Granularity Judgments in Proof Tutoring*

Marvin Schiller and Christoph Benzmüller

Dept. of Computer Science, Saarland University, Germany
{schiller, chris}@ags.uni-sb.de

1 Introduction

The SFB 378 project DIALOG [2] investigates natural tutorial dialog between a student and an assistance system for mathematics (ASM). The aim is to mechanize the tutoring of mathematical proofs. At the present time, the (simplified) approach in the DIALOG system is:

(1) The student inputs a proof via the user interface using a mixture of natural language and formulas. The analysis of this input and its conversion into a formalized representation suitable for an ASM is a challenge for Computational Linguistics.

(2) Theorem proving techniques are employed to support the tutoring of mathematical proofs. In particular, the ASM analyzes the student's formalized proof step and makes judgments on its *soundness*, its *granularity* (i.e., argumentative complexity), and its *relevance* for solving the proof problem.

(3) The analysis results of step (2) are passed on to a tutoring module which determines appropriate feedback to be presented to the student via the DIALOG system user interface.

Here we focus on the aspect of *granularity* in mathematical dialogs, which refers to the size of a proof step w.r.t. its argumentative complexity. It is well known that when proving a theorem, the same line of reasoning can be expressed in different step sizes. Compare for example the two excerpts[1] from the proof for $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ – note that (2) omits an intermediate argument:

(1) Let x be an element of $A \cap (B \cup C)$, then $x \in A$ and $x \in B \cup C$. This means that $x \in A$, and either $x \in B$ or $x \in C$. [...]

(2) Let x be an element of $A \cap (B \cup C)$. This means that $x \in A$, and either $x \in B$ or $x \in C$. [...]

Does granularity play a role in tutorial dialogs on proofs? In order to answer this question, we have collected a corpus of tutorial dialogs on mathematics in a Wizard-of-Oz experiment with four experienced human tutors (cf. [3]). The corpus indicates that granularity was indeed a relevant aspect for the tutors in

* This work has been funded by the DFG Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378 (http://www.coli.uni-saarland.de/projects/sfb378/).

[1] Excerpt (1) is taken from R. Bartle and D. Sherbert, *Introduction to Real Analysis*, Wiley, 1982. Proof fragment (2) is a constructed example.

Table 1. Average number of calculus level proof steps that constitute a student's proof step for twenty steps from the study, grouped by their granularity level as identified by the tutors. Standard deviations in the "too detailed" group, which consists of only two student proof steps, are omitted. Zero length proof steps occur when the formalization of the analyzed statement is identical to a previous statement by the user.

Tutor's rating	Avg. calculus level proof length (std. deviation in brackets)			
	PSYCOP calculus		Gentzen's ND calculus	
"too detailed"	1,00		0	
"appropriate"	5,27	(4,88)	5,00	(5,14)
"too coarse-grained"	11,67	(6,80)	10,33	(7,72)

the role of the wizard: During each experiment session the experts annotated an average of 1.92 utterances from the students as to "too detailed" or "too coarse-grained" (out of on average of 25 dialog contributions the students made during a session). More details on the reported work can be found in [7]. Even though the phenomenon of granularity has been identified in the literature, related work such as [8] generally focuses on the correctness aspect of informal proofs.

2 Framework and Calculi for Granularity Analysis

We started with the simple and plausible hypothesis that the size of the formalized proof step, i.e. the sum of the nodes in the proof tree for this step, can serve as an indicator for the granularity of the student's argument. However, which target calculus should we choose? And would it not be more appropriate to employ weighted sums over the nodes to account for the fact that not all calculus level steps may have the same cognitive status? To account for these and similar options, we have developed a generic framework for granularity analysis, described in [7], which is parameterized over the target proof calculus and the particularly selected weighting.

What are good proof calculi for granularity analysis? It is plausible to watch out for human-oriented, cognitively adequate calculi which closely reflect the actual way humans proceed when proving mathematical theorems. Two prominent candidates are the traditional natural deduction (ND) calculus by Gentzen [4] and the empirically motivated PSYCOP (shorthand for "Psychology of Proof") calculus [6]. Before considering any further options, such as assertion level proofs [5], we decided to evaluate their suitability first.

Evaluation. We instantiated our framework with these two calculi and used equal weights for each rule. Both framework instances were then applied to the student proof steps from the corpus. The resulting proof size figures were related and compared to the granularity annotations the expert tutors had provided for each proof step during the experiment. The tutors had been asked to categorize the granularity of each student step as either *appropriate*, *too coarse-grained* or *too detailed*.

The results are shown in Table 1. Those proof steps that were *appropriate* in the eyes of the tutor require on average more steps in both framework instances (with only small differences) than the average *too detailed* proof step, and less calculus level steps than the average *too coarse-grained* proof step. This provides evidence for our hypothesis that the granularity of student proof steps is reflected by the size of the formalized proofs in ND or PSYCOP. However, as indicated by the standard deviations, the sizes of the calculus level proofs vary greatly within the *appropriate* and the *too coarse-grained* group. This overlap between the three classes thus allows no rigid distinction between them based on calculus level proof length alone. Hence, further investigation is required.

3 Discussion

In the experiments, the students often used proof steps which can be characterized as rewriting steps, or assertion level resp. deep inference steps. Such reasoning steps are not modeled appropriately in either the ND or the PSYCOP calculus. Ongoing work therefore is to build a third instance of our granularity analysis framework based on the (recently stabilizing implementation of the) OMEGA-CORE calculus [1], which aims at better supporting such reasoning steps. Furthermore, in order to obtain stronger evaluation results we plan another experiment that will be more specifically designed to support the investigation of granularity phenomena.

A particular phenomenon in our current corpus is that the granularity annotations by the four expert tutors reflect personal differences in the perception of granularity, which — in contrast to correctness — is a very subjective matter. Therefore, the coupling of the presented framework for granularity analysis with student and teaching models appears useful in order to better address the context sensitivity of granularity judgments.

References

1. S. Autexier. *Hierarchical Contextual Reasoning.* PhD thesis, Saarland University, Saarbrücken, Germany, 2003.
2. C. Benzmüller et al. Tutorial dialogs on mathematical proofs. In *Proc. of the IJCAI Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems*, pages 12–22, Acapulco, 2003.
3. C. Benzmüller et al. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In *Proc. of International Conference on Language Resources and Evaluation (LREC 2006).* ELDA, 2006. To appear.
4. G. Gentzen. Untersuchungen über das logische Schliessen. *Mathematische Zeitschrift*, 39:176–210, 405–431, 1934.
5. X. Huang. *Human Oriented Proof Presentation: A Reconstructive Approach.* Phd thesis, Universität des Saarlandes, Saarbrücken, Germany, 1994.
6. L. J. Rips. *The psychology of proof.* MIT Press, Cambridge, MA, 1994.
7. M. Schiller. Mechanizing Proof Step Evaluation for Mathematics Tutoring - the Case of Granularity. Master's thesis, Universität des Saarlandes, 2005.
8. C. Zinn. *Understanding Informal Mathematical Discourse.* PhD thesis, Universität Erlangen, 2004.