It is Time to Mechanize Programming Language Metatheory*

Benjamin C. Pierce[1], Peter Sewell[2], Stephanie Weirich[1], and Steve Zdancewic[1]

[1] Department of Computer and Information Science, University of Pennsylvania
[2] Computer Laboratory, University of Cambridge

Abstract. How close are we to a world in which mechanically verified software is commonplace? A world in which theorem proving technology is used routinely by both software developers and programming language researchers alike? One crucial step towards achieving these goals is mechanized reasoning about language metatheory. The time has come to bring together the theorem proving and programming language communities to address this problem. We have proposed the POPLMark challenge as a concrete set of benchmarks intended both for measuring progress in this area and for stimulating discussion and collaboration. Our goal is to push the boundaries of existing technology to the point where we can achieve mechanized metatheory for the masses.

1 Mechanized Metatheory for the Masses

One significant obstacle to achieving the goal of verified software is reasoning about the languages in which the software is written. Without formal models of programming languages, it is impossible to even state, let alone prove, meaningful properties of software or tools such as compilers. It is therefore essential that we develop appropriate tools for modeling programming languages and mechanically checking their metatheoretic properties. This infrastructure should provide facilities for proving properties of operational semantics, program analyses (such as type checkers), and program transformations (such as optimization and compilation).

Many proofs about programming languages are straightforward, long, and tedious, with just a few interesting cases. Their complexity arises from the management of many details rather than from deep conceptual difficulties; yet small mistakes or overlooked cases can invalidate large amounts of work. These effects are amplified as languages scale: it becomes very hard to keep definitions and proofs consistent, to reuse work, and to ensure tight relationships between theory and implementations. Automated proof assistants offer the hope of significantly easing these problems. However, despite much encouraging progress in recent years and the availability of several mature tools (ACL2 [15], Coq [2], HOL [10],

* This position paper is adapted from the introduction to the POPLMark Challenge paper [1].

Isabelle [20], Lego [16], NuPRL [5], PVS [21], Twelf [22], etc.), their use is still not commonplace.

We believe that the time is right to join the efforts of the two communities, bringing developers of automated proof assistants together with a large pool of eager potential clients—programming language designers and researchers. In particular, we intend to answer two questions:

1. What is the current state of the art in formalizing language metatheory and semantics? What can be recommended as best practices for groups (typically not proof-assistant experts) embarking on formalized language definitions, either small- or large-scale?
2. What improvements are needed to make the use of tool support commonplace? What can each community contribute?

Over the past six months, we have attempted to survey the landscape of proof assistants, language representation strategies, and related tools. Collectively, we have applied automated theorem proving technology to a number of problems, including proving transitivity of the algorithmic subtype relation in Kernel $F_{\leq}$ [4, 3, 6], proving type soundness of Featherweight Java [14], proving type soundness of variants of the simply typed λ-calculus and $F_{\leq}$, and a substantial formalization of the behavior of TCP, UDP, and the Sockets API. We have carried out these case studies using a variety of object-language representation strategies, proof techniques, and proving environments. We have also experimented with lightweight tools designed to make it easier to define and typeset both formal and informal mathematics. Although experts in programming language theory, we were (and are) relative novices with respect to computer-aided proof.

Our conclusion from these experiments is that the relevant technology has developed *almost* to the point where it can be widely used by language researchers. We seek to push it over the threshold, making the use of proof tools common practice in programming language research—mechanized metatheory for the masses.

Tool support for formal reasoning about programming languages would be useful at many levels:

1. *Machine-checked metatheory.* These are the classic problems: type preservation and soundness theorems, unique decomposition properties for operational semantics, proofs of equivalence between algorithmic and declarative variants of type systems, etc. At present such results are typically proved by hand for small to medium-size calculi, and are not proved at all for full language definitions. We envision a future in which the papers in conferences such as *Principles of Programming Languages (POPL)* and the *International Conference on Functional Programming (ICFP)* are routinely accompanied by mechanically checkable proofs of the theorems they claim.
2. *Use of definitions as oracles for testing and animation.* When developing a language implementation together with a formal definition one would like to use the definition as an oracle for testing. This requires tools that can

decide typing and evaluation relationships, and they might differ from the tools used for (1) or be embedded in the same proof assistant. In some cases one could use a definition directly as a prototype.

3. *Support for engineering large-scale definitions.* As we move to full language definitions—on the scale of Standard ML [17] or larger—pragmatic "software engineering" issues become increasingly important, as do the potential benefits of tool support. For large definitions, the need for elegant and concise notation becomes pressing, as witnessed by the care taken by present-day researchers using informal mathematics. Even lightweight tool support, without full mechanized proof, could be very useful in this domain, e.g. for sort checking and typesetting of definitions and of informal proofs, automatically instantiating definitions, performing substitutions, etc.

Our goal is to stimulate progress in this area by providing a common framework for comparing alternative technologies. Our approach has been to design a set of challenge problems, dubbed the POPLMark Challenge [1], chosen to exercise many aspects of programming languages that are known to be difficult to formalize: variable binding at both term and type levels, syntactic forms with variable numbers of components (including binders), and proofs demanding complex induction principles. Such challenge problems have been used in the past within the theorem proving community to focus attention on specific areas and to evaluate the relative merits of different tools; these have ranged in scale from benchmark suites and small problems [23, 11, 7, 13, 9, 19] up to the grand challenges of Floyd, Hoare, and Moore [8, 12, 18]. We hope that our challenge will have a similarly stimulating effect.

The POPLMark problems are drawn from the basic metatheory of a call-by-value variant of System $F_{<:}$ [3, 6], enriched with records, record subtyping, and record patterns. Our challenge provides an informal-mathematics definition of its type system and operational semantics and outline proofs of some of its metatheory. This language is of moderate scale—neither a toy calculus nor a full-blown programming language—to keep the work involved in attempting the challenges manageable.[3] The intent of this challenge is to cover a broad range of issues that arise in the *formalization* of programming languages; of course there are many programming language *features*, such as control-flow operators, state, and concurrency, not covered by our sample problem, but we believe that a system capable of formalizing the POPLMark problems should be able to formalize those features as well. Nevertheless, we expect this challenge set to grow and evolve as the community addresses some problems and discovers others.

The initial POPLMark challenge has already been disseminated to a wide audience of theorem prover and programming language researchers. We are in the process of collecting and evaluating solutions. Those results, along with related information about mechanized metatheory, will be available on our web site.[4]

[3] Our challenges therefore explicitly address only points (1) and (2) above; we regard the pragmatic issues of (3) as equally critical, but it is not yet clear to us how to formulate a useful challenge problem at this larger scale.

[4] http://www.cis.upenn.edu/proj/plclub/mmm/

In the longer run, we hope that this site, and the corresponding mailing list [5] will serve as a forum for promoting and advancing the current best practices in proof assistant technology and making this technology available to the broader programming languages community and beyond. We encourage researchers to try out the POPLMark Challenge using their favorite tools and send us their solutions for inclusion in the web site.

References

1. Brian E. Aydemir, Aaron Bohannon, Matthew Fairbairn, J. Nathan Foster, Benjamin C. Pierce, Peter Sewell, Dimitrios Vytiniotis, Geoffrey Washburn, Stephanie Weirich, and Steve Zdancewic. Mechanized metatheory for the masses: The POPLmark challenge. In *Theorem Proving in Higher Order Logics, 18th International Conference*, Oxford, UK, August 2005.
2. Yves Bertot and Pierre Castran. *Interactive Theorem Proving and Program Development*, volume XXV of *EATCS Texts in Theoretical Computer Science*. Springer-Verlag, 2004.
3. Luca Cardelli, Simone Martini, John C. Mitchell, and Andre Scedrov. An extension of System F with subtyping. *Information and Computation*, 109(1–2):4–56, 1994. Summary in TACS '91 (Sendai, Japan, pp. 750–770).
4. Luca Cardelli and Peter Wegner. On understanding types, data abstraction, and polymorphism. *Computing Surveys*, 17(4):471–522, December 1985.
5. Robert L. Constable, Stuart F. Allen, Mark Bromley, Rance Cleaveland, James F. Cremer, Robert W. Harper, Douglas J. Howe, Todd B. Knoblock, Paul Mendler, Prakash Panangaden, James T. Sasaki, and Scott F. Smith. *Implementing Mathematics with the NuPRL Proof Development System*. Prentice-Hall, Englewood Cliffs, NJ, 1986.
6. Pierre-Louis Curien and Giorgio Ghelli. Coherence of subsumption: Minimum typing and type-checking in $F_{\leq}$. *Mathematical Structures in Computer Science*, 2:55–91, 1992. Also in C. A. Gunter and J. C. Mitchell, editors, *Theoretical Aspects of Object-Oriented Programming: Types, Semantics, and Language Design*, MIT Press, 1994.
7. Louise A. Dennis. Inductive challenge problems, 2000. http://www.cs.nott.ac.uk/ lad/research/challenges.
8. Robert W. Floyd. Assigning meanings to programs. In J. T. Schwartz, editor, *Mathematical Aspects of Computer Science*, volume 19 of *Proceedings of Symposia in Applied Mathematics*, pages 19–32, Providence, Rhode Island, 1967. American Mathematical Society.
9. I.P. Gent and T. Walsh. CSPLib: a benchmark library for constraints. Technical report, Technical report APES-09-1999, 1999. Available from http://csplib.cs.strath.ac.uk/. A shorter version appears in the Proceedings of the 5th International Conference on Principles and Practices of Constraint Programming (CP-99).
10. M. J. C. Gordon and T. F. Melham, editors. *Introduction to HOL: a theorem proving environment for higher order logic*. Cambridge University Press, 1993.
11. Ian Green. The dream corpus of inductive conjectures, 1999. http://dream.dai.ed.ac.uk/dc/lib.html.

[5] `poplmark@lists.seas.upenn.edu`

12. Tony Hoare. The verifying compiler: A grand challenge for computing research. *J. ACM*, 50(1):63–69, 2003.
13. Holger Hoos and Thomas Stuetzle. Satlib. http://www.intellektik.informatik.tu-darmstadt.de/SATLIB/ .
14. Atsushi Igarashi, Benjamin Pierce, and Philip Wadler. Featherweight Java: A minimal core calculus for Java and GJ. In *ACM SIGPLAN Conference on Object Oriented Programming: Systems, Languages, and Applications (OOPSLA)*, October 1999. Full version in ACM Transactions on Programming Languages and Systems (TOPLAS), 23(3), May 2001.
15. Matt Kaufmann, J. Strother Moore, and Panagiotis Manolios. *Computer-Aided Reasoning: An Approach*. Kluwer Academic Publishers, 2000.
16. Zhaohui Luo and Robert Pollack. The LEGO proof development system: A user's manual. Technical Report ECS-LFCS-92-211, University of Edinburgh, May 1992.
17. Robin Milner, Mads Tofte, Robert Harper, and David MacQueen. *The Definition of Standard ML,* Revised edition. MIT Press, 1997.
18. J. Strother Moore. A grand challenge proposal for formal methods: A verified stack. In Bernhard K. Aichernig and T. S. E. Maibaum, editors, *Formal Methods at the Crossroads. From Panacea to Foundational Support, 10th Anniversary Colloquium of UNU/IIST, Lisbon, Portugal*, volume 2757 of *Lecture Notes in Computer Science*, pages 161–172. Springer, 2002.
19. J. Strother Moore and George Porter. The apprentice challenge. *ACM Trans. Program. Lang. Syst.*, 24(3):193–216, 2002.
20. Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. *Isabelle/HOL: A Proof Assistant For Higher-Order Logic*, volume 2283 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.
21. Sam Owre, Sreeranga Rajan, John M. Rushby, Natarajan Shankar, and Mandayam K. Srivas. PVS: Combining specification, proof checking, and model checking. In *International Conference on Computer Aided Verification (CAV), New Brunswick, New Jersey*, volume 1102 of *Lecture Notes in Computer Science*, pages 411–414. Springer-Verlag, July 1996.
22. Frank Pfenning and Carsten Schürmann. System description: Twelf — A metalogical framework for deductive systems. In Harald Ganzinger, editor, *Automated Deduction, CADE 16: 16th International Conference on Automated Deduction, Trento, Italy, July 7-10, 1999, Proceedings*, volume 1632 of *Lecture Notes in Artificial Intelligence*, pages 202–206. Springer-Verlag, 1999.
23. Geoff Sutcliffe and Christian Suttner. The TPTP problem library. *Journal of Automated Reasoning*, 21(2):177–203, 1998.