

# Języki, automaty i obliczenia

## Wykład 1: Języki regularne

Sławomir Lasota

**Uniwersytet Warszawski**

28 lutego 2024

1 Słowa, języki

2 Języki regularne

- $A$  – niepusty skończony zbiór, zwany *alfabetem*.
- *Słowo* nad  $A$  to dowolny skończony ciąg elementów zbioru  $A$ . Formalnie, słowo to funkcja  $w : \{1 \dots n\} \rightarrow A$ .
- *Słowo puste*  $\varepsilon$  to pusty ciąg, albo pusta funkcja  $\emptyset \rightarrow A$ .
- Długość słowa  $|w|$ , np.  $|aba| = 3$ .

## Przykład

$A = \{a, b\}$	$w_0 = b$	$w_3 = aba$	$w_6 = abaababaabaab$
	$w_1 = a$	$w_4 = abaab$	$w_7 = abaababaabaababa$
	$w_2 = ab$	$w_5 = abaababa$	$w_8 = \dots$

## Notacja

$A^*$  - zbiór wszystkich słów nad  $A$ .

$A^+$  - zbiór wszystkich niepustych słów nad  $A$ .

$A^n$  - zbiór wszystkich słów długości  $n$  nad  $A$ .

- *Konkatenacja* dwóch słów

$$w = a_1 \dots a_n \in A^n, \quad v = b_1 \dots b_m \in A^m,$$

$$w \cdot v = a_1 \dots a_n b_1 \dots b_m, \quad |w \cdot v| = n + m.$$

- Element neutralny:

$$\varepsilon \cdot w = w \cdot \varepsilon = w$$

- Łączność:

$$(w \cdot v) \cdot u = w \cdot (v \cdot u)$$

- Monoid słów  $(A^*, \cdot, \varepsilon)$ , czyli monoid wolny nad  $A$ .

## Notacja

Będziemy pomijać symbol konkatenacji i pisać  $wv$  zamiast  $w \cdot v$ .

- Jeśli  $u = wv$ , to  $w$  nazywamy *prefiksem*  $u$ , a  $v$  *sufiksem*  $u$ .
- Relacja prefiksu (sufiksu) jest porządkiem częściowym w  $A^*$ .

- *Język* nad alfabetem  $A$  to dowolny podzbiór zbioru  $A^*$ ,  $L \subseteq A^*$ .

## Przykład

$$A = \{a, b\}$$

$$L = \{w_n : n \in \mathbb{N}\} = \text{Fib.}$$

$$L = \text{palindromy.}$$

$$A = \{a\}$$

$$L = \{w \in A^* : |w| \text{ jest liczbą pierwszą}\}.$$

$$A = \{0, 1\}$$

$$L = \{\text{bin}(n) : n \text{ jest liczbą pierwszą}\}.$$

- Każdy język jest przeliczalny, ale języków jest nieprzeliczalnie wiele.

- Konkatenacja języków:

$$LM = \{wv : w \in L, v \in M\}$$

- Konkatenacja jest rozdzielna względem sumy (ale nie względem przecięcia)

$$L(M \cup N) = LM \cup LN$$

## Pytanie

Łączność? Element neutralny?

## Notacja

Dla  $n \geq 0$ , potęga  $L^n = \underbrace{L \dots L}_{n \text{ razy}}$ . W szczególności  $L^0 = \{\varepsilon\}$ .

Definicja indukcyjna:

$$L^0 = \{\varepsilon\}, \quad L^{n+1} = L^n L$$

$A^n$  to szczególny przypadek, jeśli utożsamimy  $A$  ze zbiorem wszystkich słów długości 1.

## Operacja odwrotna do konkatencji?

Iloraz lewostronny i prawostronny:

$$M^{-1}L = \{w : \exists v \in M. vw \in L\} \quad LM^{-1} = \{w : \exists v \in M. vw \in L\}$$

Uwaga:  $M^{-1}$  nic nie oznacza.

## Przykład

$$\{a^n : n \in \mathbb{N}\}^{-1}\{a^n b^n : n \in \mathbb{N}\} = \{a^m b^n : m, n \in \mathbb{N}, m \leq n\}$$

## Pytanie

- Czy  $\{a\}\{a\}^{-1}L = L$ ?
- Czy  $\{a\}^{-1}\{a\}L = L$ ?

Operacja *iteracji* (gwiazdki):

$$L^* = \bigcup_{n \in \mathbb{N}} L^n$$

### Przykład

$A = \{a\}$	$\{aa\}^*$ słowa długości parzystej	$(aa)^*$
$A = \{a, b\}$	$\{a, ab\}^*$ słowa nie zaczynające się od $b$ , w których litery $b$ nie leżą nigdy obok siebie	$(a + ab)^*$
$A = \{a, \neg a, b, c\}$	$(\{a\}\{b\})^* \{\neg a\}\{c\}$	$(ab)^* \neg a c$

$A^*$  to szczególny przypadek, jeśli utożsamimy  $A$  ze zbiorem wszystkich słów długości 1.

### Pytanie

$\emptyset^* = ?$



Języki będziemy utożsamiać z *zadaniami obliczeniowymi*, albo *problemami decyzyjnymi*.

## Przykład

Dane wejściowe: graf skierowany  $G$

Wynik: rozstrzygnąć, czy  $G$  ma cykl Hamiltona?

Graf można opisać jako słowo nad  $A = \{0, 1, \#\}$ , np. 110#101#001

Język  $L \subseteq A^*$  można utożsamiać z następującym zadaniem obliczeniowym:

Dane wejściowe:  $w \in A^*$

Wynik: rozstrzygnąć, czy  $w \in L$ ?

- *Homomorfizm* to funkcja  $h : A^* \rightarrow B^*$ , która zachowuje konkatencję:

$$h(wv) = h(w)h(v) \quad (\text{w szczególności } h(\varepsilon) = \varepsilon).$$

Innymi słowy, to homomorfizm monoidów.

- Homomorfizm jest jednoznacznie wyznaczony przez swoje wartości dla słów jednoliterowych, czyli przez funkcję

$$A \rightarrow B^*.$$

## Przykład

$$A = \{a, b\}$$

$$B = \{c, d\}$$

$h$  wyznaczony przez  $(a \mapsto cc, b \mapsto \varepsilon)$

$$h(abaab) = cccccc$$

$$\vec{h}(\text{Fib}) = ?$$

$$\vec{h}^{-1}(\{d\}^*) = \{b\}^*$$

- *Podstawienie* to funkcja  $\widehat{h} : \mathcal{P}(A^*) \rightarrow \mathcal{P}(B^*)$  wyznaczona jednoznacznie przez

$$h : A \rightarrow \mathcal{P}(B^*)$$

- rozszerzamy  $h$  do wszystkich słów:

$$\widehat{h}(\varepsilon) = \{\varepsilon\}$$

$$\widehat{h}(aw) = h(a)\widehat{h}(w)$$

- rozszerzamy  $\widehat{h}$  do języków:

$$\widehat{h}(L) = \bigcup_{w \in L} \widehat{h}(w)$$

### Przykład

$$A = \{a, b\}$$

$$B = \{0, 1\}$$

$$h(a) = 0^*$$

$$h(b) = 1^*$$

$$\widehat{h}(aba) = 0^*1^*0^*$$

$$\widehat{h}((ab)^*) = (0^*1^*)^* = B^*$$

1 Słowa, języki

2 Języki regularne

*Języki regularne* nad  $A$  to najmniejszy zbiór języków  $\mathcal{R}$  spełniający poniższe warunki:

- $\emptyset \in \mathcal{R}$ ,
- $\{a\} \in \mathcal{R}$ , dla każdego  $a \in A$ ,
- jeśli  $L \in \mathcal{R}$  i  $M \in \mathcal{R}$ , to  $L \cup M \in \mathcal{R}$ ,
- jeśli  $L \in \mathcal{R}$  i  $M \in \mathcal{R}$ , to  $LM \in \mathcal{R}$ ,
- jeśli  $L \in \mathcal{R}$ , to  $L^* \in \mathcal{R}$ .

## Pytanie

Czy  $\{\varepsilon\}$  jest regularny?

## Przykład

$$A = \{a, b\}$$

$$\{a, ab\} = \{a\} \cup \{a\}\{b\}$$

$$\{b\}\{a\}^*\{b\} \cup \{\varepsilon\}$$

$$\{a\}^*(\{b\}\{a\}^*\{b\}\{a\}^*)^*$$

$$a + ab$$

$$ba^*b + \varepsilon$$

$$a^*(ba^*ba^*)^*$$

Wyrażenie *regularne* to wyrażenie zbudowane z:

$\emptyset$     $a$     $\_ + \_$     $\_ \cdot \_$     $\_ ^*$

## Konwencja notacyjna

- $+$  zamiast  $\cup$
- nie używamy „wąsów”  $\{ \}$ , czyli  $a$  oznacza język zawierający jedno jednoliterowe słowo  $\{a\}$
- $*$  wiąże silniej niż konkatenacja, a ta silniej niż  $+$
- $\varepsilon$  oznacza język  $\{\varepsilon\} = \emptyset^*$

## Przykład

$a$	$\{a\}$
$abc + c$	$\{a\}\{b\}\{c\} \cup \{c\}$
$\varepsilon + a^*b$	$\{\varepsilon\} \cup \{a\}^*\{b\}$
$(b + \varepsilon)(a + ab)^*$	$(\{b\} \cup \{\varepsilon\})(\{a\} \cup \{a\}\{b\})^*$

- Łączność i przemienność:

$$L + (M + N) = (L + M) + N$$

$$L(MN) = (LM)N$$

$$L + M = M + L$$

ale nie:

$$LM = ML$$

- Idempotentność:

$$L + L = L$$

- Elementy neutralne i zerowe:

$$L\varepsilon = \varepsilon L = L$$

$$L\emptyset = \emptyset L = \emptyset$$

$$L + \emptyset = \emptyset + L = L$$

- Prawo rozdzielności:

$$L(M + N) = LM + LN$$

- Prawa dla gwiazdki:

$$\emptyset^* = \varepsilon$$

$$\varepsilon^* = \varepsilon$$

$$(L + \varepsilon)^* = L^*$$

$$L(ML)^* = (LM)^*L$$

$$(L^*)^* = L^*$$

$$(L^*M^*)^* = (L + M)^*$$

- Fałszywki:

$$(L + M)^* = L^* + M^*$$

$$L + \varepsilon = L$$



$X \subseteq \mathbb{N}$  jest *prawie okresowy* jeśli

$$\exists o \in \mathbb{N}. \exists m \in \mathbb{N}. \forall n > m. n \in X \iff (n + o) \in X.$$

## Twierdzenie

Gdy  $A$  jest jednoelementowy, język  $L \subseteq A^*$  jest regularny wtw gdy

$\{|w| : w \in L\}$  jest prawie okresowy.

?

## Twierdzenie

Gdy  $A$  jest jednoelementowy i  $L \subseteq A^*$ , język  $L^*$  jest regularny.

?

## Dlaczego języki regularne?

- prosta klasa, która zawiera wiele naturalnych języków
- zastosowania praktyczne, np.:
  - wyszukiwanie wzorca (grep)
  - analiza leksykalna
- klasa zamknięta na wiele naturalnych operacji
- wiele równoważnych definicji (np. *rozszerzone wyrażenia regularne*)

Wyrażenia regularne w Unixie:

.	$a + b + \dots$
[a b c d] [a-d]	$a + b + c + d$
[:digit:]	$0 + 1 + \dots + 9$
_   _	$_ + _$
e?	$e + \varepsilon$
e+	$e^* e$
e{n}	$\underbrace{e \dots e}_{n \text{ razy}}$
...	

Analiza leksykalna:

while	return(WHILE);
[A-Za-z] [A-Za-z0-9]*	return(ID);
>=	return(GE);
...	

## Twierdzenie

*Klasa języków regularnych jest zamknięta na*

- *dopełnienie: jeśli  $L$  jest regularny to dopełnienie  $L$  też* ?
- *przecięcia: jeśli  $L$  i  $M$  są regularne to  $L \cap M$  też* ?
- *obrazy homomorficzne: jeśli  $L$  jest regularny to  $\vec{h}(L)$  też*
- *przeciwwobrazy homomorficzne: jeśli  $L$  jest regularny to  $\vec{h}^{-1}(L)$  też* ?
- *podstawienia regularne: jeśli  $L$  jest regularny to  $\widehat{h}(L)$  też*
- ...

*Dowód dla obrazów homomorficznych:*

$$\begin{array}{ll} \vec{h}(\emptyset) = \emptyset & \vec{h}(LM) = \vec{h}(L)\vec{h}(M) \\ \vec{h}(\{\varepsilon\}) = \{\varepsilon\} & \vec{h}(L \cup M) = \vec{h}(L) \cup \vec{h}(M) \\ \vec{h}(\{a\}) = \{h(a)\} & \vec{h}(L^*) = (\vec{h}(L))^* \end{array}$$

- Dodajemy operację dopełnienia  $\bar{L} = A^* - L$ .

## Przykład

$$A = \{a, b\} \quad L = \overline{a^* b a^* b a^*}$$

- Rozszerzone wyrażenia regularne są równoważne wyrażeniom regularnym, ale bardziej zwarte.

## Przykład

$$A = \{0, 1, \#\} \quad v_n = \#\text{bin}(0)\#\text{bin}(1)\#\dots\#\text{bin}(2^n)\#$$

- Dopełnienie języka  $\{v_n\}$  można opisać wyrażeniem regularnym rozmiaru  $\text{poly}(n)$ .
- Najkrótsze wyrażenie regularne opisujące  $\{v_n\}$  jest wykładnicze względem  $n$ . Dlaczego?

Głębokość gwiazdkowa wyrażenia regularnego  $e$ :

$$e = (a(a+b)^*b^*c)^* \quad \text{gg}(e) = 2$$

Głębokość gwiazdkowa języka regularnego  $L$ :

$$\text{gg}(L) = \min\{\text{gg}(e) : \text{wyrażenie } e \text{ definiuje } L\}$$

Pytanie (Eggan 1963)

- Czy języki regularne mają ograniczoną głębokość gwiazdkową?
- Czy dla danego języka regularnego można obliczyć jego głębokość gwiazdkową?

Twierdzenie (Hashiguchi 1988)

*Głębokość gwiazdkowa języka regularnego jest obliczalna.*

Obydwa pytania są otwarte dla rozszerzonych wyrażeń regularnych!

Jaki to język?

$$(aa + bb + (ab + ba)(aa + bb)^*(ab + ba))^*$$

