Unnatural language discourse: an empirical study of multimodal proof styles

Jon Oberlander Padraic Monaghan Richard Cox Keith Stenning
Richard Tobin

March 4, 1998

Abstract

Computer-based logic proofs are a form of 'unnatural' language discourse, but the structure and process of proof generation can be observed in considerable detail, and analysis is leading to a number of general insights. We have been studying how students respond to multimodal logic teaching. Performance measures have already indicated that students' pre-existing cognitive styles have a significant impact on teaching outcome. Furthermore, a large corpus of proofs has been gathered via automatic logging of proof development. This paper applies a series of techniques, including corpus statistical methods, to the proof logs. The results indicate that students' cognitive styles influence the structure of their logical discourse, via their differing methods of handling abstract information, and transferring information between modalities. As well as uncovering different thinking styles in this artificial domain, the observations raise the issue of the importance of individual differences in natural language discourses.

1 Introduction

Computer-based multimodal tools are giving people the freedom to express themselves in brand new ways. But what do people actually *do* when given these tools? Does everyone end up generating the same forms of multimodal discourse? Do multimodal systems lead to better performance than monomodal systems?

These questions arise in many areas, such as human-computer interaction, video-games, and so on; but they are particularly important in educational applications, since multimodality is believed to be especially helpful to novices (di Sessa, 1979; Schwarz and Dreyfus, 1993). Hyperproof is a program created by Barwise and Etchemendy (1994) for teaching first-order logic. It uses multimodal (graphical and sentential) methods, and is inspired by a situation-theoretic approach to heterogeneous reasoning. A distinctive feature of Hyperproof is its set of 'graphical' rules, which permit users to transfer information to and fro, between graphical and linguistic modes.

We have been carrying out a series of experiments on Hyperproof, to help evaluate its effects on students learning logic. Amongst other things, we have built up a substantial corpus of proofs. These 'hyperproofs' are an unusual form of discourse, for

two main reasons. Firstly, they are primarily used for *self*-communication: a student arranges proof steps and rules in an external representation as an aid to their individual problem-solving activities, although at another level they do involve dialogue with the machine. Secondly, hyperproofs are, of course, *multimodal* discourse: they involve both language and graphics, and are therefore in some ways more complex than text or speech. Elsewhere, we have argued that graphical systems possess a useful property—over-specificity—whereby certain classes of information must be specified (Stenning and Oberlander 1991, 1995). From this, certain hypotheses follow, concerning the utility and usability of graphical abstraction.

Although our title emphasises the unaturalness of the discourses under analysis, our results give some encouragment that the patterns of thought they reveal would apply to styles of thinking which derive directly from natural discourse processing styles. What we see are students who have been through an extensive course of teaching in how to reason, doing course exercises in the uniform ways that have been taught. However, when they are given examination problems which are more difficult, their styles of proof diverge according to their performances in the pre-course tests. This happens in the same 'unnatural environment' in which they have been taught. If all the habits of thought they exhibit were the result of teaching, we would expect continued uniformity. What we see instead is an interaction between the unnatural environment and pre-existing styles of reasoning.

A little further evidence of the generality of these findings comes from a study by Monaghan & Stenning (submitted) which collected human tutoring discourses of the teaching of syllogisms by both sentential and graphical methods. The Euler Circles graphics used are entirely different from Hyperproof graphics, and the measures were of numbers of errors and number of tutoring interventions *during* teaching. The same GRE pre-test as used here again shows clear aptitude by treatment interactions for both error and intervention measures. Interestingly, other psychometric tests (serial/holist learning styles; paper folding) also predict individual differences in this task in different substages of reasoning.

In the rest of this paper, we focus on the structures of the hyperproofs generated by students under examination conditions, and show how our empirical methods are revealing subtle patterns in multimodal discourse structure. To this end, we first introduce Hyperproof, and indicate how multimodal proofs can be considered as structured discourses. We then outline our main hypotheses, indicating the potential influence on discourse production of individual differences in cognitive style. We then introduce some aspects of the method used in our overall study, before going on to the current results emerging from our analyses of the hyperproof corpus. First, there are basic results from analysing rule usage in the logs. Secondly, a phenomenon is revealed by the use of 'proofograms', which help visualise the abstraction structures used in hyperproofs, and indicate how abstraction is subject to individual difference. Thirdly, we focus on bigram analyses of rule use in the corpus. Finally, we examine networks revealing overall differences between the types of proofs generated by different types of student. We conclude by drawing some general morals from the study, and point to plausible directions for further research.

2 Hyperproofs as multimodal discourse

Hyperproof should be viewed as a proof-checking environment designed to support human theorem proving using heterogeneous information. As can be seen in Figure 1, the interface contains two main window panes: one presents a diagrammatic view of a chess-board world containing geometric objects of various shapes and sizes; the other presents a list of sentences in predicate calculus; control palettes are also available. These window panes are used in the construction and editing of proofs. Several types of goals can be proved, involving the shape, size, location, identity or sentential descriptions of objects; in each case, the goal can involve determining some property of an object, or showing that a property *cannot* be determined from the given information. A number of rules are available for proof construction; some of these are traditional syntactic rules (such as $\wedge$-elimination); others are 'graphical', in the sense that they involve consulting or altering the situation depicted in the diagrammatic window (Figure 2 contains a subset of these rules). In addition, a number of rules check properties of a developing proof.

A proof produced using Hyperproof can be thought of as an artefact of multimodal self-communication. Consider the two differing proofs displayed in Figure 3. These are the answers produced by two students (C2 and C14) to the exam question displayed in Figure 1.

Take C14's proof. Each line corresponds to a single utterance. Some utterances are linguistic (and are represented by formulae in the calculus pane); others are graphical (and are represented by a diamond icon in the calculus pane, and a particular situation in the graphical pane). Each utterance is associated with a single rule, which specifies its functional role within the proof. The rule is therefore similar to McKeown's (1985) notion of a rhetorical predicate. Some rules require explicit dependencies to be established between the current utterance and others; these are introduced by the student, and displayed by highlighting. The dependencies are akin to anaphoric links. As well as this dependency structure, there is a hierarchical structure, reflecting the grouping of common cases in the argument. This structure is similar in kind to Grosz and Sidner's (1986) linguistic structure.

In analysing a discourse corpus, we can ask two different sorts of question. We can focus on function, and look at the various devices by which people achieve their objectives; the aim would be to show how task constraints influence the types of constructions occurring in the discourse. On the other hand, we can concentrate on style, examining the individual variations which occur independently of task constraints; the aim would be to show how the prior cognitive styles of subjects influenced their execution of standard tasks. Most computational approaches to discourse tend to address the former type of question. Our research spans both issues: the differing problems that can be tackled with Hyperproof apparently require different types of discourse to solve them; but in addition to this, we can also track the individual variations in discourse style.

For instance, C14's proof in Figure 3 is somewhat different from the type of discourse they would produce in response to a different problem. More obviously, C14's proof

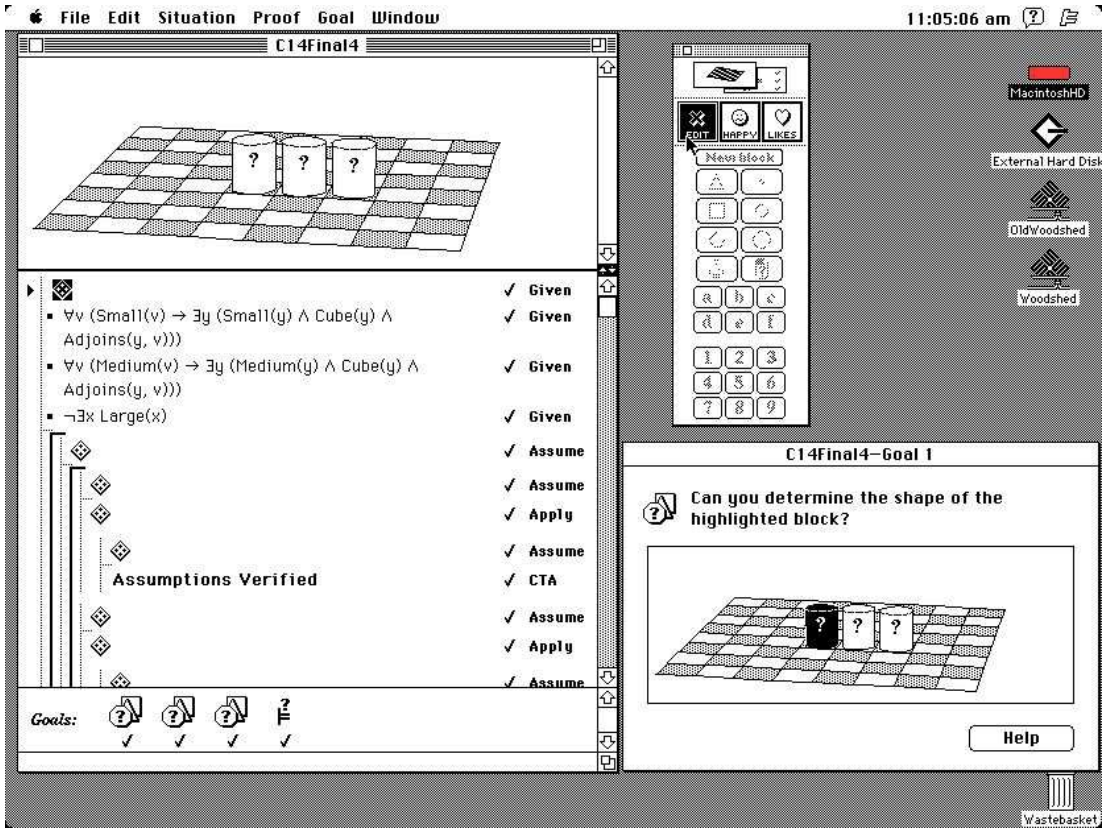

Figure 1: The Hyperproof Interface. The main window (top left) is divided into an upper graphical pane, and a lower calculus pane. The tool palette is floating on top of the main window, and the other windows reveal a set of goals which have been posed. To achieve them, a proof must be developed, by applying a set of multimodal inference rules to the graphical and calculus premises given.

Apply Extracts information from a set of sentential premises; expresses it graphically

Assume Introduces a new assumption into a proof, either graphically or sententially

Observe Extracts information from the situation; expresses it sententially

Inspect Extracts common information from a set of cases; expresses it sententially

Merge Extracts common information from a set of cases; expresses it graphically

Close Declares that a sentence is inconsistent with either another sentence, or the current graphical situation

CTA (Check truth of assumptions) Declares that all sentential and graphical assumptions are true in the current situation

Exhaust Declares that a part of a proof exhausts all the relevant cases

Figure 2: A set of relevant Hyperproof rules.

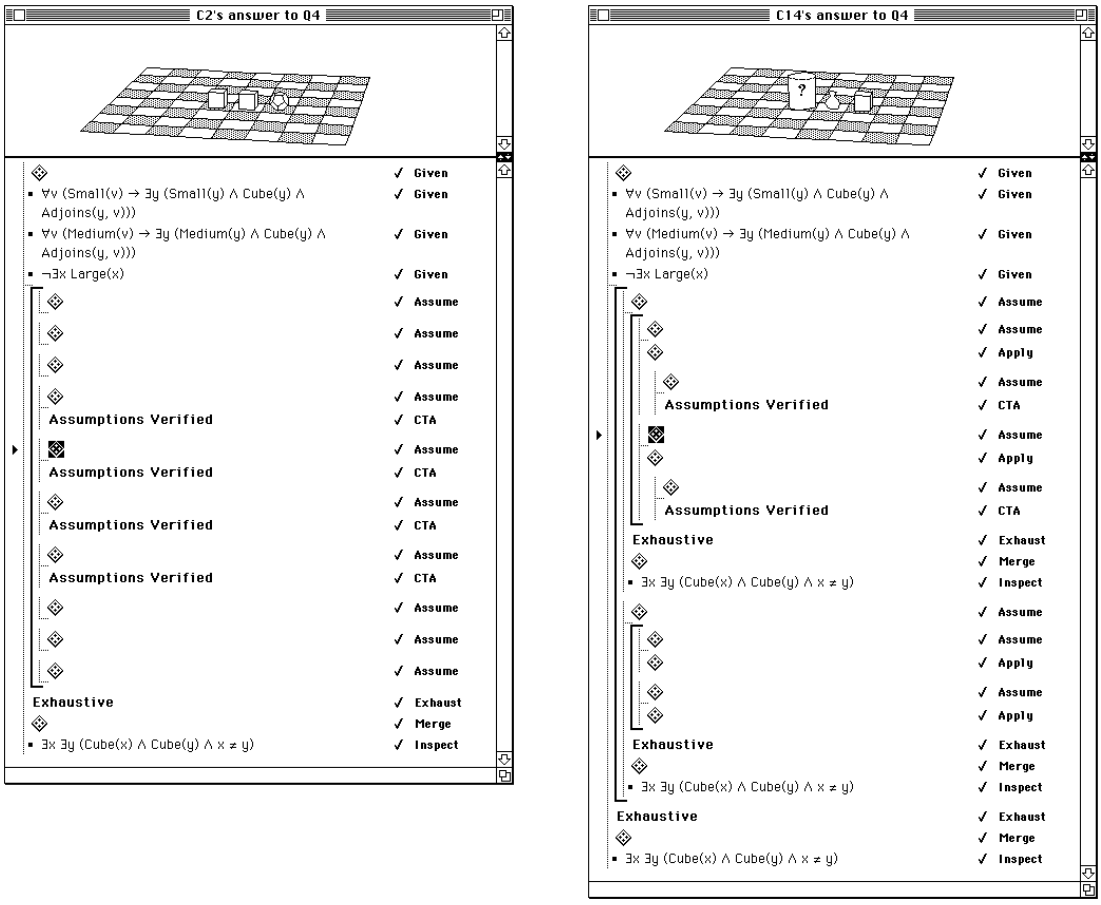

Figure 3: Two different subjects' proofs given in answer to the exam question in Figure 1. When—as here—little is fixed in the graphical situation we term a question *indeterminate* in type, and contrast it with those *determinate* questions in which all the relevant information is specified. The subjects differ in their cognitive styles; C2 (left) is a representative DetLo; C14 (right) is a DetHi.

is more hierarchically structured than C2's. The differences between their proofs are representative of broader distinctions between cognitive styles, which we discuss below.

3 Hypotheses

The observation that graphical systems require certain classes of information to be specified goes back at least to Bishop Berkeley. Elsewhere, we have termed this property 'specificity', and argued that it is useful because inference with specific representations can be very simple (Stenning and Oberlander 1991, 1995). We have also urged that actual graphical systems do allow abstractions to be expressed, and it is this that endows them with a usable level of expressive power. Thus, Hyperproof maintains a set of abstraction conventions for objects' spatial or visual attributes. As well as concrete depictions of objects, there are 'graphical abstraction symbols', which leave attributes under-specified: the *cylinder*, for instance, depicts objects of unknown size (see Figure 1); the *paper bag* depicts objects of unknown shape (see the middle object displayed on C14's proof in Figure 3). A key step, then, in mastering an actual graphical system is to learn which abstractions can be expressed, and how.

As we describe below, our pre-tests independently allowed us to divide subjects into two cognitive style groups, on the basis of their performance on a certain type of problem item. Loosely, one group is 'good with diagrams', and the other less so. The good diagrammers turned out to benefit more from Hyperproof-based teaching than the others. Our belief is that those who benefit most from Hyperproof do so because they are better able to manipulate the graphical abstractions it offers. Call this view the *abstraction ability hypothesis*.

We aim here to investigate the evidence for this hypothesis. We also intend to probe whether abstraction ability is supported by Hyperproof's visual representations—or by some other aspect of the system. One hypothesis is that the good diagrammers are simply those subjects who have a preference for the visual modality. Call this view the *visual preference hypothesis*. Another explanation would be that good diagrammers are those who are adept at translating between modalities. Call this view the *transmodal hypothesis*.

In what follows, we aim to show that the balance of evidence favours the transmodal hypothesis.

4 Distinguishing cognitive styles

Two groups of subjects were compared; one group ($n = 22$ at course end) attended a one-quarter duration course taught using the heterogeneous reasoning approach of Hyperproof. A comparison group ($n = 13$ at course end) were also taught for one quarter, but in the traditional syntactic manner supplemented with exercises using a graphics-disabled version of Hyperproof (to control for the motivational and other effects of computer-based activities). A fuller description of the method and procedure is provided elsewhere (Stenning, Cox and Oberlander, 1995).

Subjects were administered two kinds of pre- and post-course paper and pencil test of reasoning. All subjects completed the pre-tests, but 6 of the Hyperproof group did not complete the post-tests; however, this attrition does not affect the results reported. The first pre-test is the most relevant to the current discussion. It tested 'analytical reasoning' ability, with two kinds of item derived from the GRE scale of that name (Duran, Powers and Swinton 1987). One subscale consists of verbal reasoning/argument analysis. The other subscale consists of items often best solved by constructing an external representation of some kind (such as a table or a diagram). We label these subscales as 'indeterminate' and 'determinate', respectively; example items are displayed in Figure 4. Scores on the latter subscale were used to classify subjects within both Hyperproof and Syntactic groups into DetHi and DetLo sub-groups. The score reflects subjects' facility for solving a type of item that often is best solved using an external representation; DetHi scored well on these items, like the office allocation problem in Figure 4; DetLo less well. For the moment, we may consider DetHi subjects to be more 'diagrammatic', and DetLo to be less so.

Finally, we note that there is reason to believe that this categorisation, although relatively crude, is tapping into a genuine difference between subjects: Stenning, Cox and Oberlander (1995) report that in the overall study, a significant aptitude-treatment interaction (cf. Snow 1987) is found: DetHi do better on transfer tasks when they have been taught with Hyperproof, instead of Syntactically, and DetLo students exhibit opposing tendencies.

5 Analysing discourse styles

Both the Hyperproof and Syntactic groups contained DetHi and DetLo sub-groups. All subjects sat post-course, computer-based exams, although the questions differed for the two groups, since the Syntactic group had not been taught to use Hyperproof's systems of graphical rules. All 22 Hyperproof subjects completed the exams, but 2 of the 13 syntactic subjects did not. Student-computer interactions were dynamically logged, permitting a full, step-by-step, reconstruction of the process of the subject's reasoning, as well as capturing the final proof produced. Thus, the process log contains all uses of rules, and internal system responses to user input (called *manoeuvres* below), as well as system responses and feedback to the user. The final proof, on the other hand is simpler, representing a final snapshot of the development process, and encoding only the sentences, situations, and rules submitted as an answer to the exam question.

The four questions that these students were set contained two types of item: determinate and indeterminate. Here, determinate problems were taken to be those whose problem statement did not utilise Hyperproof's abstraction conventions. That is: determinate problems contained only concrete depictions of objects in their initially given graphical situation, whereas indeterminate problems—such as that in Figure 3—could contain graphical abstraction symbols in the initial situation.

In the rest of this section, we briefly discuss some tendencies observed via a relatively coarse-grained analysis of the process logs generated for the 22 Hyperproof subjects, before focusing specifically on a finer-grained analysis of their final proofs. The process

Determinate problem An office manager must assign offices to six staff members. The available offices are numbered 1–6 and are arranged in a row, separated by six foot high dividers. Therefore sounds and smoke readily pass from one to others on either side. Ms Braun's work requires her to speak on the phone throughout the day. Mr White and Mr Black often talk to one another in their work and prefer to be adjacent. Ms Green, the senior employee, is entitled to Office 5, which has the largest window. Mr Parker needs silence in the adjacent offices. Mr Allen, Mr White, and Mr Parker all smoke. Ms Green is allergic to tobacco smoke and must have non-smokers adjacent. All employees maintain silence in their offices unless stated otherwise.

- The best office for Mr White is in 1, 2, 3, 4, or 6?
- The best employee to occupy the furthest office from Mr Black would be Allen, Braun, Green, Parker or White?
- The three smokers should be placed in offices 1, 2, & 3, or 1, 2 & 4, or 1, 2 & 6, or 2, 3, & 4, or 2, 3 & 6?

Indeterminate problem Excessive amounts of mercury in drinking water, associated with certain types of industrial pollution, have been shown to cause Hobson's Disease. Island R has an economy based entirely on subsistence level agriculture with no industry or pollution. The inhabitants of R have an unusually high incidence of Hobson's' Disease.
Which of the following can be validly inferred from the above statements?

i. Mercury in the drinking water is actually perfectly safe.

ii. Mercury in the drinking water must have sources other than industrial pollution; or

iii. Hobson's Disease must have causes other than mercury in the drinking water.

- (ii) only?
- (iii) only?
- (i) or (iii) but not both?
- (ii) or (iii) but not both?

Figure 4: Examples of two types of reasoning problem. Determinate problems provide premisses which determine a (nearly) unique logical model; indeterminate problems do not. The former are closely related to what the graduate record exam (GRE) analytical test calls the *analytical reasoning* subscale; the latter to the test's *logical reasoning* subscale.

Apply Extracts information from a set of sentential premises; expresses it graphically

Assume Introduces a new assumption into a proof, either graphically or sententially

Observe Extracts information from the situation; expresses it sententially

Inspect Extracts common information from a set of cases; expresses it sententially

Merge Extracts common information from a set of cases; expresses it graphically

Close Declares that a sentence is inconsistent with either another sentence, or the current graphical situation

CTA (Check truth of assumptions) Declares that all sentential and graphical assumptions are true in the current situation

Exhaust Declares that a part of a proof exhausts all the relevant cases

Figure 5: A set of relevant Hyperproof rules.

logs obviously represent a richer source of data in the long run; however, we believe that interesting generalisations can be drawn from analysis of the final proofs on their own, and that such generalisations will serve to guide subsequent investigation of the process logs.

5.1 Analyses of variance in process logs

Preliminary analyses were performed on several parameters of the Hyperproof group's process logs. Each proof-log was coded for score (number of proof goals validated), time (time spent on proof), number of proof steps, the proof depth (the depth of nested subproofs the subjects used in their solution), and the frequency with which each of the Hyperproof logical rules was used (rule use frequency).

There was an apparent tendency for DetHi subjects to produce 'better' (that is, longer, quicker, more accurate, more nested proofs) than their DetLo counterparts within the Hyperproof group, and in fact the converse was the case within the Syntactic class. The difference between Hyperproof DetHi and DetLo subjects on the time parameter approached statistical significance ($t = -2.06, df = 14, p = .058$). No other comparisons were statistically reliable.

Interesting differences were also noted between performance on the determinate and indeterminate exam items. However, these were not differences in terms of the score, time, steps or depth parameters—the differences were in terms of rule use patterns.

As we mentioned earlier, Hyperproof supports the use of both the traditional syntactic rules of first-order logic, and special graphical rules. The most important of these are summarised in Figure 2, repeated here as Figure 5; see Barwise and Etchemendy (1994) for a full account of Hyperproof's rule system.

A two-factor ANOVA for subjects (DetHi, DetLo) and item determinacy (determinate, indeterminate) was conducted separately for each of seventeen rules, user manoeuvres and system responses, with frequency of rule use as the dependent variable. By considering the significant main effect for the determinacy factor in each analy-

sis, we found that subjects used a number of rules and manoeuvres significantly more frequently in developing proofs for the 2 indeterminate questions than for the 2 determinate questions. The rules were: Assume, Apply, CTA, create step, create subproof, move focus, cite step in support, update sentences in proof, and delete step. The Close rule was used significantly more on the *determinate* than on indeterminate questions. The remove step manoeuvre was used significantly more by DetLo subjects than by DetHi subjects.

A two-way interaction was significant in one of the analyses: the Apply rule was used more on determinate questions by DetLo subjects than by DetHi subjects. Conversely, on indeterminate questions, DetHi subjects used it more frequently than DetLo subjects.

As we have mentioned, Hyperproof's determinate questions differ from indeterminate questions in that their initial graphical situations are entirely concrete: there are no graphical variables involved. If question-type influences the types of rules being used, it must be because the relative concreteness of the graphical situation requires differing proof structures to be realized. Furthermore, the rule Apply transfers information from one modality to another, and its use is subject to individual difference.

The analyses of variance therefore offer tentative support for the abstraction ability hypothesis. It is therefore worth looking more closely at the way graphical concreteness is manipulated by subjects; and at the differing contexts in which rules occur. The rest of this section thus pursues an increasingly detailed analysis, focussing on evidence from subjects' final exam proofs.

5.2 Proofogram analysis of final proofs

How good is the evidence for the abstraction ability hypothesis? Among the Hyperproof students, do the two sub-groups—DetHi and DetLo—use graphical abstraction symbols in characteristically different ways?

We can score each step of each proof on the basis of number of concrete situations compatible with the graphical depiction; one possible scoring method is described in Oberlander, Cox and Stenning (1996). A situation's score depends on the number of objects in it, and how many of their attributes—size, shape and location—are depicted as known. A low score always indicates more abstraction; a higher score indicates more concreteness.

We can explore the way concreteness varies through the course of a proof by graphing it against the hierarchical structure of the proof. We call such graphs 'proofograms'. A proofogram plots the proof step number against the concreteness of the graphical situation at that point in the proof. At steps where a subproof is introduced, the graph branches to show both the concreteness of the new subproof and that of its parent. Thus the branching structure of the graph corresponds to the subproof structure. Figures 6 and 8 show how subjects C2 and C14 tackle an indeterminate exam question; Figures 7 and 9 give their proofograms. The visual differences between proofograms are quite striking: one group is 'spikey'—as in Figure 7; and the other is 'layered'—as in Figure 9. The differences are particularly pronounced on indeterminate questions, but also carry through to determinate items. The visual grouping of proofograms suggests

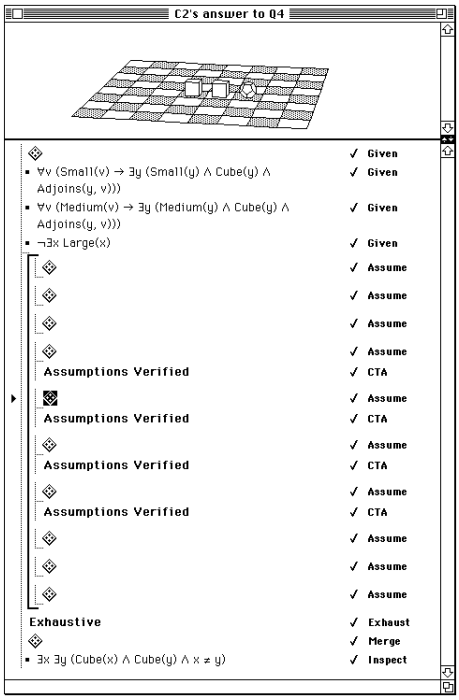

Figure 6: Submitted proof for a DetLo subject (C2) attempting an indeterminate question (Q4). The situation on view is from the 9th step of the proof. The concreteness score is 9, the maximun possible for three objects with three attributes each.

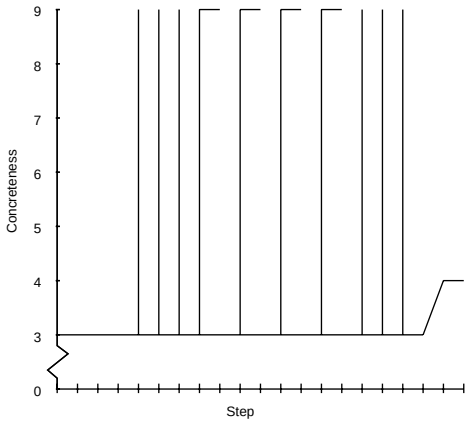

Figure 7: Proofogram for C2 attempting Q4. Proof steps are plotted on the x-axis; the concreteness of the current graphical situation is computed for each step of the proof, and is plotted on the y-axis. The branching structure of the graph corresponds to the subproof structure; vertical lines correspond to the use of Assume (which introduces a new subproof), and sloping lines correspond to the use of Apply or Merge (which increase the concreteness of an existing subproof). C2's proofogram is 'spikey', indicating a series of independent, concrete cases.

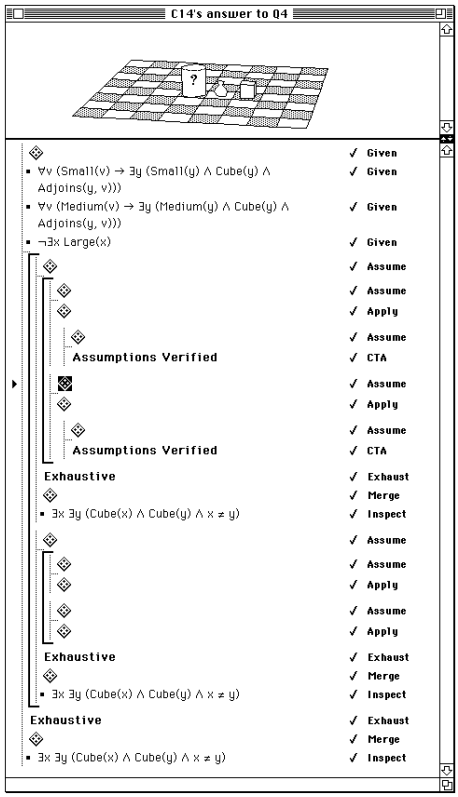

Figure 8: Submitted proof for a DetHi subject (C14) attempting an indeterminate question (Q4). The situation on view is from the 9th step of the proof. The concreteness score is 6, since only one of the objects has all three of its attributes specified; while all locations are known, the left-hand object has unknown size ad shape, and the middle has unknown shape, but known size.

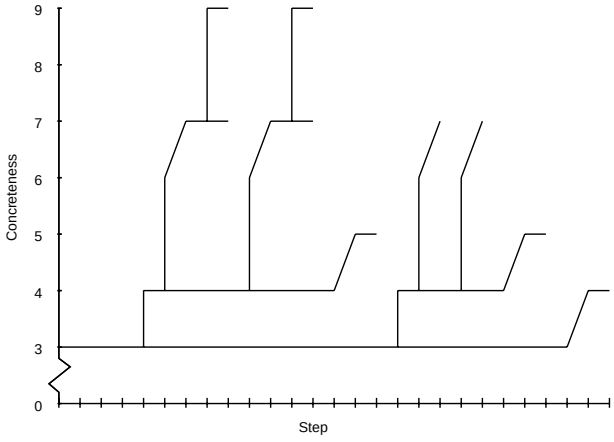

Figure 9: Proofogram for C14 attempting Q4. C14's proofogram is 'layered', indicating parallel sub-case structures with abstract superordinate cases.

the following hypothesis: DetHi subjects introduce concreteness *by stages*, whereas DetLo subjects introduce it more immediately.

To assess whether this apparent patterning was reliable, we printed the 88 proofograms (4 for each of the 22 subjects). Each proofogram was invisibly coded by subject and question. The proofograms were randomly ordered, and two prototypes (one spikey, one layered) were selected as category exemplars. The proofograms were then given to two independent raters, who were separately instructed to assign each proofogram to one category or the other, under a forced choice regime.

The results indicated a high degree of agreement between raters, with a discrepancy between the raters on only 2 of the 88 proofograms. A third observer was employed to resolve the categorisation disagreement. The raters' classification decisions for each proof were recorded.

We then analysed the concordance between proofogram category and problem determinacy level. For each of the 4 questions, 2×2 tables were produced, showing the number of items in each cell (DetLo/spikey; DetLo/layered; DetHi/spikey; DetHi/layered). A nonparametric measure of association (ϕ coefficient) was then calculated for each table.

The results indicated that the hypothesised association only held on indeterminate questions (on question 2, $\phi = .43^*$;[1] on question 4, $\phi = .28$). On questions 1 and 3 (determinate questions), both raters assigned all proofograms to the spikey category.

This confirms that there is a 'staging phenomenon': DetHi introduce concreteness *by stages*, whereas DetLo introduce it more immediately. In terms of proof structure, DetHi tend to produce structured sets of cases, with superordinate cases involving graphical abstraction; DetLo tend to produce sets of cases without such overt superordinate structure. This staging phenomenon supports the abstraction ability hypothesis: the two groups are certainly using abstractions in different ways.

5.3 Bigram analysis of final proofs

Of Hyperproof's rules, only Assume, Apply and Merge increase concreteness. We therefore examined the kind of patterns in which they occur through proof-corpus analysis. The existence of the staging phenomenon indicates that DetHi and DetLo differ in the way they handle concreteness. Since Assume is by far the most frequent means of adding concreteness, the corpus analysis distinguishes between uses of the rule which introduce totally concrete graphical situations, and those which leave some abstractness in the graphic. The term Fullassume denotes the former type of use, and assume denotes the latter.

Now, we are taking hyperproofs to be hierarchically structured discourses. It should therefore be possible to analyse them using techniques developed for the study of natural language corpora. In particular, we can carry out bigram analyses of rule use, where hierarchy and linear ordering can be taken into account. This section and the next depend on the results emerging from Monaghan's (1995) study.

Many statistical tests of natural language assume a normal distribution of units

[1] Significance at the $p < .05$ level is denoted by *.

Table 1: Bigram profile for subject C2 on question 4. C2 is a reasonable representative of the DetLo group. The first column indicates Dunning's 'log-likelihood', a measure of the significance of the distribution of a particular bigram in the corpus. $k(AB)$ is a count of the number of times the bigram AB occurs, $k(A \sim B)$ is a count of the number of times A is followed by a rule other than B, and so on. There are no significantly associated ($p < .05$) bigrams in this profile.

- -2logλ	k(AB)	k(A$\sim$B)	k($\sim$AB)	k($\sim$A$\sim$B)	A	B
7.61	1	0	0	16	Exhaust	Merge
7.61	1	0	0	16	Merge	Inspect
5.09	4	0	6	7	CTA	Assume
5.09	4	6	0	7	Assume	CTA
1.10	1	0	9	7	Given	Assume
1.10	1	9	0	7	Assume	Exhaust
0.80	5	5	5	2	Assume	Assume

within the corpus of language. For various reasons, this assumption is too strong. For a representative sample of the language, an enormous corpus is required, and even then, there is no guarantee that there will be 'standard' occurrences of all words. Of course, one could focus only on the most frequent words within the corpus. But this means that 'anomalies' are ignored, and these may be the most important elements in determining individual style, so that the data would be severely impoverished. Dunning's (1993) 'Log–Likelihood Test' addresses these problems, and unlike the more traditional Pearson's χ^2 Test, will apply effectively to the small corpus of Hyperproof logs. Ranking the bigrams according to this test provides a good *profile* of the individual's, or the group's, rule use in the proof body. Dunning's test is designed to "highlight particular A's and B's that are highly associated in text" (p.71), and it is claimed that its results accord well with intuitions regarding the naturalness (or otherwise) of the relevant bigrams.

We can compare individuals' bigram profiles for a given question, as in Tables 1 and 2, and we can also consider how significant a given bigram is, by using the χ^2 test on the log-likelihood value.[2] The most general results, of course, concern the performance of the two cognitive types on the two question types, and it is to these that we now turn.

A number of factors are immediately visible in the bigram summary statistics, given in Table 3. For instance, DetHi exhibit more distinct bigrams overall, and considerably more significant bigrams. Considering only the latter, two scores were derived for each subject. One score corresponded to the number of significant bigrams on questions 1 and 3 combined (determinate items); the other to the number of significant bigrams on questions 2 and 4 combined (indeterminate items). A 2×2 analysis of variance

[2]It should be noted that the current paper corrects an error in Oberlander et al. (1996), concerning the measure of significantly associated bigrams. Though the amended results lead to slight quantitative differences, the qualitative results supporting our hypotheses from the previous paper are maintained in the current study.

Table 2: Bigram profile for subject C14 on question 4. C14 is a reasonable representative of the DetHi group. The bigrams above the horizontal line are significantly associated ($p < .05$).

- -2logλ	k(AB)	k(A$\sim$B)	k($\sim$AB)	k($\sim$A$\sim$B)	A	B
17.81	3	0	0	20	Exhaust	Merge
17.81	3	0	0	20	Merge	Inspect
10.16	4	4	0	15	Assume	Apply
4.59	2	6	0	15	Assume	CTA
3.32	3	1	5	14	Apply	Assume
2.20	1	0	7	15	Given	Assume
1.83	1	1	2	19	CTA	Exhaust
1.83	1	1	2	19	Inspect	Exhaust
0.53	1	3	2	17	Apply	Exhaust
0.53	2	6	6	9	Assume	Assume
0.21	1	1	7	14	CTA	Assume
0.21	1	1	7	14	Inspect	Assume

Table 3: Summary of bigram statistics. The table registers Assume's split into assume and Fullassume, while distinguishing subject-style (DetHi/Lo) and question-type (det/indet).

	DetLo.det	DetLo.indet	DetHi.det	DetHi.indet
number of distinct bigrams	35	37	44	46
number of significant bigrams (p<.05)	11	10	15	15
highly significant (p<.005)	8	7	10	13

was performed, in which factor 1 was a between-subject factor (DetHi, DetLo), and factor 2 corresponded to item determinacy (determinate, indeterminate). Factor 2 (within-subjects) was treated as a repeated measure in the analysis. The ANOVA summary table revealed a significant main effect for proof determinacy ($F = 4.93, df = (1, 20), p < .05$), but only an insignificant interaction between subject type and proof determinacy ($F = 0.07, df = (1, 20), p = .79$). All subjects manifested more significant bigrams on indeterminate questions. This indicates that the trends implied by the figures in the Table 3 are reliable. However, it does not reflect any differences in rule use by the two groups. Thus, we turn to an analysis of differences between the groups in the use of particular bigrams.

To indicate the nature of the revised profiles, Tables 4 and 5 show the bigram profiles for DetHi and DetLo on indeterminate questions. Taking the profiles for the two groups, we can consider differences both between-groups and within-groups. On indeterminate questions, we find that the bigrams assume Close, Given assume, assume Fullassume, Observe assume, NegIntro Univintro, Close assume, and Apply Observe are significant in DetHi proofs, but not in DetLo ones. Conversely, only the bigrams Inspect Merge and Given Apply are significant in DetLo proofs, but not in DetHi ones. The profiles are weakly but significantly correlated ($r = 0.167^{*}$).[3] When taking into account only those bigrams that are significantly associated in the profiles, the correlation is higher, but not significant ($r = 0.443, ns$).

On determinate questions, the bigrams assume Apply, CTA Observe, Close Fullassume, Observe CTA, Apply Fullassume, Inspect Implyelim, and Fullassume Fullassume are significant in DetHi proofs, but not in DetLo ones. Conversely, as with the indeterminate questions, there are only two bigrams significant in DetLo proofs, but not in DetHi ones: Inspect Merge and Exhaust Inspect. Here, the two subject groups' profiles are significantly correlated ($r = 0.537^{**}$). The correlation between significantly-associated bigrams is even stronger and still highly significant ($r = 0.809^{**}$).

This finding accords with the proofograms' indication that it is indeterminate questions which best discriminate the two subject groups. Recall that these are the questions in which the initial graphical situation is abstract, so that all concreteness must be introduced explicitly by the subjects.

This difference may be explained in terms of the difference in response demanded by the different types of question. Indeterminate questions require a greater degree of graphical abstraction, and are actually more difficult than determinate questions. When confronted with a determinate question, DetLo subjects perform using several rules that are associated in their proofs; DetHi subjects utilise more associated rules. When confronted with an indeterminate question, DetLo subjects lose some of the structure of their proofs, in that there are now fewer associated rules. DetHi subjects, however, maintain a higher level of structure, the number of significantly associated bigrams remaining the same.

Arguably, then, DetHi subjects maintain the structure of their proofs regardless of the type of question they are confronted with. By contrast, DetLo display more

[3] Correlations reported here are non-parametric (Spearman's ρ). Significance at the $p < .05$ level is denoted by *; significance at the $p < .001$ level by **.

Table 4: Bigram profile for DetHi subjects' indeterminate questions. The bigrams above the horizontal line are significantly associated ($p < .05$).

-2logλ	k(AB)	k(A$\sim$B)	k($\sim$AB)	k($\sim$A$\sim$B)	A	B
145.43	57	22	29	355	Fullassume	CTA
123.52	26	13	5	419	Exhaust	Merge
78.72	33	85	3	342	assume	Apply
69.47	17	11	12	423	Merge	Inspect
53.63	39	43	40	341	CTA	Fullassume
36.46	2	77	116	268	Fullassume	assume
26.16	14	68	7	374	CTA	Observe
26.01	12	27	17	407	Exhaust	Inspect
25.06	15	103	4	341	assume	Close
19.56	17	9	101	336	Given	assume
19.55	6	112	73	272	assume	Fullassume
17.95	11	3	107	342	Observe	assume
14.27	1	0	0	462	Negintro	Univintro
12.54	12	7	106	338	Close	assume
8.58	6	30	15	412	Apply	Observe
7.60	5	14	26	418	Inspect	Merge
6.68	1	117	20	325	assume	Observe
6.54	8	28	33	394	Apply	Exhaust
6.44	1	0	18	444	Existintro	Close
6.44	1	18	0	444	Close	Negintro
4.41	9	10	109	335	Inspect	assume
3.55	1	0	78	384	Univintro	Fullassume
3.40	14	22	104	323	Apply	assume
2.77	4	15	37	407	Close	Exhaust
2.77	4	15	37	407	Inspect	Exhaust
2.74	1	117	0	345	assume	Existintro
2.67	28	90	58	287	assume	CTA
2.52	1	18	78	366	Inspect	Fullassume
2.52	2	26	77	358	Merge	Fullassume
2.28	24	94	94	251	assume	assume
2.09	1	38	35	389	Exhaust	Apply
1.68	7	19	72	365	Given	Fullassume
1.54	1	13	85	364	Observe	CTA
1.29	1	27	40	395	Merge	Exhaust
0.81	5	74	36	348	Fullassume	Exhaust
0.69	8	28	71	356	Apply	Fullassume
0.67	2	17	77	367	Close	Fullassume
0.67	2	77	17	367	Fullassume	Close
0.28	1	13	18	431	Observe	Close
0.15	8	20	110	325	Merge	assume
0.10	8	74	33	348	CTA	Exhaust
0.06	1	13	40	409	Observe	Exhaust
0.03	10	108	31	314	assume	Exhaust
0.02	13	66	66	318	Fullassume	Fullassume
0.00	21	61	97	284	CTA	assume
0.00	2	24	34	403	Given	Apply

Table 5: Bigram profile for DetLo subjects' indeterminate questions. The bigrams above the horizontal line are significantly associated ($p < .05$).

-2logλ	k(AB)	k(A$\sim$B)	k($\sim$AB)	k($\sim$A$\sim$B)	A	B
134.90	67	33	15	221	Fullassume	CTA
78.70	19	13	5	299	Exhaust	Inspect
36.75	10	8	11	307	Inspect	Merge
34.48	1	99	54	182	Fullassume	assume
33.09	45	35	55	201	CTA	Fullassume
27.97	11	21	10	294	Exhaust	Merge
14.58	8	72	2	254	CTA	Observe
10.52	4	14	7	311	Given	Apply
8.95	6	48	5	277	assume	Apply
8.29	5	12	19	300	Merge	Inspect
6.29	7	10	48	271	Merge	assume
5.84	4	96	28	208	Fullassume	Exhaust
5.45	15	39	40	242	assume	assume
4.95	13	67	19	237	CTA	Exhaust
4.73	1	31	0	304	Exhaust	Unknown
3.63	1	0	54	281	Unknown	assume
3.31	6	12	49	269	Given	assume
3.31	6	12	49	269	Inspect	assume
3.29	4	6	51	275	Apply	assume
3.27	2	15	98	221	Merge	Fullassume
2.16	1	31	31	273	Exhaust	Exhaust
1.85	5	5	95	231	Apply	Fullassume
1.83	12	42	88	194	assume	Fullassume
1.83	1	5	10	320	Observe	Apply
1.83	8	10	92	226	Given	Fullassume
1.14	3	14	29	290	Merge	Exhaust
0.98	26	74	74	162	Fullassume	Fullassume
0.82	7	47	25	257	assume	Exhaust
0.51	2	98	8	228	Fullassume	Observe
0.30	1	5	31	299	Observe	Exhaust
0.22	1	5	81	249	Observe	CTA
0.10	14	66	41	215	CTA	assume
0.08	14	40	68	214	assume	CTA
0.05	2	16	30	288	Inspect	Exhaust
0.04	2	4	98	232	Observe	Fullassume
0.00	1	5	54	276	Observe	assume
0.00	1	9	31	295	Apply	Exhaust

structure in answers to determinate questions than indeterminate questions, the greater degree of complexity of the latter producing a different response profile.

5.4 Network modelling of final proofs

The proofogram and corpus analyses therefore support the abstraction ability hypothesis. On questions where the subject must construct the concrete graphic, it seems that DetHi subjects exhibit staging behaviour, and build their graphics incrementally, whereas DetLo subjects are prone to construct their concrete graphics in one go. The abstraction ability hypothesis seems plausible, since the 'stagers' are exactly those whom our main study showed benefit most from teaching with Hyperproof (Stenning, Cox and Oberlander, 1995).

But why do the subject groups' generated discourses diverge under these circumstances? As we have suggested, one tempting hypothesis comes from identifying our DetHi—DetLo distinction with the traditional visualiser—verbaliser distinction. If it's a matter of visual preference, then the diagrammatically capable DetHi subjects are just the visualisers, and therefore, they prefer to produce graphically-biased discourses, when the task allows. The diagrammatically less capable DetLo are the verbalisers, and hence prefer sententially-biased discourses—or at least, they do not show a strong preference for the graphical.

The alternative transmodal hypothesis is that DetHi subjects are better at multimodal reasoning, mixing sentential and graphical information. On this account, DetLo might be perfectly happy producing graphically-biased discourses, so long as they do not have to translate information back and forth between the graphical and the sentential.

One way of testing these competing views is to compare network models of the bigram transitions, for the two subject types. The transition networks in Figures 10 and 11 represent, respectively, DetHi and DetLo behaviour on indeterminate questions. In the networks, the area of a node represents the frequency with which a rule is used, while the thickness of links represents the probability of taking that exit arc, given that one is in the state denoted by the node.

First, consider the left-hand parts of the networks. Any proof must start from a Given step; now, it is clear that there are several ways in which the use of assume and Fullassume varies between the DetHi and DetLo groups. First, DetHi make more use of assume than DetLo, while the latter make around twice as much use of Fullassume than the former. DetLo subjects' favouring of Fullassume over assume certainly confirms that they are not 'stagers'; but in a sense, it also suggests that it is *they* who exhibit a preference for the graphical modality, moving straight into it and working entirely within it, rather than gradually transferring information into it from the sentential pane.

Notice also that around two-thirds of DetHi transitions from Given are to assume, and the rest go to Fullassume. By contrast, just one-third of DetLo transitions from Given go to assume; some 22% go to Apply, and 44% go straight to Fullassume. So, as well as favouring Fullassume over assume, DetLo subjects also often commence proof contruction by the use of Apply. This also helps to reduce subsequent interaction be-

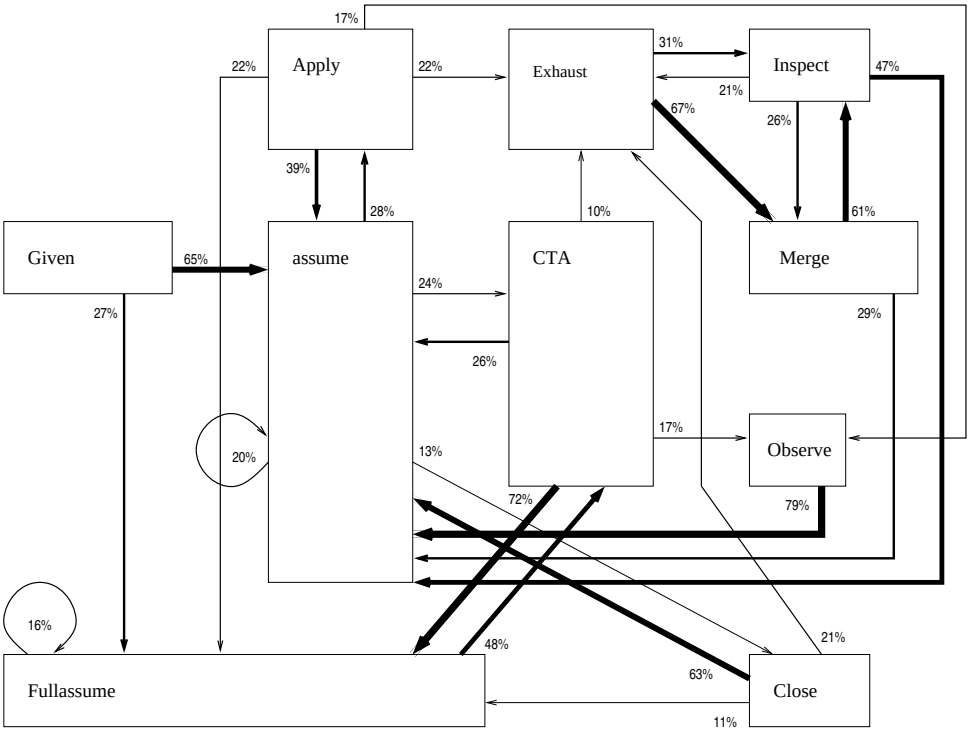

Figure 10: Transition network for DetHi behaviour on indeterminate questions. Nodes represent rules, and their areas represent the frequency at which that rule was invoked. Links represent the probability of transition from one rule to another; transitions at 10% probability and below are not shown.

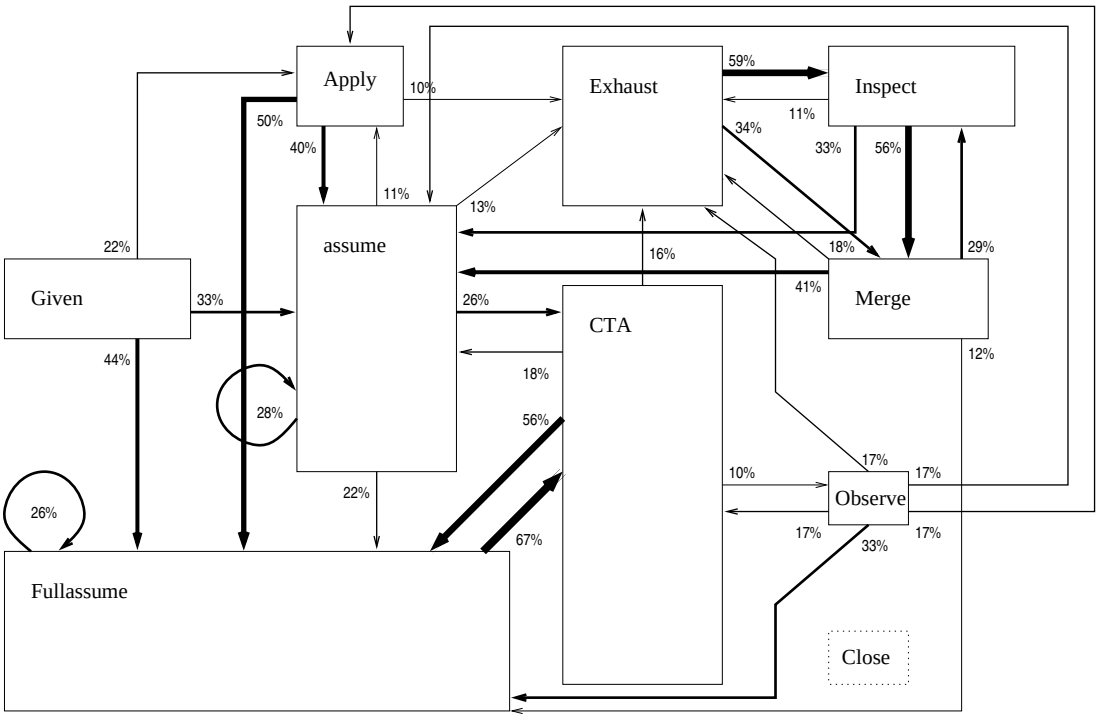

Figure 11: Transition network for DetLo behaviour on indeterminate questions. Note that Close is not visited at all.

tween the modalities, with case construction being performed only within the graphical window. Looking at Apply in more detail, it is apparent that DetHi are more likely to use it after assume (it accounts for 28% of their transitions out of assume, as opposed to just 11% amongst DetLo). And while both groups are as likely to use assume after Apply, DetLo are more than twice as likely as DetHi to go from Apply direct to Fullassume.

The assume Apply transition confirms that DetHi subjects tend to add information graphical window pane gradually, either by assumption, or by transfer from the sentential pane (via Apply).

Secondly, consider the top right portions of the networks. From Exhaust, DetLo are most likely to move first to Inspect, and from there to Merge. By contrast, DetHi are most likely to move first to Merge, and only from there to Inspect. Both Inspect and Merge find common information from the set of cases declared exhaustive by Exhaust. The difference is that Inspect provides this information sententially, and Merge does it graphically. It seems from the networks that DetHi find the graphical before the sentential, while DetLo find the sentential first, and only then carry out the graphical operation.[4]

Taking these two parts of the network together, it should be clear by now that this is not a simple matter of DetHi preferring the visual modality, not least since DetLo move to that modality more directly. Instead, the difference seems to be that the DetHi group *operate over* the graphical situations, frequently using a graphic as input, or guide, to further stages of proof construction. The DetLo, on the other hand, seem just to *output* graphics, without subsequently using them.

Finally, consider the bottom right hand corners of the networks. There is one very striking fact: DetLo subjects *barely ever* use the Close rule on these indeterminate questions. Close is used in proofs of inconsistency: students use it show that some assumptions contradict given information. And this means that while DetLo make considerable use of proofs of consistency—evidenced by the high frequency of use of CTA—they never proceed by showing that certain cases can be explicitly ruled out as inconsistent with existing premises and assumptions. Although it is specific to indeterminate questions, the aversion to proof by contradiction in these cases is intriguing. One possibility is that it may ultimately be related to findings concerning people's ability to verify or falsify general propositions (as in the four card selection problem, discussed, for instance, in Wason 1977). But another possibility is that this failure to use a strategy is only a failure in a particular context. Perhaps DetLo students are quite able at using the corresponding *reductio* strategies in sentential systems?

For the time being, however, it suffices that DetHi subjects do *not* show a simple preference for visual–graphical discourse. Rather, what distinguishes them is their greater tendency to *translate* between graphical and sentential modalities in *both* directions. It is tempting to say that they tend to produce multimodal discourses, instead of the monomodal discourses preferred by their DetLo colleagues. But of course, both groups produce 'multimodal' discourses; what distinguishes them is the relative

[4]Examination of the process logs confirms this: the order of occurrence of these rules in the final proof faithfully reflects the order in which the rules were applied in proof development.

frequency of occurrence of information from different modalities. DetHi produce discourses with frequent inter-mixing of different information types, while DetLo produce discourses, with the fewer transitions between diagrammatic and sentential modalities, which thus remain in essentially separate segments of the proof.

6 Conclusion

6.1 Summary

Computer-based logic proofs are a form of multimodal self-communication. If each line of a hyperproof is an utterance, then the proof as a whole functions as an organised discourse, possessing hierarchical structure, inter-utterance dependencies, and rhetorical structure. Proofograms provide one way of visualising the hierarchical structure, revealing patterns of abstraction use. Bigram analysis provides a more precise way of capturing varying patterns of rhetorical structure. Network analysis encapsulates the differences between discourse styles in a synoptic form.

It might seem that producing multimodal logic proofs is merely a form of 'unnatural' language processing. However, we believe that there are good reasons for studying these discourses. In the first place, results from our transfer tests indicate that the experience of being taught first-order logic generalises to other kinds of linguistic skill: from reasoning about proofs in a formal language (first-order logic) to reasoning in natural language. Logical discourse may be unnatural, but it is certainly connected to natural language discourse. In addition, computer-based protocol taking has allowed us to observe the structure and process of a type of discourse production in very considerable detail. The study therefore represents an approach that could be replicated for more natural forms of language.

We have here discussed two main questions concerning our corpus of hyperproofs. First, we have noted the extent to which the discourse structure of hyperproofs is influenced by question-type. Our findings here fit into the mainstream computational tradition of investigating how task constraints influence discourse structure. Secondly, we have noted the extent to which hyperproof discourses are influenced by individual differences in cognitive style, related to subjects' abstraction ability. We have suggested in particular that rules which handle abstraction and transmodal information transfer are used in characteristic ways by different groups of subjects. The evidence here is that abstraction ability tends to favour high frequency discourses, characteristic of transmodal reasoning, rather than the lower frequency discourses which would result from a simple preference for a single modality, visual or not.

6.2 Next steps

First, it is worth noting that the majority of work on individual differences in linguistics tends to focus on cross-linguistic comparisons, differences in language acquisition, or variations in language perception, rather than production. There have been some studies that relate to individual differences in the language production of adults (Cheung and Kemper, 1992; Suedfeld and Coren, 1992). This research reveals that there

are syntactical differences in language production, and differences in individuals' capacity to process complex sentences, with respect to the type of embedding and the amount of embedding found in the sentence. To the extent that Hyperproof sub-proof structures are analogous to embedding within individual sentences, our study of hyperproofs certainly seems to agree that the generation of structural complexity is subject to individual differences. An obvious next step, then, would be to move on to look at individual differences in more traditional language generation, using adaptations of the experimental methods and analytic tools we have deployed.

Secondly, we must bridge the gap between our DetHi/DetLo style distinction, and more traditional psychometric dimensions of individual difference, such as the visualiser–verbaliser or serial–holist dimensions. Typically, learning style studies that have investigated the visualiser–verbaliser distinction use psychometric instruments as the basis for classifying subjects. For example, the paper-folding test has been used by Mayer and Sims (1994) in a recent study of learning from computer-generated animation; and by Campagnoni and Ehrlich (1989) in a study of individual differences in hypertext navigation. However, it is currently unclear how strongly internal behaviour (as measured by paper-and-pencil psychometric tests) is related to external reasoning performance. In the current study, subjects were classified purely according to their performance on diagrammatic reasoning items. It remains to be seen how the performance measures used here map back onto the psychometric measures, and a new study is currently underway which we hope will throw some light on this issue.

Finally, however, even without a further study, we have a whole class of data which remains to be investigated. We have distinguished analysis of proof development process logs from analysis of final proofs; most of the discussion here has focussed on the latter. We therefore intend to go on to investigate the process logs, informed by the findings reported here. The key point is that the two types of subject mix their modalities to differing extents: one group produces higher frequency multimodal discourse than the other.

Our long term goal, then, is to develop a process account of diagrammatic thinking, to supercede the largely descriptive account given here. But even in the short term, a general lesson can be drawn: individual differences should prove to be a particularly exciting area of empirical study for those concerned with discourse generation.

Acknowledgements

The support of the Economic and Social Research Council for HCRC is gratefully acknowledged. The work was supported by UK Joint Councils Initiative in Cognitive Science and HCI, through grant G9018050 (Signal); and by NATO Collaborative research grant 910954 (Cognitive Evaluation of Hyperproof). The first author is supported by an EPSRC Advanced Fellowship. The paper extends and amends the results and discussion reported in Oberlander et al. (1996). Special thanks to Dave Barker-Plummer, Chris Brew, Tom Burke, John Etchemendy, and Mark Greaves.

References

Barwise, J. and Etchemendy, J. (1994). *Hyperproof*. Stanford: CSLI Publications.

Campagnoni, F. R. and Ehrlich, K. (1989). Information retreival using a hypertext-based help system. *ACM Transactions on Information Systems*, **7**, 271–291.

Cox, R., Stenning, K. and Oberlander, J. (1994). Graphical effects in learning logic: reasoning, representation and individual differences. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society*, pp237–242, Atlanta, Georgia, August.

Cheung, H. and Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics* **13**, 53–76.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**, 61–74.

Duran, R., Powers, D. and Swinton, S. (1987). Construct Validity of the GRE Analytical Test: A Resource Document. ETS Research Report 87–11. Princeton, NJ: Educational Testing Service.

Grosz, B. and Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, **12**, 175–204.

McKeown, K. (1985). *Text Generation*. Cambridge: Cambridge University Press.

Monaghan, P. (1995). A corpus-based analysis of individual differences in proof-style. MSc Thesis, Centre for Cognitive Science, University of Edinburgh.

Monaghan & Stenning (submitted) Effects of representational modality and thinking style on learning to solve reasoning problems. 20th annual meeeting of the Cognitive Science Society of America.

Oberlander, J., Cox, R. and Stenning, K. (1996). Proof styles in multimodal reasoning. In Seligman, J. and Westerståhl, D. (Eds.) *Language, Logic and Computation: Volume 1*, pp403–414. Stanford: CSLI Publications.

Oberlander, J., Cox, R., Mongahan, P., Stenning, K. and Tobin, R. (1996). Individual differences in proof structures following multimodal logic teaching. In *Proceedings of the 18th Annual Meeting of the Cognitive Science Society*, pp201–206, La Jolla, Ca., July 1996.

Schwarz, B. and Dreyfus, T. (1993). Measuring integration of information in multirepresentational software. *Interactive Learning Environments*, **3**, 177–198.

di Sessa, A. A. (1979). On 'learnable' representations of knowledge: A meaning for the computational metaphor. In Lochhead, J. and Clement, J. (Eds.) *Cognitive Process Instruction*. Philadelphia, PA: The Franklin Institute Press.

Snow, R. E. (1987). Aptitude complexes. In Snow, R. E. and Farr, M. J. (Eds.) *Aptitude, learning, and instruction, Volume 3: Conative and affective process analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stenning, K., Cox, R. and Oberlander, J. (1995). Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. *Language and Cognitive Processes*, **10**, 333–354.

Stenning, K. and Oberlander, J. (1991). Reasoning with Words, Pictures and Calculi: computation versus justification. In Barwise, J., Gawron, J. M., Plotkin, G. and Tutiya, S. (Eds.) *Situation Theory and Its Applications*, Volume 2, pp607–621. Chicago: Chicago University Press.

Stenning, K. and Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, **19**, 97–140.

Suedfeld, P. and Coren, S. (1992). Cognitive correlates of conceptual complexity. *Personality and Individual Differences* **13**, 1193–1199.

Wason, P. C. (1977). Self-contradictions. In Johnson-Laird, P. N. and Wason, P. C. (Eds.) *Thinking: Readings in Cognitive Science*, pp114–128. Cambridge: Cambridge University Press.