Initial Experiments with External Provers and Premise Selection on HOL Light Corpora

Cezary Kaliszyk
University of Innsbruck
Innsbruck, Austria
cezarykaliszyk@gmail.com

Josef Urban*
Radboud University Nijmegen
Nijmegen, The Netherlands
Josef.Urban@gmail.com

Abstract

This paper reports our initial experiments with using external ATP and premise selection methods on some corpora built with the HOL Light system. The testing is done in three different settings, corresponding to those used earlier for evaluating such methods on the Mizar/MML corpus. This is intended to provide the first estimate about the usefulness of such external reasoning and AI systems for solving problems over HOL Light and its libraries.

1 Motivation

Usage of external first-order ATPs like Vampire [22], E [24], SPASS [30], and recently also SMT solvers like Z3 [6] for ITP-based (large-theory) formalization has been developed quite significantly in the recent decade. Particularly in the Isabelle community, the Isabelle/Sledgehammer [3, 2] bridge to such external tools is getting increasingly popular. This helps to further develop various parts of the technology involved. ATPs have recently gained the ability to quickly load large theories over large signatures and work with them. Methods for automated selection of relevant knowledge and for proof guidance are actively developed, together with specialized automated systems targeted at particular mathematical domains. Formats and translation methods handling more formalization-friendly foundations are being defined, and metasystems that decide which ATP, translation method, strategy, parallelization, and premises to use to solve a given problem with limited resources are being designed. Cooperation of humans and computers over large corpora of formal knowledge is an interesting field, allowing exploration of new AI systems and combinations of different AI techniques that can attempt to encode concepts like analogy and intuition, and rigorously evaluate their usefulness. Perhaps not only Hilbert and Turing, but also the formality-opposing and intuition-oriented Poincaré[1] [21] would have been interested to learn about the new "semantic AI paradise" of such large corpora of formal and computer-understandable mathematics (from which we do not intend to be expelled).

The HOL Light [10] system is probably the first among the existing well-known ITPs which has integrated and extensively used a general ATP procedure, the MESON tactic [11]. Hurd has developed and benchmarked early bridges [12, 13] between HOL and external systems, and his Metis system [14] has also become a significant part of the Isabelle/Sledgehammer bridge to ATPs [20]. Using the very detailed Otter/Ivy [18] proof objects, Harrison also later implemented a bridge from HOL Light to Prover9 [17].

HOL Light however does not yet have a general bridge to large-theory ATP/AI methods, similar to Isabelle/Sledgehammer or MizAR [27, 28], which would attempt to automatically solve a new goal by selecting the relevant knowledge from the large library and running (possibly

*Supported by the NWO projects "MathWiki" and "Learning2Reason"

[1] 2012 is not just the year of Turing [9], but also of Poincaré, whose ideas about creativity and invention involving random, intuition-guided exploration confirmed by critical evaluation quite correspond to what AI metasystems like MaLARea [29] try to emulate in the large-theory formal setting.

several) external ATPs on such (possibly several alternative) premise selections. HOL Light seems to be a natural candidate for adopting such methods, because of the amount of work already done in this direction mentioned above, and also thanks to HOL Light's foundational closeness to Isabelle/HOL. Also, it seems that thanks to the Flyspeck project [8], HOL Light is becoming less of only a "single, very knowledgable formalizer" tool, and also getting increasingly used as a "tool for interested mathematicians" (particularly Vietnamese[2]) that know the large libraries much less and have much less experience with crafting their own targeted proof tactics. For such ITP users it is good to provide a small number of strong methods that allow fast progress.

The work reported here consists of several experiments intended to give an initial information about the usefulness of building such a bridge for HOL Light. The evaluation tries to follow the pattern introduced for Mizar/MML in [25, 26]:

1. Evaluate the ATP efficiency on simple ITP steps ("by" in Mizar, MESON in HOL Light).
2. Evaluate the ATP efficiency on re-proving whole theorems in the libraries from their (as exact as possible) proof dependencies.
3. Evaluate the ATP efficiency on proving whole theorems when the premises are chosen from the large library by AI (heuristic, learning) methods.

In general, the work is much less complete and polished than the similar work done for Mizar and Isabelle, and also in much rawer state than the finished work by Harrison and Hurd mentioned above. The first issues are now efficiency and encoding of the export from HOL Light to FOL, and also the compatibility and alignment of the data that we use for re-proving of whole theorems from their dependencies, and from trained advice. But we hope that obtaining the initial results and reporting them and the problems encountered could attract some interest and expert advice with such technical issues, so that the bridges are finished (not necessarily by us) sooner, and the HOL Light and Flyspeck large mathematical corpora become available to ATP and AI research in the same way as the Mizar and Isabelle corpora.

Apart from the attempt to inspire in this section, the rest of the paper is organized as follows. Section 2 explains how problems in the above categories were prepared, building on the work of Harrison and Adams. Section 3 reports the experiments and results, and Section 4 concludes.

2 Problem Exports

All three kinds of export described below initially rely on using parts of the MESON tactic for exporting the problems to the TPTP FOF format. MESON is based on the model elimination method invented by Loveland [15] and later combined with Prolog-like search tree [16]. The implementation of MESON in HOL-Light first applies a number of tactics that transform the HOL goal to a FOL goal (or multiple goals). The FOL goal is then passed to an ML procedure that returns a proof which is later replayed using HOL Light proof steps. This often means that multiple ATP problems are created from one such MESON call (due to pre-processing), and the formulas are already skolemized. The MESON-based export can become very slow for larger problems, probably depending on the use of higher-order features in the problems. However the export still provides a sufficient number of problems for the first evaluation. Our plan is to later switch to TFF1-based (extension of FOF with types and polymorphism) export [4], for which for example Why3 [7] already has a usable translation tool to FOF by Andrei Paskevich.

[2] `http://weyl.math.pitt.edu/hanoi2009/Participants/`

Even though MESON uses CNF, we encode it as FOF to get around some syntactic issues with TPTP, and also because large-theory systems have a longer tradition of working on the FOF level. In each problem, apart from the TPTP formulas themselves, we keep as a comment also the original HOL Light goal and assumptions, and their first-order encoding used internally by the MESON tactic. This is intended to help debugging the translation, and we invite interested readers to check that we proceed (at least for most problems) correctly. For the problems created from the full theorems, we additionally keep the name of the theorem inside the problem.

2.1 Exporting problems from the original MESON calls

We have first hooked the exporting code into the MESON (and ASM_MESON) tactic itself, to get a large number of TPTP problems corresponding to the HOL Light problems on which the MESON tactic is used. There does not seem to be any issue running this export, so we ran it fully on the core HOL Light, HOL Multivariate, and Flyspeck corpora. The problems created are available online.[3] From core HOL Light this yields 2057 problems, from HOL Multivariate 12428, and from Flyspeck 19634. Their average, minimum, and maximum sizes are shown in Table 1

Table 1: MESON problems

Corpus	Problems	Average size	Minimum size	Maximum size
Core HOL Light	2057	8.1	2	64
HOL Multivariate	12428	17.7	2	226
Flyspeck	19634	12.7	2	132
Total	34119	14.3	2	226

These problem sizes (which should additionally be considered as the CNF sizes) are lower than in the problems corresponding to the "by" (atomic justifications) in Mizar. There the dependent types with Horn-like adjective mechanisms are a very significant part of the automation, and particularly in more advanced theories their TPTP encoding can produce tens of formulas.

The HOL Light type system also does not have the additional features like type classes that complicate the problems for Isabelle, and it seems that explicit encoding of the type system in the MESON export is entirely avoided. One guard against type-related unsoundness in MESON seems to be exhaustive instantiation of different polymorphic variants into different (untyped first-order) symbols, including equality (this is probably not difficult in a tableau-based system like MESON). Our export to TPTP currently merges all polymorphic instances of equality into the one standard FOL equality, which can make some TPTP problems unsound.[4]

If this all is correct, then it is a bit surprising that the problem sizes (which on our corpora often correlates with first-order ATP difficulty) of the "full-scale ATP" MESON problems that actually appear in the HOL Light corpora look comparable to the problems originating from the Mizar's "limited-by-design" and "obvious-inference-only" [5, 23] atomic justifications. This also hints that HOL Light users might be able to do bigger steps if using external ATPs.

[3] `http://mizar.cs.ualberta.ca/~mptp/hh/tptp.tgz`

[4] We became aware of this issue thanks to the PAAR workshop reviews, and so far we have not tried to measure the influence of such unsoundness in the problems. Some related quantification is available in the work of Hurd and Meng and Paulson.

2.2 Exporting theorem problems

The second interesting set of problems is on the "theorem" level of ITP libraries. This level seems to be quite similar in the major ITPs: "theorem" is typically not corresponding to what mathematicians call a theorem, but it is rather a self-sufficient lemma with a formal proof of tens to hundreds lines that can be useful in other formal proofs and hence should be named and exported. Since the ITP proofs can be longer, proving such theorems fully automatically is typically a challenge, which makes such problems suitable for ATP benchmarks, challenges, and competitions.

In Mizar and in Isabelle (done by Blanchette in so far unpublished work) the corresponding ATP problems for theorems can be produced by collecting the dependencies (premises) from the proofs (by suitable tracking mechanisms), and then translating the $Premises \vdash Theorem$ statement to first-order logic. It seems that HOL Light does not provide (at least not out-of-the box) such high-level tracking of dependencies, however there is recent work by Adams in exporting HOL Light to HOL Zero [1] (with cross-verification as the main motivation) that does (also) high-level dependency tracking. We have used these data (dependency table) as follows:

1. First we attempted to synchronize the theorem names used by Adams with our work (there can be different naming conventions). This was an iterative process that we ended when there were 55 remaining differences out of 1782 theorems (possibly caused also by small version difference).

2. Then for each HOL Light theorem, we have replaced the default "prove" function with a function that first looks up the dependencies (in the external dependency table), filters out those that (for whatever reason) do not exist in the current environment, and calls the MESON exporting code described above for the problem $Dependencies \vdash Theorem$

This is problematic not just because of the possible dependency incompatibilities. The MESON export of some problems can take very long time (one problem that we left overnight took more than five hours), and create very large files (several megabytes). Thousands of formulas are no longer a problem for existing large-theory ATP techniques, but the processing time inside HOL Light makes experiments impractical. This was the main reason why the re-proving experiments based on the dependency information about whole proofs were limited to 1178 HOL Light theorems for which we get the TPTP translation before we encounter such slowdowns. This inefficiency seems to be caused by MESON's exhaustive treatment of polymorphism, which is not a problem for normal solving of small HOL Light tasks with MESON, but does not scale well to large numbers of premises. We either need to a more efficient implementation of this code in HOL Light,[5] or as already mentioned, we might just try to entirely switch from the MESON export to the TFF1 export which will likely avoid major optimizations (those can be done while translating from TFF1 to FOF).

The 1178 theorems give rise to 1993 problems after the MESON export, also available online.[6] The average number of formulas in them is 46, the maximum is 2469, which means that some of these problem should benefit from premise-selection methods.

2.3 Exporting theorem problems with premise selection

Given the large libraries that have been built with HOL Light, the interesting ATP/AI task is to prove new theorems without having to manually select the relevant premises. ATP problems of

[5] After reading the first version of this paper, John Harrison started to look at this issue.

[6] `http://mizar.cs.ualberta.ca/~mptp/hh/theorems.tar.gz`

this kind are created for Mizar/MML by consistent translation of the whole MML to TPTP, and then letting premise selection algorithms find the most relevant premises for a given theorem t from the large set of t-allowed premises (typically those theorems and definitions that were already available when t was being proved, expressed, e.g., as TPTP include files).

Currently, the MESON export that we use invents specific symbols for each problem, and it is not clear to us if it can be easily modified to translate each HOL Light theorem separately to FOF, so that such separate theorem translations could be later consistently combined into large ATP problems. Again, the TFF1 layer should make this possible.

So in order to do the initial test of how good ATP/AI methods can be when using the whole available HOL Light theory at each point, we put the premise selection directly inside HOL Light. This is done as follows:

1. First we train (using the SNoW system in naive Bayes mode) a premise selector on the proof dependency data by Adams. Similar to such training on MML data, HOL Light symbols are used for the input-feature representation of the proved theorems, and the necessary proof dependencies are used as the output features (labels) for the learning. The resulting standalone premise selector is now also accessible online.[7] The 10-fold cross-validation (i.e.: training on 9/10 and testing on 1/10 of the data) gives so far on average about 43% cover of the needed dependencies in the first 100 hits. For Mizar/MML this is about 70% [25]. Some of this difference can be caused by bugs in the data processing, but it is also possible that just using symbols for characterizing formulas is weaker on the HOL Light corpora, and using other features (e.g., all formula (sub)terms) will be useful.

2. Then for each HOL Light theorem, we have replaced the default "prove" function with a function that extracts the symbols from the theorem[8] and sends them as a query to the premise selector. The premise selector replies with a list of theorem names ordered by their expected relevance for the goal, from which we again filter out those that do not exist in the current environment. Then we take the N most relevant of the remaining recommended premises, and call the MESON exporting code for the problem $RecommendedPremises \vdash Theorem$.

This solution is a bit similar to how Isabelle/Sledgehammer selects premises, which is also done internally on Isabelle terms rather than externally on their TPTP representation.[9] However, Isabelle/Sledgehammer now uses a manually taylored relevance filter [19].

The MESON translation with a higher number of premises is however again a bottleneck, so we limit the number of advised premises to 60, and do the export to TPTP only for N equal to 10, 20, 30, 40, 50, and 60. For the same efficiency reasons (and to have the same set of theorems for all N) we also stop the export after producing problems for 964 theorems, resulting in about 1700 TPTP problems. The problem numbers can slightly differ for different N, for example for `NUMPAIR_INJ_LEMMA` only one problem is created by the MESON export when using 20 premises, but two problems are created when using 60 premises. Table 2 summarizes the six datasets, which are also available online.[10]

[7] `http://mws.cs.ru.nl/~urban/holdata/isab_adv_demo.html`

[8] We could have used Adams' data for this too, but this way we can use the premise selector also on conjectures that are not in Adams' data.

[9] The Isabelle TPTP export now uses consistent symbol naming, so external premise selectors can be already tested on Isabelle data. Some initial (so far unpublished) experiments with the MaLARea system have been started by Blanchette and Urban.

[10] `http://mizar.cs.ualberta.ca/~mptp/hh/advised_theorems10.tar.gz`, and so on for 20 to 60.

Table 2: Theorem problems (after MESON export) with advised premises

Premises	Theorems	Problems	Avg. size/theorem	Avg. size/problem	Max size
10	964	1649	86.32	50.46	882
20	964	1662	136.84	79.37	922
30	964	1680	196.44	112.72	1097
40	964	1683	250.80	143.65	1470
50	964	1687	339.06	193.75	1781
60	964	1687	462.47	264.27	2181

For the higher values of N the average problem sizes reach values that can further benefit from internal ATP large-theory methods developed in the past years. We should also note that in some cases there is currently a total mismatch between the symbols on which the advisor was trained, and the symbols that we extract from the theorem that is to be advised (this can be due to various omissions in the processing and synchronizing of the symbol names with Adams' data). So again, these problems should be considered as an initial experiment rather than the best of what the current premise selection techniques can achieve.

3 Experiments

We use Vampire 1.8 and E 1.4 on the problems. All ATPs are run with 5s time limit on an Intel Xeon X5650 2.67GHz server with 24GB RAM and 12MB CPU cache. Each problem is always assigned one CPU.

3.1 Using external ATPs to prove the calls to MESON

Table 3 shows the results of running Vampire and E on the MESON problems. The solutions are also online.[11] The problems turn out to be very easy, and the average number of needed Vampire premises is quite low in comparison to the average problem size. One problem (tptp19150.p) has been found countersatisfiable by E. After manually adding an extensionality axiom for functions, both E and Vampire can prove it. It is possible that some knowledge about extensionality of basic constants of HOL is hard-coded in MESON. So far we have not decided to add this axiom to the translation of every problem to FOL, because it is explicitly present in some of the problems. The low number of premises that are actually needed for the proof is also a bit suspicious, but some problems seem to be really trivial (for example, asking to prove that $c \neq c$), which might be partially due to the MESON preprocessing and splitting into multiple problems. There are some harder problems, for example multivariate/tptp13687.p took E to generate 152441 clauses and process 15889 of them.

3.2 Using external ATPs to prove theorems

Table 4 shows the results of running Vampire and E on the 1993 theorem problems. The solutions are again online.[12] Many problems are reported countersatisfiable by E, which can mean that we are missing some proof dependencies, processing them in a wrong way, or that the completeness of the MESON export is limited (note that for the MESON problems in the previous section

[11] `http://mizar.cs.ualberta.ca/~mptp/hh/meson_results.tar.gz`
[12] `http://mizar.cs.ualberta.ca/~mptp/hh/theorems_results.tar.gz`

Table 3: ATP results on problems created from the MESON calls

Corpus	Problems	Avg. size	V-proved (%)	E-proved (%)	Avg. V-premises
Core HOL Light	2057	8.1	2055 (99.9%)	2057 (100%)	2.91
HOL Multivariate	12428	17.7	12422 (100%)	12393 (99.7%)	3.40
Flyspeck	19634	12.7	19592 (99.8%)	19621 (99.9%)	2.15
Total	34119	14.3	34069 (99.9%)	34071 (99.9%)	2.60

we are only using those which succeeded in HOL Light). The average number of premises that Vampire needed for the 519 proofs went a bit higher than in the previous section, but it is still suspiciously low. The overall success rate is 26%, however as mentioned above, these are theorems from the beginning of the core HOL Light corpus, and the export (and thus also likely the problems) get harder later.

Table 4: ATP results on the 1993 theorem problems

Problems	Avg. size	V-proved (%)	E-proved (%)	Avg. V-premises	E-CounterSat (%)
1993	46	519 (26%)	517 (26%)	4.6	876 (44%)

3.3 Using external ATPs to prove theorems with premise selection

Table 5 shows the results of running Vampire and E on the six differently advised batches of theorem problems. The solutions are again online.[13] Advising more premises helps quite a lot, and particularly Vampire is good in handling the larger problems. The advised theorem problems are a subset of those from previous section, so the result of 455 proved by Vampire in the 60-advised batch compares quite well to the 519 proved in the previous section. This seems encouraging, but again, modulo all the possible bugs and imperfections that might be involved.

Table 5: ATP results on the advised theorem problems

Premises	Problems	Avg. size	V-proved (%)	E-proved (%)	Avg. V-premises	E-ContrSat (%)
10	1649	50.46	225 (13.6%)	221 (13.4%)	3.33	352 (21.3%)
20	1662	79.37	294 (17.7%)	288 (17.3%)	4.22	175 (10.5%)
30	1680	112.72	350 (20.1%)	340 (20.2%)	4.55	95 (5.7%)
40	1683	143.65	387 (23%)	354 (21%)	4.86	41 (2.4%)
50	1687	193.75	427 (25.3%)	362 (21.5%)	5.03	29 (1.7%)
60	1687	264.27	455 (27%)	367 (21.7%)	5.13	29 (1.7%)

4 Conclusion and Future Work

What we did seems straightforward. Inside HOL Light we encoded the translation to TPTP using MESON, introduced some bookkeeping to keep track of available theorems, implemented the calls to the premise advisor, and hooked these functions to suitable places. Outside HOL Light, most of the work was in researching how to use and synchronize Adams' dependency

[13] `http://mizar.cs.ualberta.ca/~mptp/hh/advised_theorems_results.tar.gz`

data. Training the basic naive Bayes premise selection and providing the trained advisor is now a standard technology done already many times.

We might have made mistakes in tweaking the HOL Light data and functions for our purpose, and one reason for this workshop paper is to expose any serious bugs to better-informed eyes. However even if there were serious issues in exporting the problems in the TPTP format, it still seems that doing what we are attempting to do is quite well-researched today, and the large-theory ATP/AI is out there, ready to be applied to the HOL Light corpora. We have not done any major sanity checking yet w.r.t. the proofs that we obtain. One issue that we became aware of (after the reviews of the first version of the paper) is our use of one global equality predicate, which together with MESON's removal of type guards can lead to unsound translations. We have not measured the influence of such unsoundness yet. However, HOL Light has methods that import MESON and Ivy proofs, and a lot of relevant work has been done on proof import with Isabelle/Sledgehammer using Metis.

Future work has been mentioned several times. Probably the lowest hanging fruit is to export all theorems from the corpora in the TFF1 format. This could solve the efficiency and symbol-consistency problems, and allow us to use premise selection externally rather than internally.

5 Acknowledgments

Mark Adams gave us his HOL Light proof export data for HOL Zero, which made it possible to attempt the reproving and learning experiments for larger proofs. Tom Hales helped to start the MESON exporting work at CICM 2011, and his interest as a leader of Flyspeck motivates us. Piotr Rudnicki has made his Mizar workstation available for the experiments. Thanks to John Harrison for providing MESON (and HOL Light) and discussing some topics. Obviously, none of these people have any responsibility for possible bugs and errors that we might have committed when exploring their work and attempting to use it. Thanks also to the PAAR 2012 referees, who provided a number of helpful comments and suggestions.

References

[1] Mark Adams. Introducing HOL Zero - (extended abstract). In Komei Fukuda, Joris van der Hoeven, Michael Joswig, and Nobuki Takayama, editors, *ICMS*, volume 6327 of *LNCS*, pages 142–143. Springer, 2010.

[2] Jasmin Christian Blanchette, Sascha Böhme, and Lawrence C. Paulson. Extending Sledgehammer with SMT solvers. In Nikolaj Bjørner and Viorica Sofronie-Stokkermans, editors, *CADE*, volume 6803 of *LNCS*, pages 116–130. Springer, 2011.

[3] Jasmin Christian Blanchette, Lukas Bulwahn, and Tobias Nipkow. Automatic proof and disproof in Isabelle/HOL. In Cesare Tinelli and Viorica Sofronie-Stokkermans, editors, *FroCoS*, volume 6989 of *LNCS*, pages 12–27. Springer, 2011.

[4] Jasmin Christian Blanchette and Andrei Paskevich. TFF1: The TPTP typed first-order form with rank-1 polymorphism. Available online at `http://www21.in.tum.de/~blanchet/tff1spec.pdf`.

[5] Martin Davis. Obvious logical inferences. In Patrick J. Hayes, editor, *IJCAI*, pages 530–531. William Kaufmann, 1981.

[6] Leonardo Mendonça de Moura and Nikolaj Bjørner. Z3: An Efficient SMT Solver. In C. R. Ramakrishnan and Jakob Rehof, editors, *TACAS*, volume 4963 of *LNCS*, pages 337–340. Springer, 2008.

[7] Jean-Christophe Filliâtre and Claude Marché. The Why/Krakatoa/Caduceus platform for deductive program verification. In Werner Damm and Holger Hermanns, editors, *CAV*, volume 4590 of *LNCS*, pages 173–177. Springer, 2007.

[8] Thomas C. Hales. Introduction to the Flyspeck project. In Thierry Coquand, Henri Lombardi, and Marie-Françoise Roy, editors, *Mathematics, Algorithms, Proofs*, volume 05021 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2005.

[9] Thomas C. Hales. Mathematics in the age of the Turing machine. *Lecture Notes in Logic*, 2012. to appear; `http://www.math.pitt.edu/~thales/papers/turing.pdf`.

[10] John Harrison. HOL Light: A tutorial introduction. In Mandayam K. Srivas and Albert John Camilleri, editors, *FMCAD*, volume 1166 of *LNCS*, pages 265–269. Springer, 1996.

[11] John Harrison. Optimizing Proof Search in Model Elimination. In M. McRobbie and J.K. Slaney, editors, *Proceedings of the 13th International Conference on Automated Deduction*, number 1104 in Lecture Notes in Artificial Intelligence, pages 313–327. Springer-Verlag, 1996.

[12] Joe Hurd. Integrating Gandalf and HOL. In Yves Bertot, Gilles Dowek, André Hirschowitz, C. Paulin, and Laurent Théry, editors, *TPHOLs*, volume 1690 of *LNCS*, pages 311–322. Springer, 1999.

[13] Joe Hurd. An LCF-style interface between HOL and first-order logic. In Andrei Voronkov, editor, *CADE*, volume 2392 of *LNCS*, pages 134–138. Springer, 2002.

[14] Joe Hurd. First-order proof tactics in higher-order logic theorem provers. In Myla Archer, Ben Di Vito, and César Muñoz, editors, *Design and Application of Strategies/Tactics in Higher Order Logics (STRATA 2003)*, number NASA/CP-2003-212448 in NASA Technical Reports, pages 56–68, September 2003.

[15] Donald W. Loveland. Mechanical theorem proving by model elimination. *Journal of the ACM*, 15(2):236–251, April 1968.

[16] Donald W. Loveland. *Automated Theorem Proving: A Logical Basis*. North-Holland, Amsterdam, 1978.

[17] William McCune. Prover9 and Mace4. `http://www.cs.unm.edu/~mccune/prover9/`, 2005–2010.

[18] William McCune and Olga Shumsky Matlin. Ivy: A Preprocessor and Proof Checker for First-Order Logic. In M. Kaufmann, P. Manolios, and J. Strother Moore, editors, *Computer-Aided Reasoning: ACL2 Case Studies*, number 4 in Advances in Formal Methods, pages 265–282. Kluwer Academic Publishers, 2000.

[19] Jia Meng and Lawrence C. Paulson. Lightweight relevance filtering for machine-generated resolution problems. *J. Applied Logic*, 7(1):41–57, 2009.

[20] Lawrence C. Paulson and Kong Woei Susanto. Source-level proof reconstruction for interactive theorem proving. In Klaus Schneider and Jens Brandt, editors, *TPHOLs*, volume 4732 of *LNCS*, pages 232–245. Springer, 2007.

[21] Henri Poincaré. *The foundations of science: Science and hypothesis, The value of science, Science and method*. The Science Press, New York, 1913.

[22] Alexandre Riazanov and Andrei Voronkov. The design and implementation of VAMPIRE. *AI Commun.*, 15(2-3):91–110, 2002.

[23] Piotr Rudnicki. Obvious Inferences. *Journal of Automated Reasoning*, 3(4):383–393, 1987.

[24] Stephan Schulz. E - A Brainiac Theorem Prover. *AI Commun.*, 15(2-3):111–126, 2002.

[25] Josef Urban. MPTP - Motivation, Implementation, First Experiments. *Journal of Automated Reasoning*, 33(3-4):319–339, 2004.

[26] Josef Urban. MPTP 0.2: Design, implementation, and initial experiments. *J. Autom. Reasoning*, 37(1-2):21–43, 2006.

[27] Josef Urban, Piotr Rudnicki, and Geoff Sutcliffe. ATP and presentation service for Mizar formalizations. *J. Autom. Reasoning*, 50:229–241, 2013.

[28] Josef Urban and Geoff Sutcliffe. Automated reasoning and presentation support for formalizing mathematics in Mizar. In Serge Autexier, Jacques Calmet, David Delahaye, Patrick D. F. Ion, Laurence Rideau, Renaud Rioboo, and Alan P. Sexton, editors, *AISC/MKM/Calculemus*, volume 6167 of *LNCS*, pages 132–146. Springer, 2010.

[29] Josef Urban, Geoff Sutcliffe, Petr Pudlák, and Jiří Vyskočil. MaLARea SG1 - Machine Learner for Automated Reasoning with Semantic Guidance. In Alessandro Armando, Peter Baumgartner, and Gilles Dowek, editors, *IJCAR*, volume 5195 of *LNCS*, pages 441–456. Springer, 2008.

[30] Christoph Weidenbach, Dilyana Dimova, Arnaud Fietzke, Rohit Kumar, Martin Suda, and Patrick Wischnewski. SPASS Version 3.5. In Renate A. Schmidt, editor, *CADE*, volume 5663 of *LNCS*, pages 140–145. Springer, 2009.