RAIRO-Inf. Theor. Appl. **39** (2005) 297-304
DOI: 10.1051/ita:2005017

# IMRE SIMON: AN EXCEPTIONAL GRADUATE STUDENT

Denis Thérien[1]

**Abstract.** This short note reviews the main contributions of the Ph.D. thesis of Imre Simon. His graduate work had major impact on algebraic theory of automata and thirty years later we are in a good position to appreciate how sensitive he was in selecting good problems, and how clever in solving them!

**Mathematics Subject Classification.** 68Q70.

## Introduction

The Ph.D. thesis written by Imre Simon in 1972 [9] was a masterpiece. One gets even more impressed when realizing that his work was predating [2] significantly, hence could not rely on the sophisticated framework introduced by Eilenberg.

It had been known for some time that aperiodic monoids were recognizing exactly the star-free languages [8]. This result clearly indicated the relevance of using algebra to get effective characterizations of languages that are combinatorially defined. Cohen and Brzozowski [1] had suggested the neat idea of parametrizing star-free languages by alternating concatenation and boolean operations, *i.e.* they introduced the dot-depth hierarchy. It remained to see if non-trivial information could be obtained on this parametrization, and this is when Simon entered the scene.

Concentrating on $\mathcal{D}_1$, the first level of the hierarchy, he contributed three important results, each of which later had major impact on the development of algebraic theory of automata. First, he had the remarkable intuition to define a systematic and manageable system of finite-index congruences to classify languages in the family. He then used his powerful mathematical skills to characterize two natural

*Keywords and phrases.* Algebraic theory of automata, piece-wise testable languages, dot-depth hierarchy.

[1] School of Computer Science, McGill University, 3480 University Street, McConnell Engineering Building, Room 318, Montreal, Québec, H3A 2A7 Canada; denis@cs.mcgill.ca

hierarchies arising in his classification, the piecewise testable languages and the locally testable languages. His thesis fell short of providing an effective characterization for the full class $\mathcal{D}_1$; this was achieved 10 years later by Robert Knast [3], strongly building on the ideas contained in Simon's thesis.

In this short paper, we will review the historical contributions made by Simon while he was a graduate student and indicate how these later led to further progress. We concentrate on ideas rather than on technicalities and refer the interested reader to the literature for details. Hopefully, we will be able to convey to the audience the importance and the intelligence of the early production of a brilliant theoretical computer scientist.

## 1. Classifying dot-depth one languages

We begin with some preliminary definitions; we will introduce more as needed. The reader should consult [5] for a more complete story. A semigroup is a set $S$ on which is defined a binary associative operation. We are interested in using finite semigroups to recognize subsets of $A^+$. The language $L \subseteq A^+$ is recognized by $S$ iff there exists a morphism $\phi : A^+ \to S$ and a subset $F$ of $S$ such that $L = \phi^{-1}(F)$. The natural ordering on semigroups is that of division: $S$ divides $T$ iff $S$ is a morphic image of a subsemigroup of $T$.

To every language $L$ can be associated a canonical semigroup in the following way. Suppose $L$ is a subset of $A^+$; define on $A^+$ the syntactic congruence of $L$ by $x \; \gamma_L \; y$ iff for all $u$ and $v$ in $A^*$ $uxv$ is in $L$ iff $uyv$ is in $L$. It is an exercise to show that $\gamma_L$ is the coarsest congruence that saturates $L$. We will write $S(L)$ for the quotient $A^+/\gamma_L$ and call it the syntactic semigroup of the language. Equivalently, $S(L)$ is the unique minimal semigroup (in the division ordering) that recognizes $L$. We are exclusively concerned with the case where the syntactic semigroup is finite, *i.e.* the language $L$ is regular. The number one paradigm in algebraic theory of automata is to relate the combinatorial definition of a language $L$ to algebraic properties of its syntactic semigroup. It turns out that the right level at which to pursue these investigations is that of *pseudovariety*, which we will simply call variety in this paper; a variety $\mathbf{V}$ is a class of finite semigroups that is closed under division and direct product. Although not yet formally present in the literature of the early seventies, all classes of semigroups studied by Simon were in fact varieties.

As a graduate student, Imre Simon got interested in the dot-depth hierarchy that had just been introduced by his supervisor Janusz Brzozowski. Let $\mathcal{D}_0$ be the boolean algebra consisting of the finite and cofinite subsets of $A^+$. One constructs a hierarchy $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \cdots$ by declaring that $L$ belongs to $\mathcal{D}_i$ iff $L$ is a boolean combination of languages that are of the form $L_1 \cdots L_t$, where each $L_j$ belongs to $\mathcal{D}_{i-1}$; in other words, starting from $\mathcal{D}_0$, the dot-depth hierarchy is obtained by alternating closure under concatenation and closure under boolean operations. Trivially, a language is star-free, *i.e.* can be obtained from the finite subsets of $A^+$ using concatenation and boolean operations, iff it belongs to $\mathcal{D}_i$ for some $i \geq 0$;

Schützenberger had proved in 1965 that $L$ is star-free iff $S(L)$ is aperiodic, *i.e.* no subset of the syntactic semigroup forms a non trivial group. The very interesting consequence of this deep result is that there exists an effective procedure to decide if a language (given by an automaton, an expression, a grammar, ...) is star-free or not.

It is easy to show that $L$ is in $\mathcal{D}_0$ iff $S(L)$ is aperiodic and contains a unique idempotent, *i.e.* an element $s$ such that $s = s^2$. Simon investigated the question of characterizing $\mathcal{D}_1$ algebraically but this proved to be quite difficult. Let us describe how he attacked the problem.

His first observation was to obtain normalized expressions for languages in $\mathcal{D}_1$, namely as boolean combinations of languages of the form $w_0A^*w_1 \cdots A^*w_s$ where each $w_j$ belongs to $A^*$. From this normal form, he got the intuition to define a 2-parameter congruence on $A^+$. We need the following notation. Let $m, k \geq 0$; for $x \in A^+$, $\mathrm{pref}_k(x)$ is the prefix of length $k$ of $x$ if $x$ has length at least $k$, and is $x$ itself otherwise; $\mathrm{suf}_k(x)$ is similarly defined in terms of the suffix. We now define on $A^+$: $x \; \gamma_{m,k} \; y$ iff $\mathrm{pref}_k(x) = \mathrm{pref}_k(y)$, $\mathrm{suf}_k(x) = \mathrm{suf}_k(y)$ and for all $s \leq m$, for all $u_1, \ldots, u_s$ where each $u_j$ has length at most $k+1$, $x \in A^*u_1A^* \ldots A^*u_sA^*$ iff $y \in A^*u_1A^* \ldots A^*u_sA^*$.

It is easily checked that every $\gamma_{m.k}$ is a congruence of finite index and that $L$ is a language of dot-depth one iff $L$ is a union of $\gamma_{m,k}$-classes for some $m$ and some $k$. The name of the game is now to find algebraic information on the various semigroups $A^+/\gamma_{m,k}$.

The next two examples are well-known: the first one is folklore, the second one is attributed to Perrin.

**Example 1.** *$L$ is a union of $\gamma_{1,0}$-classes iff $S(L)$ is commutative and idempotent.*

**Example 2.** *$L$ is a union of $\gamma_{0,k}$-classes for some $k$ iff $S(L)$ is locally trivial, i.e. it satisfies the condition that $eSe = e$ for every idempotent $e$.*

In his thesis, Imre Simon was able to characterize the $\gamma_{m,0}$-languages and the $\gamma_{1,k}$-languages; these very sophisticated results will be the focus of the next two sections. Although an effective characterization of the complete class of dot-depth one languages was beyond reach at the time, an important intermediate step was also proved in the thesis, namely a decomposition theorem in terms of the classical wreath product operation. We will briefly discuss the result in the last section.

The idea of classifying regular languages by using finite-index congruences was pushed further in [12]. The basic technique of Simon was to define the congruence class of a word by testing the *existence* of a factorization of the *simple* form $x = x_0A^*x_1A^* \cdots A^*x_s$; our extension was provided by allowing *counting* of factorizations of a *more general* type, namely of the form $x = x_0L_1x_1 \cdots L_sx_s$ where the $L_j$'s are defined recursively (Simon's definition forming the basis for the process). In this way we were able to capture all regular languages $L$ for which the syntactic semigroup $S(L)$ contains only solvable groups.

## 2. Piecewise testable languages

Perhaps the most beautiful result of the thesis of Imre Simon was his characterization of the $\gamma_{m,0}$-languages, which he called the piecewise testable languages. To this day his theorem remains a true jewel of algebraic theory of automata. We present a brief discussion of the theorem and of its modern ramifications. For reasons of convenience, in this section languages will be taken to be subsets of $A^*$ (instead of $A^+$) and the discussion is conducted in terms of monoids (rather than semigroups). The distinction between semigroups and monoids is in general crucial to make, but in the case we study here it would make no difference. A word $u = a_1 \dots a_m$ is a *subword* of a word $x$ if $x$ can be written as $x = x_0a_1x_1 \dots a_mx_m$, *i.e.* $u$ can be obtained from $x$ by erasing some letters. One readily observes that the binary relation $x \leq u$ iff $u$ is a subword of $x$ is a partial order on $A^*$ that is compatible with concatenation, and that the empty word is the maximal element in this ordering. In an algebraic perspective, we will say that a monoid $M$ is *partially ordered* iff there is a binary relation on $M$ that forms a partial order, that is compatible with the monoid operation and for which the identity is the maximal element.

Recall that the congruence $\gamma_{m,0}$ is defined precisely in terms of subwords: $x\ \gamma_{m,0}\ y$ iff $x$ and $y$ have the same subwords of length at most $m$. The monoid $A^*/\gamma_{m,0}$ is partially ordered, the order being naturally given by $[x]_{\gamma_{m,0}} \leq [y]_{\gamma_{m,0}}$ iff every subword of length at most $m$ of $y$ is also a subword of $x$. In fact, a converse to this last statement is not very difficult to prove.

**Lemma 2.1.** *Let $\phi : A^* \to M$, where $M$ is partially ordered. There exists $m \geq 0$ such that $x\ \gamma_{m,0}\ y$ implies $\phi(x) = \phi(y)$.*

*Proof.* Suppose the longest chain in $M$ has length $m+1$ and let $x$ and $y$ be $\gamma_{m,0}$-related. Write $x = x_0a_1x_1 \dots a_sx_s$, where $1 = \phi(x_0) > \phi(x_0a_1) = \phi(x_0a_1x_1) > \cdots > \phi(x_0a_1x_1 \dots a_s) = \phi(x)$; note that $s \leq m$, and that $\phi(x) = \phi(a_1 \dots a_s)$. Since $a_1 \dots a_s$ is also a subword of $y$, and because $M$ is partially ordered, we get $\phi(y) \leq \phi(a_1 \cdots a_s) = \phi(x)$. By symmetry $\phi(x) \leq \phi(y)$ so that the two are actually equal. □

Consider next the following binary relation on a monoid $M$: $s \leq_{\mathcal{J}} t$ iff $MsM \subseteq MtM$. This relation is certainly reflexive and transitive. It is also antisymmetric, *i.e.* it defines a partial order on $M$, exactly when $MsM = MtM$ iff $s = t$. Such monoids are called $\mathcal{J}$-trivial and they form a variety which is denoted by **J**. Note that the identity is always the maximal element in this order, but it is in general not compatible with the multiplication of the monoid: hence it is not true that every $\mathcal{J}$-trivial monoid is partially ordered according to our definition above, although it is easily seen that any partially ordered monoid must be $\mathcal{J}$-trivial. Nevertheless Simon was able to prove that every $\mathcal{J}$-trivial monoid is a homomorphic image of a partially ordered one. He stated it in the following form.

**Theorem 2.2.** *Let $\phi : A^* \to M$, where $M$ is $\mathcal{J}$-trivial. There exists $m \geq 0$ such that $x\ \gamma_{m,0}\ y$ implies $\phi(x) = \phi(y)$.*

The key combinatorial ingredient of his proof is a result showing that whenever $x$ and $y$ are in the same $\gamma_{m,0}$-class, there must exist a word $z$ again in the same class such that both $x$ and $y$ are subwords of $z$. In [11] a different proof was given, where ideas of Rhodes and Birget were used to directly construct a partially ordered expansion to a given $\mathcal{J}$-trivial monoid.

The idea of investigating partial orders compatible with monoid operations was revived rather spectacularly in the mid nineties. Pin and Weil carefully analyzed what algebraic framework was necessary to understand classes of languages not closed under complement [7]. The answer turned out to be a complete theory of monoids with partial orders (in which the condition $s \leq 1$ for all $s$ is replaced by arbitrary inequations); a long stream of results has appeared since then. The case where the identity is the maximal element, that was considered by Simon, is thus the most natural example of an idea that has powerful ramifications. Another magnificent example of that theory is the case of ordered monoids in which the partial order satisfies $e \leq 1$ for every idempotent $e$; it corresponds to regular languages that are open in the group topology on the free monoid [6].

## 3. Locally testable languages

A second deep result contained in Simon's thesis was dealing with *locally testable languages* that had been introduced in [4]. These sets are defined as follows: imagine a window of length $k+1$ that one can slide on an input $x$. The prefix and suffix of length $k$ can be detected, as well as the set of factors of length $k+1$ (although not their order of appearance or their multiplicity). This is exactly the idea of the congruence $\gamma_{1,k}$: $x \ \gamma_{1,k} \ y$ iff $x$ and $y$ can not be distinguished by the above apparatus.

When $k=0$, it is an easy exercise to show that $L$ is a $\gamma_{1,0}$-language iff $S(L)$ is idempotent and commutative, *i.e.* $S(L)$ belongs to the variety which is often denoted by $\mathbf{J_1}$. It is also readily checked that, for every $k$, the syntactic semigroup of a $\gamma_{1,k}$-language is always locally idempotent and commutative, *i.e.* belongs to the variety $\mathbf{LJ_1} = \{S : \text{for every idempotent } e, eSe \in \mathbf{J_1}\}$. Simon was able to show the converse; in doing so, he introduced ideas that were later developed into a very powerful and elegant framework to analyse decompositions of semigroups via wreath and block products.

Given two semigroups $S$ and $T$ we define their wreath product, denoted $S \circ T$, as the set $S^T \times T$ with the operation $(f_1,t_1)(f_2,t_2) = (f, t_1t_2)$ where $f(t) = f_1(t)f_2(tt_1)$. For two varieties $\mathbf{V}$ and $\mathbf{W}$, it is customary to write $\mathbf{V} * \mathbf{W}$ for the smallest variety that contains all semigroups of the form $S \circ T, S \in \mathbf{V}, T \in \mathbf{W}$. It is convenient to think of the computational power of $S \circ T$ in the following way. Let $x = a_1 \ldots a_n \in A^+$ be the input. The "front" machine $T$ processes $x$ normally, and it thus corresponds to a congruence $\beta$ on $A^+$. The "tail" machine $S$, at the $i$th step will have access to $a_i$ together with the $\beta$-class of $a_1 \ldots a_{i-1}$, and will compute, with $S$, on the word $(\cdot, a_1)([a_1]_\beta, a_2)...([a_1 \ldots a_{n-1}]_\beta, a_n)$ which lives in the alphabet $(T \cup \{\cdot\}) \times A$.

First, Simon made the following observation about the congruence $\gamma_{1,k}$. The semigroup $A^+/\gamma_{1,k}$ divides $S \circ A^+/\gamma_{0,k}$ where $S$ is idempotent and commutative. Indeed consider an input $x = a_1 \dots a_n$ with $n > k$. The prefix and suffix of length $k$ are determined by the front machine $A^+/\gamma_{0,k}$. Moreover, at the $i$th step of the computation the machine $S$ sees the letter $a_i$ together with the value of $a_1 \dots a_{i-1}$ in $A^+/\gamma_{0,k}$; this becomes new information only when $i > k$ and then $S$ has access to the factor of length $k+1$ ending at $a_i$. If we view $S$ as working over the "alphabet" that consists of the possible factors of length $k+1$ (*i.e.* $A^{k+1}$), then $S$ needs to know which "letters" have appeared in order to provide the information needed to identify the $\gamma_{1,k}$-class of the input $x$; by a remark previously made, detecting the set of letters that appear in a word can always be done with an idempotent and commutative semigroup. The converse also holds and this analysis can be summarized by saying that $L$ is a $\gamma_{1,k}$-language iff $S(L)$ is in the variety $\mathbf{J_1} * \mathbf{LI}$.

It remained to establish that $\mathbf{J_1} * \mathbf{LI} = \mathbf{LJ_1}$. Simon introduced an explicit formulation of this question in terms of graphs and was then able to prove the appropriate combinatorial lemma on graphs that corresponds to the above claim. We can briefly describe his method. Let $S$ be a semigroup and $E$ be the set of its idempotents. To $S$ one associates a graph as follows: the vertices are the idempotents $E$ and the edges are the triples in $E \times S \times E$, the edge $(e, s, f)$ going from vertex $e$ to vertex $f$. We define a congruence on the set of paths by declaring that two paths are related if they are coterminal (*i.e.* they start and end at the same vertex) and multiplying out the sequence of triples of each path yields the same value in $S$. By definition, the semigroup $S$ is locally idempotent and commutative iff this congruence satisfies the condition that $xx$ is congruent to $x$ for every loop $x$ and $xy$ is congruent to $yx$ whenever $x$ and $y$ are loops around the same vertex. It can also easily be shown that $S$ is in $\mathbf{J_1} * \mathbf{LI}$ iff the congruence satisfies the property that any two paths that are coterminal and contain the same set of edges are congruent. The required combinatorial result that will yield the theorem is thus that whenever two paths are related by the later condition, one can be transformed in the other by using idempotency and commutativity of loops. Once again, Simon was able to provide a proof for that and that proved the

**Theorem 3.1.** *Let $\phi : A^+ \to S$, where $S$ is locally idempotent and commutative. There exists $k \geq 0$ such that $x\ \gamma_{1,k}\ y$ implies $\phi(x) = \phi(y)$.*

## 4. More on dot-depth one languages

The analysis in the previous section described how the semigroup $A^+/\gamma_{1,k}$ could be decomposed into $S \circ A^+/\gamma_{0,k}$ where $S$ is idempotent and commutative, *i.e.* $S$ is testing for the presence of letters over a certain extended alphabet. Simon conducted exactly the same analysis on the semigroup $A^+/\gamma_{m,k}$; this semigroup can be decomposed into $S \circ A^+/\gamma_{0,k}$ where $S$ is $\mathcal{J}$-trivial, *i.e.* $S$ is testing for the presence of subwords over the same extended alphabet. In this case as well the converse also holds and this analysis can be summarized by saying that $L$ is a $\gamma_{m,k}$-language iff $S(L)$ is in the variety $\mathbf{J} * \mathbf{LI}$.

This result does not give an effective criterion to decide if a language has dot-depth one as we do not know *a priori* how to check if a semigroup decomposes as above. It is easy to see that $\mathbf{J} * \mathbf{LI}$ is contained in $\mathbf{LJ} = \{S : eSe$ is $\mathcal{J}$-trivial for all idempotents $e\}$ and Simon conjectured this was an equality. In [3] the conjecture was disproved and a subtle characterization was established for membership in $\mathcal{D}_1$.

Further work on varieties of the form $\mathbf{V} * \mathbf{LI}$ was conducted by Straubing [10] who in particular gave a post-Eilenberg presentation of the ideas used by Simon to prove his theorem on locally testable languages. Around that time, [13] obtained a difficult characterization for the variety $\mathbf{Com} * \mathbf{LI}$. In parallel, Tilson completely formalized the graph approach (which he expressed in the equivalent language of categories) to study general instances of the wreath product [14]. This point of view can also be adapted to the case of the block product, which is the two-sided incarnation of the wreath operation. The key idea is that, given $S$ and $T$, one can actually construct the "best" solution to the equation $S$ divides $X \circ T$, and this optimal value for $X$ is a category rather than a semigroup. An exact similar statement holds in the case of block product.

## 5. Conclusion

Imre Simon had powerful intuition. In retrospect, his idea of parametrizing dot-depth one languages using sequences of segments may look obvious; at the time it was providing a new point of view that proved to be most helpful in much more general situations. Simon was also able to see that his construction was naturally connected to the wreath product operation; the graph theoretical presentation he gave for his problem was an important piece in the puzzle of understanding correctly this operation which eventually led to the categorical presentation of Tilson. His theorem on $\mathcal{J}$-trivial monoids and piecewise testable languages is a gem that would deserve to be better known in the mathematical world at large. Indeed, our field was fortunate to be able to count on his ability.

## References

[1] R.S. Cohen and J.A. Brzozowski, Dot-depth of star-free events. *J. Comput. Syst. Sci.* **5** (1971) 1–15.

[2] S. Eilenberg, *Automata, Languages and Machines*, Vol. B. Academic Press, New York (1976).

[3] R. Knast, A semigroup characterisation of dot-depth one languages. *RAIRO Inform. Théor.* **17** (1984) 321–330.

[4] R. McNaughton and S. Papert, *Counter-free automata.* MIT Press, Cambridge, Massachussetts (1971).

[5] J.E. Pin, *Varieties of Formal Languages.* Plenum, London (1986).

[6] J.E. Pin, Polynomial closure of group languages and open sets of the hall topology. *Theor. Comput. Sci.* **169** (1996) 185–200.

[7] J.E. Pin and P. Weil, Polynomial closure and unambiguous product. *Theor. Comput. Syst.* **30** (1997) 1–39.

[8] M. Schützenberger, On finite monoids having only trivial subgroups. *Inform. Control* **8** (1965) 190–194.

[9] I. Simon, *Hierarchies of events with dot-depth one.* Ph.D. thesis, University of Waterloo (1972).

[10] H. Straubing, Finite semigroup varieties of the form $\mathbf{V} * \mathbf{D}$. *J. Pure Appl. Algebra* **36** (1985) 53–94.

[11] H. Straubing and D. Thérien, Partially ordered finite monoids and a theorem of I. Simon. *J. Algebra* **119** (1988) 393–399.

[12] D. Thérien, Classification of finite monoids: The language approach. *Theor. Comput. Sci.* **14** (1981) 195–208.

[13] D. Thérien and A. Weiss, Graph congruences and wreath products. *J. Pure Appl. Algebra* **36** (1985) 205–212.

[14] B. Tilson, Categories as algebra: An essential ingredient in the theory of monoids. *J. Pure Appl. Algebra* **48** (1987) 83–198.