A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material.

Christoph Benzmüller[†], **Helmut Horacek**[†], **Henri Lesourd**[†], **Ivana Kruijff-Korbayova**[*], **Marvin Schiller**[†], **Magdalena Wolska**[*]

[†]Fachrichtung Informatik [*]Fachrichtung Allgemeine Linguistik
Universität des Saarlandes, Postfach 15 11 50, D-66041 Saarbrücken, Germany
{chris,horacek,henri,schiller}@ags.uni-sb.de, {korbay,magda}@coli.uni-sb.de

Abstract

We present a new corpus of tutorial dialogs on mathematical theorem proving that was collected in a Wizard-of-Oz setup. Our study is a follow up on a previous experiment conducted in a similar simulated environment. A major difference between the current and the previous experimental setup was that in this study we varied the presentation of the study-material with which the subjects were provided. One sub-group of the subjects was presented with a highly formalized presentation consisting mainly of formulas, while the other with a presentation mainly in natural language. Our goal was to obtain more data on the kind of mixed-language that is characteristic of informal mathematical discourse. We hypothesized that the language style of the subjects' interaction with the simulated system will reflect the style of presentation of the study-material. In the paper we briefly present the experimental setup, the corpus, and a preliminary quantitative results of the corpus analysis.

1. Introduction

In the DIALOG[1] project (Benzmüller et al., 2003a), we are investigating and modeling semantic and pragmatic phenomena in student-tutor dialogs on problem solving skills in mathematics. Our goal is to empirically investigate the use of flexible natural language dialog in tutoring mathematics, and to develop a prototype tutoring system gradually embodying the empirical findings.

In (Wolska et al., 2004), we presented an annotated corpus of tutorial dialogs on theorem proving collected in a Wizard-of-Oz setup (Benzmüller et al., 2003b) in which subjects interact with a system simulated by a human "wizard" (Fraser and Gilbert, 1991; Dahlbäck et al., 1993; Maulsby et al., 1993). As noted in (Zinn, 2003; Wolska and Kruijff-Korbayová, 2004; Horacek and Wolska, 2005), the prominent property of the language of mathematical texts is that it consists of interleaved natural and (semi-) formal language: conventionalized mathematical expressions. Following up on the previous study, we present a new corpus of tutorial dialogs on mathematical theorem proving collected in a recently completed experiment conducted in a similar simulated environment. Our goal was to obtain more data on the kind of mixed-language interaction as discussed in (Horacek and Wolska, 2005).

The previous study dealt with the domain of naive set theory. For the second experiment, we chose the domain of binary relations. The reason for the choice of a different domain was, among others, to facilitate future investigation the scalability of our input interpretation component. A major difference in the experimental setup was that in the recently completed follow-up experiment we varied the presentation of the study-material with which the subjects were provided: highly formalized presentation consisting mainly of formulas vs. presentation in natural language. Our aim was to elicit the mixed-language style of informal proofs. The hypothesis was that the language style of the subjects' interaction with the simulated system will reflect the style of presentation of the study-material. In the last section of this paper, we present a preliminary data analysis with respect to this question.

Below, we present the setup of the corpus collection experiment, in particular, pointing out differences with respect to the previous study, the corpus itself, and the preliminary quantitative results of a comparative study of the dialogs obtained in the two conditions.

2. The corpus collection experiment

We invited thirty seven subjects to participate in the experiment. The subjects were Saarland University students of different educational backgrounds. A prerequisite for participation was that the candidate subjects had taken at least one mathematics course at the university level. The subjects were informed that they were interacting with a conversational system with natural language capabilities.

We provided the subjects with background reading material for the domain of binary relations (see Section 2.1.) and allowed a study time before starting the tutoring session. Next, we asked the subjects to prove four theorems using the system (R, S, and T are binary relations on a set M):

- **W.** $(R \circ S)^{-1} = S^{-1} \circ R^{-1}$
- **A.** $(R \cup S) \circ T = (R \circ T) \cup (S \circ T)$
- **B.** $(R \cup S) \circ T = (T^{-1} \circ S^{-1})^{-1} \cup (T^{-1} \circ R^{-1})^{-1}$
- **C.** $(R \cup S) \circ S = (S \circ (S \cup S)^{-1})^{-1}$
- **E.** Assume R is asymmetric. If R is not empty (i.e. $R \neq \emptyset$, then $R \neq R^{-1}$)

Exercises **W.**, **A.**, **B.**, and **C.**[2] build upon each other in that

[1] The DIALOG project is a collaboration between the Computer Science and Computational Linguistics departments of University of the Saarland as part of the Collaborative Research Center on *Resource-Adaptive Cognitive Processes*, SFB 378 (http://www.coli.uni- sb.de/sfb378).

[2] Exercise **C.** is a theorem if S is a symmetric relation, but not in the general case. The expectation was that the subjects would be able to provide an argument for this.

Sind A, B Mengen und gilt

$$\forall x (x \in A \Rightarrow x \in B),$$

so heißt A eine *Teilmenge* von B. Man schreibt dafür $A \subseteq B$.

If A and B are sets and $\forall x (x \in A \Rightarrow x \in B)$ holds, then A is called a subset of B. We write $A \subseteq B$.

Sind A, B Mengen und gilt daß jedes Element von A auch Element von B ist, so heißt A eine *Teilmenge* von B. Man schreibt dafür $A \subseteq B$.

If A and B are sets and it is the case that every element of A is also an element of B, then A is called a subset of B. We write $A \subseteq B$.

Figure 1: Presentation of the Subset definition in the “formal” (left) and “verbose” (right) material.

Theorem
Sei R eine Relation in einer Menge M. Es gilt: $R = (R^{-1})^{-1}$
Proof
Eine Relation ist definiert als eine Menge von Paaren. Die obige Gleichheit ist demnach eine Gleichung zwischen zwei Mengen. Mengengleichungen kann man nach dem Prinzip der Extensionalitaet dadurch beweisen, dass man zeigt, das jedes Element der ersten Menge auch Element der zweiten Menge ist. Sei also (a, b) ein Paar in $M \times M$, dann ist zu zeigen $(a, b) \in R$ genau dann wenn $(a, b) \in (R^{-1})^{-1}$. $(a, b) \in (R^{-1})^{-1}$ gilt nach Definition der Umkehrrelation genau dann wenn $(b, a) \in R^{-1}$ und dies gilt nach erneuter Definition der Umkehrrelation genau dann wenn $(a, b) \in R$, was zu zeigen war.

Let R be a relation on a set M. Prove: $R = (R^{-1})^{-1}$
A relation is defined as a set of pairs. The above equation expresses an equality between sets. Set equality can be proven by The Principle of Extensionality, where one shows that every element of one set is also an element of the other set. Let (a, b) be a pair on $M \times M$. We have to show that $(a, b) \in R$ if and only if $(a, b) \in (R^{-1})^{-1}$. $(a, b) \in (R^{-1})^{-1}$ holds by definition of the inverse relation if and only if $(b, a) \in R^{-1}$ and this again holds by the definition of the inverse relation if and only if $(a, b) \in R$, which was to be proven.

Figure 2: Example proof.

once solved, they may be used in the subsequent exercises. Exercise **W.** was a warm-up exercise and exercise **E.** was presented only to those subjects who had difficulties in completing the initial exercise.

We instructed the subjects to enter proof steps, rather than complete proofs at once, to encourage dialog with the system, as well as to think aloud while solving the exercises. The language of the dialogs was German. The dialogs were typed using keyboard and/or buttons available on the user interface (see Section 2.2.). Both the wizards and the subjects were free in the way they formulated their turns.

2.1. The study material

The subjects were provided with preparatory material (adapted from (Bronstein and Semendjajew, 1991)) reviewing the notions, definitions, and basic theorems in binary relations, and were allowed 25 minutes to revise before starting the tutoring session. They were divided into two groups: a sub-group of twenty subjects was provided with a formalized presentation of the study material, while the other sub-group was presented with material that avoided formalization, and used natural language verbalization instead. In Figure 1, we show the definition of a subset as it appeared in both versions to illustrate the difference. The subjects were also provided with an example proof of a simple theorem, shown in Figure 2, that used a mixture of natural language and mathematical expressions.

2.2. The interface

The interaction between the subject and the wizard was mediated by a chat environment with a Graphical User Interface (GUI) that consisted of a customized version of the TeXmacs [3] editor; a LaTeX editor operating in the *what-you-see-is-what-you-get* mode. The advantage of using TeXmacs was that the subjects were given multiple alternatives for inserting mathematical expressions: traditional GUI buttons with symbols, original LaTeX commands (e.g. \cup), as well as German translations of the LaTeX commands (e.g. \Vereinigung for the set union).[4] Additionally, the GUI supported *copy and paste* functionality with which portions of text could be copied from the prior dialog. The area of the GUI that displayed the prior dialog

[3] http://www.texmacs.org/

[4] All the available commands were printed on a handout. Prior to starting the tutoring sessions, the subjects were instructed on using the GUI (in particular, shown the different input modes for formulas) by one of the experimenters, and had a few minutes time to familiarize themselves with the GUI in a typing exercise.

	No. turns	Student turns	Wizard turns
FM-group	974	474	463
VM-group	943	463	480
Total	1917	937	980

Table 1: Corpus overview

	no math	only math	mixed
FM-group	90 (19%)	194 (41%)	190 (40%)
VM-group	82 (18%)	69 (15%)	312 (67%)

Table 2: Student turns with respect to symbolic content.

was available in a *read-only* mode.

2.3. The tutoring

We invited four tutors to play the role of wizards in the experiment. The tutors' background with respect to teaching mathematical proofs was the following: tutor 1.: senior lecturer with several years of experience in lecturing a course *Foundations of Mathematics*, tutor 2.: trained mathematics teacher with a few years of teaching experience, tutor 3.: recent graduate with a degree in teaching mathematics, tutor 4.: doctoral student in Institute of Theoretical Mathematics at Saarland University with several years of experience as a Teaching Assistant in various mathematics courses.

The tutors were given general instruction on what constitutes *socratic* tutoring, but unlike in our previous experiment (Benzmüller et al., 2003b), they were not provided with any tutoring algorithm. The tutors were free in formulating their responses using natural language and/or formulas. They were asked to annotate the students' proof contributions with "answer categories"; evaluations of the contributions as to their correctness, relevance, and granularity.[5] They were also asked to record a spoken commentary on their responses where they considered it appropriate.

3. The corpus and quantitative analysis

The collected raw corpus consists of 37 sets of dialog session logfiles. During each session, the following material was collected:

1. dialog session logfiles (in the raw ascii format as well as a TeXmacs document);
2. think-aloud audio recording of the subject;
3. video of the subject interacting with the system;
4. subject's screen recording;
5. wizards's audio commentary.

Aside from time-stamp information and the text of the dialog contributions, the logfiles contain annotations of the answer category assigned by the wizards during tutoring. Moreover, they contain information on the mode in which mathematical symbols were inserted (menu button vs. English vs. German LaTeX command) recorded by the GUI. We are planning to use this information to look at the preferences in the ways of typing mathematical expressions to gain insights into designing better interfaces for the future Wizard-of-Oz studies as well as to guide us in building a GUI for the prototype system.

	min	max	mean	median	std
FM nl	1.00	219.00	**71.05**	46.50	67.41
FM math	14.00	50.00	27.65	25.00	11.11
FM f-len	1.00	145.00	**25.23**	17.00	26.42
VM nl	15.00	308.00	**118.71**	98.00	79.14
VM math	9.00	95.00	46.47	43.00	22.36
VM f-len	1.00	110.00	**10.86**	8.00	11.19

Table 3: Distribution of natural language vs. symbolic tokens in the student turns per session in the two conditions.

Corpus size Overall, the corpus consists of 1917 dialog turns (average of 51 turns per session), of which 980 are wizard and 937 student turns. Table 1 presents the overview of the corpus; the FM and VM sub-totals refer to the subgroups of subjects presented with the Formal and Verbalized material respectively.

Analysis of the logs We first looked at the difference between the number of student turns that contained only symbolic expressions and those that contained no symbolic expressions. Recognition of the symbolic expressions was performed semi-automatically by adapting the mathematical expression tagger developed for the previous corpus to the new domain. Table 2 presents an overview of the results. As mathematical expressions we counted occurrences of formulas, terms, as well as single character tokens intended to represent relation or set symbols. This cursory analysis shows that overall, students tended to use a mixture of symbolic and natural language. However, the group presented with formalized material has a larger number of turns (41% vs. 15% of all turns in the VM-group) consisting of symbolic material alone.

Second, we looked at the distribution of the symbolic vs. natural language content per dialog session. Overall, the average number of natural language tokens was 92.95 with 75.90 standard deviation (std) and the average number of symbolic expressions 36.30 (std 19.43). The average formula length[6] (f-len) in the dialogs was 16.09 (std 32.48).

Table 3 shows the same distribution with respect to the two conditions: minimum (min), maximum (max), mean, median, and standard deviation (std) of natural language tokens (nl), symbolic expressions (math), and length of symbolic expressions (f-len). While there was little difference between the VM- and the FM-group in the number of turns that contained natural language words alone (see Table 2), the average number of natural language words per session was higher in the VM-group. Also, the VM-group tended to use fewer and shorter formulas. The large maximal formula length in both conditions, we believe, might be an artifact of the interface's *copy and paste* mechanism.

[5] The annotation was inserted during the tutoring session, however, it was not visible on the subject's end of the interface.

[6] We counted all tokens intended to form a mathematical expression, including punctuation and single character tokens for variables and constants.

	min	max	mean	median	std
FM nl	49.00	354.00	173.85	158.50	87.53
FM math	0.00	80.00	11.65	4.00	19.24
VM nl	93.00	364.00	209.88	210.00	70.32
VM math	1.00	48.00	16.82	14.00	13.74

Table 4: Distribution of natural language vs. symbolic tokens in the wizard turns per session in the two conditions.

Discussion A preliminary analysis of the corpus data reveals differences in the use of natural language vs. formulas. One of the factors contributing to this difference may be the format of the presentation of the study material having a priming effect. However, other factors may include the wizard's style of interaction and the alignment effect, or the individual mathematical skills of the student. Table 4 shows the distribution of the natural language vs. symbolic tokens per session in the wizard turns. We do not show the formula length counts because the wizards sometimes copied even lengthy subjects' formulas into their turns, for instance, while asking clarification questions. Although there is little difference between the two conditions here, a difference may be in the specifics of styles of subject-wizard pairs and have to do with the mathematical skills of the student. We plan to investigate this in the future.

4. Related work

With respect to tutorial dialog corpora, a number of collections of tutorial dialogs have been analyzed, a selection of which can be found at the CIRCLE website.[7] (Tomko and Rosenfeld, 2004) study of the effect of instructions about the system's language capabilities given to participants of a Wizard-of-Oz experiment with a speech-based dialog system. The goal was to investigate how easily the users can be persuaded to use a restricted input style. The study finds that all uses were adapting their language to the system's language. (Dahlqvist et al., 1999) presents a study on presentational formats and implications on learning in the domain of physics. The effect of linguistic alignment in interactions with dialog systems has been widely studied (Ringle and Halstead-Nussloch, 1989; Zoltan-Ford, 1991; Brennan and Ohaeri, 1994).

5. Conclusion

The paper presents a new corpus of tutorial dialogs on mathematical problem solving collected in a Wizard-of-Oz setup. In order to elicit a mixed-language style of interaction, we divided the subjects into two groups and provided them with different presentations of the material (formalized and verbose). We hypothesized that the format of the study material may influence the way the subjects would interact with the system. The analysis of the corpus revealed differences in the language used by the subjects related to the style of presentation, which confirms our hypothesis. However, investigation of other factors that may have influenced the subjects, such as individual differences in the interaction styles between subject-wizard pairs is needed.

[7] http://www.pitt.edu/~circle/Archive.htm

6. References

C. Benzmüller, A. Fiedler, M. Gabsdil, H. Horacek, I. Kruijff-Korbayová, M. Pinkal, J. Siekmann, D. Tsovaltzi, B.Q. Vo, and M. Wolska. 2003a. Tutorial dialogs on mathematical proofs. In *Proc. of the IJCAI Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems*, pp. 12–22.

C. Benzmüller, A. Fiedler, M. Gabsdil, H. Horacek, I. Kruijff-Korbayová, M. Pinkal, J. Siekmann, D. Tsovaltzi, B.Q. Vo, and M. Wolska. 2003b. A Wizard-of-Oz experiment for tutorial dialogues in mathematics. In *AIED-03 Supplementary Proc.*, Advanced Technologies for Mathematics Education, pp. 471–481.

S.E. Brennan and J.O. Ohaeri. 1994. Effects of message style on users' attributions toward agents. In *CHI -94: Conference companion on Human factors in computing systems*, pp. 281–282.

I.N. Bronstein and K.A. Semendjajew. 1991. *Taschenbuch der Mathematik*. Teubner.

N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proc. of the 1st International Conference on Intelligent User Interfaces*, pp. 193–200.

P. Dahlqvist, R. Ramberg, and Y. Waern. 1999. The effects of different presentation formats on learning. In *European Association for Research on Learning and Instruction*.

N.M. Fraser and G.N. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.

H. Horacek and M. Wolska. 2005. Interpretation of mixed language input in a mathematics tutoring system. In *Proc. of the AIED-05 Workshop on Mixed Language Explanations in Learning Environments*, pp. 27–34.

D. Maulsby, S. Greenberg, and R. Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Conference on Human Factors in Computing Systems*, pp. 277–285. ACM Press.

M.D. Ringle and R. Halstead-Nussloch. 1989. Shaping user input: a strategy for natural language dialogue design. *Interacting with Computers*, 1(3):227–244.

S. Tomko and R. Rosenfeld. 2004. Shaping spoken input in user-initiative systems. In *The 8th International Conference on Spoken Language Processing*, pp. 2825–2828.

M. Wolska and I. Kruijff-Korbayová. 2004. Analysis of Mixed Natural and Symbolic Language Input in Mathematical Dialogs. In *Proc. of the 42nd Meeting of the Association for Computational Linguistics*, pp. 25–32.

M. Wolska, B.Q. Vo, D. Tsovaltzi, I. Kruijff-Korbayová, E. Karagjosova, H. Horacek, M. Gabsdil, A. Fiedler, and C. Benzmüller. 2004. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *Proc. of International Conference on Language Resources and Evaluation*, pp. 1007–1010.

C. Zinn. 2003. A Computational Framework For Understanding Mathematical Discourse. *Logic Journal of the IGPL*, 11(4):457–484.

E. Zoltan-Ford. 1991. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34(4):527–547.