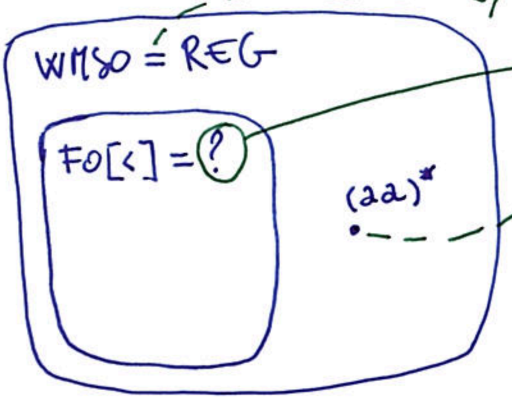
STAR-FREE LANGUAGES

So far we know:

by Büchi's theorem



The goal of this part is finding an algebraic characterisation of FO[<]-definable languages.

∉ FO by EF-theorem and winning strategy of Duplicator

Def (Star-free Languages)

Let Σ be an alphabet. The class of <u>star-free languages over Σ</u> (denoted by SF_Σ) is the smallest class so that

(1) $\emptyset \in SF_\Sigma$, $\{\varepsilon\} \in SF_\Sigma$, $\{a\} \in SF_\Sigma$ (for each $a \in \Sigma$)

(2) If L_1 and L_2 are in SF_Σ then

- $L_1 \cdot L_2 \in SF_\Sigma$
- $L_1 \cup L_2 \in SF_\Sigma$
- $\overline{L_1} \in SF_\Sigma$ ← main difference wrt REG: no Kleene star but complement

Intuition: FO[<] allows us to speak about union ($\vee$), finitely many positions at once ($L \cdot L'$) and can complement ($\neg$) but cannot speak of unboundedly many positions at the same time (L^*).

Examples The presence of complement allows us to express languages which can also be expressed using Kleene star (but not all of them)

- $\Sigma^* \in SF_\Sigma$ since $\Sigma^* = \overline{\emptyset}$
- If $L_1, L_2 \in SF_\Sigma$ then
  - $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}} \in SF_\Sigma$
  - $L_1 \setminus L_2 = L_1 \cap \overline{L_2} \in SF_\Sigma$
- Let $D \subseteq \Sigma$. Then $D^* \in SF_\Sigma$ as $D^* = \Sigma^* \setminus (\Sigma^* \cdot (\Sigma \setminus D) \cdot \Sigma^*)$
- The language $(ab)^*$ is star-free because

$$(ab)^* = \Sigma^* \setminus \underbrace{b\Sigma^*}_{\text{not starting with } b} \setminus \underbrace{\Sigma^* aa \Sigma^*}_{\text{no consecutive as}} \setminus \underbrace{\Sigma^* bb \Sigma^*}_{\text{no consecutive bs}} \setminus \underbrace{\Sigma^* a}_{\text{not ending with } a}$$

Theorem (McNaughton & Papert '71) Let $L \subseteq \Sigma^*$.

1. If L is star-free then L is FO[<]-definable
2. If L is FO[<]-definable then L is star-free

Proof of 1 as exercise

Proof of 2 requires some insights on FO[<] to handle quantifiers

- if $w \in \mathcal{L}(\exists x : \Psi)$ and $qd(\Psi) \leq k$ then

  $w = u\,\underbrace{a}_{\text{position assigned to } x \text{ to satisfy } \Psi}\,v$.

  But then for all $u' \equiv_k u$ and $v' \equiv_k v$, $u'av'$ must also be in $\mathcal{L}(\exists x : \Psi)$
- the equivalence class of u, $[u]_{\equiv_k} := \{u' \mid u' \equiv_k u\}$ can be represented as the language of a formula with $qd \leq k$
- $\equiv_k$ as finitely many classes

Notation Let $\vec{s} = s_1 \dots s_m$ and $\vec{x} = x_1 \dots x_m$.

- $\mathcal{L}_m(\varphi(\vec{x})) := \{ (S_v, \vec{s}) \mid S_v[\vec{s}/\vec{x}] \models \varphi(\vec{x}) \}$
- $\Phi_{k,m} := \{ \varphi(\vec{x}) \in FO[<] \mid qd(\varphi(\vec{x})) \leq k \}$
- $\Phi_k(S_v, \vec{s}) := \{ \varphi \in \Phi_{k,m} \mid S_v[\vec{s}/\vec{x}] \models \varphi \}$

Lemma For all $k, m \in \mathbb{N}$ the following holds

1. The relation $\equiv_{k,m}$ is an equivalence
2. For all $v \in \Sigma^*$ and positions $\vec{s} = s_1 \dots s_m$,
$$[(S_v, \vec{s})]_{\equiv_{k,m}} = \bigcap_{\varphi \in \Phi_k(S_v, \vec{s})} \mathcal{L}_m(\varphi)$$
3. There exists a finite set $\widetilde{\Phi}_{km} \subseteq \Phi_{km}$ such that
$$\Phi_{km} = \{ \varphi \mid \exists \varphi' \in \widetilde{\Phi}_{km} \text{ s.t. } \mathcal{L}_m(\varphi) = \mathcal{L}_m(\varphi') \}$$
(i.e. Φ_{km} is finite up to logical equivalence)
4. The equivalence $\equiv_{km}$ has finitely many classes, each characterised by a formula $\varphi_{[(S_v, \vec{s})]_{\equiv_k}}$ such that
$$S_w[\vec{t}/\vec{x}] \models \varphi_{[(S_v, \vec{s})]_{\equiv_{km}}} \iff (S_w, \vec{t}) \in [(S_v, \vec{s})]_{\equiv_{k,m}}$$

Proof (1) easy

(2) "$\supseteq$" immediate: every structure in $[(S_v, \vec{s})]_{\equiv_{km}}$ satisfies the same formulas $\varphi \in \Phi_{km}$ as $S_v\ \vec{s}$ by definition of $\equiv_{km}$ so it will be in every $\mathcal{L}_m(\varphi)$

"$\subseteq$" exercise

④ We need to show that, up to logical equivalence, there are only finitely many formulas with m variables an $qd \leq k$.

We proceed by induction on k.

$k=0$ For every quantifier free formula $\varphi(\vec{x}) \in FO[<]$, we can obtain an equivalent formula $\varphi'(\vec{x})$ in DNF

$$\varphi'(\vec{x}) = \bigvee_{j \in J} \underbrace{\left(\bigwedge_{i \in I} c_{ij} \right)}_{\text{disjunct}}$$

such that no disjunct is repeated and no conjunct in each disjunct is repeated.

Let us count how many different atomic propositions we can write using at most m variables:

$$P_a(x) \rightsquigarrow |\Sigma| m$$

$$x < y \rightsquigarrow m^2$$

Since each c_{ij} is either atomic or a negation of a atomic formula we get at most $2(|\Sigma| m + m^2)$ distinct c_{ij}.

Therefore we have at most $2^{2(|\Sigma| m + m^2)}$ distinct disjuncts and at most $2^{2^{2(|\Sigma| m + m^2)}}$ DNF that we need to put in $\widetilde{\Phi}_{0,m}$ to represent all possible formulas in $\Phi_{0,m}$ up to logical equivalence

$k+1$ Induction step is analogous: Count the DNF of formulas with $qd \leq k$ and $m+1$ free variables.

⑤ By ② we have $[(S_v, \vec{s})]_{\equiv_{km}} \bigcap_{\varphi \in \Phi_k(S_v, \vec{s})} \mathcal{L}_m(\varphi)$ but by ④ we know that for each $\varphi \in \Phi_{km}$ there is a formula $\varphi' \in \widetilde{\Phi}_{km}$ such that $\mathcal{L}(\varphi) = \mathcal{L}(\varphi')$

Therefore

$$[(S_v, \vec{s})]_{\equiv_{km}} = \bigcap_{\varphi \in \Phi_k(S_v, \vec{s})} \mathcal{L}_m(\varphi) = \bigcap_{\varphi \in \underbrace{\widetilde{\Phi}_k \cap \Phi_k(S_v, \vec{s})}_{\text{finite!}}} \mathcal{L}_m(\varphi) = \mathcal{L}_m\left(\bigwedge_{\varphi \in \widetilde{\Phi}_k \cap \Phi_k(S_v, \vec{s})} \varphi \right)$$

Hence there are at most as many classes in $\equiv_{km}$ as there are conjunctions of formulas in $\widetilde{\Phi}_k$ √/4

Now we can proceed with the proof of ② of McNaughton & Papert '71:

Let φ be a FO[<]-sentence. Then $\mathcal{L}(\varphi)$ is star-free

Proof We proceed by induction on the quantifier-depth k of closed formulas φ

(k=0) The only closed formulas with $qd = 0$ are, up to logical equivalence true and $\neg$true. $\mathcal{L}(\text{true}) = \Sigma^* = \overline{\emptyset} \in SF_\Sigma$, $\mathcal{L}(\neg\text{true}) = \emptyset \in SF_\Sigma$ ✓

(k+1) Assume formulas of $qd \leq k$ define star-free languages. A formula of $qd = k+1$ will be a boolean combination of formulas of the form $\varphi = \exists x : \psi$ with $qd(\psi) \leq k$. While the boolean connectives can be easily handled by using the closure properties of SF_Σ, the existential quantification requires the following characterisation

CLAIM ⊛: $$\mathcal{L}(\exists x : \psi) = \bigcup \{ [S_u]_{\equiv_k} \cdot a \cdot [S_v]_{\equiv_k} \mid S_{uav}, [{}^{|u|}/_x] \models \psi \}$$

By Lemma ④ the union is finite and there are formulas $\varphi_{[u]_{\equiv_k}}$ and $\varphi_{[v]_{\equiv_k}}$ of $qd \leq k$ such that $\mathcal{L}(\varphi_{[u]_{\equiv_k}}) = [S_u]_{\equiv_k}$ and $\mathcal{L}(\varphi_{[v]_{\equiv_k}}) = [S_v]_{\equiv_k}$

Since they have $qd \leq k$ we can apply our induction hypothesis to obtain that $R_u = \mathcal{L}(\varphi_{[u]_{\equiv_k}})$ and $R_v = \mathcal{L}(\varphi_{[v]_{\equiv_k}})$ are star-free languages.

Then $$\mathcal{L}(\exists x : \psi) = \bigcup \{ R_u \cdot a \cdot R_v \mid S_{uav}, [{}^{|u|}/_x] \models \psi \}$$

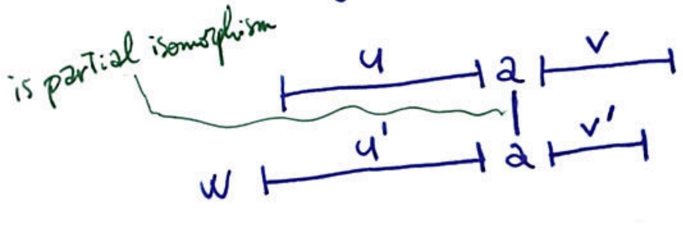
which is a finite union of concatenations of star-free languages, and hence star-free.

All is left to prove is the CLAIM ⊛

Proof of ⊛

"$\subseteq$" Let $S_w \models \exists x{:}\,\psi$. Then there is a position i with $w(i) = a$ for some $a \in \Sigma$ such that $w = uav$ $|u| = i$ and $S_{uav}, [i/x] \models \psi$, proving that $w \in [S_u]_{\equiv_k} \cdot a \cdot [S_v]_{\equiv_k}$ and hence in the union

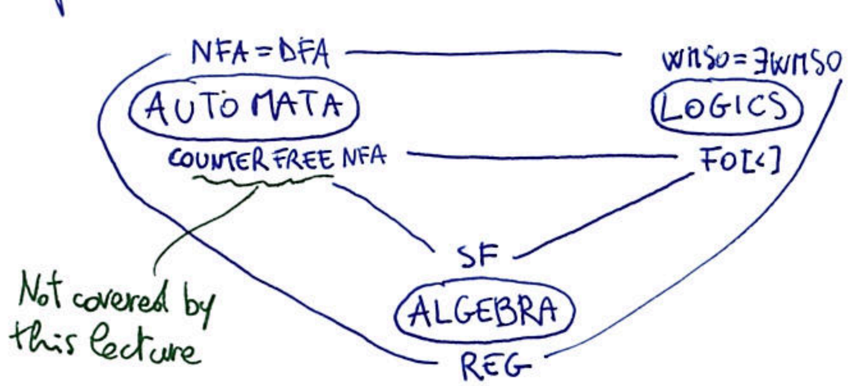
"$\supseteq$" Let $w \in [S_u]_{\equiv_k} \cdot a \cdot [S_v]_{\equiv_k}$ for some $u, v \in \Sigma^*$.
Then there are u' and v' such that $S_u \equiv_k S_{u'}$ and $S_v \equiv_k S_{v'}$ and $w = u'av'$. By the EF-theorem, Duplicator wins the games
$G_k((S_u, S_{u'})$ and $G_k(S_v, S_{v'})$.
Consider the game $G_k((S_{uav}, |u|), (S_{u'av'}, |u'|))$



We can win the game by playing winning strategy of $G_k(S_u, S_{u'})$ in first half and the one of $G_k(S_v, S_{v'})$ on second half.

Therefore $(S_{u'av'}, |u'|) \equiv_{k,1} (S_{uav}, |u|)$ by EF-theorem.
Since by assumption $S_{uav}, [|u|/x] \models \psi$ we get $S_{u'av'}, [|u'|/x] \models \psi$ and therefore $S_w \models \exists x{:}\,\psi$ which proves $w \in \mathcal{L}(\exists x{:}\,\psi)$ as desired $\square$

Overall picture:



Techniques

- REG → NFA Closure properties
- NFA → REG Arden's Lemma (Algebraic view)
- COMPLEMENTATION VIA DETERMINISATION
- NFA ≡ DFA (Automata view) Powerset
- NFA → WMSO Encoding runs using second order quantification (Logics view)
- WMSO → NFA Encoding $\exists x$ with Σ_V and projection
- $(aa)^* \notin FO[<]$ Quantifier depth and EF-theorem
- SF → FO[<] easy
- FO[<] → SF by studying $\equiv_k$ classes and EF theorem.