Reguláris kifejezések, környezetfüggetlen nyelvek

1. Legyen $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ és álljon L azokból a szavakból, melyekben az `a` és a `b` betűk száma megegyezik. Reguláris-e ez az L nyelv?

   Megoldás: Nem. Indirekt bizonyítunk, az előadáson az $\{a^n b^n\}$ nyelvre elmondott bizonyítás itt is bizonyít.

   Fontos: általában az nem igaz, hogy egy nem reguláris nyelvet tartalmazó nyelv sem reguláris! Miért?

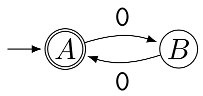
2. Álljon az ábécé a nyitó és a csukó zárójelből. Igazolja, hogy a helyes zárójelsorozatokból álló nyelv nem reguláris!

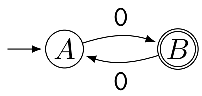
   Megoldás: Az előző bizonyítás itt is érvényes, csak az `a` helyett nyitó, a `b` helyett csukó zárójellel.

3. Reguláris-e az a nyelv, ami az olyan, csupa `0` sorozatból áll, amelyeknek a hossza

   (a) páros szám? (b) páratlan szám?
   (c) négyzetszám? (d) kettő hatvány?

   Megoldás:

   (a) Igen: 

   (b) Igen: 

   A (c) és (d) esetén a válasz nem: gondoljuk meg, hogyan néz ki egy DVA ha csak egy elemű az ábécé! Minden állapotból egyetlen nyíl (átmenet) indul ki. A kezdőállapotból a gráfban van egy mondjuk t hosszú út, aminek utolsó csúcsán vagy egy hurok van vagy innen az él visszamutat egy korábbi állapotba. Tehát a gráf egy kezdeti útból és egy az út végén levő körből áll (az út lehet 0 hosszú, a kör meg 1 hosszú, utóbbi ha hurok van). Ha nincs elfogadó állapot a körön, akkor csak véges sok különböző szót tud elfogadni. Ha van (lehet akár több is), akkor végtelen sok szót. Méghozzá, ha c jelöli a kör hosszát, és $\mathtt{0}^k \in L$ egy a körön levő elfogadó állapotban ér véget, akkor kőrbe érve $\mathtt{0}^{k+c} \in L$ is teljesül. Ezért az nem fordulhat elő, hogy egy adott nyelvnél az elfogadott szavak hosszai között tetszőlegesen nagy ugrás előforduljon.

   A korábbi bizonyítási technika is működik, például így ha t állapota van DVA-nak, akkor vegyük a nyelv $t+1$ darab legrövidebb szavát. Biztos van közöttük kettő, ami ugyanabban az (elfogadó) állapotban ér véget. Tegyük fel, hogy $|w_1| = k^2$ és $|w_2| = \ell^2$ ugyanabban a q állapotban végződik és $k < \ell$ Ha q-ból még $2k+1$ lépést teszünk, akkor elfogadó állapothoz kell jussunk, hiszen ha a w_1-et folytatjuk, akkor egy $k^2 + 2k + 1 = (k+1)^2$ hosszú szót kapunk. Viszont ha a w_2-t folytatjuk, akkor is elfogad az automata, pedig a szó hossza $\ell^2 + 2k + 1 < \ell^2 + 2\ell + 1 = (\ell+1)^2$, és mivel ugyanakkor nagyobb mint ℓ^2, ezért nem négyzetszám.

   A (d)-re ugyanez az ötlet (de más számolás) működik.

4. Legyen az ábécé $\Sigma = \{\mathtt{0}, \mathtt{1}\}$. Határozza meg az alábbi reguláris kifejezésekhez tartozó nyelveket!

   (a) $(\mathtt{0}+\mathtt{1})^*\mathtt{011}(\mathtt{0}+\mathtt{1})^*$ (b) $\mathtt{1}(\mathtt{0}+\mathtt{1})^*\mathtt{0}$ (c) $((\mathtt{0}+\mathtt{1})(\mathtt{0}+\mathtt{1}))^*$

   Megoldás: (a) A `011`-et tartalmazó szavak. (b) Az `1`-gyel kezdődő és `0`-ra végződő szavak. (c) A páros hosszú szavak (tetszőleges kettő hosszú szavakból rakunk egymás után valahány, akár nulla darabot).

5. Adjon reguláris kifejezést azokra a nyelvekre, amelyek a $\{\mathtt{0}, \mathtt{1}\}$ ábécé felett a következő szavakból állnak!

   (a) páratlan hosszú szavak;

   (b) páros hosszú nem üres szavak melyeknek első és utolsó karaktere is `1`;

   (c) legalább 3 db `0`-t tartalmazó szavak;

   (d) páros sok `0`-t tartalmazó szavak;

   (e) a `0`-val kezdődő és páratlan hosszú, valamint az `1`-gyel kezdődő és páros hosszú szavak;

   (f) a `00` részszót tartalmazó páratlan hosszú szavak.

   Megoldás: (a) Az üres szó nem jó. Tetszőleges első karakter után egy tetszőleges páros hosszú szó következik:

$(0+1)((0+1)(0+1))^*$ (vagy persze az utolsó karaktert is levághatjuk az első helyett).
(b) Az első és utolsó karakter között tetszőleges páros hosszú szó állhat: $1((0+1)(0+1)^*1$
(c) A három kiválasztott 0 karakter előtt, után és között is bármi állhat: $(0+1)^*0(0+1)^*0(0+1)^*0(0+1)^*$
(d) A 0-kat párosával tesszük le, két szomszédos között, előttük, utánuk tetszőleges számú 1 állhat: $(1^*01^*01^*)^*1^*$, vagy pl. az utánuk álló 1-eket elhagyhatjuk, a következő pár elején úgyis van 1^*, de ilyenkor a legvégén kell gondoskodni arról, hogy végződhessen valahány egyesre is: $(1^*01^*0)^*1^*$.
(e) Azt, hogy a két lehetőség legalább egyike teljesül a két reguláris kifejezés összege fejezi ki:
$0((0+1)(0+1))^* + 1(0+1)((0+1)(0+1))^*$.
(f) Vagy páros sok karakter van a 00 előtt, és akkor utána páratlan vagy előtte van páratlan és utána páros, azaz $((0+1)(0+1))^*00(0+1)((0+1)(0+1))^* + (0+1)((0+1)(0+1))^*00((0+1)(0+1))^*$ vagy kicsit másként csoportosítva, valamivel rövidebben: $((0+1)(0+1))^*(00(0+1) + (0+1)00)((0+1)(0+1))^*$.

6. Adjon olyan reguláris kifejezéseket, amelyek rövidebbek az itt szereplőknél, de ugyanazt a nyelvet írják le!

(a) $(0+\varepsilon)^*$ (b) $((0+\varepsilon)(0+\varepsilon))^*$ (c) $(0+1)^*01(0+1)^* + 1^*0^*$

Megoldás: (a) A kifejezés tetszőleges számú 0-t generál, a 0^* is pont ezt csinálja.
(b) A két zárójel együtt nulla, egy vagy kettő 0 karaktert ad, ezt lehet tetszőleges számszor ismételni, azaz minden, 0-kból álló szót generál, ami leírható a 0^* kifejezéssel is.
(c) A leírt szavak: van benne 01 vagy előbb 1-ek és utána 0-k állnak, azaz bármilyen szó lehet, tehát a $(0+1)^*$ jó.

7. Adjon reguláris kifejezést arra a nyelvre, ami az összes, az 110 részszót nem tartalmazó $\{0,1\}$ feletti szóból áll!

Megoldás: A reguláris kifejezésnél nincs művelet a kivonásra, azt kell kitalálnunk, hogyan néznek ki a megengedett szavak. Egy ilyen szó állhat csupa 0-ból, vagy kezdődhet tetszőleges számú 0 karakterrel (akár nulla darabbal is). Ha van benne 1, akkor két 1 között kell legyen 0, kivéve, ha a szó végén vagyunk, ott akárhány 1 lehet egymás után. Ha nincs két 1 egymás után, akkor a végén még lehetnek 0-k. Ezek alapján egy lehetséges megoldás: $0^*(\varepsilon + 1(0^*01)^*(0^* + 1^*))$ vagy egy elegánsabb: $(0+10)^*1^*$

8. Határozza meg az $S \to A \mid B \quad A \to 0A1 \mid 01 \quad B \to 1B0 \mid 10$ nyelvtan által generált nyelvet!

Megoldás: a nyelv az A-ból, illetve a B-ből generálható nyelvek uniója, és ezért
$L = \{0^n1^n : n \geq 1\} \cup \{1^n0^n : n \geq 1\}$

9. Adjon környezetfüggetlen nyelvtant 4. feladatban szereplő nyelvekre!

Megoldás: Sok helyes nyelvtan van, mutatunk egyet-egyet, ami a reguláris kifejezés szerkezetét tükrözi.
(a) $S \to A011A,\ A \to 0A \mid 1A \mid \varepsilon$ (A-ból a $(0+1)^*$ generálható, előadáson is volt)
(b) $S \to 1A0,\ A \to 0A \mid 1A \mid \varepsilon$
(c) $S \to 00S \mid 01S \mid 10S \mid 11S \mid \varepsilon$ de például $S \to 0S0 \mid 0S1 \mid 1S0 \mid 1S1 \mid \varepsilon$ is jó

10. Adjon környezetfüggetlen nyelvtant a jó zárójelezések nyelvéhez!

Megoldás: A zárójelsorozatot elképzelhetjük úgy, hogy vannak a külső szintű zárójelek (ezekből akárhány), és minden ilyen külső szintű zárójelpáron belül is egy jó sorozatnak kell lenni. Ebből a következő nyelvtant kaphatjuk:
$Z \to ZZ \mid (Z) \mid \varepsilon$
Az első szabállyal legyárthatjuk a tetszőleges számú külső szintű zárójel helyét, a második szabály kirakja a zárójelpárokat, és lehetőséget ad, hogy a belsejükben folytassuk az eljárást.

Kezdhetjük az első külső zárójelpárral is, aminek a belsejében, és utána is jó sorozatnak kell lenni:
$Z \to (Z)Z \mid \varepsilon$

Vagy kezdhetjük egy tetszőleges külső zárójelpárral, akkor előtte, benne és utána is helyes sorozat kell álljon:
$Z \to Z(Z)Z \mid \varepsilon$

(Ráadás: a fentiek közül melyik nyelvtan egyértelmű és melyik nem?)

11. Határozza meg az alábbi környezetfüggetlen nyelvtanok által generált nyelveket!

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

(b) $R \to T\mathtt{a}T \qquad T \to TT \mid \mathtt{a}T\mathtt{b} \mid \mathtt{b}T\mathtt{a} \mid \mathtt{a} \mid \varepsilon$

Megoldás: (a) Az világos, hogy a keletkezett szóban, ha nem az üres szó, akkor legalább annyi a betű lesz mint b. (Ez utóbbi tulajdonság valójában az üres szóra is teljesül.) Megmutatjuk, hogy minden ilyen szó levezethető, ezt a hossz szerinti teljes indukcióval csináljuk. Tekintsünk egy ilyen w szót. Ha a hossza $|w| \leq 1$, akkor vagy $w = \varepsilon$ vagy $w = \mathtt{a}$, és mindkettő valóban levezethető. Tegyük fel, hogy minden n-nél rövidebb, legalább annyi a betűt mint b betűt tartalmazó szó levezethető.

Legyen most $w = x_1 x_2 \cdots x_n$ egy $n \geq 2$ hosszú szó, amiben legalább annyi a van, mint b.

Legyen i a legkisebb olyan pozitív szám, melyre teljesül, hogy az $x_1 \cdots x_i$ részszóban ugyanannyi a van mint b.

Ha nincs ilyen i, akkor minden kezdőszeletben, és az egész szóban is több az a, mint a b. Tehát biztos, hogy $x_1 = \mathtt{a}$ és ha ezt a betűt levágjuk, akkor is a nyelvben maradunk, ezért az indukciós feltevés miatt a $T \Rightarrow TT \Rightarrow \mathtt{a}T$ kezdés folytatható úgy, hogy a végén jó levezetést kapjunk.

Vegyük észre, hogy ha van jó i, akkor $x_1 \neq x_i$ (mert az i-ediknél lesz pont ugyanannyi a mint b). Ezért ha a levezetés első lépése után az első T-ből a 2. vagy 3. szabállyal megkapjuk az x_1 és x_i karaktereket, középük a többi (itt is ugyanannyi a van mint b) a T-ből generálható. A szó végében is legalább annyi a van mint b, ezért ez megkapható az első lépésben kapott második T-ből.

Másik megfontolás ("alulról felfelé"): kiindulunk a szóból, és alkalmazzuk a következő átírási szabályokat: tetszőleges ab vagy ba részszót helyettesítsünk T-vel (a $T \Rightarrow \mathtt{a}T\mathtt{b} \Rightarrow \mathtt{ab}$ vagy a $T \Rightarrow \mathtt{b}T\mathtt{a} \Rightarrow \mathtt{ba}$ lépéssorozatok megfordításai); TT-t helyettesítsünk T-vel; $\mathtt{a}T\mathtt{b}$ és $\mathtt{b}T\mathtt{a}$ szintén helyettesíthető T-vel. Mit kapunk, amikor ezek egyike sem alkalmazható: b betű biztos nem marad (mert akkor a is kell, hogy maradjon, és lesz, esetleg T-vel elválasztott, a és b is). Ha már csak a és T maradt, helyettesítsük az a betűket T-vel, a szomszédos T-ket meg egyetlen T-vel. Így a végén egyetlen T marad csak, és ekkor megkaptunk (visszafele) egy levezetést.

(b) A nyelv azokból a szavakból áll, amelyekben több a van mint b.

Vegyük észre, hogy a T-re az előző szabályok maradtak, azaz T-ből azok a szavak vezethetők le, amelyekben legalább annyi a van mint b. A kezdő szabállyal még egy a betűt hozzáteszünk, tehát biztos, hogy a levezetett szavakban több a lesz mint b. Azt kell még megindokolni, hogy minden ilyet megkapunk: Egy ilyen w szót bontsunk úgy fel, hogy $w = w_1 \mathtt{a} w_2$, ahol w_1-ben ugyanannyi a van mint b, w_2-ben meg legalább ugyanannyi. Ez a w_1 és w_2 is levezethető T-ből, tehát az egész szó is.

Ilyen felbontás mindig van, mert tekintsük a legrövidebb kezdőszeletet, amiben az a-k száma több mint a b-k száma. Ennek az utolsó karaktere a, az ez előtti rész legyen w_1, az utána levő pedig w_2. (Lehet, hogy $w_1 = \varepsilon$ vagy $w_2 = \varepsilon$.)

12. Határozza meg a következő nyelvtan által generált nyelvet!

$$R \to XRX \mid S$$
$$S \to \mathtt{a}T\mathtt{b} \mid \mathtt{b}T\mathtt{a}$$
$$T \to XTX \mid X \mid \varepsilon$$
$$X \to \mathtt{a} \mid \mathtt{b}$$

Megoldás: Ez a nem palindromokból álló nyelv. Egy szó pontosan akkor nem palindrom, ha van olyan i, hogy az elölről és a hátulról számított i-edik karaktere eltérő. (Több ilyen i is lehet, de most válasszunk egyet.) Egy ilyen szó úgy vezethető le a nyelvtanból, ha $(i-1)$-szer alkalmazzuk az 1. szabályt, amivel megcsináljuk a helyet az első és utolsó $i-1$ karakternek, utána a 2. szabály segítségével kapunk egy S-et, amivel a harmadik, illetve negyedik szabály segítségével létrehozzuk az i-edik pozíciókba az eltérő karaktereket, utána az 5-9 szabályokkal tetszőlegesen kitölthetjük a többi helyet.

Azt is könnyű látni, hogy csak ilyen szavak vezethetők le a nyelvtanból, mert X-ből az $(\mathtt{a}+\mathtt{b})$, T-ből az $(\mathtt{a}+\mathtt{b})^*$ reguláris kifejezéssel leírható nyelvek vezethetők le, S-ből az olyanok, amiknek az első és utolsó betűje különbözik, e köré rak R ugyanannyi karaktert előre, mint hátulra.