LANGAGES FORMELS

ENS Paris-Saclay
DER informatique

2ème semestre 2025-26

Semaine 4

9. Star-free languages

What can we do without Kleene star?

Theorem (Schützenberger; McNaughton & Papert) A language $L \subseteq \Sigma^*$ is starfree if, and only if, L is first-order definable.

We thus have three equivalent conditions on a language $L \subseteq \Sigma^*$:

1) L is starfree
2) L is first-order definable
3) M_L is finite and aperiodic.

We will only prove $(1) \Leftrightarrow (3)$ and $(1) \Rightarrow (2)$ here. (We may do $(2) \Rightarrow (3)$ in the Logic course.)

The proof will take us on a little tour of typical techniques in the theory of monoids, automata, and logic, of which we will only see the tip of the iceberg here.

Any starfree language is first-order definable. We prove this by induction on the expression.

- If $e = a \in \Sigma$, we can take $\varphi := \exists x \, (a(x) \wedge \forall y \, (y = x))$.
- If $e = \varepsilon$, we can take $\varphi := \forall x \, (a(x) \wedge \neg a(x))$ → this is only true if there are no positions.
- If $e = e_1 + e_2$, pick φ_i such that $\mathcal{L}(\varphi_i) = \mathcal{L}(e_i)$ for $i = 1, 2$. Then $\varphi := \varphi_1 \vee \varphi_2$ defines $\mathcal{L}(e)$.
- If $e = f^c$, and φ defines $\mathcal{L}(f)$, then $\neg\varphi$ defines $\mathcal{L}(e)$.
- If $e = \emptyset$, take $\varphi := \bot$.
- If $e = e_1 \cdot e_2$, pick φ_i such that $\mathcal{L}(\varphi_i) = \mathcal{L}(e_i)$ for $i = 1, 2$.

  Let x be a variable not occurring and not quantified in φ_1 or φ_2.

  Define the formula ψ_1 by replacing in φ_1, from the outside to the inside, each '$\forall y \, \theta$' by '$\forall y \, (y \leq x \to \theta)$'

  —— " —— ψ_2 —— in φ_2 —— " —— $>$ ——.

  Take $\varphi' := \exists x \, (\psi_1 \wedge \psi_2)$, and define $\varphi := \begin{cases} \varphi' & \text{if } \varepsilon \notin \mathcal{L}(\varphi_1) \\ \varphi' \vee \varphi_2 & \text{if } \varepsilon \in \mathcal{L}(\varphi_1) \end{cases}$.

  notation: '$y > x$' means '$x \leq y \wedge \neg(x = y)$'.

  Then $\mathcal{L}(\varphi) = \mathcal{L}(\varphi_1) \cdot \mathcal{L}(\varphi_2) = \mathcal{L}(e_1) \cdot \mathcal{L}(e_2) = \mathcal{L}(e)$.

  ↳ We do not prove this in detail, but give an example below.

$\square$

Example. Consider $\varphi_1 := \exists \ell\,(a(\ell) \wedge \forall y\,(y \leq \ell))$ and $\varphi_2 := \exists f\,(b(f) \wedge \forall y\,(f \leq y))$.

Then $\mathcal{L}(\varphi_1) = \Sigma^* a$ and $\mathcal{L}(\varphi_2) = b\,\Sigma^*$.

The formula ψ_1 of the above proof is: $\exists \ell\,(\ell \leq x \wedge a(\ell) \wedge \forall y\,(y \leq x \to y \leq \ell)) \equiv a(x)$

ψ_2 : $\exists f\,(f > x \wedge b(f) \wedge \forall y\,(y > x \to y \geq f)) \equiv b(Sx)$

The formula φ is $\exists x\,(\psi_1 \wedge \psi_2)$.

For $w \in \Sigma^*$, we have

$w \models \varphi \iff$ there is $p \in \{0, \dots, |w|-1\}$ such that

$$\underbrace{w[0..p]}_{\text{prefix of } w \text{ of length } p+1} \models \varphi_1 \text{ and } \underbrace{w]p..|w|[}_{\text{suffix of } w \text{ of length } |w|-(p+1)} \models \varphi_2$$

$\iff$ there is $p \in \{0, \dots, |w|-1\}$ such that $w[p] = a$ and $w[p+1] = b$.

$\iff w \in \Sigma^* a \cdot b\,\Sigma^*$.

Recall:

Let M be a monoid. A subset G of M is a **group contained in M** if:

- G is closed under multiplication: for all $m_1, m_2 \in G$, $m_1 \cdot m_2 \in G$
- G has a unit 1_G: for all $m \in G$, $1_G \cdot m = m = m \cdot 1_G$
- for every $x \in G$, there exists $y \in G$ such that $xy = 1_G = yx$.

groups contained in M $\supsetneq$ subgroups of M

NB: We do not require that $1_G = 1_M$, and it is not the case in general.

A monoid M is **aperiodic** if every group contained in M is trivial.

For any finite monoid M, we defined:

k_x := the smallest k such that there exists $0 \leq \ell < k$ with $x^k = x^\ell$, and

ℓ_x := the smallest $\ell \geq 0$ such that $x^\ell = x^{k_x}$, and $p_x := k_x - \ell_x$.

Proposition Let M be a finite monoid. The following are equivalent:

(1) M is aperiodic; (2) for all $x \in M$, $p_x = 1$; (3) there exists $\ell \in \mathbb{N}$ such that $x^\ell = x^{\ell+1}$ for all $x \in M$.

Any starfree language has aperiodic syntactic monoid.

Let L be a starfree language. Observe that M_L is certainly finite, since L is regular.

Lemma. M_L is aperiodic if, and only if, there exists $\ell \in \mathbb{N}$ such that, for all $u, x, y \in \Sigma^*$:

$$xu^\ell y \in L \iff xu^{\ell+1} y \in L.$$

Proof. $M_L = \Sigma^*/\sim_L$, use the definition of $\sim_L$ and the characterization (3) of aperiodicity. $\square$

If M_L is aperiodic, define the **index** of L, $i(L) := \min\{\ell \in \mathbb{N} \mid \text{for all } u \in \Sigma^*,\ u^\ell \sim_L u^{\ell+1}\}$.

For $L \in \mathrm{Rec}(\Sigma^*)$, we also say L is **aperiodic** if M_L is aperiodic.

Lemma. Let $K, L \in \mathrm{Rec}(\Sigma^*)$ be aperiodic. Then $K \cup L$, $K \cdot L$ and $\Sigma^* \setminus L$ are aperiodic, and

$$i(K \cup L) \leq \max(i(K), i(L)), \quad i(K \cdot L) \leq i(K) + i(L) + 1, \quad i(\Sigma^* \setminus L) = i(L).$$

Moreover, $\emptyset$, $\{\varepsilon\}$, and $\{a\}$ are aperiodic ($a \in \Sigma$), with indices $0, 1, 2$, respectively.

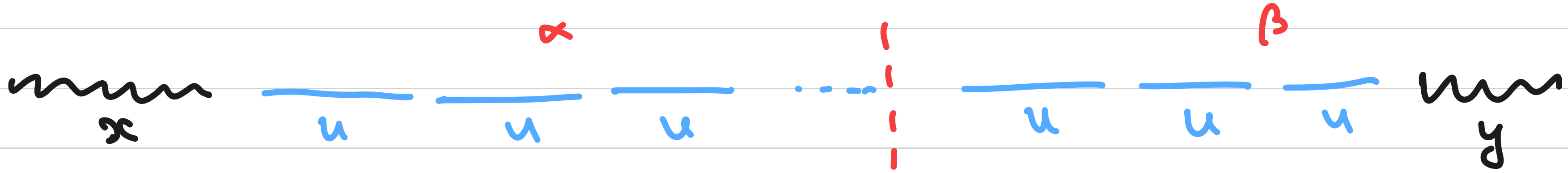
Proof. We show $\cdot$ and leave the other statements as exercises (useful for understanding $\sim_L$!)

Proof of $i(K\cdot L) \le i(K) + i(L) + 1$.

Let $l := i(K) + i(L) + 1$, and suppose $xu^l y \in K\cdot L$. Pick $\alpha \in K$, $\beta \in L$ such that $xu^l y = \alpha\beta$.

We must have either (a) there are $\geq i(K)$ copies of u in α, or (pigeon-hole principle)

(b) there are $\geq i(L)$ copies of u in β.



In case (a), we can write $\alpha = xu^{i(K)}y'$ for some $y' \in \Sigma^*$. By definition of $i(K)$, we also have $\alpha' := xu^{i(K)+1}y' \in K$. Now $\alpha'\beta = xu^{l+1}y \in K\cdot L$, as required.

In case (b), the proof is the same, using $i(L)$ and β.

This concludes the proof that $xu^l y \in K\cdot L \Rightarrow xu^{l+1}y \in K\cdot L$. The proof of the converse direction is similar, this time defining α' or β' by removing a copy of u. $\square$

We conclude from the lemma that any starfree language is aperiodic, by induction, Proof of the direction aperiodic $\Rightarrow$ starfree. On the blackboard.

Corollary. The membership problem for the class of starfree languages is decidable.

Proof. Given a regular language L, compute M_L and check whether or not it is aperiodic. $\square$

Remarks & pointers to research problems.

- Schützenberger's Theorem is part of a general correspondence theory

  classes of regular languages $\longleftrightarrow$ classes of finite monoids.

  The classes of monoids involved are called varieties of finite monoids, and are defined using a special kind of "equation" called profinite equations. E.g. $x^\omega = x^{\omega+1}$ for aperiodic, $x^\omega = 1$ for groups.

  $\rightarrow$ See, e.g., the MPRI course notes of Jean-Éric Pin https://www.irif.fr/~jep/PDF/MPRI/MPRI.pdf

- More general problems than membership are considered, e.g.,

  Starfree Separation Problem. Given regular languages L_1, L_2, does there exist a starfree language K such that $L_1 \subseteq K$ and $K \cap L_2 = \emptyset$?

  Decidable by Henckell (1988), using more involved techniques for aperiodic finite monoids.

Star-height problem. For a regular expression e, write $h(e)$ for the maximum nesting depth of $()^*$ in the expression e. $h(e)$ is called the star height of e.

For a regular language L, define $h(L) := \min\{ h(e) \mid \mathcal{L}(e) = L \}$

Fact. For every $n \in \mathbb{N}$, there exists L with $h(L) = n$.

For example, define, for $n \in \mathbb{N}$, $L_n := \{ |w|_a - |w|_b$ is divisible by $2^n \}$ has star height n.

(Exercise: find an expression of star height n for L_n. For the proof that one cannot do better, see e.g. J. Sakarovitch, Elements of Automata Theory, §6.3.)

Theorem. (Hashiguchi, 1988) The function h is computable.

(Improved algorithms by D. Kirsten 2005, T. Colcombet & C. Löding 2008).

A generalized regular expression allows $()^c$ in addition to $\emptyset, \cup, \cdot, ()^*, \varepsilon, \{a\}$.

So generalized star height 0 = starfree.

Open Problem. Does there exist any regular language of generalized star height > 1?

- **Simon's Theorem.** A regular language is piecewise testable if, and only if, its syntactic monoid is $\mathcal{J}$-trivial.

Here, L is piecewise testable if it is a Boolean combination of languages of the form,

for $u \in \Sigma^*$, $\quad \uparrow_{sub} u := \{ w \in \Sigma^* \mid u \text{ is a subword of } w \}$.

↳ recall: this means "scattered", not factor!

Equivalently, L is piecewise testable iff it is definable by an FO-sentence without quantifier alternations.
"$B\Sigma_1$"

- **k^{th} Quantifier alternation problem** Given a regular language L, does there exist an FO-sentence with at most $k-1$ quantifier alternations that defines L? "$B\Sigma_k$"

Decidable for $k=1$ by Simon's Theorem, for $k=2$ by Place and Zeitoun 2014,
for $k=3$ by Place and Zeitoun 2024, OPEN for $k>3$.

Equivalent to the Straubing-Thérien dot-depth problem: define $\mathcal{C}_0 := \{\emptyset, \Sigma^*\}$ and, for any $k \geq 0$,

$$\mathcal{C}_{k+1} := \{ L \subseteq \Sigma^* \mid L \text{ is a Boolean combination of } L_0 a_1 L_1 \dots a_n L_n \text{ where } a_1, \dots, a_n \in \Sigma,\ L_1, \dots, L_n \in \mathcal{C}_k \}$$

Open Problem (for $k>3$) Is membership in $\mathcal{C}_k$ decidable?

Krohn-Rhodes complexity. Let $A = (Q_A, \Sigma, \delta_A)$ and $B = (Q_B, \Sigma \times Q_A, \delta_B)$ be (semi-) DFAs. ↳ no I and F

Define the cascade product $A \circ B := (Q_A \times Q_B, \Sigma, \delta)$, where,

for $(q_1, q_2) \in Q_A \times Q_B$ and $a \in \Sigma$, $(q_1, q_2) \cdot a := q_2 \cdot_B (a, q_1 \cdot_A a)$.

We define $A_1 \circ \cdots \circ A_n := ((A_1 \circ A_2) \circ A_3) \circ \cdots \circ A_n$, associate on the left.

A DFA A is prime if, for every letter $a \in \Sigma$, the function $\delta_a : Q \to Q$ is either constant or bijective.

Theorem (Krohn-Rhodes, 1962) For any DFA A, there exists a DFA $B = B_1 \circ \cdots \circ B_n$ and a homomorphism $B \to A$, such that each B_i is prime.

This is also called the "prime decomposition theorem" for DFA's (or finite monoids).

Problem. (Krohn-Rhodes complexity) Given a DFA A, compute the minimum n such that a decomposition of length n exists.

OPEN for > 50 years. A solution is claimed in Margolis, Rhodes, Schilling 2024. arXiv:2406.18477