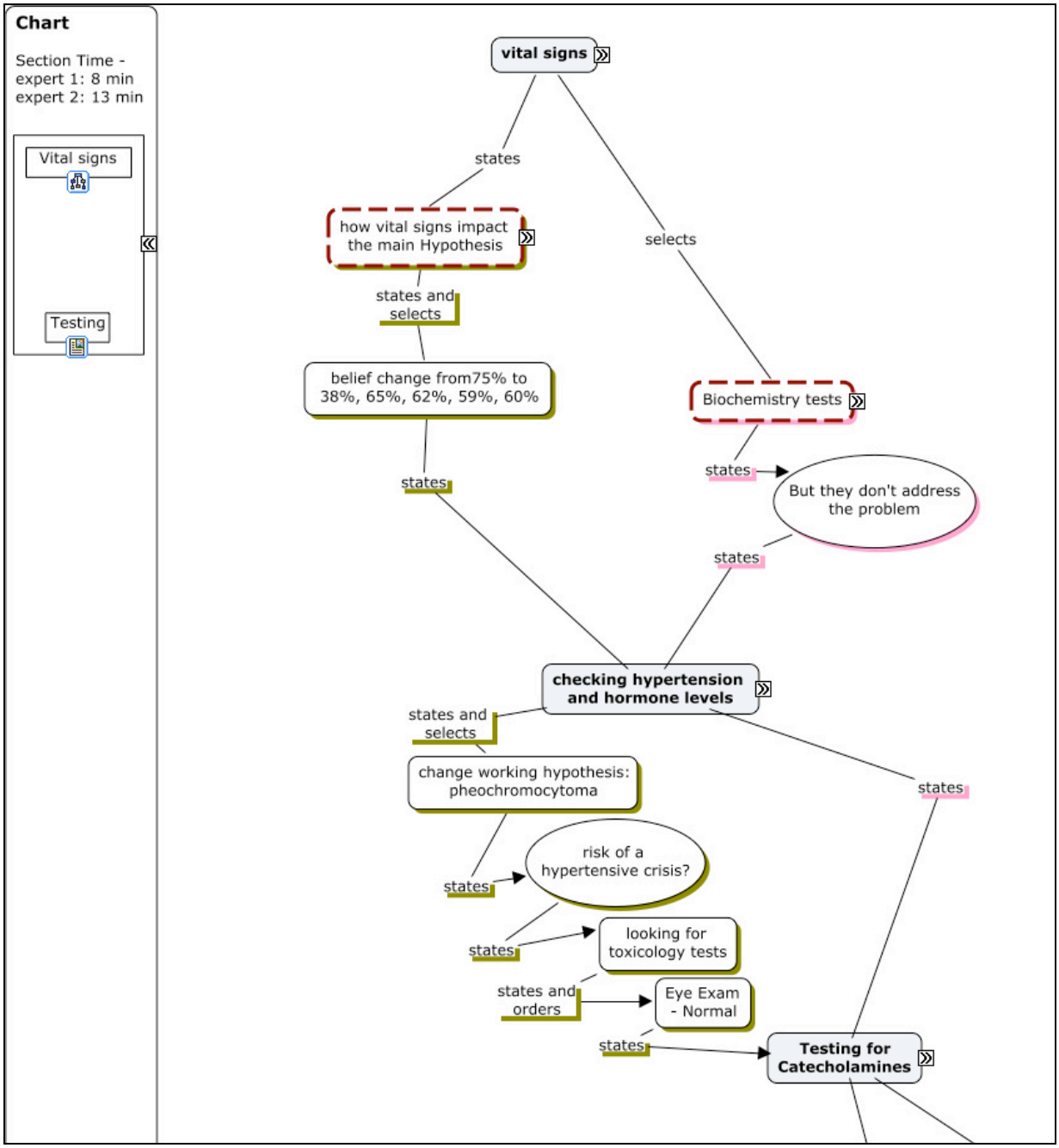


Figure 3: Section of the combined solution path

Preliminary Results

Visual representation of the problem space for each case

The use of multi-layers diagram allows for relatively simple overview of the reasoning process and it shows relationships amongst important elements of a case. It is particularly useful to capture details in the inner layers of the map as it is flexible in presenting peripheral information related to the case.

Exploring and explaining the variability of solution path

To explore and try to document and understand the variability of the solution path we will present sample of the results and analysis related to the case of Pheochromocytoma. The categorization of elements from the case resolution was used to explore consensus. Elements from the categorized diagram were sample from both experts into each of the three categories (red, yellow, blue – see annex I for detailed categorization tables). Table 3 combines the sum of elements for each category by each expert. As you can see the number of elements selected is similar (20 for expert 1 and 18 for expert 2). However when the weights are applied to the categorized elements as shown in table 4, the difference between our two experts is a lot more important (46 for

expert 1 and 72 for expert 2). Expert 2 uses a higher weighting for most of the elements he selected which might be due to his experience teaching similar cases to students.

	Expert 1	Expert 2
Absolutely necessary	3	12
Necessary	7	3
Useful information	10	3
Total elements	20	18

Table 3: Comparison of the number of key elements from categorization tables

	Expert 1	Expert 2
Absolutely necessary (+5)	15	60
Necessary (+3)	21	9
Useful information (+1)	10	3
Total scoring	46	72

Table 4: Comparison of weigted elements from categorization tables

In table 5 we have combined element without taking categories into consideration to calculate the percentage of consensus between our two experts. This consensus rate is coherent with the literature in medicine.

Similar elements with similar weighting	8
Similar elements overall	13
Total of elements	38
Percentage of similar identified elments	34.21%

Table 5: Summary table

Exploring the resolution process

To better understand and categorize the reasoning process of our participant we use a coding scheme that was adapted from previous work in medical reasoning (Faremo, 2004). Elements of interest from our coding scheme fall into six main categories. We are analyzing time, diagnostic tests, interpretation of tests, hypothesis and confidence level, evidences and use of metacognitive skills throughout our sources of data. The three main sources of data are computer log and report, the verbal transcript and the categorization tables. We also have observational and questionnaire data but we consider them as secondary relevance for this research. In brief, our initial analysis show a consistency in the number and specific evidence collected, the list of hypothesis generated (verbally and with computer log) and the time required for the formulation of the correct hypothesis. However we found differences in the number and list of tests ordered as well as the length and level of explanations.

Limitations and lessons learned

As this pilot study was used to test material and procedures we are confident that the real study will address some of the limitations identified throughout this experimentation. Careful selection of participants seems to be key since we are testing the pedagogical model and that our data show that one of our participant only had textbook kind of experience with these cases. Other lesson learned concerned the reduction of data manipulation and the inclusion of screen captures to improve the transcription phase and make our data more reusable. The use of pen and paper method with participants has been restricting so we will have our participants directly interact with the Cmap software in our next experimentation.

Conclusion

Ill-defined problems do not have clear-cut answers but contrasting optimal and not so optimal solutions might improve participants' fragmentation of the problem representation and meta-cognition. By carefully documenting the resolution of cases by participants we hope to gain an understanding of how scripts or schema develop for diagnostic reasoning (Charlin, Tardif, & Boshuizen, 2000) and explain the variability found in cases' solutions. We want to further explore if visual representations of solution paths can provide a meaningful framework to synthesize, structure and communicate different levels of knowledge, reasoning strategies and metacognition.

As more data are collected we will test the robustness of the methodology and add a developmental perspective to the problem space representation for each of our cases. The coding scheme developed needs to

be validated with other participants. This step is a pre-requisite to allow for comparison of diagrams from multiple participants as described in other studies using visual knowledge representation tools (Henderson et al., 2003; Johnson, 2005).

Building cognitive model by consensus building

The main goal in sampling expert collective solution paths for each case is not only to explain variability but to reach consensus on what are the key elements for leading to successful or acceptable solutions. Building cognitive models will also include errors and strategies in the context of specific case resolution. The visual representation aims at being a source of information for researchers and experts for analysing diagnostic reasoning. Additionally the representation also provides learners and instructors in medicine with a meaningful tool to record and build on the case presentation practice. We are spending a great amount of time on expert pedagogical models to enable their use and test their utilities with cohort of students. The next step in our research will be to test these models and verify the accuracy of pedagocical models as captured by this activity.

Reference

Barrow, H. (1994). *Practice-based learning: Problem-based learning applied to medical education*. Springfield, IL: SIU School of Medicine.

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-Based Learning: An Approach to Medical Education*. New York, NY: Springer Pub. Co.

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. *Acad Med, 75*(2), 182-190.

Cox, K. (2001). Stories as case knowledge: case knowledge as stories. *Med Educ, 35*(9), 862-866.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (Rev. ed.). Cambridge, Mass.: MIT Press.

Faremo, S. (2004). Examining Medical Problem Solving in a Computer-Based Learning Environment. Unpublished doctoral dissertation. McGill University.

Greenhalgh, T., & Hurwitz, B. (1999). Narrative based medicine: Why study narrative? *BMJ, 318*(7175), 48-50.

Henderson, C., Yerushalmi, E., Heller, K., Heller, P., & Kuo, V. H. (2003). *Multi-Layered Concept Maps for the Analysis of Complex Interview Data.* Paper presented at the Physics Education Research Conference, Madisson, WY.

Johnson, T. (2005). *Analysis-Constructed Shared Mental Model Methodology: Using Concept Maps as Data for the Measurement of Shared Understanding in Teams.* Paper presented at the Extending Cognitive Load Theory and Instructional Design to the Development of Expert Performance, Open University of the Netherlands.

Jonassen, D. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research & Development, 45*(1), 65-94.

Koschman, T., Kelson, A. C., Feltovich, P. J., & Barrow, H. S. (1996). Computer-supported problem-based learning: A principled approach to the use of computer in collaborative learning. In T. Koschman (Ed.), *CSCL: Theory and Practice of an Emerging Paradigm*. Mahwah, NJ: Lawrence Erlbaum Association.

Lajoie, S. P., Lavigne, N., Guerrera, C. P., & Munsie, S. D. (2001). Constructing knowledge in the context of BioWorld. *Instructional Science, 29*(2), 155-186.

Novak, J. D., & Cañas, A. J. (2006). *The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools 2006-01*: Florida Institute for Human and Machine Cognition.

Silver, E. A. (1994). On mathematical problem posing. *For the Learning of Mathematics, 14*(1), 19-28.

Woods, D. K., & Fassnacht, C. (2007). Transana (Version 2.20). Madison, WI: The Board of Regents of the University of Wisconsin System.

Annex I : Categorization Tables of Expert for the Case of Pheochromocytoma

E1	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	**Urinary Catecholamines / Norepinephrine**	**high blood pressure**	**37 yrs old**
2	**Urinary Catecholamines / Total (Epinephrine + Norepinephrine)**	**extremely anxious**	not a new problem
3	**Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr**	**palpitation, profuse sweating, and flushing**	Fasting Blood Glucose Level - normal
4		more frequent in the past little while	Serum Electrolytes / Anion Gap (Na-(Cl+HCO3))
5		**weight loss**	Serum Electrolytes / Magnesium (Mg)
6		Ultrasound / Abdominal Scan	Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT)
7		CT / Body	Aldosterone
8			Adrenocorticotropin homrone (ACTH)
9			Cortisol
10			Dehydroepiandosterone Sulfate (DHEA-S)

Table 1: Categorization of key elements of expert 1

E2	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	headache, palpitations, sweating and flushing; makes me think of secondary causes of hypertension	medication	**37 yrs old**
2	pheochromocytoma is rare so you need to keep other causes in mind	**10 pounds in the last 4 months and (evidence)**	Dizzy
3	**high blood pressure**	checking toxicology tests	eye exam test
4	frequent headaches		
5	**periods of time during which she feels "extremely anxious" with palpitation, profuse sweating, and flushing.**		
6	hypothesis 1: Grave's disease		
7	hypothesis 2: pheochrocytoma		
8	hypothesis 3: essential hypertension with reaction to medication		
9	hypothesis 4: drug abuse		
10	pulse of 98 a minute		
11	**one of the 3 followint tests: a)Urinary catecholamines, b) Urinary Metabolites VMA, c) Urinary catecholamines**		
12	submit hypothesis pheochromocytoma with high belief		

Table 2: Categorization of key elements of expert 2

Argument diagramming as focusing device: does it scaffold reading?

Collin Lynch[1], Kevin Ashley[2], Niels Pinkwart[3] and Vincent Aleven[4]

[1]*Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA (collinl@cs.pitt.edu)*

[2]*Intelligent Systems Program and School of Law, University of Pittsburgh, Pittsburgh, PA, 15260 USA (ashley@pitt.edu)*

[3]*Clausthal University of Technology, Julius-Albert-Str. 4, 38678 Clausthal-Zellerfeld, Germany (niels.pinkwart@tu-clausthal.de)*

[4]*Carnegie Mellon University, HCI Institute, 5000 Forbes Avenue, Pittsburgh PA 15213, USA (aleven@cs.cmu.edu)*

Abstract: In this paper we report on a study of attention and student recall in our ITS LARGO. The system was employed in a study of graphical markup in legal education. Students in the study were divided into two groups, one employing the graphical tutoring environment, and the other traditional text notes and highlighting. Post-test comparisons between the two showed gains among the incoming students who had scored lower on a standardized Law School Admissions Test (LSAT). We argue that the system and its graphical prompts were effective in guiding the students to the relevant textual portions and that they showed some gains in focus of attention.

Keywords: Ill-defined domains, note-taking, attention, self-explanation.

Introduction

An ill-defined problem-solving task is one in which (1) the problem does not have a definitive answer, (2) the way in which the problem-solver solves the problem depends on how he conceptualizes it, and (3) problem-solving involves identifying relevant concepts and mapping them onto the situation to be solved [10]. Deciding how to resolve a legal dispute is an ill-defined task. Reasoning with hypotheticals is a strategy for dealing with that. Each participant (i.e., the contending advocates, the deciding judges) may propose a different, perhaps inconsistent but often reasonable solution. The alternatives often evidence differences in the ways in which the participants conceptualized the problem or applied those legal concepts to the problem's facts. In applying the concepts, legal reasoners often draw analogies between the problem's facts and past or hypothetical cases; these analogies map legal concepts that apply in the hypotheticals or precedents onto the present case's facts to help draw and justify conclusions.

This work focuses on legal problem-solving at the Supreme Court of the United States (SCOTUS). A feature of problem-solving at this level are oral arguments before the Court. Each side in a case gets thirty minutes to address the issues before the Court; the arguments are recorded and later published. In it an advocate for one side proposes a rule or test for deciding the case in favor of his client. Justices in turn pose hypotheticals in order to probe the proposed rule. The hypotheticals help the Justices to understand what the proposed test means, whether it is consistent with past decisions, and how well it implements and reconciles the conflicting legal policies and principles. Legal reasoning with hypotheticals is one of the tools Justices have for mapping legal concepts from past decisions and applicable statutory and constitutional provisions onto the problem's facts and adjusting the mappings to account for underlying legal policies and principles.

As such SCOTUS oral arguments provide good examples of reasoning with hypotheticals for law students to study. Law students are exposed to, and sometimes participate in, Socratic dialogues in classes from which they should learn to reason about legal rules with cases and hypotheticals. The SCOTUS oral arguments are potentially a pedagogically valuable source of examples of this kind of reasoning. They are realistically complex, often highly dramatic, and they are written down which

facilitates studying them at some length. On the other hand, they are an underutilized pedagogical resource. Law professors may employ SCOTUS oral arguments to teach lessons about the substantive law of an area, but they do not generally use them as examples of argumentation methods. While traditional legal education encourages students to make and respond to arguments, it does not provide much explicit support for reflecting on the process.

The LARGO program attempts to redress that failing by helping law students reflect on SCOTUS oral arguments as examples of legal argumentation. An intelligent tutoring system (ITS), it teaches legal reasoning with hypotheticals by helping students to represent selected elements of these examples of expert legal arguments in diagrams (Other legal ITSs include CATO and CATO-Dial [2,6].) The elements include an advocate's proposed test for deciding a legal case, Justices' hypothetical examples posed to probe the test, and the advocate's responses to the hypotheticals. Students identify these elements in the text, represent them in a diagram, providing their own reformulations of the text, and link the elements graphically indicating certain dialectical relationships among them [12]. Given the ill-definedness of the task, and the subjectivity of interpreting the textual argument, LARGO cannot simply teach by identifying "right" and "wrong" answers. Instead it provides feedback based upon expert markup and an understanding of common dialectical patterns. This hint mechanism will be described below.

The value of note-taking has long been recognized in legal education [9] but the focus has always been on text notes. Graphical notes have been shown to be beneficial through their ability to focus the student's attention on relevant portions of the text [13]. A similar effect has been noted for ITS feedback [1]. Graphical argument representations have been studied in philosophy [15] and legal education [7]. Unfortunately the results were inconclusive.

In an experiment, we compared the LARGO program with a more traditional text-highlighting-and-note-taking word-processing environment that focused students on the same elements and relationships of hypothetical legal reasoning but without the diagramming or feedback. We found evidence that students with lower LSAT scores benefited the most from LARGO and its support of graphically diagramming arguments [12]. These students using LARGO learned some targeted skills of hypothetical legal reasoning better than comparable students in the control group (the Text-only group).

We have begun to attempt to explain why LARGO has benefited such students in the Diagram group. This paper reports the results for our initial hypotheses in explaining the data, that (1) students in the Diagram group, with LARGO's support, are more successful in finding and attending to pedagogically-relevant portions of the text than students in the Text-only group. In particular, (2) students in the Diagram group with lower LSAT scores, which may indicate lower reading skills, benefit more from LARGO's support in finding and attending to important portions of the text than higher LSAT students.

In the next section, we describe LARGO's instruction about hypothetical reasoning and illustrate it with an example of a student's diagram of excerpts of an oral argument. Following that we describe the former experiment and illustrate the output of a student in the Text-only group. In the Empirical Evaluation of Attention section we describe our current empirical evaluation comparing the portions of the text students attended to in the Text group vs. the Diagram group, including the way we operationalized that comparison. In the subsequent sections, we present the results, discussion, and our conclusions.

LARGO Instruction.

In our study, law students read oral argument transcripts from SCOTUS. Figure 1 contains an example of the tests and hypotheticals encountered in such arguments drawn from the oral argument in *Burnham v. Superior Court of California*, 495 U.S. 604 (1990). The left column contains the text of the argument with line numbers. Mr. Sherman makes arguments on behalf the "petitioner" in the case, Dennis Burnham; "QUESTION:" indicates a Justice's question.

Here are the facts of the case. After Burnham and his wife decided to separate, she moved to California with their two children. In January, 1988, Mrs. Burnham filed suit in California for divorce. Later that month, Burnham visited California on business and to visit his children. Upon returning one of them to Mrs. Burnham's home, he was served with her divorce petition. Later that year, he appeared in California Superior Court to assert that the courts there lacked *personal jurisdiction* over him. Personal jurisdiction, a technical legal concept first year law students encounter in

their "Legal Process" course, means a court's power to require a person or corporation to appear in court and defend against a lawsuit. Burnham argued that his contacts with California, consisting only of a few short visits to conduct business and visit his children, were insufficient to grant the courts there jurisdiction of his person under the Due Process Clause of the Fourteenth Amendment, which guarantees certain minimum procedural safeguards against the arbitrary exercise of government power. Conflicting with that principle is the principle that a state may redress wrongs committed within or affecting residents of the state. The California Superior Court denied his motion, and the SCOTUS agreed to review that decision. The Court affirmed the lower court decision, but could not agree on a majority opinion.

Oral argument excerpts	**Argument Move According to Model of Hypothetical Reasoning**
5. The issue presented here is whether a state can exercise personal jurisdiction over a nonresident defendant who was personally served while present in the state if that defendant does not otherwise have sufficient contacts with the state to satisfy the minimum contacts test announced in International Shoe. 11. We're here today to ask you to instruct the courts of this land otherwise, to give effect to what the Court said in Shaffer, that personal jurisdiction in all cases must be tested by the minimum contacts test.	**→ Proposed test of Mr. Sherman for Petitioner Burnham**
15, 17. QUESTION: Mr. Sherman, even if you are correct that some minimum contact is necessary for personal jurisdiction, wouldn't the transitory presence within the state of someone meet that test -- in a good many instances?	**← J.'s hypo**
18. MR. SHERMAN: I think not, Your Honor. And it's important to distinguish --	**→ Response: distinguish cfs/hypo**
19, 21. QUESTION: I would have thought so and that perhaps someone who voluntarily enters a state to transact some business or to visit there might well meet whatever minimum contacts are -- required.	**← J.'s hypo**
22, 24. MR. SHERMAN: On that -- on those facts, yes. If he were just passing through momentarily, say, stopping over on his way to Hawaii, not conducting any business or -- classically flying over --	**→ Response: distinguish cfs/hypo**
25, 27, 29. QUESTION: You have a different situation if someone is flying over the state, overhead -- and is served in mid-air than you do with someone in your client's --- position.	**← J.'s hypo**
30. MR. SHERMAN: But the question that your hypothetical poses is what kinds of contacts would be sufficient under the minimum contacts test for somebody who was not in the state very long. And the answer to that would depend upon applying the minimum contacts test and typically the cause of action has to be related to or rise out of contacts that the defendant has.	**→ Response: distinguish cfs/hypo; modify test to exclude hypo.**

Figure 1: Oral Argument Excerpts and Argument Moves from Burnham.

The right column of Figure 1 lists the argument moves in our model that correspond to the assertions. For more on the model see [3,4,5]. Mr. Sherman proposes a test deciding the issue in favor of Mr. Burnham, namely there is no personal jurisdiction without a showing of adequate minimum contacts. The Justices challenge that test with hypotheticals, asserting that there *are* adequate minimum contacts in this case. Mr. Sherman distinguishes the hypotheticals and finally asserts a meaningful distinction with a corresponding test modification. Even if minimum contacts exist, he argues, the cause of action (the subject of the suit) must arise from them to have personal jurisdiction.

Figure 2 shows an example of the LARGO interface. The argument transcript is on the top left. On the right is a workspace for creating the diagram using the palette of representation elements at the bottom left. Students create graphs representing an argument exchange in the transcript by dragging the elements from the palette to the workspace. Elements exist for representing the current fact situation, proposed tests (and modifications), hypotheticals, and various relations among them (e.g., modifying a test, distinguishing or analogizing a hypothetical, and a general relation). Students also link the elements in their diagrams to passages in the transcript via a highlighting feature.

The diagram in the workspace at the right of Figure 2 is a student's representation of some of the excerpts of the *Burnham* argument (see Figure 1.) The student has represented the Justice's "served-while-flying-over" hypothetical and linked it via the highlight function to a portion of the transcript including lines 25, 27, and 29. The student has also identified a version of Mr. Sherman's test (top) which gets modified to a version that roughly accounts for the additional contact-centric limitation imposed in line 30.

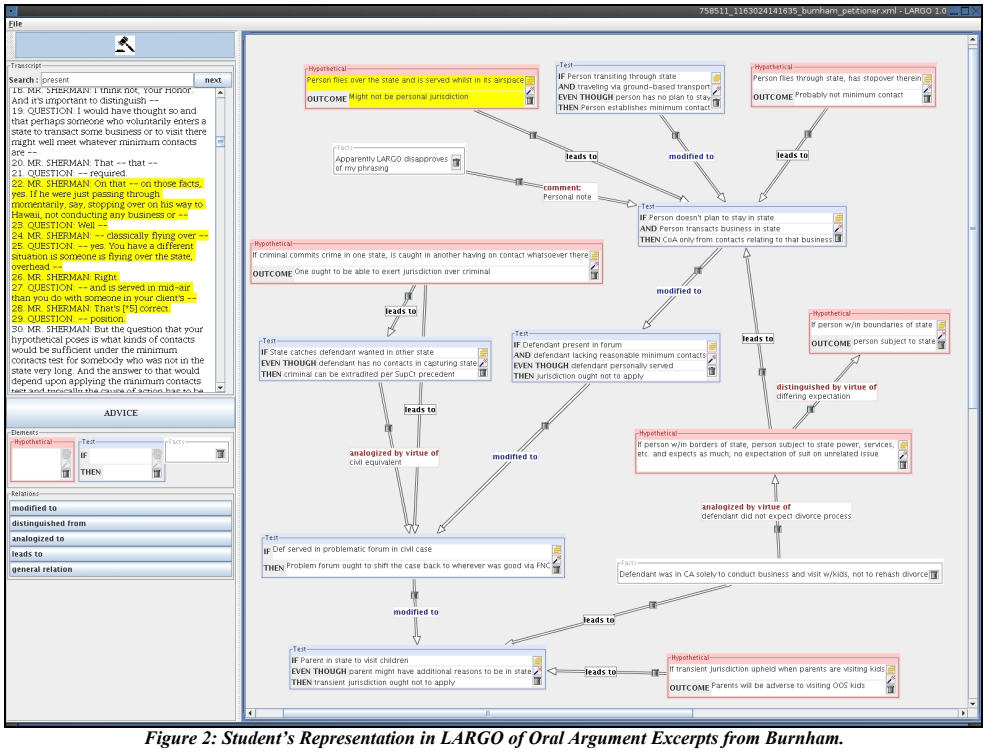


Figure 2: Student's Representation in LARGO of Oral Argument Excerpts from Burnham.

Students may obtain feedback on their developing diagram by clicking on the Advice button. This brings up a palette with up to three hint choices with titles such as "Reflect on the role of hypotheticals in the transcript". Clicking on a title brings up a more detailed message. All of LARGO's help is provided by request only.

The feedback leads students to review the text of the oral argument in at least four ways: (1) Some feedback identifies regions in the text where the student's graph lacks elements corresponding to those in an expert's markup of the argument transcript. Prior to a transcript's use in LARGO, an expert marks passages of interest such as those shown in Figure 1. This type of hint points the students to a larger region than the actual element and informs them that an item of interest is present in it. (2) Other feedback identifies parts of the diagram where the relations among elements do not correspond to the "standard" model. For instance, hypotheticals are commonly analogized to or distinguished from one another and the current fact situation A student's graph may fail to show such relationships or may indicate uncommon relationships (e.g., analogizing or distinguishing a test and a hypothetical.) This type of feedback may lead students to reexamine the portions of the text that embody the elements and their relations. (3) Some system advice asks students to compare their test formulations to examples from other students or the professor; this may lead students to reexamine the text containing an advocate's test as they consider which conditions to include and how abstractly to characterize them. (4) Finally, some LARGO advice identifies a standard dialectical pattern or node configuration in the students' model and asks them to reflect on how the pattern bears on the argument's merits and whether a different decision might have been more appropriate. The student may be pointed to a proposed test that led to a hypothetical which in turn prompted the advocate to modify the test (e.g. top of figure 2). Such self-explanation prompts [8] lead them to reread the initial text and, on occasion, to modify their representations. This has been shown to be effective when studying examples [14].

Fall 2006 Experiment.

In fall 2006, we conducted an experiment to investigate to what extent LARGO can lead to better learning than a traditional purely text-based alternative. The alternative tool simulates the

"traditional" process of examining the argument transcript with a notepad alone by allowing students to highlight selected portions of the transcript text and enter their notes in a text pane. Figure 3 shows a screenshot of the tool as it has been used by one of the study participants to annotate the transcript of the oral argument in Burnham.

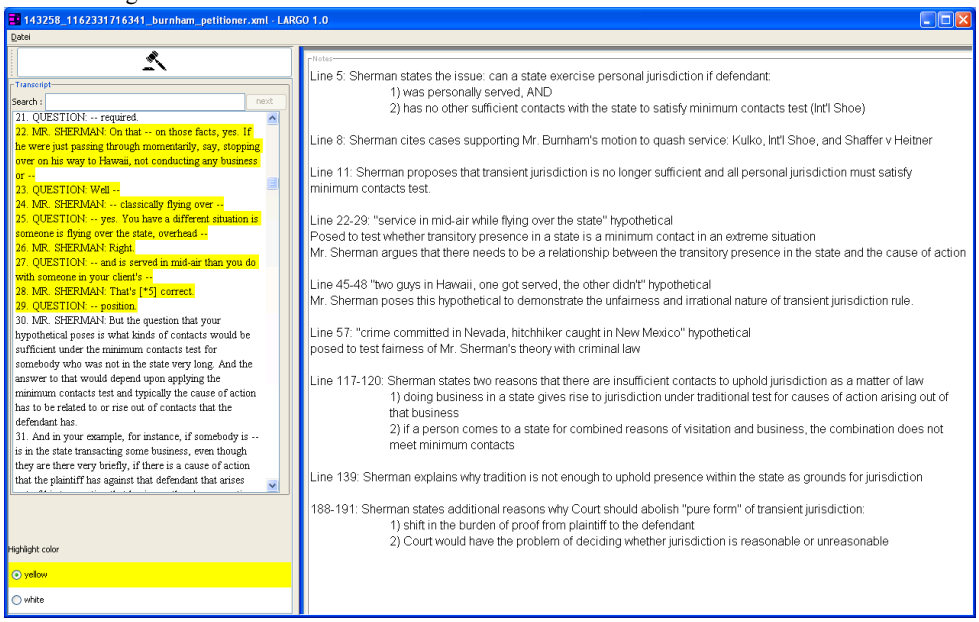


Figure 3. Screenshot of control condition (text tool).

The experiment was conducted with first-year students at the University of Pittsburgh's School of Law. All were volunteers and were paid $80 on completion. All cases examined in the study centered on questions of personal jurisdiction and were part of their coursework. The students were assigned randomly to the conditions. 38 students began the study, 28 completed.

The experiment consisted of four 2 hour sessions over a single one month. The first involved a pre-test and an introduction, with example, to the dialogue model and the appropriate system. In session 2 students used the systems to examine extracts of the Burnham oral arguments. In Session 3 they repeated the process with *Burger King Corp v. Rudzewicz* 471 U.S. 462 (1985). Experimental students used LARGO for note taking while the control subjects employed the text system. Both groups were instructed to take notes on the argument using the tool.

Our main hypothesis was that LARGO's graphical and advice tools would help students better identify and mark up the argument components, leading to better learning of argument skills. A first analysis of the results has been published in [12]. On average, the experimental group did better in the post-test than the Control group, yet the difference was not statistically significant ($t(1,26)=.92$, $p>0.1$). Dividing the students into three groups according to their Law School Admission Test (LSAT) scores, revealed that the "Low" experimental students benefited most from LARGO. This group scored significantly higher than their Control counterparts in several categories of post-test items (although not overall) including argumentation about a near-transfer problem, questions on a novel personal-juristiction case and questions asking them to evaluate argument components [12]. The results support our hypothesis (though they perhaps fall short of decisive confirmation). For students who do not (yet) have the ability to learn argumentation skills based on reading alone, using LARGO led to significantly better learning of argumentation skills than the traditional note-taking techniques. For the more advanced/skilled students, LARGO was neither better nor worse than traditional methods.

Empirical Evaluation of Attention.

Our post-test results, while intriguing, do not provide a definite measure of LARGO's utility. We thus undertook a more detailed attention analysis. This analysis was performed on the data files and logs

produced by the students during the study in order to understand what aspects of LARGO (visual representation, linking of graph to transcript, or advice) led to the observed difference between conditions. We investigated whether students using LARGO were more successful in finding and attending to the important portions of the text (i.e., relevant test and hypothetical formulations in the argument) than students who use the text-tool. If there is a difference between the conditions in terms of finding and locating relevant textual items it could partially explain the post-test differences: Students who, assisted by the tool, focus their attention on important parts of the long oral argument transcript, are likely to learn the important parts better.

We began our analysis by determining how much of the students work was *relevant* (that is, forwarded the goals of their analysis) and how much of it was not. Our particular focus was on the students' identification of the relevant tests and hypotheticals within the transcript. We did not consider their facility at identifying the relevant legal issues or relationships between these elements. As we noted above the students were assigned to examine three cases during the study. Their examination of the intro case (California v. Carney, 105 S. Ct. 2066 (1985)) was guided by a document which presented the domain model as well as a step-by-step sequence of appropriate analyses. The remaining two cases, *Burnham* and *Burger King* were analyzed without such guidance. And we focus on them here.

In preparation for this study an expert legal instructor marked up the redacted transcripts of each argument. This individual identified a set of important Tests and Hypotheticals in each oral argument. This markup did not include defining an ideal graph or summary statements of each element, only the identification of relevant regions. This resulted in a list of 33 regions over the two cases. Sixteen of these regions (*Core* set) were encoded into Largo for use in providing hints of type 1 (see above). The remaining 17 were reserved for this study and designated as the *Test* set.

Our goal was to provide both a basis for hinting (Core), and a baseline (Test) for comparison. For the Graph condition these two sets form a test-train split of a type commonly employed in Machine Learning (ML) [11]. As noted in the introduction this process of markup is an ill-defined one. We hold no expectations that the students overall analysis will match ours. However we feel that it is neither illogical nor extreme to expect them to locate the same statements of tests and hypotheticals as a legal expert, and we expect good students to perform better. While there are complicating factors, we argue that this is an appropriate methodology and one that draws on the relevant literature.

(Entries) In order to effectively compare the two groups we defined a standard baseline unit of student work. We therefore defined the *Note* as a single atomic reference or notation made by the students. For LARGO students a note is a single graph node or relation. The test and hypothetical nodes may be linked to the graph. A node is *location-relevant* if it is linked to one of the Core or Test locations irrespective of type. It is *type-relevant* if it links to the location and is of the correct type.

The graph shown in figure 2 contains 29 notes. Including relational links and fact nodes in the class of `notes' penalizes the graph subjects for making those entries as they cannot increase the success measures only decrease them. We opted to include them for three reasons: (1) the students' task was to markup the transcript including relations and discounting that effort would skew the counting toward minimal graphs. (2) the relationship structure has value and should be a part of any reasonable assessment. (3) dropping the edges unilaterally from the graph condition would bias the results in their favor as no viable standard was available for discounting text notes in the same way.

A Text note is defined as a single paragraph entry that may be accompanied by a highlight. Such a note is *location-relevant* if the text explicitly references some key transcript portion by line number or via a highlight. It is *type-relevant* if it explicitly identifies the type of the location in text. Figure 3 contains 9 textual notes. The 4th, 5th, and 6th all specify a type. The structure of these notes is similar to those given as examples in Session 1. We defined note in this way to ensure that the text and graph subjects employed roughly the same amount of cognitive effort when making each note.

(Measurement) We will focus the remainder of our discussion on three Data measurements (Time, Help & Work), and three Success measures (Efficiency, Precision, & Recall) commonly employed in ML applications. We will discuss all the measurements on a case basis (e.g. Time spent on Burnham) and overall. *Time*, is a measure of the time spent on task. *Help* reflects the number of advice requests the student made and the number of times they followed up with specific advice (applicable only to LARGO students). *Work* is a measure of the number of notes made by each student. For the LARGO students we counted each node and edge. For the text students we approximated the total number of notes by using the number of highlights or text notes whichever was largest. Thus we kept the count linked to distinct note-taking acts. While this may undercount

slightly we think that it is a viable choice.

The success measures reflect the extent to which the student did or did not focus on the key elements. *Recall* is defined as the number of relevant notes that were located by the student out of the total number. *Efficiency* is the rate at which the students located the relevant notes. *Precision* is the number of relevant elements located versus the amount of work done. These definitions vary somewhat from those typically used in ML but we find them more appropriate here. We calculated each measurement with respect to the Core and Test sets and overall. We present the results below.

$$\frac{(Key\ Spots\ Found)}{(Total\ Key\ Spots)} \qquad \frac{(Key\ Spots\ Found)}{(Total\ Notes\ Made)} \qquad \frac{(Key\ Spots\ Found)}{(Time\ on\ Task)}$$

Recall Precision Efficiency

Figure 4: Success Measures.

(Hypotheses) Extant research on the benefits of graphical notes asserts that the students using them will be better able to 'focus in' on the relevant material. As such we have the following hypotheses: *(h0)*: students in the graph condition will have higher *recall* than their textual counterparts. *(h1)*: students in the graph condition will be more *efficient.* And *(h2)* students in the graph condition will have higher *precision* than their text counterparts. We discuss data relating to these hypotheses below.

Results

During the study we controlled for time on task and there was no significant difference between the two conditions either in terms of total study time or time spent on each case. The only variation occurred within the High student pool. There the Text students spent significantly more time overall ($t(5.09)=33.67$ $p < .0.0069$[1]) and on a per-case basis ($t(13.2)=3.71$ $p < 0.002$ for Burnham and Burger King). We discuss to the significance of these differences below.

There was no overall difference between the conditions in terms of the work done. There was, however, a case-specific difference. On Burnham there was a trend ($t(12.3)=1.75$ $p < 0.05$) in favor of the Text condition indicating that they did more work. This same pattern appeared in Burger King but was very significant ($t(12.43)=2.7$ p 0.008). Unlike Time there was a within-condition trend with the graph condition doing more work on Burger King than on Burnham ($t(27.87)=-1.7$ $p < 0.05$). As before the High Text students did more overall ($t(5)=3.71$ $p < 0.01$) and on a case basis ($t(5)=4.33$ $p < 0.01$ and $t(5)=3.9$ $p < 0.01$ for both cases).

Table 1 shows an overall comparison between the text and graph groups. For this and other tables p-values are not shown. Regular text indicates $p < 0.025$ while italics indicates $0.025 < p < 0.1$. As you can see in Table 1 the Graph condition outperformed the Text condition in terms of Efficiency, Precision, and Recall on the *Core* set both in terms of the location and type-relevance standards. They were also were more precise at locating the relevant elements and showed greater efficiency, precision and recall about typing them. This pattern continued on the full set. Interestingly, the Text condition performed better at locating the relevant entries but not at assigning types to them on the *Test* set. This pattern was observable on Burnham alone (Table 2) but not for Burger King. Here the graph condition was dominant and, while not outperforming the Text condition in terms of Efficiency and Recall, did outperform them in terms of Precision (Table 3).

[Table 1]	Core	Test	All
Eff Located	T<G t(25.8)=-5.3[2]	T>G t(25.3)=3.34	
Typed	T<G t(25.17)=-7.6		T<G t(23.7)=-5.56
Prec Located	T<G t(25.3)=-4.3	T>G t(26)=2.5	*T<G t(26)=-1.8*
Typed	T<G t(25.4)=-6.7		T<G t(26)=-4.9
Rec Located	T<G t(12)=-6.8	T>G t(24.4)=3	
Typed	T<G t(12)=-11.1		T<G t(17.3)=-7.3

Table 1: Overall Condition Comparison.

1 Unless otherwise stated all test values are from Welch's Two-sample 1-sided t-test.

2 T<G means that the text students scored below the graph students.

	[Table 2] Burnham.			[Table3] Burger King.		
	Core	Test	All	Core	Test	All
Eff L.	T<G t(25.6)=-4.7	T>G t(20.5)=4.4		T<G t(26)=-5.1		*T<G t(25.3)=-1.8*
T.	T<G t(23.5)=-6.7		T<G t(19)=-5.2	T<G t(24.3)=-7.5		T<G t(25.8)=-4.8
Prec L.	T<G t(18.5)=-5.9	T>G t(25.2)=2.3	T<G t(23.7)=-5	T<G t(17.6)=-7.6	T<G t(18.6)=-2.4	T<G t(18)=-5.1
T.	T<G t(18)=-7.1	*T<G t(22)=-1.6*	T<G t(21.5)=-7.8	T<G t(16.8)=-9.1	T<G t(20.8)=-3.2	T<G t(18)=-6.4
Rec L.	T<G t(12)=-5.9	T>G t(20.5)=4		T<G t(12)=-5.7		T<G t(22.3)=-2.2
T.	T<G t(12)=-9.5		T<G t(15.6)=-6.3	T<G t(12)=-9.8		T<G t(21.1)=-6

Tables 2 & 3: Case by case comparison of condition.

Further analysis of within-condition variations revealed that neither group performed better overall on either case. While both groups were more efficient and precise on Burnham than Burger King on the Core set, this was not reliably the case for the Test set or the full sets. Interestingly both groups performed better on the later case with respect to the Test set (Tables 4 and 5).

When analyzing our study results we split the students into Low, Medium, and High groups based upon their LSAT scores [12]. In that analysis, the Low LSAT Graph students gained more than their Low Text counterparts while the Medium and High students showed no across the board distinctions. We further analyzed the overall variation between the groups with respect to the five measures. The Low groups showed no significant difference in terms of time-on-task and amount of work done while the High Text group both took significantly more time than their Graph counterparts overall ($p < 0.0069$) and on a case basis (Burnham $p < 0.001$, Burger King $p < 0.01$). This was also true for work (Overall $p < 0.01$; Burnham p0.001; Burger King $p < 0.006$).

	[Table 4] Text.			[Table 5] Graph.		
	Core	Test	All	Core	Test	All
Eff L.	B>G t(21.4)=2.1[3]	B<G t(22.8)=-2.2		B>G t(26.4)=3.2	B<G t(20.8)=-4.5	
T.		B < G t(22.2)=-3.7		B>G t(26.4)=3.2	B<G t(21.6)=-4.7	
Prec L.	B > G t(16.9)=3.5		B>G t(20.9)=2.1	B>G t(21)=3.4	B<G t(20.5)=-3.7	
T.	B > G t(16.3)=2.7	B<G t(18.7)=-2.4		B>G t(21)=3.4	B<G t(20.5)=-3.7	
Rec L.						
T.						

Tables 4 & 5: Between case comparison for the Text and Graph Conditions.

Comparisons between the Low groups on the three success measures (Table 6) closely parallels the overall breakdown between the groups. The High students showed more consistent variation (Table 7) in favor of the Graph condition with the graph students outperforming their text counterparts across the board on the Core set and having higher type performance on all three measures. They did not, however, display the same variation on the Test set. There the only variation was the text students' higher recall of locations but not types.

	[Table 6] Low LSAT.			[Table 7] High LSAT.		
	Core	Test	All	Core	Test	All
Eff. Fnd.	T<G t(7.8)=-3.4	T>G t(7)=2.6		T<G t(5.3)=-9.7		T<G t(5.8)=-3.3
Typ.	T<G t(7.6)=-6.6		T<G t(7.5)=-4.9	T<G t(5.2)=-11.7		T<G t(4.6)=-8.1
Prec. Fnd.	*T<G t(7.6)=-1.6*	*T>G t(7.9)=2*		*T<G t(5)=-17.6*		T<G t(2.24)=-5.4
Typ.	T<G t(6.7)=-3.5		T<G t(7.7)=-2.6	T<G t(5)=-16.3		T<G t(4)=-7.6
Rec. Fnd.	T<G t(4)=-3.6			T<G t(5)=-7.3	*T>G t(2.1)=2.9*	
Typ.	T<G t(4)=-9.8		T<G t(5.6)=-5.3	T<G t(5)=-9.8		T<G t(2.4)=-6.8

Tables 6 & 7: Cross-condition comp for the Low and High LSAT students.

3 For this and Table 4 B<G means that the measure was higher for Burger King than Burnham.

Analysis of the help usage revealed no significant variation in help usage from case to case. Nor was there any significant difference in help usage between either the Low or High students. Indeed the only notable variation detected was in the amount of help selection between the High group and the remaining students (p 0.03). That is, the High Graph students clicked on the help button as often as their peers but followed up on that by selecting one of the choices less often.

Discussion

The lack of clear overall differences between the conditions in terms of time on task indicates that the graphical tools imposed no additional cognitive load. Despite every law student's unfamiliarity with graphical representations they took no more time to utilize the graphical tools. If the tools were overwhelmingly complicated, we would expect some students, especially in the low group, to perform worse and this was not the case. Similarly the equality of work performed suggests that the students were, for better or worse, expending the same amount of effort in either condition. Thus any gains attributable to the system are due not to load reductions but better use of time and effort.

This hypothesis is supported by the success measures. Our original hypotheses that the LARGO condition would dominate in the success measures held in part. When measured both overall and case-by-case, the LARGO condition was clearly dominant on the Core set. This was true both in terms of location-relevance and the higher type-relevance standard. This suggests that the advice was effective though it did not explicitly state the missing tests or hypotheticals only a region of interest.

This dominance did not hold when measuring against the Test set, consisting of elements that LARGO did not point students to. There the Text students were dominant with respect to location-relevance and the two conditions were equal in terms of type-relevance. This was true both overall and for Burnham save for the LARGO condition's increased type precision. This reversal was not present on Burger King where the two groups were equal in terms of Efficiency and Recall and the LARGO condition was dominant in terms of Precision.

In our opinion this can be explained by three related factors. Firstly we believe that the students within the graph condition may initially have engaged in some form of help-dependence and relied overmuch on the system to point out all essential components on Burnham and less so on Burger King. This would explain their clear success on the Core and mixed success on the Test sets. Further analysis will be necessary to confirm this.

Secondly the distinction between the Core and Test sets was not random as is the case in most ML applications. The Core elements were arguably more important to the dialogue than those of the Test set. It is possible that the students, making the same subjective assessment, focused more effort on the Core components. Moreover, the two sets were unequally distributed in terms of Tests and Hypotheticals with the bulk of the tests located in the Core set. Thus measurement of overall improvement was confounded somewhat with set measurement.

Thirdly, in examining the text students' notes we observed that a high proportion of their focus was on "action" or "concept" notes rather than relations. Thus while they were informed of the value of distinctions and other relationships they took few notes about them and focused instead on identifying relevant tests, hypotheticals and legal concepts. While we have not yet fully coded the notes we believe that the text students spent more time "making dots" rather than connecting them. Thus they have a higher proportion of candidate tests and hypotheticals to actual tests and hypos than their LARGO counterparts. This gave them an initial boost when it came to location-relevance but not type-relevance. By Burger King this gap had been removed or even reversed.

This hypothesis is somewhat complicated by the between case comparison for the conditions. Both conditions performed equally or better on Burnham with respect to the Core set and equally or better on Burger King with respect to the test set. Clearly in both cases the students were gaining familiarity with the model and were more willing to move beyond the most "central" elements. While this might be taken to suggest that the conditions learned equally, the lack of improvement in location-relevant precision by the text students as compared to the graph students and the overall dominance of the LARGO students on Burger King suggests otherwise. More data is required.

We believe that the results are consistent with our low-LSAT versus high-LSAT post-test results. While the low students clearly followed the same overall pattern of the group, the high students did not. There the LARGO condition was dominant on the Core and full sets and equal (save for location-relevant recall) on the Test set. This in spite of the fact that in this case the high text students both performed more work and spent more time than their graph counterparts. We believe that this demonstrates effective use of the system by the high LARGO students and a wiser

recognition of where to focus their efforts. In our opinion the post-test results suffered from somewhat of a ceiling effect thus washing out any apparent variation save between the low students. We further note that our "low" LSAT scores are in fact in the middle or upper middle segment of the average law school population. As such they are not representative of what truly "low LSAT" students might do. At present we are planning to conduct such a test this summer.

Conclusions.

The results that we noted above are positive and support our position that LARGO is beneficial for the students. The gains observed in the study can be partially explained by the findings presented in this paper – by means of reflective prompts, LARGO is able to help the students focus their attention on the important elements of the argument and helps them better to recognize tests and hypotheticals. Since thinking in terms of tests and hypotheticals is central to the argument model, this would mean that they had an important foundation in place on which to build further understanding of the model. Far from requiring a great deal of initial ramp-up, the students were able to adapt to the tools fairly quickly and showed improvements contrary to some expectations. While some potential system improvements were suggested by this study the system was largely successful.

As we noted above there were several complicating factors in this study that we expect to address in future studies. We will be conducting a study with genuinely 'low LSAT' students this summer. We plan to retool our post-test, provide more study cases, and to strive for a more appropriate distinction between the test and train sets. The last is of course the most difficult to address. Unlike individual utterances in a text-to-speech scenario, the tests and hypotheticals in our cases are not entirely independent. They often refer to one another, are modified, reappear and so on. Some may genuinely be considered more or less important than others although debate rages over which. Additionally some cases favor a large proportion of hypotheticals to tests, and some only a few or even none. Thus we will not be able to achieve a truly random split.

The fact that this method works in an ill-defined domain is intriguing. In such domains the challenge for an ITS is often to balance between providing too much structure and too little. While researchers agree that this problem exists there is little agreement on where the sweet spot may be found. One mechanism commonly used to address such a question is the fading of help from explicit to general hints, to no hints over time. Our system by contrast provides solely general information which proved useful to low, medium and high students without undue constraint. We plan further investigations along these lines as part of our subsequent work.

References

[1] Anderson, J.; Corbett, A.; Koedinger, K.; Pelletier, R.: "Cognitive tutors: Lessons learned." *The Journal of Learning Sciences* 4 (1995) 167-207.
[2] Aleven, V. (2003). Using Background Knowledge in Case-Based Legal Reasoning: A Computational Model and an Intelligent Learning Environment. Artificial Intelligence 150, 183-237.
[3] Ashley, K.D. (1990) *Modeling Legal Argument: Reasoning with Cases and Hypotheticals.* MIT Press. Cambridge.
[4] Ashley, K.D. (2006). "Hypothesis Formation and Testing in Legal Argument." Invited paper. *Inst. de Investig. Jurídicas 2d Intl Meet. on AI and Law*, UNAM, Mexico City. April
[5] Ashley, K.D. (2007). "Interpretive Reasoning with Hypothetical Cases." In *Proc. 20th Int'l FLAIRS Conference, Special Track on Case-Based Reasoning,* Key West, May.
[6] Ashley, K.D., Desai, R. and Levine, J.M. (2002) "Teaching Case-Based Argumentation Concepts using Dialectic Arguments vs. Didactic Explanations". In Proc., ITS '02 (S.A. Cerri, G. Gouardères, F. Paraguaçu, ed.) pp. 585-595. Springer: Berlin.
[7] Carr, C. (2003). "Using Computer Supported Argument Visualization to Teach Legal Argumentation." In *Visualizing Argumentation*, 75-96. London, Springer.
[8] Chi, M. (2000) "Self-explaining expository texts: The dual process of generating inferences and repairing mental models." In *Advances in Instructional Psychology.* Glaser R. (Ed.) Lawrence Erlbaum
[9] Llewellyn, K.N. (1960) *The Bramble Bush; on Our Law and its Study.* Oceana Publications. New York.
[10] Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). "Defining Ill-Defined Domains; A literature survey." In *Proc. of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.), NCU, Jhongli, Taiwan. 1-10.
[11] Mitchell Tom M. (1997). *Machine Learning*, McGraw Hill, Singapore.
[12] Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2007): "Evaluating Legal Argument Instruction with Graphical Representations using LARGO." To appear in *Proceedings of AIED 2007.*
[13] Robinson et al. (2006) "Increasing Text Comprehension and Graphical Note Taking using a Partial Graphical Organizer." *Journal of Educational research* 100(2) 103-111.
[14] Schworm, S.; Renkl, A. (2002) "Learning by solved example problems: Instructional explanations reduce self-explanation activity." In *Proc. of the Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum, Mahwah, 816-821
[15] Van Gelder, T. (2002) "Argument Mapping with Reason!Able." *The American Philosophical Assoc. Newsletter on Philosophy and Computers.* 85-90.

Resolving Ambiguity in German Adjectives

Amanda NICHOLAS and Brent MARTIN
Intelligent Computer Tutoring Group (ICTG),
University of Canterbury, Christchurch New Zealand.

Abstract. One problem in ill-defined domains is accurately identifying the source of errors. Obtaining sufficient information about the error can be difficult because doing so may interfere with the learning task.

In this paper we present the results of an experiment in the domain of German adjectives. We trialed a modified student interface that gathers more data during problem solving by requiring the student to perform a related subtask. There is evidence that the students who performed the subtask outperformed the control group on a post-test despite the extra task slowing them down, suggesting the extra effort required by the students to overcome ambiguity was worth the intervention.

Keywords. Student Modeling, Language learning, Ambiguity

Introduction

Dealing with ambiguity is a serious problem in developing Intelligent Tutoring Systems for foreign languages [1]. Natural language processing has not yet reached the point where we can process an unconstrained statement made by a student and accurately identify the source of any errors [2]. By constraining the scope of statements made by the student, it is possible to mark an answer as correct or incorrect. However, although the system can detect that the student has made an error, the source of this error may be difficult to determine. Menzel defines four sources of ambiguity: limited observability, polysemy, alternative conceptualizations of domain knowledge and structural uncertainty. In a domain with high ambiguity, feedback messages can be difficult to determine. Good feedback should refer the student to the underlying domain principle [3]. If it is not possible to determine which domain principle has been broken, correctly targeted feedback cannot be given.

One approach to avoid ambiguity is to require the student to specify the intermediate steps they carry out mentally. This approach is not popular; “such an interaction renders the exercise somehow unnatural.” [1]. Requiring the student to specify intermediate steps also raises the issue of transference [4]. When developing an ITS, the interface is generally designed to stay as close to the real world as possible, in order to ensure that the skills learnt on the computer will transfer to the real situation. By requiring the student to specify additional information the transference of skills may be weakened. This research compares two constraint-based (CBM) tutors; one that matches the real world more closely, and one that decreases ambiguity. Two Intelligent Tutoring Systems were

developed for the domain of German adjective endings, a domain where errors have a high level of ambiguity. An error in an adjective ending could be caused by a number of factors: mistaking the case, mistaking the gender, mistaking the article, or simply not knowing the correct ending in this situation. The experimental tutor required the student to specify the gender of the noun, the case of the noun, and the type of article. These three factors specify exactly the ending required. By considering these factors, the tutor could determine if it was the ending of the adjective that the student had specified incorrectly, or if they had some misconceptions about the sentence. The tutoring system could then provide targeted feedback based on the student's misconceptions. In contrast, the control only asked the student to specify the adjective ending. This matches what is required in the real world, but means that the tutoring system has to make assumptions about the error the student has made.

In the next section we further describe the problem of ambiguity in the domain of German adjectives, and constraint-based modeling (CBM) is summarized in Section 2. Sections 3 and 4 describe the experiment and present the results. Finally, we conclude in Section 5.

1. Ambiguity in German Adjectives

Adjective endings are a difficult topic for students to master. This is due to the number of endings that must be memorized, and the amount of knowledge required of the sentence to get the ending correct. Rogers studied the main areas of weakness in students with more than four years of experience learning German [5]. She states "...much anecdotal 'evidence' from teachers of German as a foreign language emphasizes morphology as a major areas of weakness (e.g. adjective endings...)". Her study showed that approximately 5% of errors made by advanced learners of German were errors in adjective endings. The number one error was in selecting gender, which could also affect the choice of adjective ending. Each error was only classified once, so if the student mistook the gender, it would not also appear as a mistaken adjective ending. The number of errors in adjective endings is therefore likely to be much higher than 5% when all reasons are considered. Further, Juozulynas studied students with two years of experience learning German and found that "The biggest problem in the students' writing seems to be syntax ...inflectional morphology with its much-feared endings takes second place. Syntax and morphology together make up 53% of the errors in the corpus." [6] Note that adjective endings are contained in inflectional morphology.

Case	Masculine	Feminine	Neuter	Plural
Nominative	-e	-e	-e	-en
Accusative	-en	-e	-e	-en
Genitive	-en	-en	-en	-en
Dative	-en	-en	-en	-en

Table 1. Adjective ending when preceded by the definite article

In German, adjectives must agree with the nouns they modify. This means that the ending of an adjective varies based on the gender and the case of the noun, and whether

the noun is preceded by the definite article, indefinite article, or no article. Table 1 lists the endings for the case where an adjective is preceded by the definite article. For example, take the sentence “Das graue Haus ist neu”. (The gray house is new). Here “Haus” is the noun, and its gender is neuter. The house is the subject of the sentence, and so it is in the nominative case. The article is “das”, and it is the direct article. The adjective is “grau”, and it takes the ending “e” because, by consulting Table 1, we see that adjectives preceding a neuter noun in the nominative case must end in “e”. If we change only the article in this sentence, so that it now read “Ein graues Haus ist neu”. (A gray house is new), the ending on the adjective changes also, from “e” to “es”. It is important to note that the endings are not unique; the ending “e” appears in a number of situations, as does “en”. This is one reason why these endings are ambiguous.

Menzel identified four major sources of ambiguity that should be considered when creating CBM tutors, particularly for foreign languages [1]. These are: a limited observability of internal variables of the problem domain; polysemy of symbols used in the problem domain (symbols with multiple meanings); alternative conceptualizations of domain knowledge; uncertainty about the intended structure of the students solution. He further suggests that because of this constraints alone are not sufficient to provide enough information to respond to students appropriately. German adjective endings suffer from three of the four defined sources of ambiguity. Limited observability and polysemy are both present in the multiple possible meanings of a single ending. For example, a student could correctly give an adjective requiring a nominative, masculine, definite article ending, the ending “e”. However, it is also possible that the student thought that the adjective required a nominative, feminine, definite article ending, for which the ending is also “e”. The student might even believe that the adjective requires a dative, masculine, definite article ending, which should be “en”, and might have given it the ending “e” incorrectly, based on their (incorrect) knowledge. Without awareness of the student’s thought processes, the tutor is unable to determine if the student has answered the question correctly on purpose, or by mistake. This problem also encompasses that of alternative conceptualizations of domain knowledge. When the student incorrectly gives an adjective ending, it could be due to either a rule error or a fact error. If the student does not know the gender or the case of the noun, they have made a fact error. If the student has correctly determined the case, gender and article, and still gives the adjective ending incorrectly, they have made a rule error; they do not know the underlying grammatical principle that determines the adjective ending. It is also possible for a student to make a rule error and a fact error simultaneously.

2. Constraint-Based Modeling and German Adjectives

CBM[7] is a relatively new approach to domain and student modeling, based on the theory of learning from performance errors [8]. It models the domain as a set of state constraints, where each constraint represents a declarative concept that must be learned and internalized before the student can achieve mastery of the domain. Constraints represent restrictions on solution states, and take the form:

IF <relevance condition> is true for the student’s solution,
THEN <satisfaction condition> must also be true

The relevance condition of each constraint checks whether the student's solution is in a pedagogically significant state. If so, the satisfaction condition is checked. If it succeeds, no action is taken; if it fails, the student has made a mistake, and appropriate feedback is given.

The student model consists of the set of constraints, and information about whether or not each constraint has been successfully applied each time it is relevant. Thus the student model is a trace of the performance of each individual constraint over time. Constraints may be grouped together, giving the average performance of the constraint set as a whole over time, which can then be plotted as a learning curve [9,11].

3. Experiment Design

We hypothesized that forcing the students to supply information about their problem-solving process and providing feedback based on that information would enable the system to give them better instruction, and thus they would be better able to learn the domain. We tested this hypothesis by building two versions of an ITS for German adjectives, where the two systems differed in the interface used and the underlying domain/student model (constraints).

The tutors were developed using WETAS [10]. WETAS is a shell that can be quickly adapted to provide basic functionality for an ITS. It provides student modeling, student management, and other features. The developer must supplement this with the problem set, the necessary constraints and, if desired, an interface. The problem set comprised of 55 problems, which was identical for both tutors. Some were obtained from existing sources [12,13], however, most problems were written especially for this ITS. An example of one of the problems in the tutor is

"Die ? Blumen gefallen mir. (bunt)" (I like the colorful flowers)

The two tutors shared a very similar interface. In the center of the screen was an area for the student to answer the question. Below the problem, a selection box allowed the student to choose the desired feedback level, and a button to submit their answer for feedback. Feedback messages appeared at the bottom of the screen. The problem was displayed in the form of a sentence. A gap was left where the adjective should be, and the adjective to be inserted was given in brackets at the end of the sentence. This was a format the students were familiar with, because it had been used during class and quizzes.

Students using the experimental system were asked to fill in the gender and case of the noun, the article type, and the adjective with its ending. The possible answers for gender, case and article were all given in combo boxes. This ensured that there would not be problems with students referring to the same concept by a different name, or misspelling names. Below the combo boxes, there was a text field for the student to fill in the adjective. Students using the control were only asked to fill in the correct adjective and ending. A textbox for the student to fill in was placed in the correct location in the sentence.

Domain constraints were sourced from a number of German textbooks [13,14,15,16, 17], which contain advice on how students can remember the endings more easily. They typically explain a pattern in the endings, for example that every adjective after the direct article ends in either "e" or "en" (see Table 1). The resulting constraints can be divided

into three groups. The first set of constraints is used for error checking, ensuring that the student has answered the question and used the appropriate adjective. The second set occurs only in the experimental tutor and checks whether the student has specified the gender, case and article correctly. The third set of constraints is the group that checks the validity of the adjective ending.

The experimental tutor has 33 constraints. Six are for error checking; ten are for checking that the student has specified the case, gender and article correctly; the remaining constraints check the adjective ending. The adjective ending is checked for validity with respect to the case, gender and article the student has used; incorrect values for case, gender and article will trigger other feedback messages. In this manner, the system determines whether the student has made an error because they have inaccurate knowledge about the sentence, or because they do not know their adjective endings, i.e. whether they have made a fact error or a rule error.

The control tutor had twelve constraints. Three were for error checking, and the remaining nine checked the ending the adjective has been given. Because the only information available to the tutor is the ending the student has given the adjective, the tutor provides feedback relative to the correct gender, case and article. It is assumed that the student knows this information, but may be unaware of the ending that matches correctly. This means that the tutor considers all mistakes to be rule errors, not fact errors. An example of one such constraint is:

```
(10
; FEEDBACK
"When they are preceded by a 'der-word', all adjectives end
in either -e or -en."
; RELEVANCE CONDITION
(and
(match IS ARTICLE ("D"))
(match SS ANSWER (?something ?*)))
; SATISFACTION CONDITION
(or-p
(match SS ANSWER (?*w2 "e" "n"))
(match SS ANSWER (?*w1 "e")))
"ANSWER")
```

The relevance clause of this constraint checks that the sentence contains a definite article (“D”), and that the student has answered the question. If this is true, the student’s answer must end in “e” or “en”, as all adjectives end in “e” or “en” after the definite article. If the student’s answer does not end in “e” or “en”, the system assumes that they have forgotten the rule, not that they have not realized that the sentence contains a definite article.

An evaluation study of the two tutors was conducted on the 6th of September 2006 at the University of Canterbury, Christchurch. Students enrolled in GRMN115, a beginning German course, used one of the two systems over one class period. The students had been taught adjective endings previously in class, however there was a two-week holiday period between when the topic was taught and when the study was carried out. The class was divided into two even groups. This was done alphabetically by last name. The evaluation took place during one lecture period, a time span of 50 minutes. The students

were first asked to complete a pre-test. They then used the tutoring system for as long as time permitted, or until they finished all 55 questions. Afterwards they completed a post-test. Each test contained six questions. All questions contained sentences of the form:

"Die ? Jacke ist preiswert. (gelb)" (The yellow jacket is affordable)

The student was expected to transfer the adjective (here 'gelb') into the gap in the sentence, and give it the appropriate ending. The final three questions also asked the student to specify the gender and case of the noun present in the sentence, and the type of article preceding the noun. The experiment was carried out in two streams. The control and experimental tutor were used by students from both streams. To allow for any difference in the difficulty of the pre- and post-tests, Test 1 was used as the pre-test for Stream A, and the post-test for Stream B; Test 2 was used as the post-test for Stream A and pre-test for Stream B.

4. Results

23 students took part in the evaluation. 12 students used the experimental tutor and 11 students used the control. Statistics about the system usage can be seen in Table 2. We can see that students using the control system solved more problems with fewer attempts than those using the experimental tutor. This result is unsurprising, because students using the control only had fill in one answer correctly, whereas students using the experimental tutor had to fill in answer values. Students using the experimental tutor also saw more messages. This is also unsurprising; their task was larger so there were more opportunities to make mistakes.

Measure	Control	Experiment
Attempted Problems	52	22
Solved Problems	49	21
Attempts per Problem	2.0	4.0
Seen Messages per Problem	1.5	5.0

Table 2. System usage statistics

Unfortunately, the pre- and post-test were not of comparable difficulty. Over all students, irrespective of which tutor the student used or whether the test was taken as a pre- or post-test, the average score for Test 1 was 83%, and the average score for Test 2 was 65%. This means that the scores for the pre-and post-test are not directly comparable. The reason for the difference in difficulty is that Test 2 contained two questions where the gender of the noun could not be unambiguously determined from the rest of the sentence; the student either knew the gender of the word or they did not. To overcome this, we compared the results for Test 1 only, and compared the outcome for pre- and post-test regardless of which stream the students belonged to. This is not strictly valid because the samples are different; it relies on the assumption that the students in the two streams (and using the same tutor) were comparable, and this cannot be easily measured. Using this assumption, a t-test of the score for producing the correct adjective ending showed

no significant difference between the Test 1 pre-test scores for the two tutors (mean = 4.8 and 4.6 for the control and experimental groups respectively, SD = 0.8 and 1.6, $p > 0.7$). When Test 1 was used as a post-test however, there was a larger difference between the two groups, with the experimental tutor achieving a score of 5.7 compared to 5.0 for the experimental group, although the result is not statistically significant ($p > 0.15$).

We also compared the performance of the two groups in terms of their ability to perform the subtask (determine case and gender). Again there was no significant difference on pre-test score between the control and experimental groups (5.0 versus 4.9). For the post-test, the experimental group again outperformed the control group, scoring an average of 5.7 compared to 4.8 for the control group. The result was statistically significant ($p < 0.05$).

Another method of comparing student performance is via learning curves [9,11]. If the units being measured are being learned by the students, we expect to see a "power law of practice". Learning curves therefore give an indication of the relative performance of samples of students and the quality of the model. Fig. 1 shows the learning curves for the two groups for just those constraints that test for the correct adjective endings (Tutor1 is the experimental group, Tutor2 is the control). The power law fit for the curves for both groups is only average, although it is better for the experimental group (R^2 = 0.63 versus 0.45). Fig. 2 shows the corresponding learning curve for the only those constraints that test the subtask. This latter curve is for the experimental group only since these constraints did not exist for the control group. In this case we see a slightly better power law (R^2 = 0.71), suggesting that the constraints form a fairly good model of what is actually being learned.

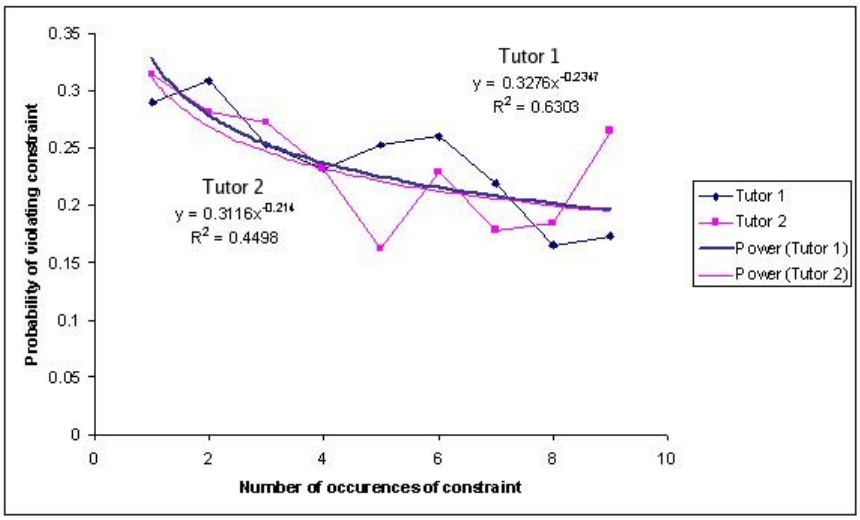


Figure 1. Learning curves for the main task (shared) constraints

From both the learning curves and the pre-/post-tests, there is evidence that the experimental group learned the task of choosing the correct adjective endings better than the control group in the time available. This is despite the fact that the experimental group were slowed by the need to perform the subtask, and so they completed far fewer prob-

lems. This strongly suggests that the additional effort required to perform the subtask was worth it because it allowed the feedback given to better target the current misconception.

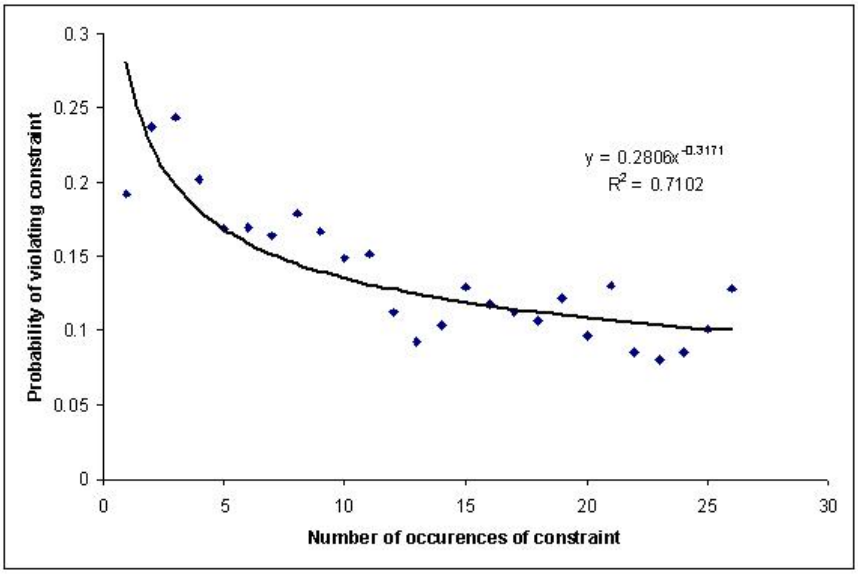


Figure 2. Learning curve for the subtask constraints (experimental tutor)

An alternative explanation is that the subtask itself proved to be useful for learning (and modeling) the main task. Recall that the constraints that were common to both tutors (and the only ones for the control) assumed that the student had made a rule error, i.e. that they knew the gender, case and article, but selected the wrong ending for that situation. (For the experimental group the constraints were subtly different in that they compared the adjective ending to the student's answer for the gender, case and article, i.e. they definitively determined that the problem was a rule error). However, selecting the correct case for an adjective in a sentence requires both that the correct ending be supplied for the situation, and that the situation be correctly interpreted in the first place. The constraints in the control therefore only represents part of the model for this domain, while the model for the experimental tutor is more complete.

Finally, the students were asked to fill in a subjective survey at the end of the study. Responses from were overwhelmingly positive to both versions of the tutor. Comments included "It was good that the mistakes were explained + the grammer rules were also explained." "I liked it and found very useful". Further, the staff from the German department indicated they would like to pursue this technology further, because the students had reacted so positively. They also commented that the results for the formal adjectives test were considerably higher than in previous years, which they attributed to the tutoring systems.

5. Conclusions

Tutoring systems that teach natural languages are susceptible to the problem of ambiguity in student answers, making it difficult to apportion blame appropriately, and thus provide effective feedback. Even a highly constrained domain such as German adjectives exhibits this problem. Requiring the student to supply additional information is often frowned upon because it reduces the correspondence to "real world" problems and may thus negatively affect transfer.

We examined this problem in the domain of German adjectives by providing two versions of a simple ITS; the control required the students to complete the original task only (and thus suffered from ambiguity) while the experimental group forced them to also complete a subtask that disambiguated their response. The results were not conclusive because of problems with the pre- and post-test difficulties. However, there was evidence from these tests that the experimental group performed better on both the original task and the subtask despite having solved considerably fewer problems because of the additional time needed to complete the subtask. This suggests that far from detracting from the students' ability to complete the main task, the extra disambiguation benefited their learning.

Several questions remain unanswered. First, this study was conducted for a highly constrained problem domain; further investigation is needed on more open-ended domains. The German department at the University of Canterbury has indicated that they would like to pursue the technology further, so it is likely we will conduct further studies for other parts of the German curriculum in 2007. Second, the study made several assumptions that require further exploration. In particular, the assumption that students in the control group always make rule errors (i.e. they know the situation but choose the wrong ending) is highly likely to be invalid; if this were the case, we would expect the students in the control group to perform the subtask flawlessly during the pre- and post-tests, which they clearly did not. In fact, the reverse assumption (that mistakes are caused by misinterpreting the situation) has greater supporting evidence since the learning curve for the associated constraints was stronger. One way to test this assumption might be to have the same constraints, but alter the feedback; instead of telling the student how to work out the ending for a particular set of situations, it could indicate the situations for which the ending they have chosen is correct. This warrants further investigation. Finally, the experimental tutor gave feedback for rule errors if the student submitted an ending that contradicted their supplied case, article and gender combination, even if the ending was correct for the problem. In general it is unclear whether or not feedback about the ending should be provided at all if the subtask has not been completed.

This study has shown that adding extra task requirements to overcome ambiguity in language learning is not always a bad thing, and can in fact be advantageous. This is a positive outcome that encourages us to further explore how constraint-based models may support language learning.

References

[1] Menzel, W., Constraint-based modeling and ambiguity. International Journal of Artificial Intelligence in Education, 2006. 16(1): p. 29-63.

[2] Menzel, W. and Schroeder, I., Constraint-based Diagnosis for Intelligent Language Tutoring Systems. Proceedings IT&KNOWS, IFIP World Congress. 1998. p.484-497.
[3] Zakharov, K., Mitrovic, A., and Ohlsson, S. Feedback Micro-engineering in EER-Tutor. in Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005. Amsterdam: IOS Press.
[4] Anderson, J.R., Corbett, A.T., Koedinger, K.R., and Pelletier, R., Cognitive Tutors: Lessons Learned. Journal of the Learning Sciences, 1995. 4(2): p. 167-207.
[5] Rogers, M., On major types of written error in advanced students of german. International Review of Applied Linguistics in Language Teaching, 1984. 22(1): p. 1-39.
[6] Juozulynas, V., Errors in the compositions of second-year german students: an empirical study for parser-based icali. CALICO Journal, 1994. 12(1): p. 5-17.
[7] Ohlsson, S., Constraint-Based Student Modeling, in Student Modeling: The Key to Individualized Knowledge-Based Instruction, J. Greer and G. McCalla, Editors. 1994, Springer-Verlag: New York. p. 167-189.
[8] Ohlsson, S., Learning from Performance Errors. Psychological Review, 1996. 3(2): p. 241-262.
[9] Newell, A. and Rosenbloom, P.S., Mechanisms of skill acquisition and the law of practice, in Cognitive skills and their acquisition, J.R. Anderson, Editor. 1981, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 1-56.
[10] Martin, B. and Mitrovic, A. Authoring web-based tutoring systems with WETAS. in International conference on computers in education. 2002. Auckland.
[11] Martin, B., Koedinger, K.R., Mitrovic, A., and Mathan, S. On Using Learning Curves to Evaluate ITS. in Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005. Amsterdam: IOS Press.
[12] Werner, G., Langenscheidts Grammatik-training Deutsch. 2001: Langenscheidt KG.
[13] Kahlen, L., Interactive German Made Easy. 2006: McGraw-Hill.
[14] Terell, T.D., Tschirner, E., and Nikolai, B., Kontakte: a comunicative approach. 2004: McGraw-Hill.
[15] Webster, P., Schwarz, rot, gold: the German handbook: a practical grammar guide. 1987: Cambridge University Press.
[16] Sparks, K. and Vail, V.H., German in review. 1986: Harcourt Brace Jovanovich.
[17] Dreyer, H. and Schmitt, R., A practice grammar of German. 2001: Max Hueber Verlag.