Graphical revelations: Comparing students' translation errors in graphics and logic

Richard Cox[1], Robert Dale[2], John Etchemendy[3], and Dave Barker-Plummer[3]

[1] Department of Informatics, University of Sussex, Falmer, BN1 9QJ, UK,
richc@sussex.ac.uk
[2] Center for Language Technology, Macquarie University, Sydney, NSW 2019, Australia
rdale@ics.mq.edu.au
[3] CSLI, Stanford University, Stanford, CA, 94305, USA
{dbp,etch}@csli.stanford.edu

Abstract. We are interested in developing a better understanding of what it is that students find difficult in learning logic. We use both natural language and diagram-based methods for teaching students the formal language of first-order logic. In this paper, we present some initial results that demonstrate that, when we look at how students construct diagrammatic representations of information expressed in natural language (NL) sentences, the error patterns are different from those observed when students translate from NL to first-order logic (FOL). In the NL-to-diagram construction task, errors associated with the interpretation of the expression *not a small dodecahedron* were manifested much more frequently with respect to the object's size than with respect to its shape. In the NL-to-FOL task, however, no such asymmetry was observed. We hypothesize a number of possible factors that might be implicated here: differences between the NL-to-diagram and NL-to-FOL tasks; the reduced expressivity of diagrams compared to language; scoping errors in participants' NL parsing; and the visuospatial properties of the blocks-world domain. In sum, constructing a diagram requires the student to provide an instantiated representation of the meaning of a natural language sentence; this tests their understanding in a way that translation into first-order logic does not, by ensuring that they are not simply carrying out a symbol manipulation exercise.

Keywords: errors; natural language; graphical representations; first-order logic

1 Introduction

In this paper we present findings concerning the errors that students make when learning first order logic. In our teaching we use both natural language and diagrammatic representations to aid the students in learning this material. Here we consider exercises in which students had to construct a blocks world (a particular kind of diagrammatic representation) in which twelve English sentences are true, in addition to translating each of the sentences into first-order logic. We find that the consideration of the diagrammatic modality and the ability to compare two channels of information flow through deduction (natural language into diagrams and natural language into first-order logic) identifies problems in students' understanding that are not obvious simply from their first-order logic translations.

Our data are derived from student-generated solutions to exercises in *Language, Proof and Logic* (LPL) [4], a courseware package consisting of a textbook together with desktop applications which students use to complete exercises presented in the text.[4] Students may submit answers to 489 of LPL's 748 exercises[5] to The Grade Grinder (GG), a robust automated assessment system that has assessed approximately 1.8 million submissions of work by more than 38,000 individual students over the past eight years. These submissions form an extremely large corpus of high ecological validity which we wish to exploit in order (*inter alia*) to gain insights into cognitive processes during formal reasoning and to extend our research on individual differences in reasoning (e.g. [8]), to improve logic teaching.

Some exercises require the student to translate an English sentence, such as **d** *is a small dodecahedron unless* **a** *is small*, into first-order logic (FOL). Tarski's World (TW)[6] allows the student to enter candidate solutions, and then to manipulate diagrams of worlds containing blocks on a checkerboard to see whether the situations in which the English statement is true also make the proposed translation true. Blocks in TW can have one of three shapes: tetrahedron, cube and dodecahedron, represented by the predicates Tet, Cube and Dodec; and one of three sizes: small, medium and large, represented by the predicates Small, Medium and Large.[7]

If the truth of the student's translation matches the truth of the English statement in the worlds under consideration, then the student has evidence that the translation is a good candidate for the correct answer. However, the student's answer may yet be incorrect, perhaps because they have not considered a relevant situation. The student can only obtain a definitive answer by submitting the proposed solution to GG.

In recent work we analysed students' errors in translating natural language (NL) sentences into first-order logic [2]. In that work, we demonstrated that students had particular difficulties with distinguishing the conditional from the biconditional, were sensitive to source-sentence word-order effects during translation, and were sensitive to factors associated with the naming of constants. In [2], we noted that students had particular problems in translating sentences of the form *P unless Q* into FOL on a sentence-by-sentence basis. In the present paper, we once again find this same form to be a significant source of errors, this time in a 'deeper' reasoning context—one which requires students to engage in a chain of inference steps in order to build a diagrammatic representation.

2 The Focus of This Study

For the exploration described in this paper, we chose to focus on LPL Exercise 7.15, which, like the exercise discussed in [2], addresses conditionals, and involves the translation of sentences from NL to FOL. However, it also requires the student to make inferences from the sentences (all of which concern the sizes and shapes of objects **a**

[4] See `http://lpl.stanford.edu`.

[5] The other exercises require that students submit their answers on paper to their instructors.

[6] TW is one of LPL's three desktop applications.

[7] Blocks also have a position on the checkerboard, leading to predicates such as LeftOf, but these are unused in the work described in this paper.

"…translate the following English sentences (into FOL) …(build) a world in which the 12 English sentences are true. Verify that your translations are true in this world as well. Submit both your sentence file and your world file."

1. *If* **a** *is a tetrahedron then* **b** *is also a tetrahedron.*
2. **c** *is a tetrahedron if* **b** *is.*
3. **a** *and* **c** *are both tetrahedra only if at least one of them is large.*
4. **a** *is a tetrahedron but* **c** *isn't large.*
5. *If* **c** *is small and* **d** *is a dodecahedron, then* **d** *is neither large nor small.*
6. **c** *is medium only if none of* **d**, **e**, *and* **f** *are cubes.*
7. **d** *is a small dodecahedron unless* **a** *is small.*
8. **e** *is large just in case it is a fact that* **d** *is large if and only if* **f** *is.*
9. **d** *and* **e** *are the same size.*
10. **d** *and* **e** *are the same shape.*
11. **f** *is either a cube or a dodec, if it is large.*
12. **c** *is larger than* **e** *only if* **b** *is larger than* **c**.

Fig. 1. The natural language sentences in Exercise 7.15

through **f**) and then to build a blocks world in which these sentences are true. In order to complete the exercise, students are required to submit both their FOL sentences and the constructed world.

#	N	% incorrect	Correct FOL translation
1	183	1.7	$\mathsf{Tet(a)} \rightarrow \mathsf{Tet(b)}$
2	755	7.3	$\mathsf{Tet(b)} \rightarrow \mathsf{Tet(c)}$
3	2739	26.5	$(\mathsf{Tet(a)} \wedge \mathsf{Tet(c)}) \rightarrow (\mathsf{Large(a)} \vee \mathsf{Large(c)})$
4	865	8.4	$(\mathsf{Tet(a)} \wedge \neg\mathsf{Large(c)}$
5	2093	20.2	$(\mathsf{Small(c)} \wedge \mathsf{Dodec(d)}) \rightarrow (\neg\mathsf{Large(d)} \wedge \neg\mathsf{Small(d)})$
6	3762	36.4	$\mathsf{Medium(c)} \rightarrow (\neg\mathsf{Cube(d)} \wedge \neg\mathsf{Cube(e)} \wedge \neg\mathsf{Cube(f)})$
7	3258	31.6	$\neg\mathsf{Small(a)} \rightarrow (\mathsf{Small(d)} \wedge \mathsf{Dodec(d)})$
8	4055	39.2	$\mathsf{Large(e)} \leftrightarrow (\mathsf{Large(d)} \leftrightarrow \mathsf{Large(f)})$
9	224	2.2	$\mathsf{SameSize(d,e)}$
10	236	2.3	$\mathsf{SameShape(d,e)}$
11	1175	11.4	$\mathsf{Large(f)} \rightarrow (\mathsf{Cube(f)} \vee \mathsf{Dodec(f)})$
12	2477	24.0	$\mathsf{Larger(c,e)} \rightarrow \mathsf{Larger(b,c)}$

Fig. 2. FOL translations of the sentences in Figure 1

A translation for a sentence (which we refer to here as a **solution**) is considered correct if it is equivalent to a **reference solution** known to GG.[8] Figure 1 shows the sentences in this exercise; example correct FOL translations are presented in Figure 2.

[8] There are infinitely many correct answers for any sentence, so GG employs a theorem prover to determine equivalence.

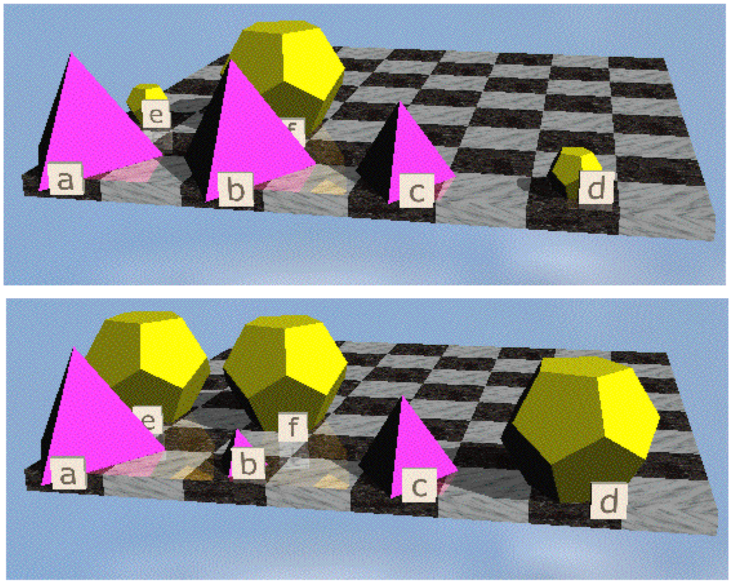

Fig. 3. Most frequent above-median (upper) & below-median (lower) diagrams.

A submission for the NL-to-FOL translation task is considered correct if all twelve of the student's FOL sentences in the submission are equivalent to their corresponding reference sentences.

The twelve sentences uniquely determine the sizes and shapes of the blocks with names **a** through **f**. The submitted world is correct if the blocks have these sizes and shapes, or, equivalently, if all of the reference solutions evaluate to true in the submitted world. The blocks world shown in the upper part of Figure 3 is an example of a (correct) world in which all of the NL sentences in Figure 1 are true.

The corpus of submissions for Exercise 7.15 made by students during the calendar years 2000–2007 contains more than 29,500 submissions, of which 18,609 submissions (61%) were erroneous. Of the erroneous submissions, 7,918 were missing the world file, and 372 were lacking the sentence file. We discarded these from the analysis, leaving 10,319 submissions. These submissions were made by 5,176 different students.

3 Information Flow through Deduction

Unlike the NL-to-FOL translation task, in which the sentences may be translated independently of one another, building the blocks world requires the use of the sentences in concert. We can trace the information flow required to complete the task by looking at the sentences and determining the inferences that can be drawn. This is a heterogeneous deduction of the kind implemented in our Hyperproof program [3].

There is only one place to start in the deduction, namely with Sentence 4, the only sentence that contains unconditional information: **a** *is a tetrahedron but* **c** *isn't large*.

N	FOL	Error type
350	$(\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d})) \to \neg\mathsf{Small}(\mathsf{a})$	ACREV
235	$\neg\mathsf{Small}(\mathsf{a}) \to \mathsf{Dodec}(\mathsf{d})$	Missing conjunct
219	$\mathsf{Small}(\mathsf{a}) \to (\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d}))$	Missing negation
195	$\mathsf{Small}(\mathsf{a}) \to \neg(\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d}))$	Moved negation
189	$(\mathsf{Dodec}(\mathsf{d}) \land \mathsf{Small}(\mathsf{d})) \to \neg\mathsf{Small}(\mathsf{a})$	ACREV
137	$\mathsf{Small}(\mathsf{a}) \to \neg(\mathsf{Dodec}(\mathsf{d}) \land \mathsf{Small}(\mathsf{d}))$	Moved negation
117	$(\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d})) \to \mathsf{Small}(\mathsf{a})$	ACREV; missing negation
104	$\mathsf{Small}(\mathsf{a}) \to (\mathsf{Dodec}(\mathsf{d}) \land \mathsf{Small}(\mathsf{d}))$	Missing negation
80	$\mathsf{Small}(\mathsf{a}) \to (\neg\mathsf{Small}(\mathsf{d}) \land \neg\mathsf{Dodec}(\mathsf{d}))$	Moved negation
79	$(\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d})) \leftrightarrow \neg\mathsf{Small}(\mathsf{a})$	ACREV; Biconditional

Fig. 4. 10 highest-frequency FOL errors on S7. ACREV = antecedent–consequent reversal.

It is a conjunction of two facts, one of which tells us that **a** is a tetrahedron. This fact can be combined with the conditional information in Sentence 1 (*If* **a** *is a tetrahedron then* **b** *is also a tetrahedron*) to infer the shape of **b**, and once the shape of **b** is known, it can be combined with the conditional information in Sentence 2 (**c** *is a tetrahedron if* **b** *is*) to infer that **c** is also a tetrahedron. A more complex inference is now required, one which uses Sentence 3: **a** *and* **c** *are both tetrahedra only if at least one of them is large.* We know that **a** and **c** are indeed tetrahedra, and from Sentence 4 (again) we know that **c** is not large, and so we conclude that **a** is large. These four results—that **a**, **b** and **c** are tetrahedra, and that **a** is large—are obtained in 8,650 (84%) of all of the (erroneous) submitted worlds. In other words, up to this point, most students do not have any problems in carrying out the task.

The next inference, however, is the focus of this paper. In a correct deduction, Sentence 7 (**d** *is a small dodecahedron unless* **a** *is small*) should now be used to infer that **d** is a small dodecahedron (since the only occasion when it might not be is when **a** is small, and we know that it is not). Students find this inference relatively difficult. 2,552 (29%) of the submitted worlds show an error concerning the size and/or shape of **d**.

4 FOL translations

Figure 2 shows the NL-to-FOL translation error rate for each of the 12 sentences. Six of the twelve sentences account for 85% of all errors (S3, S5, S6, S7, S8, S12). The length of the natural language sentences (number of words per sentence) is significantly positively correlated with percentage error rate ($r = .71, p = .01, n = 12$). Sentences containing conditional or biconditional connectives (*i.e.* Sentences 3, 5-8, 11 and 12) are also associated with high rates of translation error.

Sentence 7—**d** *is a small dodecahedron unless* **a** *is small*—is the third most error-prone FOL translation (Figure 2). It is a relatively simple sentence, and the only one in this exercise that includes the term *unless*. The LPL textbook suggests that *P unless Q* is best translated into FOL as $\neg\mathsf{Q} \to \mathsf{P}$; by this rule, the corresponding translation into FOL is $\neg\mathsf{Small}(\mathsf{a}) \to (\mathsf{Small}(\mathsf{d}) \land \mathsf{Dodec}(\mathsf{d}))$, although any (propositionally) equivalent formula is accepted by GG.

Sentence 7 results in 372 different forms of FOL translation error, for a total of 3,258 errors. Figure 4 shows the ten most frequently occurring forms (accounting for 52% of the errors) and that, of the 10 highest-frequency errors, four involve antecedent–consequent reversals.

The fact that Sentence 7 presents a problem to students is, perhaps, not surprising. We found in our earlier work that the *unless* form is problematic for students. One contributing feature is that the translation of Sentence 7 (**d** *is a small dodecahedron unless* **a** *is small*) into $\neg\mathsf{Small(a)} \rightarrow (\mathsf{Small(d)} \wedge \mathsf{Dodec(d)})$ does not preserve word order. When word order is not preserved, antecedent–consequent errors are more likely [2]. Six error forms involved misplacement or omission of the negation symbol. Errors involving negation ranked as the fourth most common (accounting for 9.2% of all errors) in a different translation exercise involving conditional sentences [2].

An important feature of almost all of the mistranslations (and the correct translation) is that the sentences are **symmetrical** with respect to size and shape for **d** (the exception being $\neg\mathsf{Small(a)} \rightarrow \mathsf{Dodec(d)}$), which is to say that the two atoms appear in the same configuration: typically as two conjuncts; sometimes with the conjunction negated, and sometimes with each conjunct negated. Only in the FOL error type $\neg\mathsf{Small(a)} \rightarrow \mathsf{Dodec(d)}$ do we see any evidence that the two atoms are being treated differently from one another; but note that this error type represents only a small percentage (4.3%) of the large variety of other forms of FOL error, almost all of which refer to both the shape and size of **d**.[9]

5 Translation errors and TW diagrams

The observations above lead us to conjecture that the students (correctly) understand the noun phrase *a small dodecahedron* as concerning two properties of an object both applying the the same way to the same object. However, it turns out that the errors students make in the graphical domain tell a different story: the graphical products of students' reasoning reveal more errors related to block size compared to block shape.

By analyzing the flow of information through deductions across the Exercise 7.15 sentences, the source of the size/shape asymmetry can be tracked down to Sentence 7. Of the submitted worlds that contain an error resulting from a mistaken inference using this sentence, 2,346 (92%) of the errors concern the size of **d**, while only 623 (24%) concern the shape of **d**; 417 (16%) make an error on both properties. Thus the evidence based on the graphical data indicates that the information in Sentence 7 concerning the size of **d** is handled differently than that concerning the shape of **d**.

The properties of the objects in the correct blocks world are shown in the top row of Figure 5. The ten most popular incorrect blocks worlds, accounting for 3,155 (42%) of the 7,489 erroneous worlds, are also presented in Figure 5 (those with count values), with incorrect values shown in bold.

Students' blocks-world diagrams vary in terms of their **diagram accuracy**, which we assess by scoring each diagram according to the number of the twelve correct properties (size and shape for each of six blocks) it possesses (scores can therefore range

[9] Note that the sentence $\mathsf{Small(d)} \rightarrow \neg\mathsf{Small(a)}$ *does* appear in the incorrect translation set with a smaller frequency (19 occurrences).

Count	a		b		c		d		e		f	
	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	Dodec	Large
607	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	**Tet**	**Small**
493	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	Dodec	**Small**
423	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	**Tet**	**Medium**
365	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	**Cube**	Large
351	Tet	Large	Tet	**Small**	Tet	Medium	Dodec	**Large**	Dodec	**Large**	Dodec	Large
256	Tet	Large	Tet	**Small**	Tet	**Small**	Dodec	**Medium**	Dodec	**Medium**	**Cube**	Large
175	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	Dodec	**Medium**
168	Tet	Large	Tet	**Small**	Tet	**Small**	Dodec	Small	Dodec	Small	**Cube**	Large
162	Tet	Large	Tet	Large	Tet	Medium	Dodec	**Large**	Dodec	**Large**	Dodec	Large
155	Tet	Large	Tet	Large	Tet	Medium	Dodec	Small	Dodec	Small	**Tet**	Large

Fig. 5. Ten most popular incorrect blocks worlds.

from 0–12). Students whose blocks-world diagrams scored below the 50th percentile tended to make many more errors (an average of .29 errors per student) in translating Sentence 7 than students whose diagrams scored above the median (.18 errors per student). For example, Sentence 7 evaluates as true in the upper diagram in Figure 3 (as do all 12 sentences), but in the lower (below-median-score) diagram it is the only sentence of the twelve to evaluate as false.[10]

The size of **d** must be inferred from Sentence 7. Students infer that **d** is not a small dodecahedron, but they are much more likely to infer that it is not small than infer that it is not a dodecahedron. In other words their inference is asymmetrical, and impacts the size dimension much more than the shape dimension.

6 Discussion

We have explored the translation of NL sentences into two different modalities, first-order logic and graphical. As the preceding discussion demonstrates, this provides a much clearer picture of the nature of scoping errors during NL interpretation than the study of, say, translations from one linguistic modality (natural language) into another one (FOL).

Errors in the translation into graphics turn out to be consistent with a wrongly-scoped reading of Sentence 7 (**d** *is a small dodecahedron unless* **a** *is small*) which leads these students to conclude that **d** is not a small dodecahedron; further, this is incorrectly scoped as akin to **d** *is (not a small) dodecahedron* rather than **d** *is (not a small dodecahedron)*. However, there is no evidence of this misunderstanding in the results of the NL-to-FOL task.

The absence of shape/size-asymmetry in the NL-to-FOL case is possibly due to the cognitively 'shallower' sentence-by-sentence nature of the NL-to-FOL translation task (referred to as **immediate inference** by Newstead [6]). In other words, the scoping error

[10] The lower diagram in Figure 3 corresponds to the row of Figure 5 with count = 351.

is associated only with sentence comprehension in the context of *inference across sentences*, plus the need to recall from working memory parameters of the blocks (*e.g.* the shape of **a**) established from earlier sentences (in particular, Sentence 4) to supplement new information (*i.e.* the size of **a** given in Sentence 7). It appears that the expression *small dodecahedron* is understood as a whole in the FOL case, but increased working memory load causes it to become fractured when building a blocks-world diagram, resulting in *small* being treated differently from *dodecahedron*).

The student's focus on only the size of **d** rather than on the size *and* shape of **d** may be due to size being the only aspect of **a** on the right hand side of the sentence that is explicitly referred to. The shape of **a** is unstated in Sentence 7, and so has to either be extracted by the student from her blocks-world-in-progress, or retrieved from visuospatial working memory. It should also be noted that, in terms of the flow of information through deduction, the shape of **a** is the very first block parameter to be established (from Sentence 4) and is therefore the piece of information that has been longest in memory by the time Sentence 7 becomes the focus of the student's reasoning.

Several factors may interact to produce the effects we report. The first concerns differences between the tasks. The NL-to-FOL task differs from the NL-to-diagram task in that the former requires the translation of one sentence at a time, whereas the latter requires (a) the integration of information from several sentences used in concert and (b) the active construction of a blocks world diagram. The difference between the tasks, together with the observation that graphics are more limited than linguistic representations in terms of their ability to express abstraction [9], may go some way towards explaining the finding. The NL-to-FOL translation process occurs on a sentence-by-sentence basis involving the translation or mapping of an English sentence's subject, object, nouns, verbs, adjectives, adverbs, and so on into antecedent, consequent clauses, atoms, constants, connectives, negation symbols, and so on. To some extent this translation task can be seen as one which only requires the application of NL-to-FOL transformation rules; however, as we have shown in [2], NL features such as word order, types of connective and the labelling of constants can have systematic negative effects upon translation accuracy. Constructing a blocks world, on the other hand, requires information to be deduced from a collection of sentences, a task that imposes a considerably greater load on working memory's phonological, visuospatial and central executive (*e.g.* attention management) components [1]. The graphical task requires the production of a blocks world in which twelve NL sentences, describing the sizes and shapes of objects **a** through **f**, are true. Analysing the flow of information through the deduction process shows that whereas the shape of block **a** is explicitly given (in Sentence 4), other blocks' size and shape parameters entail quite lengthy chains of reasoning. For example, the size of **d** involves inference across five sentences (1, 2, 3, 4 and 7). Determining the shape of **f** seems to be the most taxing: it requires inference across eleven sentences (1–7 and 9–12).[11]

The findings we report here are from a blocks-world construction task that involves *visual* or *visuospatial* block parameters such as *size* (for example, $\mathsf{Large(a)}$) and *shape* (for example, $\mathsf{Tet(a)}$) and which do not involve *spatial* parameters (board positions). Size is a dimension that possesses a natural commensurability and ordinality (smaller–

[11] Note that the block **f** is the one on which the most shape errors are observed (Figure 5).

larger) whereas shape doesn't have this property (it is nominal). In the context of the exercise reported here, however, the natural commensurability of size is irrelevant to the logical reasoning. These semantic factors contribute to the tendency of visual images to be either 'not critical' or 'interfering' in deductive reasoning (whereas spatial representations help), as argued by [5]. This effect may also contribute to the differences in error patterns observed for size *vs* shape on the graphical task. Mental imagery effects such as these are not predicted by rule-based theories of deductive reasoning (e.g. [7]).

At this stage of our research, we can only speculate about which of several competing explanations account for these effects. We aim to more clearly delineate their relative effects in future work. However, an important pedagogical implication is already clear from these preliminary results: the ambiguity of NL and scoping during NL interpretation are topics that should be given more attention in the logic curriculum, and the problems evident here may not be so clear if we only consider NL-to-FOL translation exercises. The insights into students' reasoning that these analyses provide are also useful for improving the Grade Grinder and for enriching the type of feedback that it can provide to students.

7 Acknowledgements

RC gratefully acknowledges support provided by an ESRC TLRP (UK) & SSRC (US) Visiting Americas Fellowship; RD gratefully acknowledges the support of the Australian Research Council. Albert Liu assisted with the collection of the data. Thanks too to three anonymous reviewers for their constructive suggestions.

References

1. A. Baddeley. *Working memory, thought, and action.* Oxford University Press, 2007.
2. D. Barker-Plummer, R. Cox, R. Dale, and J. Etchemendy. An empirical study of errors in translating natural language into logic. In V. Sloutsky, B. Love, and K. McRae, editors, *Proceedings of the 30th Annual Cognitive Science Society Conference*. Lawrence Erlbaum Associates, 2008.
3. J. Barwise and J. Etchemendy. *Hyperproof.* CSLI Publications and University of Chicago Press, September 1994.
4. J. Barwise, J. Etchemendy, G Allwein, D. Barker-Plummer, and A. Liu. *Language, Proof and Logic.* CSLI Publications and University of Chicago Press, September 1999.
5. M. Knauff and P.N. Johnson-Laird. Visual imagery can impede reasoning. *Memory and Cognition*, 30:363–371, 2002.
6. S. Newstead. Interpretational errors in syllogistic reasoning. *Journal of Memory and Language*, 28:78–91, 1989.
7. L.J. Rips. *The Psychology of Proof.* Cambridge, MA: MIT Press., 1994.
8. K. Stenning and R. Cox. Reconnecting interpretation to reasoning through individual differences. *The Quarterly Journal of Experimental Psychology*, 59:1454–1483, 2006.
9. K. Stenning and J. Oberlander. A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, 19:97–140, 1995.