

HOL(y)Hammer: Online ATP Service for HOL Light

Cezary Kaliszyk and Josef Urban

Abstract. HOL(y)Hammer is an online AI/ATP service for formal (computer-understandable) mathematics encoded in the HOL Light system. The service allows its users to upload and automatically process an arbitrary formal development (project) based on HOL Light, and to attack arbitrary conjectures that use the concepts defined in some of the uploaded projects. For that, the service uses several automated reasoning systems combined with several premise selection methods trained on all the project proofs. The projects that are readily available on the server for such query answering include the recent versions of the Flyspeck, Multivariate Analysis and Complex Analysis libraries. The service runs on a 48-CPU server, currently employing in parallel for each task 7 AI/ATP combinations and 4 decision procedures that contribute to its overall performance. The system is also available for local installation by interested users, who can customize it for their own proof development. An Emacs interface allowing parallel asynchronous queries to the service is also provided. The overall structure of the service is outlined, problems that arise and their solutions are discussed, and an initial account of using the system is given.

1. Introduction and Motivation

HOL Light [14] is one of the best-known interactive theorem proving (ITP) systems. It has been used to prove a number of well-known mathematical theorems¹ and as a platform for formalizing the proof of the Kepler conjecture targeted by the Flyspeck project [13]. The whole Flyspeck development, together with the required parts of the HOL Light library consisted of about 14000 theorems as of June 2012, growing to about 19000 theorems as of August 2013. Motivated by the development of large-theory automated theorem proving [17, 29, 36, 42] and its growing use for ITPs like Isabelle [30] and Mizar [40, 41], we have recently implemented translations from HOL Light to ATP (automated theorem proving) formats, developed a number of premise-selection techniques² for HOL Light, and experimented with the strongest and most orthogonal combinations of the premise-selection methods and various ATPs. This initial work, described in [25], has shown that 39% of the (June 2012) 14185 Flyspeck theorems could be proved in a push-button mode (without any high-level advice and user interaction) in 30 seconds of real time on a fourteen-CPU workstation. More recent work on the AI/ATP methods have raised this performance to 47% [24].

The experiments that we did emulated the Flyspeck development (when the user always knows all the previous proofs³ at a given point, and wants to prove the next theorem), however they were all done in an offline mode which is suitable for such experimentally-driven research. The ATP problems were created in large batches using different premise-selection techniques and different ATP

¹<http://www.cs.ru.nl/~freek/100/>

²Premise selection [3, 27] is the problem of selecting suitable premises (theorems, definitions, lemmas, etc.) from a large formal library for proving a new conjecture over such library.

³The Flyspeck processing order is used to define precisely what “previous” means. See [25] for details.

encodings (untyped first-order [33], polymorphic typed first-order [5], and typed higher-order [12]), and then attempted with different ATPs (17 in total) and different numbers of the most relevant premises. Analysis of the results interleaved with further improvements of the methods and data have gradually led to the current strongest combination of the AI/ATP methods.

This strongest combination now gives to a HOL Light/Flyspeck user a 47% chance (when using 14 CPUs, each for 30s) that he will not have to search the library for suitable lemmas and figure out the proof of the next toplevel theorem by himself. For smaller (proof-local) lemmas such likelihood should be correspondingly higher. To really provide this strong automated advice to the users, the functions that have been implemented for the experiments need to be combined into a suitable AI/ATP tool. Our eventual goal (from which we are of course still very far) should be an easy-to-use service, which in its online form offers to formal mathematics (done here in HOL Light, over the concepts defined formally in the libraries) what services like Wolfram Alpha offer for informal/symbolic mathematics. Some expectations, linked to the recent success of the IBM Watson system, are today even higher⁴. Indeed, we believe that developing stronger and stronger AI/ATP tools similar to the one presented here is a necessary prerequisite providing the crucial semantic understanding/reasoning layer for building larger Watson-like systems for mathematics that will (eventually) understand (nearly-)natural language and (perhaps reasonably semanticized versions/alternatives of) \LaTeX . The more user-friendly and smarter such AI/ATP systems become, the higher also the chance that mathematicians (and exact scientists) will get some nontrivial benefits⁵ from encoding mathematics (and exact science) directly in a computer-understandable form.

This paper describes such an AI/ATP service based on the formal mathematical corpora like Flyspeck developed with HOL Light. The service – HOL(y)Hammer⁶ (HH) – is now available as a public online system⁷ instantiated for several large HOL Light libraries, running on a 48-CPU server spawning for each query by default 7 different AI/ATP combinations and four decision procedures. We first describe in Section 2 the static (i.e., not user-updatable) problem solving functions developed in the first simplified version of the service for the most interesting example of Flyspeck. This initial version of the service allowed the users to experiment with ATP queries over the fixed June 2012 version of Flyspeck for which the AI/ATP components had been gradually developed over several months in the offline experiments described in [25]. Section 3 then discusses the issues and solutions related to running the service for multiple libraries and their versions at once, allowing the users also to submit a new library to the server or to update an existing library and all its AI/ATP components. Section 4 shows examples of interaction with the service, using web, Emacs, and command-line interfaces. The service can be also installed locally, and trained on user’s private developments. This is described in Section 5. Section 6 concludes and discusses future work.⁸

2. Description of the Problem Solving Functions for Flyspeck

The overall problem solving architecture without the updating functions is shown in Figure 1. Since Flyspeck is the largest and most interesting corpus on which this architecture was developed and tested, we use the Flyspeck service as a running example in this whole section. The service receives a query (a conjecture to prove, possibly with local assumptions) generated by one of the clients/front-ends (Emacs, web interface, HOL session, etc.). If the query produces a parsing (or type-checking) error, an exception is raised, and an error message is sent as a reply. Otherwise the parsed query is processed in parallel by the (time-limited) AI/ATP combinations and the native HOL Light decision

⁴See for example Jonathan Borwein’s article: <http://theconversation.edu.au/if-i-had-a-blank-checke-id-turn-ibms-watson-into-a-maths-genius-1213>

⁵Formal verification itself is of course a great benefit, but its cost has been so far too high to attract most mathematicians.

⁶See [44] for an example of future where AIs turn into deities.

⁷<http://colol2-c703.uibk.ac.at/hh/>

⁸This paper is an extended version of [20].

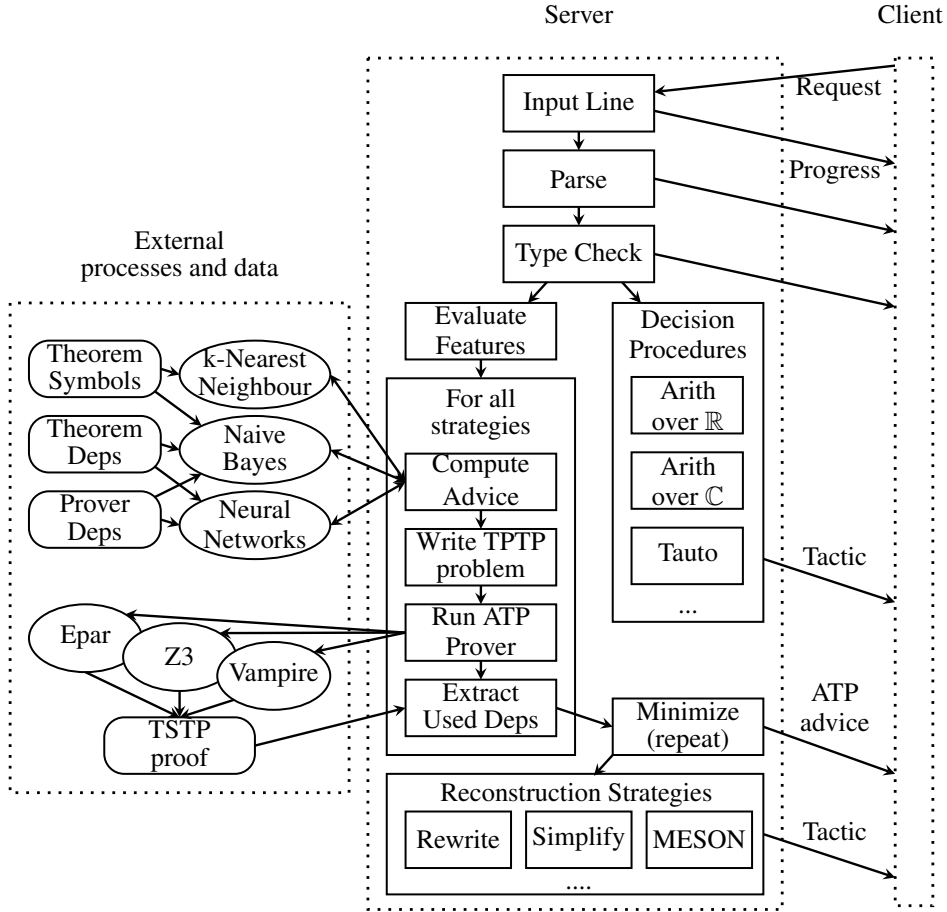


FIGURE 1. Overview of the problem solving functions

procedures (each managed by its forked HOL Light process, and terminated/killed by the master process if not finished within its global time limit). Each of the AI/ATP processes computes a specific feature representation of the query, and sends such features to a specific instance of a premise advisor trained (using the particular feature representation) on previous proofs. Each of the advisors replies with a specific number of premises, which are then translated to a suitable ATP format, and written to a temporary file on which a specific ATP is run. The successful ATP result is then (pseudo-)minimized, and handed over to the combination of proof-reconstruction procedures. These procedures again run in parallel, and if any of them is successful, the result is sent as a particular tactic application to the frontend. In case a native HOL Light decision procedure finds a proof, the result (again a particular tactic application) can be immediately sent to the frontend. The following subsections explain this process in more detail.

2.1. Feature Extraction and Premise Selection

Given a (formal) mathematical conjecture, the selection of suitable premises from a large formal library is an interesting AI problem, for which a number of methods have been tried recently [24, 27, 36]. The strongest methods use machine learning on previous problems, combined in various ways with heuristics like SInE [17]. To use the machine learning systems, the previous problems

have to be described as training examples in a suitable format, typically as a set of (input) features characterizing a given theorem, and a set of labels (output features) characterizing the proof of the theorem. Devising good feature/label characterizations for this task is again an interesting AI problem (see, e.g. [41]), however already the most obvious characterizations like the conjecture symbols and the names of the theorems used in the conjecture's proof are useful. This basic scheme can be extended in various ways; see [25] for the feature-extraction functions (basically adding various subterm and type-based characteristics) and label-improving methods (e.g., using minimized ATP proofs instead of the original *Flyspeck* proofs whenever possible) that we have so far used for *HOL Light*. For example, the currently most useful version of the characterization algorithm would describe the *HOL* theorem `DISCRETE_IMP_CLOSED`:⁹

```
Vs:real^N→bool e.
  &0 < e ∧ (∀x y. x IN s ∧ y IN s ∧ norm(y - x) < e ⇒ y = x)
  ⇒ closed s
```

by the following set of strings that encode its symbols and normalized types and terms:

```
"real", "num", "fun", "cart", "bool", "vector_sub", "vector_norm",
"real_of_num", "real_lt", "closed", "_0", "NUMERAL", "IN", "=", "&0",
"&0 < Areal", "0", "Areal", "Areal^A", "Areal^A - Areal^A",
"Areal^A IN Areal^A→bool", "Areal^A→bool", "_0", "closed Areal^A→bool",
"norm (Areal^A - Areal^A)", "norm (Areal^A - Areal^A) < Areal"
```

Here, `real` is a type constant, `IN` is a term constructor, `Areal^A→bool` is a normalized type, `Areal^A` its component type, `norm (Areal^A - Areal^A) < Areal` is a normalized atomic formula, and `Areal^A - Areal^A` is its normalized subterm.

On average, for each of our feature-extraction methods there are in total about 30000 possible conjecture-characterizing features extracted from the theorems in the *Flyspeck* development. The output features (labels) are in the simplest setting just the names of the *Flyspeck* theorems¹⁰ extracted from the proofs with a modified (proof recording [19]) *HOL Light* kernel. These features and labels are (for each extraction method) serially numbered in a stable way (using hashtables), producing from all *Flyspeck* proofs the training examples on which the premise selectors are trained. The learning-based premise selection methods currently used are those available in the *SNoW* [8] sparse learning toolkit (most prominently sparse naive Bayes), together with a custom implementation [24] of the distance-weighted k -nearest neighbor (k -NN) learner [10]. Training a particular learning method on all (14185) characterizations extracted from the *Flyspeck* proofs takes from 1 second for k -NN (a lazy learner that essentially just loads all the 14185 proof characterizations) and 6 seconds for naive Bayes using labels from minimized ATP proofs, to 25 seconds for naive Bayes using the labels from the original *Flyspeck* proofs.¹¹ The trained premise selectors are then run as daemons (using their server modes) that accept queries in the language of the numerical features over which they have been trained, producing for each query their ranking of all the labels, corresponding to the available *Flyspeck* theorems.

Given a new conjecture, the first step of each of the forked *HOL Light* AI/ATP managing process is thus to compute the features of the conjecture according to a particular feature extraction method, compute (using the corresponding hashtable) the numerical representation of the features, and send these numeric features as a query to the corresponding premise-selection daemon. The daemon replies within a fraction of a second with its ranking, the exact speed depending on the

⁹http://mws.cs.ru.nl/~mtp/hol-flyspeck/trunk/Multivariate/topology.html#DISCRETE_IMP_CLOSED

¹⁰In practice, the *Flyspeck* theorems are further preprocessed to provide better learning precision, for example by splitting conjunctions and detecting which of the conjuncts are relevant in which proof. Again, see [25] for the details. The number of labels used for the June 2012 *Flyspeck* version with 14185 theorems is thus 16082.

¹¹The original *Flyspeck* proofs are often using theorems that are in some sense redundant, resulting in longer proof characterizations (and thus longer learning). This is typically a consequence of using larger building blocks (e.g., decision procedures, drawing in many dependencies) when constructing the ITP proofs.

learning method and the size of the feature/label sets. This ranking is translated back (using the corresponding table) to the ranking of the HOL Light theorems. Each of the AI/ATP combinations then uses its particular number (optimized so that the methods in the end complement each other as much as possible) of the best-ranked theorems, passing them together with the conjecture to the function that translates such set of HOL Light formulas to a suitable ATP format.

2.2. Translation to ATP Formats and Running ATPs

As mentioned in Section 1, several ATP formalisms are used today by ATP and SMT systems. However the (jointly) most useful proof-producing systems in our experiments turned out to be E [32] version 1.6 (run under the Epar [39] strategy scheduler), Vampire [26] 2.6, and Z3 [9] 4.0. All these systems accept the TPTP untyped first-order format (FOF). Even when the input formalism (the HOL logic [31] - polymorphic version of Church’s simple type theory) and the output formalism (TPTP FOF) are fixed, there are in general many methods [4] to translate from the former to the latter, each method providing different tradeoffs between soundness, completeness, ATP efficiency, and the overall (i.e., including HOL proof reconstruction) efficiency. The particular method chosen by us in [25] and used currently also for the service is the polymorphic tagged encoding [4]. To summarize, the higher-order features (such as lambda abstraction, application) of the HOL formulas are first encoded (in a potentially incomplete way) in first-order logic (still using polymorphic types), and then type tags are added in a way that usually guarantees type safety during the first-order proof search.

This translation method is in general not stable on the level of single formulas, i.e., it is not possible to just keep in a global hashtable the translated FOF version for each original HOL formula, as done for example for the MizAR ATP service [22, 40]. This is because a particular optimization (by Meng and Paulson [28]) is used for translating higher-order constants, creating for each such constant c a first-order function that has the minimum arity with which c is used in the particular set of HOL formulas that is used to create the ATP (FOF) problem. So once the particular AI/ATP managing process advises its N most-relevant HOL Light theorems for the conjecture, this set of theorems and the conjecture are as a whole passed to the translation function, which for each AI/ATP instance may produce a slightly different FOF encoding on the formula level. The encoding function is still reasonably fast, taking fractions of a second when using hundreds of formulas, and still has the property that both the FOF formula names and the FOF formulas (also those inferred during the ATP proof search) can typically be decoded back into the original HOL names and formulas, allowing later HOL proof reconstruction.

Each AI/ATP instance thus produces its specific temporary file (the FOF ATP problem) and runs its specific ATP system on it with its time limit. The time limit is currently set globally to 30 seconds for each instance, however (as usual in strategy scheduling setups) this could be made instance-specific too, based on further analysis of the time performance of the particular instances. Vampire and Epar already do such scheduling internally: the current version of Epar runs a fixed schedule of 14 strategies, while Vampire runs a problem-dependent schedule using for each problem a varied number of strategies. Assuming one strategy for Z3 and on average eight strategies for Vampire, this now means that using 10-CPU parallelization results in about 100 different proof-data/feature-extraction/learning/premise-slicing/ATP-strategy instantiations tried by the online service within the 30 seconds of the real time allowed for each query. Provided sufficient complementarity of such instantiations and enough CPUs, this significantly raises the overall power of the service [24, 25].

2.3. The AI/ATP Combinations Used

An example of the 25 initially used combinations of the machine learner, proof data, number of top premises used, the feature extraction method, and the ATP system is shown in Table 1. The proof data are either just the data from the ATP proofs, or a combination of the ATP proofs with

the original HOL proofs. The ATP proofs (ATP0, ..., ATP3) are created by a particular MaLAREa-style [42] (i.e., re-using the proofs found in previous iteration for further learning) iteration of the experimenting, possibly preferring either the Vampire or Epar proofs (V_pref, E_pref). The HOL proofs are obtained by slightly different versions of the HOL proof recording. The HOL/ATP combinations typically use the HOL proof only when the ATP proof is not available, see [25] for details. The `standard` feature extraction method combines the formula’s symbols, standard-normalized subterms and normalized types into its feature vector. The standard normalization here means that each variable name is in each formula replaced by its normalized HOL type. Types are normalized by renaming their polymorphic variables with de Bruijn indices. The `all-vars-same` and `all-vars-diff` methods respectively just rename all formula variables into one common variable, or keep them all different. This obviously influences the concept of similarity used by the machine learners (see [25] for more discussion). The 40-NN and 160-NN learners are k -nearest-neighbors, run with $k = 40$ and $k = 160$. The particular combination of the AI/ATP is chosen by computing in a greedy fashion the set of methods with the greatest coverage of the solvable `Flyspeck` problems. This changes often, whenever some of the many components of this AI architecture get improved. For example, after the more recent strengthening of the premise-selection and ATP components described in [24], and the addition of multiple developments and functions for their dynamic update described in Section 3, the number of AI/ATP combinations run for a single query was reduced to 7.

TABLE 1. The 25 AI/ATP combinations used by the initial `Flyspeck` service

Learner	Proofs	Premises	Features	ATP
Bayes	ATP2	0092	standard	Vampire
Bayes	ATP2	0128	standard	Epar
Bayes	ATP2	0154	standard	Epar
Bayes	ATP2	1024	standard	Epar
Bayes	HOL0+ATP0	0512	all-vars-same	Epar
Bayes	HOL0+ATP0	0128	all-vars-diff	Vampire
Bayes	ATP1	0032	standard	Z3
Bayes	ATP1_V_pref	0128	all-vars-diff	Epar
Bayes	ATP1_V_pref	0128	standard	Z3
Bayes	HOL0+ATP0	0032	standard	Z3
Bayes	HOL0+ATP0	0154	all-vars-same	Epar
Bayes	HOL0+ATP0	0128	standard	Epar
Bayes	HOL0+ATP0	0128	standard	Vampire
Bayes	ATP1_E_pref	0128	standard	Z3
Bayes	ATP0_V_pref	0154	standard	Vampire
40-NN	ATP1	0032	standard	Epar
160-NN	ATP1	0512	standard	Z3
Bayes	HOL3+ATP3	0092	standard	Vampire
Bayes	HOL3+ATP3	0128	standard	Epar
Bayes	HOL3+ATP3	0154	standard	Epar
Bayes	HOL3+ATP3	1024	standard	Epar
Bayes	ATP3	0092	standard	Vampire
Bayes	ATP3	0128	standard	Epar
Bayes	ATP3	0154	standard	Epar
Bayes	ATP3	1024	standard	Epar

2.4. Use of Decision Procedures

Some goals are hard for ATPs, but are easy for the existing decision procedures already implemented in HOL Light. To make the service more powerful, we also try to directly use some of these HOL Light decision procedures on the given conjecture. A similar effect could be achieved also by mapping some of the HOL Light symbols (typically those encoding arithmetics) to the symbols that are reserved and treated specially by SMT solvers and ATP systems. This is now done for example in Isabelle/Sledgehammer [29], with the additional benefit of the combined methods employed by SMTs and ATPs over various well-known theories. Our approach is so far much simpler, which also means that we do not have to ensure that the semantics of such special theories remains the same (e.g., $1/0 = 0$ in HOL Light). The HOL Light decision procedures might often not be powerful enough to prove whole theorems, however for example the `REAL_ARITH`¹² tactic is called on 2678 unique (sub)goals in Flyspeck, making such tools a useful addition to the service.

Each decision procedure is spawned in a separate instance of HOL Light using our parallel infrastructure, and if any returns within the timeout, it is reported to the user. The decision procedures that we found most useful for solving goals are:¹³

- `TAUT`¹⁴ — Propositional tautologies.
 $(A \implies B \implies C) \implies (A \implies B) \implies (A \implies C)$
- `INT_ARITH`¹⁵ — Algebra and linear arithmetic over \mathbb{Z} (including \mathbb{R}).
 $\&2 * \&1 = \&2 + \&0$
- `COMPLEX_FIELD`¹⁶ — Field tactic over \mathbb{C} (including multivariate \mathbb{R}).
 $(Cx (\&1) + Cx (\&1)) = Cx (\&2)$

Additionally the decision procedure infrastructure can be used to try common tactics that could solve the goal. One that we found especially useful is simplification with arithmetic (`SIMP_TAC[ARITH]`), which solves a number of simple numerical goals that the service users ask the server.

2.5. Proof Minimization and Reconstruction

When an ATP finds (and reports in its proof) a subset of the advised premises that prove the goal, it is often the case that this set is not minimal. By re-running the prover and other provers with only this set of proof-relevant premises, it is often possible to obtain a proof that uses fewer premises. A common example are redundant equalities that may be used by the ATP for early (but unnecessary) rewriting in the presence of many premises, and avoided when the number of premises is significantly lower (and different ordering is then used, or a completely different strategy or ATP might find a very different proof). This procedure is run recursively, until the number of premises needed for the proof no longer decreases. We call this recursive procedure *pseudo/cross-minimization*, since it is not exhaustive and uses multiple ATPs. Minimizing the number of premises improves the chances of the HOL proof reconstruction, and the speed of (re-)processing large libraries that contain many such reconstruction tactics.¹⁷

Given the minimized list of advised premises, we try to reconstruct the proof. As mentioned in Section 2.1, the advice system may internally use a number of theorem names (now mostly produced

¹²http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/REAL_ARITH.html

¹³The reader might wonder why the above mentioned `REAL_ARITH` is not among the tactics used. The reason is that even though `REAL_ARITH` is used a lot in HOL Light formalizations, `INT_ARITH` is simply more powerful. It solves 60% more Flyspeck goals automatically without losing any of those solved by `REAL_ARITH`. As with the AI/ATP instances, the usage of decision procedures is optimized to jointly cover as many problems as possible.

¹⁴<http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/TAUT.html>

¹⁵http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/INT_ARITH.html

¹⁶http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/REAL_FIELD.html

¹⁷Premise minimization has been for long time used to improve the quality and refactoring speed of the Mizar articles. It is now also a standard part of Sledgehammer.

by splitting conjunctions) not present in standard HOL Light developments. It is possible to call the reconstruction tactics with the names used internally in the advice system; however this would create proof scripts that are not compatible with the original developments. We could directly address the theorem sub-conjuncts (using, e.g., “nth (CONJUNCTS thm) n”) however such proof scripts look quite unnatural (even if they are indeed faster to process by HOL Light). Instead, we now prefer to use the whole original theorems (including all conjuncts) in the reconstruction.

Three basic strategies are now tried to reconstruct the proof: REWRITE¹⁸ (rewriting), SIMP¹⁹ (conditional rewriting) and MESON [15] (internal first-order ATP). These three strategies are started in parallel, each with the list of HOL theorems that correspond to the minimized list of ATP premises as explained above. The strongest of these tactics – MESON – can in one second reconstruct 79.3% of the minimized ATP proofs. While this is certainly useful, the performance of MESON reconstruction drops below 40% as soon as the ATP proof uses at least seven premises. Since the service is getting stronger and stronger, the ratio of MESON-reconstructable proofs is likely to get lower and lower. That is why we have developed also a fine-grained reconstruction method – HH_RECON [23], which uses the quite detailed TPTP proofs produced by Vampire and E. This method however still needs an additional mechanism that maintains the TPTP proof as part of the user development: either dedicated storage, or on-demand ATP-recreation, or translation to a corresponding fine-grained HOL Light proof script. That is why HH_RECON is not yet included by default in the service.

2.6. Description of the Parallelization Infrastructure

An important aspect of the online service is its parallelization capability. This is needed to efficiently process multiple requests coming in from the clients, and to execute the large number of AI/ATP instances in parallel within a short overall wall-clock time limit. HOL Light uses a number of imperative features of OCaml, such as static lists of constants and axioms, and a number of references (mutable variables). Also a number of procedures that are needed use shared references internally. For example the MESON procedure uses list references for variables. This makes HOL Light not thread safe. Instead of spending lots of time on a thread-safe re-implementation, the service just (in a pragmatic and simple way, similar to the Mizar parallelization [38]) uses separate processes (Unix fork), which is sufficient for our purposes. Given a list of HOL Light tasks that should be performed in parallel and a timeout, the managing process spawns a child process for each of the tasks. It also creates a pipe for communicating with each child process. Progress, failures or completion information are sent over the pipe using OCaml marshalling. This means that it is enough to have running just one managing instance of HOL Light loaded with Flyspeck and with the advising infrastructure. This process forks itself for each client query, and the child then spawns as many AI/ATP, minimization, reconstruction, and decision procedure instances as needed.

2.7. Use of Caching

Even though the service can asynchronously process a number of parallel requests, it is not immune to overloading by a large number of requests coming in simultaneously. In such cases, each response gets less CPU time and the requests are less likely to succeed within the 30 seconds of wall-clock time. Such overloading is especially common for requests generated automatically. For example the Wiki service that is being built for Flyspeck [34] may ask many queries practically simultaneously when an article in the wiki is re-factored, but many of such queries will in practice overlap with previously asked queries. Caching is therefore employed by the service to efficiently serve such repeated requests.

Since the parallel architecture uses different processes to serve different requests, a file-system based cache is used (using file-level locking). For any incoming request the first job done by the forked process handling the request is to check whether an identical request has already been served,

¹⁸http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/REWRITE_TAC.html

¹⁹http://www.cl.cam.ac.uk/~jrh13/hol-light/HTML/SIMP_TAC.html

and if so, the process just re-sends the previously computed answer. If the request is not found in the cache, a new entry (file) for it is created, and any information sent to the client (apart from the progress information) is also written to the cache entry. This means that all kinds of answers that have been sent to the client can be cached, including information about terms that failed to parse or typecheck, terms solved by ATP only, minimization results and replaying results, including decision procedures. The cache stored in the filesystem has the additional advantage of persistence, and in case of updating the service the cache can be easily invalidated by simply removing the cache entries.

3. Multiple Projects, Versions, and Their Online Update

The functions described in Section 2 allowed the users to experiment with ATP queries over the fixed June 2012 version of *Flyspeck*. If *Flyspeck* already contained all of human mathematics in a form that is universally agreed upon, such setting would be sufficient. However, *Flyspeck* is not the only library developed with HOL Light, and *Flyspeck* itself has been updated considerably since June 2012 with a number of new definitions, theorems and proofs. In general, there is no final word on how formal mathematics should be done, and even more stable formalization libraries may be updated, refactored, and forked for new experiments.

To support this, the current version of HOL(y)Hammer also allows online addition of new projects and updating of existing projects (see Figure 2). This leads to a number of issues that are discussed in this section. A particularly interesting and important issue is the transfer and re-use of the expensively obtained problem-solving knowledge between the different projects and their versions.

Another major issue is the speed of loading large projects. *Checkpointing* of OCaml instances is used to save the load time, after HOL Light was bootstrapped. Checkpointing software allows the state of a process to be written to disk, and restore this state from the stored image later. We use DMTCP²⁰ as our checkpointing software: it does not require kernel modifications, and because of that it is one of the few checkpointing solutions that work on recent Linux versions.

3.1. Basic Server Infrastructure for Multiple Projects

Instead of just one default project, the server allows multiple projects identified by a unique project name such as “Ramsey”, “Flyspeck” and “Multivariate Analysis”. A new project can be started by an authorized user performing a password-protected upload of the project files via a HTTP POST request. In the same way, an existing project can be updated.²¹ The server data specific for each project are kept in its separate directory, which includes the user files, checkpointed images, features and proof dependencies used for learning premise selection, and the heuristically HTML-ized (hyperlinked) version of the user files. An overview of these project-specific data is given in Table 2.

Apart from the project-specific files, the service also keeps a spare checkpointed core HOL Light image and additional files that typically contain the reusable information from various projects. The core HOL Light image is used for faster creation of images for new projects. A new project can also be cloned from an existing project. In that case, instead of starting with the core HOL Light image, the new project starts with the cloned project’s image, and loads the new user files into them. This saves great amount of time when updating large projects like *Flyspeck*. The server processing of a new or modified project is triggered by the appropriate HTTP POST request. This starts the internal project creator which performs on the server the actions described by Algorithm 1. The data sizes and processing times for seven existing projects are summarized in Table 3 and Table 4.

²⁰<http://dmtcp.sourceforge.net>

²¹Git-based interface to the projects already exists and will probably also be used for updating the projects with the standard `git-push` command from users’ computers. This still requires installation of the Gitolite authentication layer on our server and implementing appropriate Git hooks similar to those developed for the Mizar wiki in [2].

FIGURE 2. The HOL(y)Hammer web with a query over Multivariate Analysis

HOL(y) Hammer
Learning-assisted automated reasoning for HOL Light

Request Advice:
Input the HOL Light formula to prove and select HOL Light session:

- polyhedron p ==> convex (relative_interior p)
- Multivariate Analysis

```
(cache:OK)(session:OK)(parse:OK)SSSSAWAAWAW
Result (3.81s): CONVEX_RELATIVE_INTERIOR POLYHEDRON_IMP_CONVEX
Replaying: SUCCESS (0.29s)SIMP_TAC[POLYHEDRON_IMP_CONVEX,CONVEX_RELATIVE_INTERIOR]
```

Examples:

- Core HOL Light: $&1 + &1 = &2$ or $ODD\ x \vee ODD\ (x + 1)$
- Model: $IS_RESULT\ r \ \ \ IS_CLASH\ r$
- Flyspeck: polyhedron p ==> convex (relative_interior p)

Upload Own Development:
Choose either a single HOL Light ".ml" file, or a ".tgz" file with a "make.ml" in top directory:

- File: - Session name:
- Password:
- Upload against: HOL Light Base
-

Uploaded Developments

- HOL Light Base
- Complex Analysis
- Flyspeck
- Gödel's incompleteness theorem
- Infinite Ramsey's theorem
- Jordan Curve theorem
- Model of HOL
- Multivariate Analysis

3.2. Safety

Since HOL Light is implemented in the OCaml toplevel, allowing users to upload their own development is equivalent to letting them run arbitrary programs on our server.²² This is inherently insecure. A brief analysis of the related security issues and their possible countermeasures has been done in the context of the WorkingWiki [45] collaborative platform.²³ The easiest practical solution is to allow uploads only by authorized users, i.e., users who are sufficiently trusted to be given shell

²²And indeed, the basic infrastructure could be also used as a platform for interacting with any OCaml project.

²³<http://lalashan.mcmaster.ca/theobio/projects/index.php/WorkingWiki/Security>

TABLE 2. The data maintained for each HOL(y)Hammer project.

Data	Description
User files	User-submitted ML files. These data are additionally Git-managed in this directory.
Image1	Checkpointed HOL Light image preloaded with the user files and the HH functions.
Image2	An analogous image that uses proof recording to extract HOL proof dependencies.
Features	Several (currently six) feature characterizations (see Section 2.1) of the project’s theorems.
HOL deps	The theorem dependencies from the original HOL proofs obtained by running the modified proof recording kernel on the user files.
ATP deps	The theorem dependencies obtained by running ATPs in various ways and minimizing such proofs. These data may be expensive to obtain, see 3.3 for the current re-use mechanisms.
Cache	The request cache for the project.
Auxiliary	Auxiliary files useful for bookkeeping and debugging.
HTML	Heuristically HTML-ized version of the user files, together with index pages for the files and theorems. These files are available for browsing and they are also linked to the Gitweb web interface, which presents the project and file history, allows comparison of different versions, regular expression search over the versions, etc.

Algorithm 1 Project creation stages

- 1: Set up the directory structure for new projects.
- 2: Open a copy of the checkpointed core HOL Light image (or another project’s cloned image) and load it with the user files and the HOL(y)Hammer functions.
- 3: Export the typed and variable-normalized statements of named theorems together with their MD5 hashes.
- 4: Export the various feature characterizations of the theorems.
- 5: Checkpoint the new image.
- 6: Re-process the user files with a proof-recording kernel that saves the (new) HOL proof dependencies.
- 7: Checkpoint the proof-recording image.
- 8: Add further compatible proof dependencies from related projects.
- 9: Run ATPs on the problems corresponding to the HOL dependencies, and minimize such proof data by running the ATPs further.
- 10: Run the heuristic HTML-izer and indexer, and push the user files to Git.

TABLE 3. The processing times for seven HOL(y)Hammer projects in seconds.

	Core	Ramsey	Model	Gödel	Complex	Multivariate	Flyspeck
Proof checking (min)	3	6	193	166	267	2716	21735
Proof recording (min)	10	14	225	215	578	3751	52002
Writing data	26	27	33	47	53	139	758
Writing ATP problems	38.56	45.35	51.14	73.37	72.12	139.12	650.15
Solving ATP problems	1582.8	1622.4	1882.2	2173.8	2284.8	9286.2	12034.2
HTML and Git	4	2	2	3	2	19	61
Image Restart	1.98	2.08	2.37	2.15	3.00	3.66	6.78

access. Asking queries to existing projects can still be done by anybody; the query is then just a string processed by a time-limited function that always exits.

We have also briefly considered sandboxing for allowing anonymous user uploads, however it adds a significant overhead to managing the server (HOL(y)Hammer currently runs in user mode), while offering little protection in the case of HOL Light. Combination of chroot jail, an iptables firewall, and disallowing users to write files, has been previously used by us in ProofWeb for multiple proof assistants [18]. This offers a sufficient level of security for a number of proof assistants where

TABLE 4. The data sizes for seven HOL(y)Hammer projects.

	Core	Ramsey	Model	Gödel	Complex	Multivariate	Flyspeck
Normal image size (kB)	33892	40952	37584	38244	55424	77292	152460
Recording image size (kB)	50960	52692	48148	46000	58368	247848	365496
Unique theorems	2482	2544	2951	3408	3582	6798	22336
Unique constants	234	234	337	367	333	466	1765
Avrg. HOL proof deps.	12.13	12.27	11.09	14.44	17.96	12.26	21.86
ATP-proved theorems	1546	1578	1714	1830	2042	4126	8907
Usable ATP proofs	6094	6141	6419	6644	6885	11408	21733
Avrg. ATP proof deps.	6.86	6.86	6.77	6.67	6.94	6.36	6.52
Total distinct features	3735	3759	4693	5755	5964	11599	43858
Avrg. features/formula	24.81	24.61	26.05	35.61	39.05	38.15	67.61

Usable ATP proofs. Vampire, Epar and Z3 are used, and we keep all the different minimal proofs. This means that the total number of ATP proofs can be higher than the number of theorems.

the ML access can be disabled, but it is not sufficient for HOL Light. Therefore also in ProofWeb, running HOL Light was restricted to the users that are allowed to use a shell on the server [16].

3.3. Re-use of Knowledge from Related Projects

It has been shown in [25] that learning premise selection from minimized ATP proofs is better than learning from the HOL proofs, and also that the two approaches can be productively combined, resulting in further improvement of the overall ATP performance. However, obtaining the data from ATP runs is expensive. For example, just running Vampire, Epar and Z3 on all Flyspeck problems for 30 seconds takes (assuming 70% unsolved problems for each ATP) about 500 CPU hours. Even with 50-fold parallelization, this takes 10 hours of wall-clock time. And this is just the initial ATP pass. In [25] we also show that further MaLAREa-style learning from such ATP data and re-running of the ATPs with the premises proposed by the learning grows the set of ATP solutions by about 20%. Obviously, such additional passes cost a lot of further CPU time. One option is to sacrifice the ATP data for speed, and only learn from the HOL data, sacrificing the final ATP performance on the queries. However, there is a relatively efficient way how to re-use a lot of the expensive data that were already computed.

Suppose that the user only updates an existing large project by adding a new file. Then it is quite sufficient to (relatively quickly) obtain the minimized ATP proofs of the (ATP-provable) theorems in the file that was added. Such ATP proofs are then added to the existing training data used for the premise selectors. In general, the project can however be modified and updated in a more complicated way, for example by adding/changing some files “in the middle”, modifying symbol definitions, theorems, etc. Or it can be a completely new project, that only shares some parts with other projects, restructuring some terminology, theorem names, and proofs. The method that we use to handle such cases efficiently is *recursive content-based encoding* of the theorem and symbol names [37]. This is the first practical deployment and evaluation of this method, which in HOL(y)Hammer is done as follows:

1. The name of every defined symbol is replaced by the content hash (we use MD5) of its variable-normalized definition containing the full types of the variables. This definition already uses content hashes instead of the previously defined symbols. This means that symbol names are no longer relevant in the whole project, neither white space and variable names.
2. The name of each theorem is also replaced by the content hash of its (analogously normalized) statement.

3. The proof-dependency data extracted in the content encoding from all projects are copied to a special “common” directory.
4. Whenever a project P is started or modified, we find the intersection of the content-encoded names of the project’s theorems with such names that already exist in other projects/versions.
5. For each of such “already known” theorems T in P , we re-use all its “already known” proofs D that are *compatible* with P ’s proof graph. This means, that the names of the proof dependencies of T in D must also exist in P (i.e., these theorems have been also proved in P , modulo the content-naming), and that these theorems precede T in P in its chronological order (otherwise we might get cyclic training data for P).

There are two possible dangers with this approach: collisions in MD5 and dealing with types in the HOL logic. The first issue is theoretical: the chance of unintended MD5 collisions is very low, and if necessary, we can switch to stronger hashes such as SHA-256. The second issue is more real: there is a choice of using content-encoding also for the HOL types, or just using their original names. If original names are used, two differently defined types can get the same name in two different projects, making the theorems about such types incompatible. If content encoding is used, all types with the same definition will get the same content name. However, the HOL logic rejects such semantic equality of the two types already in its parsing layer: two differently named types are always completely different in the HOL logic.²⁴ We currently use the first method (keeping the original type names), however the second method might be slightly more correct. In both cases, it probably would not be hard to add guards against the possible conflicts. In all cases, these issues only influence the performance of the premise-selection algorithms. The theorem proving (and proof reconstruction) is always done with the original symbols.

3.4. Analysis of the Knowledge Re-use for Flyspeck Versions

It is interesting to know how much knowledge re-use can be obtained with the content-encoding method. We analyze this in Table 5 on the theorems (or rather unique conjuncts) coming from three different Flyspeck SVN versions: 2887, 3006, and 3400. Note that the last version (3400) has not been subjected to several learning/ATP passes. Such passes raised the number of ATP-proved theorems in the earlier versions by about 20%. The table shows that the number of reusable theorems and proofs from the previous version is typically very high. This also means that more expensive AI/ATP computations (e.g., use of higher time limits, MaLAREa-style looping, and even BliStr-style strategy evolution [39]) could be in the future added to the tasks done on the server in its idle time, because the results of such computations will typically improve the success rates of all the future versions of such large projects.

TABLE 5. The re-use of theorems and ATP proofs between four Flyspeck SVN versions

Version	Unique thms	In previous (%)	ATP-proved (%)	ATP proofs	Reusable proofs (%)
2887	13647	N/A	7176 (53%)	20028	N/A
3006	13814	13480 (98%)	7235 (52%)	20081	19977 (99%)
3400	18856	12866 (93%)	8914 (47%)	21780	21320 (97%)

In previous. Theorems (conjuncts) that exist already in the previous version, and their percentage.

ATP proof. Vampire, Epar and Z3 are used, and we keep all the different minimal proofs. This means that the total number of ATP proofs can be higher than the number of theorems.

Re-usable ATP proofs. The proofs from the previous version that are valid also in the current version.

²⁴The second author could not resist pointing out that this issue disappears in set theory with soft types.

A by-product of the content encoding is also information about symbols that are defined multiple times under different names. For the latest version of Flyspeck there are 39 of them, shown in Table 6.

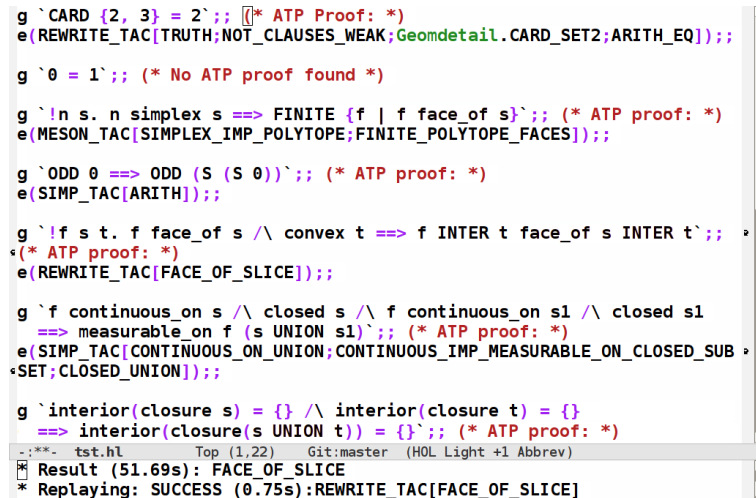
TABLE 6. 39 symbols with the same content-based definition in Flyspeck SVN 3400

face_path / face_contour	reflect_along / reflection
zero6 / dummy6	UNIV / predT
CROSS / *_c	node3_y / rotate3
EMPTY / pred0	APPEND / cat
func / FUN	set_components / set_part_components
ONE_ONE / injective	triple_of_real3 / vector_to_pair
supp / SUPP	is_no_double_joins / is_no_double_joints
dirac_delta / delta_func	unknown / NONLIN
o / compose	node2_y / rotate2
I / LET_END / mark_term	

4. Modes of Interaction with the Service

The standard web interface (Figure 2) displays the available projects, links to their documentation, allows queries to the projects, and provides an HTML form for uploading and modifying projects. Requests are processed using asynchronous DOM modification (AJAX): a JavaScript script makes the requests in the background and updates a part of the page that displays the response. Each request is first sent to the external PHP request processor, which communicates with the HOL(y)Hammer server. A prototype of a web editor interacting both with HOL Light and with the online advisor is described in [34].

FIGURE 3. Parallel asynchronous calls of the online advisor from Emacs.



```

g `CARD {2, 3} = 2`;; ([* ATP Proof: *)
e(REWRITE_TAC[TRUTH;NOT_CLAUSES_WEAK;Geomdetail.CARD_SET2;ARITH_EQ]);;

g `0 = 1`;; (* No ATP proof found *)

g `!n s. n simplex s ==> FINITE {f | f face_of s}`;; (* ATP proof: *)
e(MESON_TAC[SIMPLEX_IMP_POLYTOPE;FINITE_POLYTOPE_FACES]);;

g `ODD 0 ==> ODD (S (S 0))`;; (* ATP proof: *)
e(SIMP_TAC[ARITH]);;

g `!f s t. f face_of s /\ convex t ==> f INTER t face_of s INTER t`;;
(* ATP proof: *)
e(REWRITE_TAC[FACE_OF_SLICE]);;

g `f continuous_on s /\ closed s /\ f continuous_on s1 /\ closed s1
==> measurable_on f (s UNION s1)`;; (* ATP proof: *)
e(SIMP_TAC[CONTINUOUS_ON_UNION;CONTINUOUS_IMP_MEASURABLE_ON_CLOSED_SUB
SET;CLOSED_UNION]);;

g `interior(closure s) = {} /\ interior(closure t) = {}
==> interior(closure(s UNION t)) = {}`;; (* ATP proof: *)

```

-:***- tst.hl Top (1,22) Git:master (HOL Light +1 Abbrev)
Result (51.69s): FACE_OF_SLICE
* Replaying: SUCCESS (0.75s):REWRITE_TAC[FACE_OF_SLICE]

Figure 3 shows an Emacs session with several HOL Light goals.²⁵ The online advisor has been asynchronously called on the goals, and just returned the answer for the fifth goal and inserted the

²⁵A longer video of the interaction is at <http://mws.cs.ru.nl/~urban/ha1.mp4>

corresponding tactic call at an appropriate place in the buffer. The relevant Emacs code (customized for the HOL Light mode distributed with *Flyspeck*) is available online²⁶ and also distributed with the local HOL(y)Hammer install. It is a modification of the similar code used for communicating with the MizAR service from Emacs.

The simplest option (useful as a basis for more sophisticated interfaces) is to interact with the service in command line, for example using *netcat*, as shown for two following two queries. The first query is solved easily by *INT_ARITH*, while the other requires nontrivial premise and proof search.

```
$ echo 'max a b = &1 / &2 * ((a + b) + abs(a - b))'
| nc colo12-c703.uibk.ac.at 8080
.....
* Replaying: SUCCESS (0.25s): INT_ARITH_TAC
* Loadavg: 48.13 48.76 48.49 52/1151 46604

$ echo '!A B (C:A->bool).((A DIFF B) INTER C=EMPTY) <=> ((A INTER C) SUBSET B) '
| nc colo12-c703.uibk.ac.at 8080
* Read OK
.....
* Theorem! Time: 14.74s Prover: Z Hints: 32 Str:
  allt_notrivsyms_m10u_all_atponly
* Minimizing, current no: 9
.* Minimizing, current no: 6
* Result: EMPTY_SUBSET IN_DIFF IN_INTER MEMBER_NOT_EMPTY SUBSET SUBSET_ANTISYM
```

5. The Local Service Description

The service can be also downloaded,²⁷ installed and used locally, for example when a user is working on a private formalization that cannot be included in the public online service.²⁸ Installing the advisor locally proceeds analogously to the steps described in Algorithm 1. Two passes are done through the user's repository. In the first pass, the names of all the theorems available in the user's repository are exported, together with their features (symbols, terms, types, etc., as explained in Section 2.1). In the second pass, the dependencies between the named theorems are computed, again using the modified proof recording HOL Light kernel that records all the processing steps. Given the exported features and dependencies, local advice system(s) (premise selectors) are trained outside HOL Light. Using the fast sparse learning methods described in Section 2.1, this again takes seconds, depending on the user hardware and the size of the development. The advisors are then run locally (as independent servers) to serve the requests coming from HOL Light. While the first pass is just a fast additional function that can be run by the user at any time on top of his loaded repository, the second pass now still requires full additional processing of the repository. This could be improved in the future by checkpointing the proof-recording image, as we do in the online server.

The user is provided with a tactic (*HH_ADVICE_TAC*) which runs all the mechanisms described in the Section 2 on the current goal locally. This means that the functions relying on external premise selection and ATPs are tried in parallel, together with a number of decision procedures. The ATPs are expected to be installed on the user's machine and (as in the online service) they are run on the goal translated to the TPTP format, together with a limited number of premises optimized separately for each prover. By default *Vampire*, *Eprover* and *Z3* are now run, using three-fold parallelization.

²⁶<https://raw.githubusercontent.com/JUrban/hol-advisor/master/hol-advice.el>

²⁷<http://cl-informatik.uibk.ac.at/users/cek/hh/>

²⁸The online service can already handle private developments that are not shown to the public.

The local installation in its simple configuration is now only trained using the naive Bayes algorithm on the training data coming from the HOL Light proof dependencies and the features extracted with the standard method. As shown in [25], the machine learning advice can be strengthened using ATP dependencies, which can be also optionally plugged into the local mode. Further strengthening can be done with combinations of various methods. This is easy to adjust; for example a user with a 24-CPU workstation can re-use/optimize the parallel combinations from Table 1 used by the online service.

5.1. Online versus Local Systems

The two related existing services are MizAR and Sledgehammer. MizAR has so far been an online service (accessible via Emacs or web interface), while Sledgehammer has so far required a local install (even though it already calls some ATPs over a network). HOL(y)Hammer started as an online service, and the local version has been added recently to answer the demand by some (power)users. The arguments for installing the service locally are mainly the option to use the service offline (possibly using one's own large computing resources), and to keep the development private. As usual, the local install will also require the tools involved to work on all kinds of architectures, which is often an issue, particularly with software that is mostly developed in academia.

As described in Section 2, the online service now runs 7 different AI/ATP instances and 4 decision procedures for each query. When counting the individual ATP strategies (which may indeed be very orthogonal in systems like Vampire and E), this translates to about 70 different AI/ATP attempts for each query. If the demands grows, we can already now distribute the load from the current 48-CPU server to 112 CPUs by installing the service on another 64-CPU server. The old resolution-ATP wisdom is that systems rarely prove a result in higher time limits, since the search space grows very fast. A more recent wisdom (most prominently demonstrated by Vampire) however is that using (sufficiently orthogonal) strategy scheduling makes higher time limits much more useful.²⁹ And even more recent wisdom is that learning in various ways from related successes and failures further improves the systems' chances when given more resources. All this makes a good case for developing strong online computing services that can in short bursts focus a lot of power on the user queries, which are typically related to many previous problems. Also in some sense, the currently used AI/ATP methods are only scratching the surface. For example, further predictive power is obtained in MaLAREa [42] by computing thousands of interesting finite models, and using evaluation in them as additional semantic features of the formulas. ATP prototypes like MaLeCoP [43] can already benefit from accumulated fine-grained learned AI guidance at every inference step that they make. The service can try to make the best (re-)use of all smaller lemmas that have been proved so far (as in [21, 35]). And as usual in machine learning, the more data are centrally accumulated for such methods, the stronger the methods become. Finally, it is hard to overlook the recent trend of light-weight devices for which the hard computational tasks are computed by large server farms (cloud computing).

6. Conclusion and Future Work

We believe that HOL(y)Hammer is one of the strongest AI/ATP services currently available. It uses a toolchain of evolving large-theory methods that have been continuously improved as more and more AI/ATP experiments and computations have been recently done, in particular over the Flyspeck corpus. The combinations that jointly provide the greatest theorem-proving coverage are employed to answer the queries with parallelization of practically all of the components. The parallelization factor is probably the highest of all existing ATP services, helping to focus the power of many different AI/ATP methods to answer the queries as quickly as possible. The content-encoding

²⁹In [25], the relative performance of Vampire in 30 and 900 seconds is very different.

mechanisms allow to re-use a lot of the expensive theorem-proving knowledge computed over earlier projects and versions. And the checkpointing allows reasonably fast update of existing projects.

At this moment, there seems to be no end to better premise selection, better translation methods for ATPs (and SMTs, and more advanced combined systems like MetiTarski [1]), better ATP methods (and their AI-based guidance), and better reconstruction methods. Future work also includes broader updating mechanisms, for example using git to not just add, but also delete files from an existing project. A major issue is securing the server for more open (perhaps eventually anonymous) uploads, and maybe also providing encryption/obfuscation mechanisms that guarantee privacy of the non-public developments.³⁰ An interesting future direction is the use of the service with its large knowledge base and growing reasoning power as a semantic understanding (connecting) layer for experiments with tools that attempt to extract logical meaning from informal mathematical texts. Mathematics, with its explicit semantics, could in fact pioneer the technology of very deep parsing of scientific natural language writings, and their utilization in making stronger and stronger automated reasoning tools about all kinds of scientific domains.

References

- [1] Behzad Akbarpour and Lawrence C. Paulson. MetiTarski: An automatic theorem prover for real-valued special functions. *J. Autom. Reasoning*, 44(3):175–205, 2010.
- [2] Jesse Alama, Kasper Brink, Lionel Mamane, and Josef Urban. Large formal wikis: Issues and solutions. In James H. Davenport, William M. Farmer, Josef Urban, and Florian Rabe, editors, *Calculemus/MKM*, volume 6824 of *LNCS*, pages 133–148. Springer, 2011.
- [3] Jesse Alama, Tom Heskes, Daniel Kühlwein, Evgeni Tsivtsivadze, and Josef Urban. Premise selection for mathematics by corpus analysis and kernel methods. *J. Autom. Reasoning*, 52(2):191–213, 2014.
- [4] Jasmin Christian Blanchette, Sascha Böhme, Andrei Popescu, and Nicholas Smallbone. Encoding monomorphic and polymorphic types. In Nir Piterman and Scott A. Smolka, editors, *TACAS*, volume 7795 of *Lecture Notes in Computer Science*, pages 493–507. Springer, 2013.
- [5] Jasmin Christian Blanchette and Andrei Paskevich. TFF1: The TPTP typed first-order form with rank-1 polymorphism. In Bonacina [6], pages 414–420.
- [6] Maria Paola Bonacina, editor. *Automated Deduction - CADE-24 - 24th International Conference on Automated Deduction, Lake Placid, NY, USA, June 9-14, 2013. Proceedings*, volume 7898 of *Lecture Notes in Computer Science*. Springer, 2013.
- [7] Jacques Carette, David Aspinall, Christoph Lange, Petr Sojka, and Wolfgang Windsteiger, editors. *Intelligent Computer Mathematics - MKM, Calculemus, DML, and Systems and Projects 2013, Held as Part of CICM 2013, Bath, UK, July 8-12, 2013. Proceedings*, volume 7961 of *Lecture Notes in Computer Science*. Springer, 2013.
- [8] Andy Carlson, Chad Cumby, Jeff Rosen, and Dan Roth. The SNoW Learning Architecture. Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Department, 5 1999.
- [9] 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.
- [10] Sahibsingh A. Dudani. The distance-weighted k-nearest-neighbor rule. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-6(4):325–327, 1976.
- [11] Ulrich Furbach and Natarajan Shankar, editors. *Automated Reasoning, Third International Joint Conference, IJCAR 2006, Seattle, WA, USA, August 17-20, 2006, Proceedings*, volume 4130 of *LNCS*. Springer, 2006.
- [12] Allen Van Gelder and Geoff Sutcliffe. Extending the TPTP language to higher-order logic with automated parser generation. In Furbach and Shankar [11], pages 156–161.

³⁰The re-use performance obtained through content encoding suggests that just name obfuscation done by the client is not going to work as a privacy method.

- [13] Thomas C. Hales. Introduction to the Flyspeck project. In Thierry Coquand, Henri Lombardi, and Marie-Françoise Roy, editors, *Mathematics, Algorithms, Proofs*, number 05021 in Dagstuhl Seminar Proceedings, pages 1–11, Dagstuhl, Germany, 2006. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.
- [14] 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.
- [15] 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 *LNAI*, pages 313–327. Springer-Verlag, 1996.
- [16] Maxim Hendriks, Cezary Kaliszyk, Femke van Raamsdonk, and Freek Wiedijk. Teaching logic using a state-of-the-art proof assistant. *Acta Didactica Napocensia*, 3(2):35–48, June 2010.
- [17] Krystof Hoder and Andrei Voronkov. Sine qua non for large theory reasoning. In Nikolaj Bjørner and Viorica Sofronie-Stokkermans, editors, *CADE*, volume 6803 of *LNCS*, pages 299–314. Springer, 2011.
- [18] Cezary Kaliszyk. Web interfaces for proof assistants. In S. Autexier and C. Benzmüller, editors, *Proc. of the Workshop on User Interfaces for Theorem Provers (UITP’06)*, volume 174[2] of *ENTCS*, pages 49–61, 2007.
- [19] Cezary Kaliszyk and Alexander Krauss. Scalable LCF-style proof translation. In Sandrine Blazy, Christine Paulin-Mohring, and David Pichardie, editors, *Proc. of the 4th International Conference on Interactive Theorem Proving (ITP’13)*, volume 7998 of *LNCS*, pages 51–66. Springer Verlag, 2013.
- [20] Cezary Kaliszyk and Josef Urban. Automated reasoning service for HOL Light. In Carette et al. [7], pages 120–135.
- [21] Cezary Kaliszyk and Josef Urban. Lemma mining over HOL Light. In Kenneth L. McMillan, Aart Middeldorp, and Andrei Voronkov, editors, *LPAR*, volume 8312 of *Lecture Notes in Computer Science*, pages 503–517. Springer, 2013.
- [22] Cezary Kaliszyk and Josef Urban. MizAR 40 for Mizar 40. *CoRR*, abs/1310.2805, 2013.
- [23] Cezary Kaliszyk and Josef Urban. PRocH: Proof reconstruction for HOL Light. In Bonacina [6], pages 267–274.
- [24] Cezary Kaliszyk and Josef Urban. Stronger automation for Flyspeck by feature weighting and strategy evolution. In Jasmin Christian Blanchette and Josef Urban, editors, *PxTP 2013*, volume 14 of *EPiC Series*, pages 87–95. EasyChair, 2013.
- [25] Cezary Kaliszyk and Josef Urban. Learning-assisted automated reasoning with Flyspeck. *Journal of Automated Reasoning*, 2014. <http://dx.doi.org/10.1007/s10817-014-9303-3>.
- [26] Laura Kovács and Andrei Voronkov. First-order theorem proving and Vampire. In Natasha Sharygina and Helmut Veith, editors, *CAV*, volume 8044 of *Lecture Notes in Computer Science*, pages 1–35. Springer, 2013.
- [27] Daniel Kühlwein, Twan van Laarhoven, Evgeni Tsivtsivadze, Josef Urban, and Tom Heskes. Overview and evaluation of premise selection techniques for large theory mathematics. In Bernhard Gramlich, Dale Miller, and Uli Sattler, editors, *IJCAR*, volume 7364 of *LNCS*, pages 378–392. Springer, 2012.
- [28] Jia Meng and Lawrence C. Paulson. Translating higher-order clauses to first-order clauses. *J. Autom. Reasoning*, 40(1):35–60, 2008.
- [29] Lawrence C. Paulson and Jasmin Blanchette. Three years of experience with Sledgehammer, a practical link between automated and interactive theorem provers. In *8th IWIL*, 2010. Invited talk.
- [30] 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.
- [31] Andrew Pitts. The HOL logic. In M. J. C. Gordon and T. F. Melham, editors, *Introduction to HOL: A theorem proving environment for higher order logic*. Cambridge University Press, 1993.
- [32] Stephan Schulz. E - A Brainiac Theorem Prover. *AI Commun.*, 15(2-3):111–126, 2002.
- [33] Geoff Sutcliffe, Stephan Schulz, Koen Claessen, and Allen Van Gelder. Using the TPTP language for writing derivations and finite interpretations. In Furbach and Shankar [11], pages 67–81.

- [34] Carst Tankink, Cezary Kaliszyk, Josef Urban, and Herman Geuvers. Formal mathematics on display: A wiki for Flyspeck. In Carette et al. [7], pages 152–167.
- [35] Josef Urban. MoMM - fast interreduction and retrieval in large libraries of formalized mathematics. *Int. J. on Artificial Intelligence Tools*, 15(1):109–130, 2006.
- [36] Josef Urban. An Overview of Methods for Large-Theory Automated Theorem Proving (Invited Paper). In Peter Höfner, Annabelle McIver, and Georg Struth, editors, *ATE Workshop*, volume 760 of *CEUR Workshop Proceedings*, pages 3–8. CEUR-WS.org, 2011.
- [37] Josef Urban. Content-based encoding of mathematical and code libraries. In Christoph Lange and Josef Urban, editors, *Proceedings of the ITP 2011 Workshop on Mathematical Wikis (MathWikis)*, number 767 in *CEUR Workshop Proceedings*, pages 49–53, Aachen, 2011.
- [38] Josef Urban. Parallelizing Mizar. *CoRR*, abs/1206.0141, 2012.
- [39] Josef Urban. BliStr: The Blind Strategymaker. *CoRR*, abs/1301.2683, 2013.
- [40] Josef Urban, Piotr Rudnicki, and Geoff Sutcliffe. ATP and presentation service for Mizar formalizations. *J. Autom. Reasoning*, 50:229–241, 2013.
- [41] 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/Calculus*, volume 6167 of *LNCS*, pages 132–146. Springer, 2010.
- [42] 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.
- [43] Josef Urban, Jiří Vyskočil, and Petr Štěpánek. MaLeCoP: Machine learning connection prover. In Kai Brünner and George Metcalfe, editors, *TABLEAUX*, volume 6793 of *LNCS*, pages 263–277. Springer, 2011.
- [44] Vernor Vinge. *A Fire Upon the Deep*. Tor Books, 1992.
- [45] Lee Worden. WorkingWiki: a MediaWiki-based platform for collaborative research. In Christoph Lange and Josef Urban, editors, *ITP Workshop on Mathematical Wikis (MathWikis)*, number 767 in *CEUR Workshop Proceedings*, pages 63–73, Aachen, 2011.

Cezary Kaliszyk
University of Innsbruck, Austria, supported by FWF grant P26201

Josef Urban
Radboud University, Nijmegen, funded by NWO grant *Knowledge-based Automated Reasoning*