S. Aitken, P. Gray, T. Melham and M. Thomas, 'Interactive Proof Discovery: An Empirical Study of HOL Users', in *User Interface Design for Theorem Proving Systems: An International Workshop*, edited by P. Gray, Department of Computing Science, University of Glasgow (1995).

Interactive Proof Discovery: An Empirical Study of HOL Users

Stuart Aitken, Philip Gray, Tom Melham and Muffy Thomas
Department of Computing Science, University of Glasgow, Scotland
Email: {stuart, pdg, tfm, muffy}@dcs.gla.ac.uk

July 14, 1995

1 Introduction

One commonly cited obstacle to the widespread use of theorem provers is the poor quality of the user interface. Research projects which have addressed this problem have had mixed results. In the majority of cases, the designers of new interfaces to theorem provers have failed to draw upon the most relevant discipline: Human Computer Interaction. The design of these interfaces is carefully considered, but is generally informed by the personal experience of the designer and their intuitions.

The ITP project aims to apply the methods of HCI to the problem of designing interfaces to interactive theorem provers. Of particular relevance are task modelling and analysis (Norman, 1988) which can identify the gulfs of execution and evaluation which hinder interaction. Nielsen (1986) presents a linguistic account of interaction by asserting that the different levels at which user-system interaction may be described correspond to the lexical, syntactic and semantic levels of linguistic activity. Such multi-level views of interaction are relevant to our study of user-prover interaction. Theories about the usability and effectiveness of representations (Green, 1989, 1991) and studies of the nature of programming (Wiedenbeck, 1985)(Davis, 1991) are also important.

2 Interacting with HOL

The current study aims to produce a broad characterisation of the behaviour of experienced users of the HOL proof system (Gordon and Melham, 1993). HOL is an interactive prover in the LCF style. Theorem proving usually takes place in a backwards fashion, i.e. the top goal is broken down into subgoals by the application of tactics. Suboals may then be proven by existing theorems or may be further decomposed. The standard interface to HOL, the subgoal package, supports the recording of the proof state and its history. The user may execute a new tactic, generating a new proof state, or may back up to a previous proof state.

After either of these actions the subgoal package prints the new goal and assumptions. The HOL session is usually run in one window and a second window is used to edit a text file which contains the text of the tactics which have been entered into HOL.

In attempting to describe the user-computer interaction with HOL we can identify several levels at which the description can be made:

- The logical level, where the description is in terms of logical concepts.
- The abstract interaction level, where the description is in terms of the information objects which enable the user to communicate with the system. This level includes executing tactics and editing tactic texts.
- The concrete interaction level, where the interaction is described in terms of actions on input devices, for example, keypresses.

The model we propose describes a "proof-as-programming" view. Other views are possible, proof might be viewed as structure editing in a system such as ALF or proof might be viewed as process of selection. In these cases the contents of each layer of the model will differ (Aitken, Gray, Melham and Thomas, 1995).

3 An Experimental Investigation

An experiment was carried out in order to obtain an objective account of the activities of HOL users. Seven subjects were asked to prove a conjecture and to think-aloud while doing so. Their speech was recorded on audio tape. A record of all inputs and outputs to the HOL prover was kept. After the proof attempt subjects were asked a series of questions about each step of the proof. The questions concerned the cues used in selecting a proof step and the structure of the proof beyond the proof step level.

This investigation provided information on the number of interactions with HOL, the time taken to complete the proof, the number of proof steps, the number of tactics in each proof step, and the logical organisation of the proof as viewed by the subject. The importance of the current goal, current assumptions and past subgoals, as viewed by the subjects, was also documented.

4 Results

The results show a large variation in the time required to find a proof (7 min 36 sec to 33 min 32 sec). There was also a large variation in the number of interactions with HOL and in the rate of interaction. The average number of tactics in a proof step (i.e. in one interaction with HOL) was 1.2 and the average number of tactics in a logical context was 2.92. The current goal and the current assumptions were stated to be cues in 79% and 68% of instances respectively, while past subgoals were stated to be cues in 12% of instances (averaged across all trials).

The results show the user should not be restricted to inputing single tactics, and further, they indicate that an interface might support the larger grained organisation of proofs. The number of interactions with HOL does not appear to be a good measure of performance. The time taken to find the proof seems to be a more reliable measure.

5 Error analysis

The records of user input to HOL were analysed to identify input errors. Errors were classified into syntax errors and retrieval errors. Syntax errors were further broken down into type errors (accounting for 6.3% of all interactions) and parse errors (2.4%). Retrieval errors included input of a non-existent, or incorrect, tacic name (2.4%) and the input of a non-existent theorem name (4.3%). Input errors occurred in 15.4% of all user interactions with HOL. Here also, a large variation in errors between subjects was apparent, with some subjects making only one or two errors.

A further type of retrieval error was the failure to find or to recall an already proven theorem. This error led to several subjects re-proving a theorem which they could have found in a HOL library. For three subjects who did this, the extra actions accounted for 14.1% of the total interactions and 8% of the total time required to find the proof.

6 The Interaction Model

Analysis of the think-aloud protocol yielded a set of twelve activities which subjects engaged in during the proof attempts. These could be classified as occurring at the logic level or at the interaction level, as proposed in the three layer proof as programming model. The empirical study suggests one important modification to the original model: subjects manipulated the tactic script in order to modify the proof. This suggests that editing the tactic script must be an activity at the abstract interaction level. There may also be further implications for our description of both the abstract and concrete interaction levels.

7 Conclusions

The empirical investigation reported here has shown the wide range of user performance. It has allowed us to identify and quantify the sources of error which occur during proof attempts. The study has enabled us to fill in the model we have of how users interact with HOL in greater detail. We believe these results can be used to inform the design of better interfaces to HOL and other tactic based theorem provers.

References

Aitken, S., Gray, P., Melham, T., and Thomas, M. (1995) A Study of User Activity in Interactive Theorem Proving *Submitted to the Journal of Symbolic Computation*

Davis, S.P. (1991) Characterising the program design activity: neither strictly top-down nor globally opportunistic. *Behaviour and Information Technology* 1991, Vol. 10, No. 3, pp. 173-190

Gordon, M. J. C., Melham, T. F., eds. (1993) *Introduction to HOL: A theorem proving environment for higher order logic.* Cambridge University Press 1993.

Green, T.R.G. (1989) Cognitive dimensions of notations. *in People and Computers V*, eds. Sutcliffe, A. and Macaulay, L. Cambridge University Press 1989 pp. 443-460

Green, T. R. G. (1991) Describing information artifacts with cognitive dimensions and structure maps. in *People and Computers IV.* eds. Daiper, D. and Hammond, N., Cambridge University Press 1991 pp. 297-315.

Nielsen, J. (1986) A virtual protocol model for human-computer interaction. *International Journal of Man-Machine Studies* Vol. 24 pp. 31-312

Norman, D. A. (1988) *The Psychology of Everyday Things.* Basic Books 1988.

Wiedenbeck, S. (1985) Novice/expert differences in programming skill. *International Journal of Man-Machine Studies* Vol. 23, No. 4, pp. 383-390