Környezetfüggetlen nyelvek

Kiegészítő anyag az Algoritmuselmélet tárgyhoz VI.

(a Rónyai–Ivanyos–Szabó: Algoritmusok könyv mellé)

Friedl Katalin
BME SZIT
friedl@cs.bme.hu

2017. február 20.

A reguláris nyelveket véges automatákkal vagy reguláris kifejezésekkel adtuk meg. Van további lehetőség is, egy ilyen a *formális nyelvtan* vagy röviden *nyelvtan.*

A formális nyelvtanok nem egészen olyanok, mint például a magyar nyelvtan. Bár eredetileg a természetes nyelvek szabályainak leírására készültek, de könnyebben használhatóak mesterséges nyelvek, például programozási nyelvek pontos megadására, mint egy beszélt nyelv helyes mondatainak tökéletes leírására.

A formális nyelvtanok lényegében átírási szabályokat adnak meg, amelyekkel egy kezdő szimbólumból kiindulva szavakat tudunk előállítani.

Ezeknek itt csak egy speciális, talán a leggyakrabban használt fajtájával foglalkozunk.

1. Definíció. *Egy* környezetfüggetlen nyelvtan *(röviden* CF nyelvtan*) alatt egy olyan* $G = (V, \Sigma, S, P)$ *rendszert értünk, ahol*

- V *egy véges, nem üres halmaz, a* változók *(vagy* nemterminálisok*) halmaza,*
- Σ *egy ábécé, amire* $V \cap \Sigma = \emptyset$*, a karakterek (vagy* terminálisok*) halmaza*
- $S \in V$ *a* kezdő változó*,*
- P *egy véges halmaz, az ún.* levezetési (vagy produkciós, illetve átírási) szabályok *halmaza.*

P *elemei* $A \to \alpha$ *alakúak, ahol* $A \in V$ *egy változó,* $\alpha \in (V \cup \Sigma)^*$ *egy változókból és a* Σ *elemeiből álló tetszőleges, véges hosszú sorozat.*

1. Példa. *Legyen* $V = \{A\}$, $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, *a kezdő változó természetesen az* A, *a levezetési szabályok halmaza pedig*

$$P = \{A \to \mathtt{a}A\mathtt{b}, \quad A \to \varepsilon\}$$

A nyelvtanok megadásánál sokszor nem írjuk ki az összes paramétert, csak a levezetési szabályokat soroljuk fel. Ha mást nem mondunk, akkor a szabályokban szereplő kisbetűk a Σ elemei, a nagybetűk a változók, az első szabály bal oldala a kezdő változó. Továbbá azok a szabályok, melyeknek a bal oldalán ugyanaz áll, összevonhatóak, függőleges vonallal elválasztva a különböző jobb oldalakat, Ebben a formában az előző nyelvtan így néz ki:

$$A \to \mathtt{a}A\mathtt{b} \mid \varepsilon$$

2. Definíció. *Egy* $G = (V, \Sigma, S, P)$ *nyelvtannál* levezetés *alatt egy olyan*

$$\gamma_0 \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \cdots \Rightarrow \gamma_n$$

véges hosszú sorozatot értünk ($n \geq 0$), *melyben* $\gamma_0 = S$, *továbbá* $\gamma_i \in (V \cup \Sigma)^*$, *és mindegyik* γ_{i+1} *megkapható* γ_i*-ből egy levezetési szabály alkalmazásával. Ez azt jelenti, hogy minden* $0 \leq i < n$ *esetén* γ_i *felírható* $\gamma_i = \delta_1 A \delta_2$ *alakban, ahol* $\delta_1, \delta_2 \in (V \cup \Sigma)^*$ *és* $A \in V$ *úgy, hogy* $\gamma_{i+1} = \delta_1 \alpha \delta_2$ *ahol* $A \to \alpha$ *egy* P*-hez tartozó levezetési szabály.*

2. Példa. *Az előző nyelvtan esetén egy levezetés pl. az alábbi*

$$A \Rightarrow \mathtt{a}A\mathtt{b} \Rightarrow \mathtt{aa}A\mathtt{bb} \Rightarrow \mathtt{aaa}A\mathtt{bbb} \Rightarrow \mathtt{aaa}\varepsilon\mathtt{bbb}$$

Az így levezetett szó az aaabbb. *Könnyű látni, hogy ebből a nyelvtanból pontosan azok a szavak vezethetők le, amelyek* $\mathtt{a}^n\mathtt{b}^n$ *alakúak* ($n \geq 0$).

A levezetések közül azok lesznek számunkra érdekesek, amelyekben a kezdő változóból indulva végül egy olyan sorozathoz jutunk, amiben már nincsenek változók.

3. Definíció. *A* $G = (V, \Sigma, S, P)$ *nyelvtan által* generált $L(G)$ nyelv *azokból a* $w \in \Sigma^*$ *szavakból áll, melyekhez valamilyen* $n \geq 0$ *számra van olyan* $S \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \cdots \Rightarrow \gamma_n$ *levezetés, amiben* $\gamma_n = w$

1. Megjegyzés. *Vegyük észre, hogy ha egy levezetés során eljutunk egy* $w \in \Sigma^*$ *szóhoz, akkor a levezetés tovább már biztos nem folytatható, mivel* w *nem tartalmaz változót, ilyenkor már egyetlen szabály sem alkalmazható.*

3. Példa. *Tekintsük az alábbi nyelvtant!*

$$S \to \mathtt{a}S \mid \mathtt{b}S \mid \mathtt{a} \mid \mathtt{b}$$

Itt tehát az egyetlen változó az S, *az ábécé az* $\{\mathtt{a}, \mathtt{b}\}$. *A nyelvtanból levezethető pl. az* ab *szó:*

$$S \Rightarrow \mathtt{a}S \Rightarrow \mathtt{ab}$$

Az is látszik, hogy a nyelvtan által generált nyelv $\{\mathtt{a},\mathtt{b}\}^*$-ból az összes nem üres szót tartalmazza, azaz $L = \{\mathtt{a},\mathtt{b}\}^* \setminus \{\varepsilon\}$. Ez azért igaz, mert egy tetszőleges, legalább 1 hosszú $w \in \{\mathtt{a},\mathtt{b}\}^*$ szónak elölről kezdve egymás után tudjuk generálni a karaktereit: amíg nem az utolsó karakterről van szó, addig az vagy az első vagy a második szabállyal, az utolsó karakter pedig a 3. vagy a 4. szabállyal állítható elő.

1. Feladat. *Mely szavakból áll az alábbi nyelvtan által generált nyelv?*

$$S \to \mathtt{a}T\mathtt{a} \mid \mathtt{b}T\mathtt{b} \mid \mathtt{a} \mid \mathtt{b}$$
$$T \to \mathtt{a}T \mid \mathtt{b}T \mid \varepsilon$$

Megoldás: Világos, hogy S-ből csak olyan szó vezethető le, ami nem üres, továbbá az első és az utolsó karaktere megegyezik. Megmutatjuk, hogy a generált nyelv az összes ilyen szóból áll.

Ehhez vegyük észre, hogy T-ből minden $\mathtt{a}$ és $\mathtt{b}$ betűkből álló sorozat előállítható az előző példához hasonlóan. Az S szabályai lehetővé teszik, hogy a szó első és utolsó karakterét generáljuk. Ha 1-nél hosszabb szót akarunk, akkor ezek közé a T-ből tetszőleges karaktersorozatot előállíthatunk. □

2. Feladat. *Legyen $\Sigma = \{\mathtt{a},\mathtt{b}\}$ és L álljon azokból a szavakból, melyekben az $\mathtt{a}$ betűk száma megegyezik a $\mathtt{b}$ betűk számával. Adjunk olyan G nyelvtant, amire $L(G) = L$.*

Megoldás: Egy lehetséges megoldás: $S \to \mathtt{a}S\mathtt{b}S \mid \mathtt{b}S\mathtt{a}S \mid \varepsilon$

Ahhoz, hogy ez valóban jó nyelvtan, először is vegyük észre, hogy minden esetben, amikor valamelyik szabályt alkalmazzuk, ugyanannyi $\mathtt{a}$-t generálunk, mint ahány $\mathtt{b}$-t ezért $L(G) \subseteq L$.

Azt kell még megmutatni, hogy minden $w \in L$ szó levezethető a nyelvtanból. Ezt a w hossza szerinti indukcióval látjuk be. Nyilván ez igaz a 0 hosszú $w = \varepsilon$ szóra. Egy hosszú szó nincs az L nyelvben. A kettő hosszú szavakra is könnyű látni, mert vagy az első vagy a második szabály egyszeri alkalmazása után a harmadik szabályt kétszer használva megkaphatjuk a w szót.

Tegyük fel, hogy L legfeljebb k hosszú szavairól már tudjuk, hogy levezethetők és legyen $|w| = k+1$. Több eset lehetséges: amennyiben $w = \mathtt{a}w'\mathtt{b}$, akkor $w' \in L$ és ebben az esetben az $S \Rightarrow \mathtt{a}S\mathtt{b}S$ kezdés után S első előfordulásából, az indukciós feltevés szerint w' levezethető, miután a második S betűre az $S \to \varepsilon$ szabályt alkalmazva megkapjuk a w szót.

Hasonlóan járhatunk el, amennyiben $w = \mathtt{b}w'\mathtt{a}$.

Ha viszont w első és utolsó betűje megegyezik, és ez a betű mondjuk $\mathtt{a}$, akkor vegyük w-nek egy legrövidebb kezdőszeletét, amiben ugyanannyi $\mathtt{a}$ van mint $\mathtt{b}$, legyen ez x és $w = xy$. Ekkor egyrészt $x, y \in L$, másrészt x szükség szerűen $\mathtt{b}$-re végződik, azaz $x = \mathtt{a}z\mathtt{b}$ alakú, ahol $z \in L$. Egy ilyen w-re jó levezetést kapunk, ha az $S \Rightarrow \mathtt{a}S\mathtt{b}S$ lépés után az első S-ből az x-et, a másodikból az y-t vezetjük le (ami az indukciós feltevés miatt lehetséges).

Hasonlóan járhatunk el akkor is, amikor w első betűje b, csak ilyenkor a levezetés az $S \Rightarrow \mathtt{b}S\mathtt{a}S$ szabállyal indul. □

Nézzünk egy kicsit bonyolultabb nyelvtant!

4. Példa.

$$R \to XRX \mid S \tag{1-2}$$
$$S \to \mathtt{a}T\mathtt{b} \mid \mathtt{b}T\mathtt{a} \tag{3-4}$$
$$T \to XTX \mid X \mid \varepsilon \tag{5-7}$$
$$X \to \mathtt{a} \mid \mathtt{b} \tag{8-9}$$

Ez is környezetfüggetlen nyelvtan, ahol a kezdő változó az R. Az alábbi levezetésben az aláhúzott rész jelöli a következőként alkalmazott szabály bal oldalát, a nyíl feletti szám a szabály sorszámát.

$$\begin{aligned} R &\overset{1}{\Rightarrow} X\underline{R}X \overset{1}{\Rightarrow} XX\underline{R}XX \overset{2}{\Rightarrow} \underline{X}XSXX \overset{8}{\Rightarrow} \mathtt{a}\underline{X}SXX \overset{9}{\Rightarrow} \mathtt{ab}S\underline{X}X \\ &\overset{8}{\Rightarrow} \mathtt{ab}S\mathtt{a}\underline{X} \overset{8}{\Rightarrow} \mathtt{ab}\underline{S}\mathtt{aa} \overset{3}{\Rightarrow} \mathtt{aba}\underline{T}\mathtt{baa} \overset{5}{\Rightarrow} \mathtt{aba}X\underline{T}X\mathtt{baa} \overset{7}{\Rightarrow} \mathtt{aba}\underline{X}X\mathtt{baa} \\ &\overset{9}{\Rightarrow} \mathtt{abab}\underline{X}\mathtt{baa} \overset{9}{\Rightarrow} \mathtt{ababbbaa} \end{aligned}$$

Tehát a kapott szó $\mathtt{ababbbaa} \in L(G)$.

A fenti levezetés során több választásunk is volt, hogy melyik változót melyik szabály alapján helyettesítsük.

Egy levezetés sokszor jobban áttekinthető ha a lépéseket egy fába rendezzük.

4. Definíció. *Legyen G egy környezetfüggetlen nyelvtan és x egy szó. Az x* levezetési fája *G-ben egy gyökeres fa, melyben*

- *a gyökér a kezdő változóval,*
- *minden nem levél csúcs egy-egy változóval,*
- *minden levél pedig Σ egy-egy elemével (vagy ε-nal) van címkézve.*
- *Ha egy A csúcs gyerekei balról jobbra olvasva $B_1, B_2, \ldots, B_k$, akkor a nyelvtannak van $A \to B_1B_2 \ldots B_k$ szabálya. (Itt $B_i \in \Sigma \cup V \cup \{\varepsilon\}$.)*
- *A levelek balról jobbra olvasva éppen az x szót adják.*

A definícióból világos, hogy egy $x \in L(G)$ szó tetszőleges levezetéséből lehet levezetési fát készíteni, és a levezetési fából is kiolvasható legalább egy levezetés.

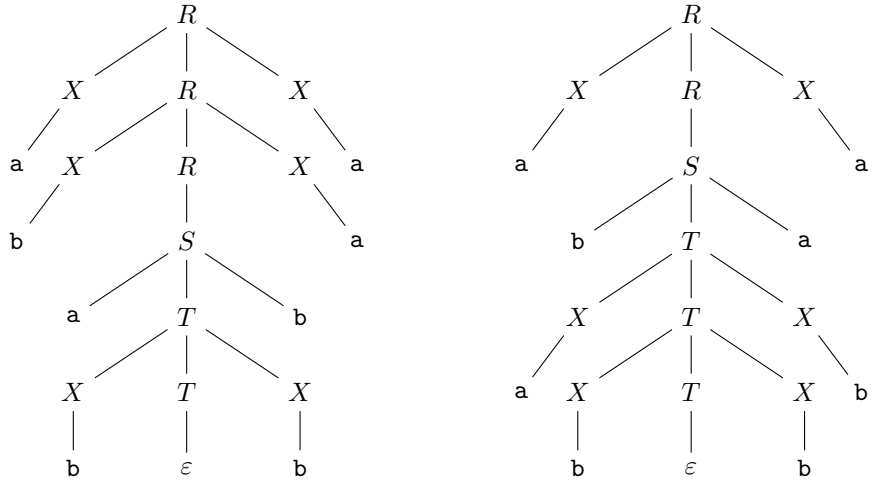
Fontos azonban megjegyezni, hogy míg a levezetés egyértelműen meghatározza a fát, visszafelé ez nem igaz, általában egy levezetési fából ugyanannak a szónak több levezetése is kiolvasható.

5. Definíció. *Egy $x \in L(G)$ szó* bal-levezetése *egy olyan levezetés, amikor minden lépésben a γ_i elejéhez legközelebbi változót helyettesítjük egy megfelelő nyelvtani szabály alapján.*

Erre már igaz, hogy egy levezetési fából egyetlen bal-levezetés olvasható ki.

5. Példa. *Az előző példában leírt levezetéshez tartozó levezetési fa. Ebből többféle levezetés is leolvasható, a bal-levezetés szabályai sorrendben: 1, 8, 1, 9, 2, 3, 5, 9, 7, 9, 8, 8.*

Jobb oldalt egy ugyanehhez a szóhoz tartozó másik levezetési fa látható.



6. Definíció. *Egy $w \in \Sigma^*$ szó* egyértelműen levezethető *a G nyelvtanból, ha G-ben csak egy levezetési fája van.*

A G nyelvtan egyértelmű*, ha G-ből minden $w \in L(G)$ szó egyértelműen levezethető.*

Az L nyelv egyértelmű*, ha létezik egyértelmű nyelvtana.*

Ezek szerint az előző példabeli szó nem egyértelműen levezethető, hiszen két különböző levezetési fája is van. Így persze a nyelvtan sem egyértelmű.

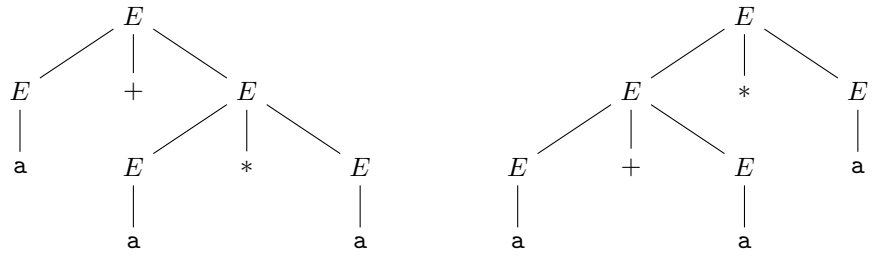
2. Megjegyzés. *Az egyértelműen levezethetőség fenti definíciója ekvivalens azzal, hogy a szó bal-levezetése egyértelmű.*

Lássunk most egy fontos példát, az aritmetikai kifejezések nyelvét. Az egyszerűség kedvéért csak összeadást és szorzást fogunk benne használni, de kiegészíthető további műveletekkel is.

$$E \to E + E \mid E * E \mid (E) \mid \mathtt{a} \tag{1}$$

Itt E az egyetlen változó, az ábécé elemei pedig $+$, $*$, $\mathtt{a}$, valamint a nyitó és csukó zárójel.

Ez egy nem egyértelmű nyelvtan, hiszen például az $\mathtt{a} + \mathtt{a} * \mathtt{a}$ kifejezéshez két különböző levezetési fa is tartozik,



3. Megjegyzés. *Ha erre a két fára nem mint levezetési fákra, hanem mint a kifejezés kiértékelésének módját megadó fákra gondolunk, akkor látszik, hogy míg az első megfelel az aritmetikai kifejezések szokásos kiértékelésének (előbb a szorzást végezzük el, utána az összeadást) a második ,,rossz sorrendben" végzi a műveleteket.*

1. Tétel. *Az aritmetikai kifejezésekre adott fenti G egyszerű (1) nyelvtan nem egyértelmű, de az általa generált $L(G)$ nyelv egyértelmű nyelv.*

Bizonyítás vázlat: Az előbb már láttuk, hogy a nyelvtan nem egyértelmű. A nyelv egyértelműségéhez mutatnunk kell egy G' egyértelmű nyelvtant, amire $L(G') = L(G)$. Legyen G' a következő:

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

Világos, hogy a G' nyelvtannal levezethető aritmetikai kifejezések levezethetők az eredeti nyelvtanból is.

Azt kell megmutatni, hogy ha $w \in L(G)$, akkor $w \in L(G')$ is teljesül, sőt a G'-beli levezetési fája egyértelmű.

Ezt a w hossza szerinti indukcióval mutathatjuk meg. Ha $|w| = 1$, akkor csak $w = \mathtt{a}$ lehet, és ez egyedül az $E \Rightarrow T \Rightarrow F \Rightarrow \mathtt{a}$ lépésekkel kapható meg a G' nyelvtanban.

Hosszabb szavakra azt kell észrevenni, hogy ha vannak zárójelen kívüli $+$ jelek, akkor először ezeket kell generálni (sorrendben visszafelé) az első szabály segítségével, utána a zárójelen kívüli $*$ jeleket, majd a zárójelekben levő kifejezéseket.

□

4. Megjegyzés. *Vegyük észre, hogy ebben a módosított nyelvtanban ha a levezetési fát kiértékelési fának tekintjük, akkor a műveletek sorrendje is a szokásos lesz.*