DEPARTMENT OF
COMPUTER
SCIENCE

UNIVERSITY OF
OXFORD

From: Assoc. Prof. Andrzej Murawski
andrzej.murawski@cs.ox.ac.uk

30 August 2023

Report on the PhD thesis of Mohnish Pattathurajan

Commutative images of languages over infinite alphabets

Context

Mohnish Pattathurajan's doctoral thesis is a contribution to automata theory over infinite alphabets. Typical automata models of this kind do not recognize specific input letters from the alphabet, but merely compare them with a selection of letters encountered by the automaton earlier. As a consequence, languages accepted by such automata, called *data languages*, are closed under permutations of the alphabet. At the technical level, this makes them akin to Fraenkel-Mostowski sets with *atoms* considered in set-theoretic research in the 1930s. Many standard concepts of automata theory can be revisited through the lens of sets with atoms. In order to make them finitely representable and amenable to automated verification, the research efforts have often focussed on the study of the so-called *orbit-finite* objects, which - although infinite - still admit finitary representations. More broadly, in the last two decades, the area of automata over infinite alphabets has gained prominence in connection with database theory, markup languages and software verification, as well as having been the source of compelling foundational problems.

The present thesis is an exploration into how one can generalize classic automata-theoretic results about Parikh images to infinite alphabets. The *Parikh image of a language*, also known as its commutative image, is the set of vectors representing the number of occurrences of each letter across all words from the language. In standard automata theory over finite alphabets, it has been known since the 1960s that Parikh images of regular languages coincide with the so-called *semilinear sets* and, what is more, Parikh images of context-free languages are also semilinear. The thesis investigates to what extent the results can be recovered for models of computation over atoms, notably for nondeterministic register automata (NRA) and register context-free grammars. The answer turns out to be more involved than one would first expect. In short, the results do not admit an immediate generalization. However, the author was able to propose and prove a series of highly non-trivial partial results, which inform further conjectures for future work.

Content

The thesis consists of nine chapters and the bibliography.

- Chapter 1 outlines the contributions of the thesis. The next two chapters introduce background material on sets with atoms and automata respectively. In particular, the latter introduces the classes of nondeterministic register automata (NRA), one-register NRA (1-NRA) and one-register context-free grammars (1-CFG), which will be the main acceptors of data languages discussed in the thesis.

- Chapter 4 defines *rational data languages* inductively, by appealing to classic regular operations and *orbit-finite unions* (instead of finite ones). Parikh images of such languages can be characterized inductively in a similar way, and are called the *rational sets of data vectors* (Lemma 4.2.4). They will play a central role in the thesis, as all the major results will state that, for certain classes of languages, the associated Parikh images are rational. Although one might hope that rational languages coincide with languages accepted by NRA, this is not the case: there exists a deterministic 1-NRA whose language is not rational (Proposition 4.1.3). This mismatch is an illustration of a common phenomenon over infinite alphabets: equi-expressivity results are scarce and the relationships between natural computational models may be surprisingly intricate.
  On a technical note, Chapter 4 also establishes an important closure property for rational languages and rational sets of data vectors, namely the Substitution Lemma (Lemma 4.3.2), which will be useful in technical arguments.
- Chapter 5 generalizes semilinear sets to the infinite-alphabet setting analogously to how the previous chapter handled rational data languages, i.e. via orbit-finite unions. This makes it possible to ask whether Parikh images of NRA languages are semilinear in the extended sense, which would amount to recovering the classic result for finite automata in its most obvious reformulation. Unfortunately, this turns out not to be the case: there is a 1-NRA whose language is not semilinear (Theorem 5.3.1).
  The failure of semilinearity makes the author investigate the hypothesis that Parikh images of NRA are rational rather than semilinear. As the two concepts coincide over finite alphabets, results that link Parikh images to rational sets of data vectors could also be viewed as extensions of classic results. The following three chapters validate the change of focus to rational sets with three positive results.
- Chapter 6 shows that Parikh images induced by 1-NRA are rational (Theorem 6.1.1). The proof of the fact requires a surprising amount of technical effort and relies on several delicate steps before the Substitution Lemma can be invoked to imply the required result. In particular, the author employs run decompositions, ingenious changes of alphabets and graph-theoretic results about Hamiltonian cycles.
- The question whether the result from Chapter 6 generalizes to all NRA is left unanswered in the thesis. However, Chapter 7 extends the result (Theorem 7.1.1) to a new subclass of NRA called *hierarchical register automata* (HRA), which involve arbitrarily many registers that must be used in a restricted way: when the ith register is modified, the content of the lower ones must be preserved but the higher ones may change arbitrarily. As expected, this makes HRA strictly weaker than NRA (Theorem 7.4.1). Theorem 7.1.1 is proved by induction using Theorem 6.1.1 as the base case. As before, the argument is far from trivial but capitalises on the experience gained from the previous chapter.
  In Chapter 7, HRA languages are also shown to contain rational languages (Theorem 7.5.1), which means that their Parikh images coincide with rational sets of data vectors.
- Chapter 8 contains the final rationality result for Parikh images, this time for one-register context-free grammars (Theorem 8.1.1). Its proof exploits a notion of width for derivation trees and exhibits a transformation that decreases width while preserving the Parikh image of its yield. This enables a reduction to 1-CFGs of bounded width.

Overall, this result could be thought as a partial generalization of the classic Parikh's theorem for context-free languages, due to the use of a single register only.

- Chapter 9 summarizes the results and outstanding problems. The main problems left open are whether the rationality results initiated in the thesis can be extended to more registers, both for register automata and context-free grammars. The second direction for further research concerns decision procedures for rational sets of data vectors. These go beyond the scope of the thesis, but the results presented in the thesis provide strong motivation for their study.

Assessment

The thesis demonstrates a very good understanding of the state of the art and features several original contributions. The most challenging results concern the rationality of Parikh images for three models of computation (1-NRA, HRA and 1-CFL), and constitute the first attempt to generalize classic theorems on Parikh images to automata models with atoms. The strength of the results has already been recognized by acceptance at high-quality conferences such as LICS (1-NRA and 1-CFL) and FSTTCS (HRA). Although the models rely on restrictions on the number of registers or manner of their use, it must be stressed that the results were not easy to establish. Indeed, their proofs exploit a combination of diverse techniques, blending automata theory, computation with atoms and elements of combinatorics. The author has shown a high level of competence when applying them to the problems at hand.

Another contribution made by the thesis consists in proposing and collecting numerous examples and counterexamples, which illustrate differences in expressiveness between various formalisms. In several cases, quite elaborate technical arguments were needed to validate the examples, and this aspect of the thesis should not go unnoticed.

Exposition

The material has been presented in a lucid and scholarly fashion. Apart from occasional typographical and grammatical mistakes, the quality of presentation is good and up to the standard of a PhD thesis. The technical developments have been discussed in logical order and in a systematic way. The author has also highlighted the main results and definitions in each chapter, which helps the reader to navigate the content. Each chapter also contains an informative summary.

In conclusion, it is my opinion that the quality of the research results discussed in the thesis and the way they were presented clearly merit the award of a PhD.

Andrzej Murawski