The abstract notion of recognition: algebra, logic and topology

Mai Gehrke\textsuperscript{1}, Serge Grigorieff\textsuperscript{2} and Jean-Éric Pin\textsuperscript{2}

\textsuperscript{1}Radboud University Nijmegen  \textsuperscript{2}LIAFA, CNRS and Université Paris Diderot

September 2010, Amiens

The authors acknowledge support from the AutoMathA programme of the European Science Foundation and from the programme Research in Paris.
A new approach to the notion of recognition

Specific features

(1) It does not rely on automata.
(2) It applies to any lattice of languages rather than to individual languages.
(3) Topology is the key ingredient.

Key properties

(1) A well-defined notion of syntactic space.
(2) A purely topological characterization of the lattices of recognizable sets.
(3) Equational theory for any lattice of languages.
Lattices of subsets

Let $M$ be a monoid. A lattice of subsets of $M$ is a set of subsets of $A^*$ containing $\emptyset$ and $M$ and closed under finite intersection and finite union.

Given $u \in M$ and a subset $L$ of $M$, the quotients of $L$ by $u$ are $u^{-1}L = \{x \in M \mid ux \in L\}$ and $Lu^{-1} = \{x \in M \mid xu \in L\}$

A lattice of subsets $\mathcal{L}$ is closed under quotients if for each $L \in \mathcal{L}$ and $u \in M$, $u^{-1}L, Lu^{-1} \in \mathcal{L}$. 

The traditional algebraic approach to recognition

Let $L$ be a subset of a monoid $M$. A monoid $N$ recognizes $L$ if there exists a surjective morphism $\varphi : M \to N$ such that $L = \varphi^{-1}(\varphi(L))$.

A subset $L$ of $M$ is said to be recognizable if it is recognized by some finite monoid.

For $A^*$, this notion coincide with the standard notion of regular sets recognized by finite automata.
The traditional syntactic monoid of a subset

Let \( L \) be a subset of a monoid \( M \). The syntactic congruence of \( L \) is the relation \( \sim_L \) defined on \( M \) by 
\[
 u \sim_L v \text{ iff, for all } x, y \in M, 
\]
\[
 xuy \in L \iff xvy \in L 
\]

The quotient monoid \( M/\sim_L \) is called the syntactic monoid of \( L \).

Universal property. A monoid \( N \) recognizes \( L \) iff the syntactic monoid of \( L \) is a homomorphic image of \( N \).
Let $\mathcal{L}$ be a Boolean algebra of subsets of $M$. Let $\sim_\mathcal{L}$ be the relation on $M$ defined by $u \sim_\mathcal{L} v$ iff, for all $L \in \mathcal{L}$, for all $x, y \in M$,

$$xuy \in L \iff xvy \in L$$

Then $\sim_\mathcal{L}$ is the syntactic congruence of $\mathcal{L}$ and the monoid $M/\sim_\mathcal{L}$ is called the (algebraic) minimum recognizer of $\mathcal{L}$ since it satisfies the same kind of universal property as the syntactic monoid.
Examples

Let $S_\leq = \{ u \in \{ a, b \}^* \mid |u|_a = |u|_b \}$. Its syntactic monoid is $\mathbb{Z}$.

Let $S_\geq = \{ u \in \{ a, b \}^* \mid |u|_a \geq |u|_b \}$. Its syntactic monoid is also $\mathbb{Z}$.

The minimum recognizer of the set of finite or cofinite languages of $A^*$ is $A^*$.

The minimum recognizer of the set of context sensitive languages of $A^*$ is also $A^*$.

Obviously, this algebraic notion does not suffice. . .
Part I

Duality
Stone duality in a nutshell

The dual space of a distributive lattice with 0 and 1 is the set of its prime filters.

- Elements $\leftrightarrow$ Prime filters
- Boolean algebras $\leftrightarrow$ Topological spaces
- Distributive lattices $\leftrightarrow$ Ordered topological spaces
- Sublattices $\leftrightarrow$ Quotient spaces
- $n$-ary operations $\leftrightarrow$ $(n + 1)$-ary relations
Prior applications to regular languages

[Almeida ↑89] **Duality** between varieties of regular languages and clopen sets of **profinite** monoids.

[Pippenger 97] Slightly more general results: **Stone duality** explicitly mentioned.

[GGP 08] **Extended Stone duality** for any lattice of regular languages. The product is the dual of the residuation operations.

\[ XY \subseteq Z \iff Y \subseteq X \setminus Z \iff X \subseteq Z / Y \]

**Equational theory** for lattices of regular languages.
Prime filters

Let $\mathcal{L}$ be a lattice of subsets of $M$.

A prime filter on $\mathcal{L}$ is a subset $\mathcal{F}$ of $\mathcal{L}$ such that

1. $M \in \mathcal{F}$, $\emptyset \notin \mathcal{F}$,
2. if $X \in \mathcal{F}$ and $X \subseteq Y$, then $Y \in \mathcal{F}$,
3. if $X, Y \in \mathcal{F}$, then $X \cap Y \in \mathcal{F}$,
4. if $X \cup Y \in \mathcal{F}$, then either $X \in \mathcal{F}$ or $Y \in \mathcal{F}$.

A valuation is a lattice morphism $\nu$ from $\mathcal{L}$ to the Boolean lattice $\{0, 1\}$.

1. $\nu(\emptyset) = 0$, $\nu(M) = 1$
2. $\nu(X \cup Y) = \nu(X) + \nu(Y)$
3. $\nu(X \cap Y) = \nu(X) \nu(Y)$
Dual space

If \( v : \mathcal{L} \to \{0, 1\} \) is a valuation, then the set \( v^{-1}(1) \) is a prime filter. If \( p \) is a prime filter, the map

\[
v(x) = \begin{cases} 
1 & \text{if } x \in p \\
0 & \text{otherwise}
\end{cases}
\]

is a valuation.

The dual space of a lattice of subsets is the set of its prime filters, or, equivalently, the set of its valuations.
Let $S(\mathcal{L})$ be the (Stone) dual space of $\mathcal{L}$. The map $e : \mathcal{L} \to \mathcal{P}(S(\mathcal{L}))$ given by

$$e(d) = \{ \text{prime filters containing } d \}$$

defines an injective lattice morphism (not necessarily surjective).

One defines a topology on $S(\mathcal{L})$ for which the sets of the form $e(d)$ ($d \in \mathcal{L}$) and their complement form a basis of clopen sets. The space $S(\mathcal{L})$ is a compact ordered topological space (not necessarily Hausdorff).
Main definitions

**Definition 1.** The dual of a lattice of subsets of a monoid is called its syntactic space.

**Definition 2.** If $L$ is a subset of a monoid, its syntactic space is the dual space of the Boolean algebra $B(L)$ generated by the quotients of $L$.

**Theorem (GGP2008)**

If $L$ is recognizable, its dual space is the usual syntactic monoid.
Dual of a finite distributive lattice

In this case, all prime filters are principal.
The language \( S_\equiv = \{ u \in \{a, b\}^* \mid |u|_a = |u|_b \} \)

Let \( \pi : A^* \to \mathbb{Z} \) defined by \( \pi(u) = |u|_a - |u|_b \). Then \( S_\equiv = \pi^{-1}(0) \). It follows that the syntactic space of \( S_\equiv \) in \( A^* \) is equal to the syntactic space of \( \{0\} \) in \( \mathbb{Z} \).

The quotients of \( \{0\} \) are the singletons \( \{n\} \). They generate the lattice of all finite or full subsets of \( \mathbb{Z} \) and the Boolean algebra of all finite or cofinite subsets of \( \mathbb{Z} \).
Subsets of $\mathbb{Z}$ (1)

Let $\mathcal{L} = \{\text{finite or full subsets of } \mathbb{Z}\}$. Filters:

1. The principal ones
   $$p_n = \{\text{finite or full subsets containing } n\}.$$  
2. The principal filter $\{\mathbb{Z}\}$.

Let $\mathcal{L} = \{\text{finite or cofinite subsets of } \mathbb{Z}\}$. Filters:

1. The principal ones
   $$p_n = \{\text{finite or cofinite subsets containing } n\}.$$  
2. The nonprincipal filter of all cofinite subsets.

Syntactic space: $\mathbb{Z} \cup \{\infty\}$. Also the syntactic space of $S_\infty = \{u \in \{a, b\}^* \mid |u|_a = |u|_b\}$. 
The language \( S_{\geq} = \{ u \in \{a, b\}^* \mid |u|_a \geq |u|_b \} \)

Similarly, the syntactic space of \( S_{\geq} \) in \( A^* \) is equal to the syntactic space of \([0, +\infty[\) in \( \mathbb{Z} \).

The quotients of \([0, +\infty[\) are the sets \([n, +\infty[\). A subset of \( \mathbb{Z} \) belongs to the Boolean algebra \( \mathcal{L} \) generated by these sets iff it has a finite symmetric difference with one of the sets \( \emptyset, \mathbb{Z}, ]-\infty, 0] \) or \([0, +\infty[\). Filters:

(1) The principal ones: \( \{ L \in \mathcal{L} \mid L \text{ contains } n \} \),

(2) The filter of cofinite subsets of \([0, +\infty[\),

(3) The filter of cofinite subsets of \([ ]-\infty, 0] \).

Syntactic space: \( \mathbb{Z} \cup \{ -\infty, +\infty \} \).
Part II

The profinite world
Separating words

A deterministic finite automaton (DFA) separates two words if it accepts one of the words but not the other one.

A monoid $M$ separates two words $u$ and $v$ of $A^*$ if there exists a monoid morphism $\varphi : A^* \rightarrow M$ such that $\varphi(u) \neq \varphi(v)$.

Proposition

One can always separate two distinct words by a finite automaton (respectively by a finite monoid).
The profinite metric

Let $u$ and $v$ be two words. Put

$$r(u, v) = \min \{|M| \mid M \text{ is a finite monoid that separates } u \text{ and } v\}$$

$$d(u, v) = 2^{-r(u,v)}$$

Then $d$ is an ultrametric, that is, for all $x, y, z \in A^*$,

1. $d(x, y) = 0$ iff $x = y$,
2. $d(x, y) = d(y, x)$,
3. $d(x, z) \leq \max\{d(x, y), d(y, z)\}$
Another profinite metric

Let

\[ r'(u, v) = \min \{ \# \text{ states}(A) \mid A \text{ is a finite DFA separating } u \text{ and } v \} \]

\[ d'(u, v) = 2^{-r'(u, v)} \]

The metric \( d' \) is uniformly equivalent to \( d \):

\[ 2^{-\frac{1}{d'(u, v)}} \leq d(u, v) \leq d'(u, v) \]

Therefore, a function is uniformly continuous for \( d \) iff it is uniformly continuous for \( d' \).
Main properties of $d$

Intuitively, two words are close for $d$ if one needs a large monoid to separate them.

A sequence of words $u_n$ is a **Cauchy sequence** iff, for every morphism $\varphi$ from $A^*$ to a finite monoid, the sequence $\varphi(u_n)$ is ultimately constant.

A sequence of words $u_n$ **converges** to a word $u$ iff, for every morphism $\varphi$ from $A^*$ to a finite monoid, the sequence $\varphi(u_n)$ is ultimately equal to $\varphi(u)$. 
The free profinite monoid

The completion of the metric space \((A^*, d)\) is the free profinite monoid on \(A\) and is denoted by \(\hat{A}^*\). It is a compact space, whose elements are called profinite words.

The concatenation product is uniformly continuous on \(A^*\) and can be extended by continuity to \(\hat{A}^*\).

Any morphism \(\varphi : A^* \to M\), where \(M\) is a (discrete) finite monoid extends in a unique way to a uniformly continuous morphism \(\hat{\varphi} : \hat{A}^* \to M\).
A nonfinite profinite word

For each $u \in A^*$, the sequence $u^n$ is a Cauchy sequence and hence converges in $\hat{A}^*$ to a limit, denoted by $u^\omega$. If $\varphi$ is a morphism from $A^*$ onto a finite monoid, $\varphi(u^\omega)$ is the unique idempotent $x^\omega$ of the semigroup generated by $x = \varphi(u)$. 

\[
\begin{align*}
&1 \quad x \quad x^2 \quad x^3 \\
&\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
&x^i + 1 \quad x^i + 2 \quad x^i + p \quad x^i + p - 1 \quad x^\omega
\end{align*}
\]
Recognizable subsets of $A^*$

Given a prime filter $p$ of $\text{Rec}(A^*)$, there is a unique profinite word $u$ such that, for every morphism $\varphi$ from $A^*$ onto a finite monoid $M$, $\hat{\varphi}(u)$ is the unique element $m \in M$ such that $\varphi^{-1}(m) \in p$.

If $u$ is a profinite word, the set

$$p_u = \{ L \in \text{Rec}(A^*) \mid \varphi^{-1}(\hat{\varphi}(u)) \subseteq L \text{ for some morphism } \varphi \text{ from } A^* \text{ onto a finite monoid } \}$$

is a prime filter of $\text{Rec}(A^*)$.

Thus the syntactic space of $\text{Rec}(A^*)$ is the free profinite monoid.
Lattice of recognizable subsets of $A^*$

More generally, if $\mathcal{L}$ is a lattice of recognizable languages of $A^*$ closed under quotients, one can define a free pro-$\mathcal{L}$ monoid, which is exactly the syntactic space of $\mathcal{L}$.

Since $\mathcal{L}$ is a subspace of $\text{Rec}(A^*)$, its syntactic space is a quotient of the dual space of $\text{Rec}(A^*)$, namely the free profinite monoid. Quotients are obtained by identifying some profinite words $u = v$, leading to an equational description of $\mathcal{L}$.

Well-known example: star-free languages are defined by the identity $x^\omega = x^{\omega+1}$. 
Recognizable subsets of $\mathbb{Z}$

Let $\mathcal{L} =$ the set of recognizable subsets of $\mathbb{Z}$ (finite unions of sets of the form $\{a + n\mathbb{Z} \mid n > 0, a \in \mathbb{Z}\}$). Its syntactic space is the one-generated profinite free group $\hat{\mathbb{Z}}$.

Let $\mathcal{L} =$ the Boolean algebra generated by the finite and the recognizable subsets of $\mathbb{Z}$. Its syntactic space is the disjoint union $\mathbb{Z} \cup \hat{\mathbb{Z}}$: $\mathbb{Z}$ corresponds to the principal prime filters of $\mathcal{L}$, $\hat{\mathbb{Z}}$ corresponds to the nonprincipal ones.
Rational subsets of $\text{Rat}(\mathbb{Z})$

Let $\mathcal{L} = \text{Rat}(\mathbb{Z})$, the rational subsets of $\mathbb{Z}$. Its syntactic space is

$$(\hat{\mathbb{Z}} \times \{-\}) \cup \mathbb{Z} \cup (\hat{\mathbb{Z}} \times \{+\})$$

$\mathbb{Z}$ corresponds to the principal filters.
$\hat{\mathbb{Z}} \times \{-\}$ corresponds to the nonprincipal filters containing $]-\infty, 0]$.
$\hat{\mathbb{Z}} \times \{+\}$ corresponds to the nonprincipal filters containing $[0, +\infty[$.
The syntactic space of the Boolean algebra of all languages is $\beta A^*$, the Stone-Čech compactification of $A^*$.

One can define $\beta A^*$ as the closure of the range of $A^*$ in $\prod_\varphi \varphi(A^*)$, where $\varphi$ is any function from $A^*$ into a compact space. $\beta A^*$ is compact, but it is not a compact monoid...
Equations

Theorem (GGP ICALP 08)

A set of recognizable languages of $A^*$ is a Boolean algebra closed under quotients iff it can be defined by a set of equations of the form $u = v$, where $u, v$ are profinite words.

Theorem

A set of languages of $A^*$ is a Boolean algebra closed under quotients iff it can be defined by a set of equations of the form $u = v$, where $u, v \in \beta A^*$. 
Let $\mathcal{L}$ be a Boolean algebra of subsets of $M$.

One can equip the minimum recognizer of $\mathcal{L}$ with a uniform space (= abstract version of metric space capturing the notion of proximity).

**Theorem (GGP 2010)**

The completion of the minimum recognizer is the syntactic space of $\mathcal{L}$.

The syntactic space satisfies a suitable universal property.
An interesting question

The minimum recognizer is a semiuniform monoid: for a fixed $s \in M$, the translations $x \mapsto sx$ and $x \mapsto xs$ are uniformly continuous.

But in general, the product $(x, y) \mapsto xy$ is not uniformly continuous.

When is the product uniformly continuous?
When is the product uniformly continuous?

**Theorem**

Let $\mathcal{L}$ be a Boolean algebra of subsets of $M$ closed under quotients. TFCAE:

1. its minimum recognizer is a **uniform monoid**, 
2. the **closure of the product** of its minimum recognizer is **functional**, 
3. its syntactic space is a **compact monoid**, 
4. the elements of $\mathcal{L}$ are all **recognizable**.
Back to the examples (1)

The minimum recognizer of $S_-$ is $\mathbb{Z}$. Its syntactic space is $\hat{\mathbb{Z}} = \mathbb{Z} \cup \{\infty\}$. The closure of the addition on $\mathbb{Z}$ is the relation $\hat{+}$ given by

<table>
<thead>
<tr>
<th>$\hat{+}$</th>
<th>$i$</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j$</td>
<td>{i + j}</td>
<td>{\infty}</td>
</tr>
<tr>
<td>$\infty$</td>
<td>{\infty}</td>
<td>$\hat{\mathbb{Z}}$</td>
</tr>
</tbody>
</table>
Let $\text{Rec}(\mathbb{Z})$ be the set of \textit{recognizable subsets of $\mathbb{Z}$}, that is, the finite unions of subsets of the form \[ \{a + n\mathbb{Z} \mid n \geq 1, 0 \leq a < n\}. \]

Its \textit{minimum recognizer} is $\mathbb{Z}$ and its \textit{syntactic space} $\hat{\mathbb{Z}}$ is the one-generated \textit{profinite free group}. The closure of the addition is here the addition of $\hat{\mathbb{Z}}$.

We denote by $i \mapsto \bar{i}$ the natural embedding of $\mathbb{Z}$ into $\hat{\mathbb{Z}}$ and by $+$ the addition on $\hat{\mathbb{Z}}$. 
Let $\mathcal{L}$ be the Boolean algebra generated by the finite subsets and the recognizable subsets of $\mathbb{Z}$. Its minimum recognizer is $\mathbb{Z}$ and its syntactic space is the disjoint union $\mathbb{Z} \cup \hat{\mathbb{Z}}$: $\mathbb{Z}$ corresponds to the principal ultrafilters of $\mathcal{L}$ and the profinite group $\hat{\mathbb{Z}}$ corresponds to the nonprincipal ultrafilters.

The closure $\hat{+}$ of the addition on $\mathbb{Z}$ is commutative but nonfunctional. It extends $+$ on $\mathbb{Z}$ and, for $i \in \mathbb{Z}$ and $u, v \in \hat{\mathbb{Z}}$, one has $i \hat{+} u = i + u$ and

$$u \hat{+} v = \begin{cases} \{k, k\} & \text{if } u + v = k \text{ with } k \in \mathbb{Z} \\ \{u + v\} & \text{otherwise.} \end{cases}$$
Part III

Applications to logic
Logic on words

To each nonempty word $u = a_1 \cdots a_n$ is associated a structure

$$M_u = (\{1, 2, \ldots, n\}, (a)_{a \in A})$$

where $a$ is a predicate symbol interpreted as the set of positions $i$ such that the $i$-th letter of $u$ is an $a$.

If $u = abbaab$, then $\text{Dom}(u) = \{1, 2, 3, 4, 5, 6\}$, $a = \{1, 4, 5\}$ and $b = \{2, 3, 6\}$.

We also use the relation symbol $<$ with its usual interpretation on the integers.
Some examples

The language defined by a sentence $\varphi$ is

$$L(\varphi) = \{u \in A^* \mid \mathcal{M}_u \text{ satisfies } \varphi\}$$

For instance the sentence $\exists x \ ax$ defines the language $A^* a A^*$.

The formula $\exists x \ \exists y \ (x < y) \land ax \land by$ defines the language $A^* a A^* b A^*$.

The formula $\exists x \ \forall y \ (x < y) \lor (x = y) \land ax$ defines the language $a A^*$. 
Characterization of some logical fragments

**Theorem** [Büchi 1960, Elgot 1961]
A language is **MSO**[<, a]-definable iff it is recognizable.

**Def.** If $x$ is a profinite word, then the sequence $x^n$ is Cauchy and converges to a profinite word $x^\omega$.

**Theorem** [Schützenberger 65 + McNaughton 71]
A language is **FO**[<, a]-definable iff its syntactic monoid satisfies the profinite equation $x^{\omega+1} = x^\omega$.

**Corollary.** One can effectively decide whether a given recognizable language is **FO**[<, a]-definable.
Let $\text{MOD}$ be the set of modular predicates, e.g. $x \equiv 1 \text{ mod } 6$.

**Theorem (Barrington et al. 1992)**

A language is $\text{FO}[^{<, \text{MOD}, a}]$-definable iff its syntactic monoid satisfies the profinite equation $(x^{\omega-1}y)^\omega = (x^{\omega-1}y)^{\omega+1}$ for all words $x, y$ of the same length.
Logic and circuit complexity

Let $\mathcal{N}$ be the class of all numerical predicates. Then the $\text{FO}[\mathcal{N}]$-definable languages of $A^*$ form a Boolean algebra, whose syntactic space is $\beta\mathbb{N}$.

It is known that $\text{FO}[\mathcal{N}, a]$ defines $\text{AC}^0$, the class of languages computed by unbounded fanin, polynomial size, constant-depth Boolean circuits.

What is the syntactic space of the Boolean algebra of all $\text{FO}[\mathcal{N}, a]$-definable languages?
Beyond recognizable languages

It is also known that

$$\text{FO}[\mathcal{N}, a] \cap \text{Rec}(A^*) = \text{FO}[<, \text{MOD}, a]$$

Is it possible to prove this result by using syntactic spaces?

This would permit to attack difficult conjectures in circuit complexity.
Perspectives and conclusion

We proved the existence of a minimum recognizer for any Boolean algebra of subsets of a monoid. It is a uniform space whose completion is the Stone dual of the original Boolean algebra.

For recognizable sets, one recovers the notions of syntactic monoid and free profinite monoid. For nonrecognizable sets, the syntactic space is no longer a monoid but is still a compact space with a ternary relation in place of the multiplication.
Perspectives and conclusion

Every Boolean algebra of languages admits an **equational** description. **Equations** suffice to **separate** classes of languages. Finding them is another matter. . .

The theory can be extended to **lattices** of subsets (for instance, classes of languages defined by **logical fragments**)

It also extends to **more general algebras** than monoids.