# 6. Algoritmikus kérdések, CYK algoritmus

1. Adott egy $L \subseteq \{0, 1\}^*$ reguláris nyelv. Az a kérdés, hogy tartalmaz-e minden 0-val kezdődő szót. Adjon eljárást, ami ezt a kérdést eldönti, ha a nyelv megadásának módja
   (a) véges automata
   (b) reguláris kifejezés
   (c) reguláris nyelvtan

*Megoldás:* Jelölje $L_0$ a 0-val kezdődő szavak nyelvét. Ekkor a kérdés: Igaz-e, hogy $L \cap L_0 = L_0$ ?

(a) Vegyük $L_0$-hoz a minimálautomatát. Ebből és az $L$-et megadó véges automatából elkészítjük a metszethez tartozó véges automatát, majd ezt minimalizáljuk. Ha az eredmény izomorf az $L_0$ minimálautomatájával, akkor a válasz igen, különben a válasz nem.

(b) Készítsünk a reguláris kifejezésből véges automatát. Innen ugyanaz, mint az (a). (Előzőleg elhagyhatjuk a reguláris kifejezésből az 1-gyel kezdődő tagokat, így a kezdő automatánk kisebb lehet.)

(c) Ha elhagyjuk az $S \to 1A$ típusú szabályokat, az nem elég, mert az is kell, hogy minden $L_0$-beli generálható. De az automatára áttérés itt is működik – ezt akár csinálhatjuk a teljes nyelvtanból vagy az $S \to 1A$ alakú szabályok elhagyása után.

2. Adott egy $M$ nemdeterminisztikus véges automata és egy $R$ reguláris kifejezés. Vázoljon algoritmust annak eldöntésére, hogy
   (a) az $M$ által elfogadott $L(M)$ nyelv megegyezik-e a reguláris kifejezés $L(R)$ nyelvével,
   (b) a két nyelv csak véges sok szóban különbözik-e!

*Megoldás:*

(a) Készítsünk a reguláris kifejezésből a tanult algoritmussal véges automatát. Ezt determinizáljuk és minimalizáljuk. Most már csak azt kell ellenőrizni, hogy az $M$ determinizálása és minimalizálása után kapott automatával izomorf DVA-t kaptunk-e.

(b) Kezdjük, mint az előbb: készítsünk a reguláris kifejezésből a tanult algoritmussal véges automatát. Ezt determinizáljuk és minimalizáljuk, legyen ez $M_R$. Determinizáljuk az $M$ automatát is, ez legyen $M'$. Az a kérdés, hogy az $L = (L(M_R) - L(M')) \cup (L(M') - L(M_R))$ nyelv véges-e. Ennek eldöntéséhez készítsük el az $L$-hez tartozó minimálautomatát, felhasználva a különbséghez és az unióhoz tartozó konstrukciót (amiből kapott automatát aztán determinizálni, minimalizálni kell). Most már csak azt kell ellenőrizni, hogy ez a DVA nem tartalmaz olyan kört, amiből elérhető elfogadó állapot.

Az utóbbi rész (minimalizálás, kör keresés) helyettesíthető annak ellenőrzésével, hogy a kapott automata nem fogad el egyetlen olyan szót sem, melynek hossza $p$ és $2p$ közé esik. (Itt a pumpálási hossz helyett választhatunk annál nagyobb $p$ értéket is amire pl. a determinisztikus teljes automata állapotainak száma megfelelő.)

3. Az $L \subseteq \{0, 1\}^*$ nyelvben csak véges sok szó van, és ezek közé tartozik a 10 db 0-ból álló szó is. Igazolja, hogy a nyelv reguláris de a minimálautomatája legalább 11 állapotú!

*Megoldás:* Jelölje $s$ a minimálautomata állapotszámát.

**1. változat – használjuk a tanultakat**: Minden véges nyelv reguláris (lehet pl. NVA-t készíteni, amiben minden nyelvbeli szóhoz egy-egy külön út tartozik, amikben csak a kezdőállapot közös. Hasonlóan lehet egyszerű reguláris nyelvtant is adni). Tudjuk, hogy a nyelv pontosan akkor véges, ha minden szavának a hossza legfeljebb $p$, ami a pumpálási hosszt jelöli. Tehát most $10 < p$. A pumpálási lemma bizonyításában $p$ egy determinisztikus teljes véges automata állapotszáma, azaz lehet $p = s$. Ezért $10 < s$.

**2. változat – elemi**: A 10 hosszú szóhoz egy 11 állapotot tartalmazó elfogadó számítási út tartozik. Ha ennek állapotai között van ismétlődés, akkor van egy kör a számítási útban, ami akárhányszor ismételhető, azaz végtelen a nyelv. Tehát minden VA legalább 11 állapotú.

4. Hozza a következő nyelvtanokat Chomsky-normálformájúra! Milyen nyelvet generálnak a felsorolt nyelvtanok?

   (a) $S \to \mathtt{a}S\mathtt{a} \mid \mathtt{ab}$

   (b) $S \to \mathtt{a}S\mathtt{a} \mid \mathtt{b}S\mathtt{a} \mid \varepsilon$

   (c) $S \to \mathtt{a}A\mathtt{b}B\mathtt{c} \mid \mathtt{a}C\mathtt{b}D\mathtt{c}$, $A \to \mathtt{a}A\mathtt{b} \mid \mathtt{ab}$, $B \to B\mathtt{c} \mid \mathtt{c}$, $C \to \mathtt{a}C \mid \mathtt{a}$, $D \to \mathtt{b}D\mathtt{c} \mid \mathtt{bc}$

   (d) $E \to E + E \mid E * E \mid \mathtt{a}$

*Megoldás:*

(a) A nyelv: $\{\mathtt{a}^{k+1}\mathtt{ba}^k : k \geq 0\}$

1. rész (minden legalább 2 hosszú jobb oldal csak változókból áll):

$$S \to X_aSX_a \mid X_aX_b \qquad X_a \to \mathtt{a} \qquad X_b \to \mathtt{b}$$

2. rész (feldarabolás):

$$S \to X_aY \mid X_aX_b \qquad Y \to SX_a \qquad X_a \to \mathtt{a} \qquad X_b \to \mathtt{b}$$

(b) A nyelv: páros hosszú szavak, a felétől kezdve csupa a betű.

Ez nem egy szabályos CF nyelvtan, előbb az $\varepsilon$-szabálytól meg kell szabadulni. Ennek eredménye:

$$S' \to \varepsilon \mid S \qquad S \to \mathtt{a}S\mathtt{a} \mid \mathtt{aa} \mid \mathtt{b}S\mathtt{a} \mid \mathtt{ba}$$

A keletkezett egyszeres szabályt is meg kell szüntetni:

$$S' \to \varepsilon \mid \mathtt{a}S\mathtt{a} \mid \mathtt{aa} \mid \mathtt{b}S\mathtt{a} \mid \mathtt{ba} \qquad S \to \mathtt{a}S\mathtt{a} \mid \mathtt{aa} \mid \mathtt{b}S\mathtt{a} \mid \mathtt{ba}$$

Most jöhet az 1. rész:

$$S' \to \varepsilon \mid X_aSX_a \mid X_aX_a \mid X_bSX_a \mid X_bX_a \qquad S \to X_aSX_a \mid X_aX_a \mid X_bSX_a \mid X_bX_a \qquad X_a \to \mathtt{a} \qquad X_b \to \mathtt{b}$$

2. rész (az algoritmust követve, nem optimalizálunk!)

$$S' \to \varepsilon \mid X_aY \mid X_aX_a \mid X_bZ \mid X_bX_a \qquad S \to X_aR \mid X_aX_a \mid X_bT \mid X_bX_a \qquad X_a \to \mathtt{a} \qquad X_b \to \mathtt{b}$$

$$Y \to SX_a \qquad Z \to SX_a \qquad R \to SX_a \qquad T \to SX_a$$

(A 4 legutóbb bevezetett változót lehetne egyetlen változóval helyettesíteni.)

(c) Az $A$ változóból levezethetők: $L_A = \{\mathtt{a}^k\mathtt{b}^k : k \geq 1\}$, a $B$ változóból levezethetők: $L_B = \{\mathtt{c}^\ell : \ell \geq 1\}$, a $C$ változóból levezethetők: $L_C = \{\mathtt{a}^m : m \geq 1\}$, a $D$ változóból levezethetők: $L_D = \{\mathtt{b}^n\mathtt{c}^n : n \geq 1\}$. Ezek alapján a nyelv: $\mathtt{a}L_a\,\mathtt{b}\,L_B\,\mathtt{c} \cup \mathtt{a}\,L_C\,\mathtt{b}\,L_D\,\mathtt{c} = \{\mathtt{a}^p\mathtt{b}^q\mathtt{c}^r : p, q, r \geq 2, p = q \text{ vagy } q = r\}$

Átalakítás 1. rész:

$$S \to X_aAX_bBX_c \mid X_aCX_bDX_c$$

$$A \to X_aAX_b \mid X_aX_b \qquad B \to BX_c \mid \mathtt{c} \qquad C \to X_aC \mid \mathtt{a} \qquad D \to X_bDX_c \mid X_bX_c$$

$$X_a \to \mathtt{a} \qquad X_b \to \mathtt{b} \qquad X_c \to \mathtt{c}$$

2. rész

$$S \to X_aY_1 \mid X_aZ_1 \qquad Y_1 \to AY_2 \qquad Y_2 \to X_bY_3 \qquad Y_3 \to BX_c \qquad Z_1 \to CZ_2 \qquad Z_2 \to X_bZ_3 \qquad Z_3 \to DX_c$$

$$A \to X_aR \mid X_aX_b \qquad R \to AX_b \qquad B \to BX_c \mid \mathtt{c} \qquad C \to X_aC \mid \mathtt{a} \qquad D \to X_bT \mid X_bX_c \qquad T \to DX_c$$

$$X_a \to \mathtt{a} \qquad X_b \to \mathtt{b} \qquad X_c \to \mathtt{c}$$

(d) A nyelv az (egyszerű, zárójel nélküli) aritmetikai kifejezések nyelve.

1. rész:

$$E \to EX_+E \mid EX_*E \mid \mathtt{a} \qquad X_+ \to + \qquad X_* \to *$$

2. rész

$$E \to EY \mid EZ \mid \mathtt{a} \qquad Y \to X_+E \qquad Z \to X_*E \qquad X_+ \to + \qquad X_* \to *$$

5. Egy $A \to \alpha$ nyelvtani szabály hossza legyen $1 + |\alpha|$. Egy CF nyelvtan hossza jelentse a benne levő szabályok hosszainak összegét.

   Legyen $\Sigma = \{\mathtt{0}, \mathtt{1}\}$ és $G$ egy egyszeres szabályok nélküli CF nyelvtan, amiben $n$ változó van és a nyelvtan hossza $N$. A $G$-ből a tanult módon elkészítjük a Chomsky-normálformájú $G'$ nyelvtant. Az $n$ és $N$ paraméterek segítségével adjon felső becslést $G'$ változóinak számára és a $G'$ nyelvtan hosszára!

   *Megoldás:* A régi változók mellett $\Sigma$ minden karakteréhez tartozhat új változó és egy új szabály. Egy $A \to \alpha$ szabályból pedig $|\alpha| - 2$ új változó keletkezik és ezt $|\alpha| - 1$ új 3 hosszú szabállyal helyettesítjük. Tehát $h = 1 + |\alpha|$ hosszból $3(|\alpha| - 1) = 3(h - 2) \leq 3h$ hossz lesz. Vegyük észre, hogy ez $|\alpha| = 2$ esetében is igaz, és a végső becslés az $A \to \mathtt{a}$ alakúaknál is helyes (az utóbbi esetekben a hossz nem változik).

   A CNF-ben a változók száma $n' \leq n + |\Sigma| + N - 3 = n + N - 1$.

   Az új nyelvtan hossza $N' \leq 2|\Sigma| + 3N \leq 3N + 6$.

6. A Cocke-Younger-Kasami algoritmussal elemezzük az `aaab` szót a következő nyelvtan alapján.

$$S \to XY \mid YX \qquad X \to AZ \mid \mathtt{a} \qquad Z \to XA \qquad Y \to AT \mid AA \mid \mathtt{b} \qquad T \to AY \qquad A \to \mathtt{a}$$

   Az alábbi táblázatban már kitöltöttük a 2. és 3. sorokat.

   (a) Töltse ki az első sort!

   (b) Mit jelent az, hogy két $S$ szimbólum került a 3. sor elsõ mezőjébe?

   (c) Egészítse ki a megfelelő indexekkel a táblázatban szereplő változókat!

   (d) Mi lesz a legfelső mező tartalma?

   (e) A táblázat alapján állapítsa meg, hogy a megadott szó levezethető-e a nyelvtanból!

| | | | | |
|---|---|---|---|---|
| 4. | | | | |
| 3. | $S$ $S$<br>$X$ $T$ | $Y$ | | |
| 2. | $Z$<br>$Y$ | $Z$<br>$Y$ | $S$<br>$T$ | |
| 1. | | | | |
| | a | a | a | b |

   *Megoldás:* A nyelvtan Chomsky-normálformában van, a CYK algoritmust átalakítás nélkül lehet rá alkalmazni,

   (a) Ide azok a változók jönnek amiknek van az adott betűből álló jobb oldala, lásd a táblázat.

   (b) Azt, hogy $S$-ből az adott szó 3 hosszú kezdőszelete kétféleképpen is levezethető.

   (c-d) Lásd a táblázat. A nem egy betűs szabályokat számoztuk meg:

$$S \to \overset{1}{X}Y \mid \overset{2}{Y}X \qquad X \to \overset{3}{A}Z \mid \mathtt{a} \qquad Z \to \overset{4}{X}A \qquad Y \to \overset{5}{A}T \mid \overset{6}{A}A \mid \mathtt{b} \qquad T \to \overset{7}{A}Y \qquad A \to \mathtt{a}$$

| | | | | |
|---|---|---|---|---|
| 4. | $S_{1,1}$ $T_{7,1}$ $S_{2,3}$ | | | |
| 3. | $S_{1,1}$ $S_{2,2}$ $X_{3,1}$ $T_{7,1}$ | $Y_{5,1}$ | | |
| 2. | $Z_{4,1}$ $Y_{6,1}$ | $Z_{4,1}$ $Y_{6,1}$ | $S_{1,1}$ $T_{7,1}$ | |
| 1. | $X$ $A$ | $X$ $A$ | $X$ $A$ | $Y$ |
| | `a` | `a` | `a` | `b` |

(e) Mivel $S$ szerepel a legfelső mezőben, a szó levezethető. (Sőt, mivel kétszer is szerepel, kétféle levezetési fa is van. Az indexek mutatják, hogy az egyik az 1., a másik a 2. szabállyal indul.)

7. A CYK-algoritmussal elemezze az alábbi nyelvtant használva az `abbbba` és az `abbba` szavakat! Rajzolja fel a kapott levezetési fákat is!

$$S \to AX \mid BY \mid AA \mid BB \qquad X \to SA \qquad Y \to SB \qquad A \to \texttt{a} \qquad B \to \texttt{b}$$

*Megoldás:* A nyelvtan Chomsky-normálformában van, tehát az algoritmus a nyelvtan további átalakítása nélkül alkalmazható.

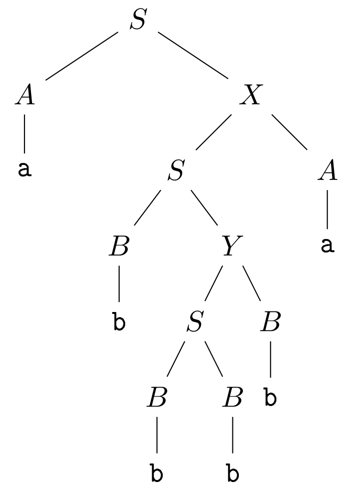
Beszámozzuk a szabályokat:

$$S \to \overset{1}{A}X \mid \overset{2}{B}Y \mid \overset{3}{A}A \mid \overset{4}{B}B \qquad X \to \overset{5}{S}A \qquad Y \to \overset{6}{S}B \qquad A \to \texttt{a} \qquad B \to \texttt{b},$$

majd kitöltjük a táblázatot.

| | | | | | | |
|---|---|---|---|---|---|---|
| 6. | $S_{1,1}$ | | | | | |
| 5. | — | $X_{5,4}$ | | | | |
| 4. | — | $S_{2,1}$ | — | | | |
| 3. | — | $Y_{6,2}$ | $Y_{6,2}$ | $X_{5,2}$ | | |
| 2. | — | $S_{4,1}$ | $S_{4,1}$ | $S_{4,1}$ | — | |
| 1. | $A$ | $B$ | $B$ | $B$ | $B$ | $A$ |
| | `a` | `b` | `b` | `b` | `b` | `a` |

Látszik, hogy egyetlen levezetési fa van a szóhoz:



A másik, rövidebb szóra:

| | | | | | |
|---|---|---|---|---|---|
| 5. | — | | | | |
| 4. | — | — | | | |
| 3. | — | $Y_{6,2}$ | $X_{5,2}$ | | |
| 2. | — | $S_{4,1}$ | $S_{4,1}$ | — | |
| 1. | $A$ | $B$ | $B$ | $B$ | $A$ |
| | `a` | `b` | `b` | `b` | `a` |

Mivel a felső mezőbe nem került be az $S$, ezért a szó nem vezethető le, nincs levezetési fája.

8. Adjon meg egy 2 szalagos, determinisztikus Turing-gépet (az átmeneti függvény leírásával) az $\{\texttt{a}^n\texttt{b}^n\texttt{c}^n : n \geq 0\}$ nyelvhez!

*Megoldás:* A működés elve: a 2. szalagra egy, a szalag elejét mutató $X$ karakter kiírása után $q_a$ állapotban lemásoljuk az `a`-kat. Amikor az 1. szalagon a `b`-khez érünk, egy új $q_b$ állapotban összehasonlítjuk a `b`-k számát a 2. szalagon levő `a`-kkal. Ha egyszerre érünk az 1. szalagon az első `c`-hez és a 2-on az $X$-hez, akkor a $q_c$ állapotban az első szalag `c`-inek számát hasonlítjuk a 2. szalag `a`-ihoz. Elfogadunk ($q_+$), ha egyszerre érünk az első $*$-hoz mindkét szalagon.

| | $(*,*)$ | (a, $*$) | (b,$*$) | (b,a) | (c,$X$) | (c,a) |
|---|---|---|---|---|---|---|
| $q_0$ | $(q_+,*,*,H,H)$ | ($q_a$,a, $X$, $H$, $J$) | | | | |
| $q_a$ | | ($q_a$,a, a, $J$, $J$) | ($q_b$,b, $*$,$H$,$B$) | | | |
| $q_b$ | | | | ($q_b$,b,a,$J$,$B$) | ($q_c$, c,$X$,$H$,$J$) | |
| $q_c$ | $(q_+,*,*,H,H)$ | | | | | ($q_c$,c,a,$J$,$J$) |

Elfogadó állapot: $q_+$ – itt a számítás véget ér. Vagy az üres bemenetnél léphetünk ide $q_0$-ból, vagy ha sikeresen feldolgoztuk az egész szót.