Formális Nyelvek - 2.

Csuhaj Varjú Erzsébet

Algoritmusok és Alkalmazásaik Tanszék
Informatikai Kar
Eötvös Loránd Tudományegyetem
H-1117 Budapest
Pázmány Péter sétány 1/c

E-mail: csuhaj@inf.elte.hu

Nyelvek sokféle módon előállíthatók, például **logikai formulával, strukturális rekurzióval, algoritmussal, matematikai gépekkel, vagy produkciós rendszerekkel**.

A produkciós rendszerekkel való előállítás egyik módja a **nyelvek generálása grammatikával**.

Generatív grammatika - Definíció

Egy G **generatív grammatikán** (grammatikán vagy (generatív) nyelvtanon) egy (N, T, P, S) négyest értünk, ahol

- N és T diszjunkt ábécék, a **nemterminális** és a **terminális** szimbólumok ábécéi;
- $S \in N$ a **kezdőszimbólum** (axióma),
- P **véges halmaza** (x, y) alakú rendezett pároknak, ahol $x, y \in (N \cup T)^*$ és x legalább egy nemterminális szimbólumot tartalmaz.

  A P halmaz elemeit **átírási szabályoknak** (röviden szabályoknak) vagy **produkcióknak** nevezzük.

  Az (x, y) jelölés helyett használhatjuk az $x \rightarrow y$ jelölést is, ahol a $\rightarrow$ szimbólum nem eleme az $(N \cup T)$ halmaznak.

Példa Generativ Grammatikára

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol

$N = \{S\}$ a nemterminálisok ábécéje,

$T = \{a, b\}$ a terminálisok ábécéje, és

$$P = \{S \to aSb, \quad S \to ab, \quad S \to ba\}$$

a szabályok halmaza.

Közvetlen levezetési lépés - Definíció

Legyen $G = (N, T, P, S)$ egy generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó **közvetlenül** vagy **egy lépésben levezethető** az u szóból G-ben és ezt

$$u \Longrightarrow_G v$$

módon jelöljük, ha $u = u_1 x u_2$, $v = u_1 y u_2$, $u_1, u_2 \in (N \cup T)^*$ és $x \to y \in P$.

Példa közvetlen levezetésre

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$ a nemterminálisok ábécéje, $T = \{a, b\}$ a terminálisok ábécéje és $P = \{S \to aSb, S \to ab, S \to ba\}$ a szabályok halmaza.

Legyen $u = aaaSbbb$.

Akkor $v = aaaaSbbbb$ közvetlenül (egy lépésben) levezethető u-ból, azaz

$$u \Longrightarrow_G v,$$

ugyanis $u_1 = aaa$, $u_2 = bbb$, $x = S$, $y = aSb$ és $S \to aSb \in P$.

Levezetés - Definíció

Legyen $G = (N, T, P, S)$ egy generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó k **lépésben levezethető** az u szóból G-ben, $k \geq 1$, ha létezik olyan $u_1, \ldots, u_{k+1} \in (N \cup T)^*$ szavakból álló sorozat, amelyre $u = u_1$, $v = u_{k+1}$, valamint $u_i \Longrightarrow_G u_{i+1}$, $1 \leq i \leq k$ teljesül.

A v szó **levezethető** az u szóból G-ben, ha vagy $u = v$, vagy létezik olyan $k \geq 1$ szám, hogy a v szó az u szóból k lépésben levezethető.

Levezetés

Legyen $G = (N, T, P, S)$ egy tetszőleges generatív grammatika és legyen $u, v \in (N \cup T)^*$.

Azt mondjuk, hogy a v szó levezethető az u szóból G-ben és ezt

$$u \Longrightarrow_G^* v$$

módon jelöljük, ha vagy $u = v$ vagy valamely $z \in (N \cup T)^*$ szóra fennáll, hogy $u \Longrightarrow_G^* z$ és $z \Longrightarrow_G v$ teljesül.

$\Longrightarrow^*$ a $\Longrightarrow$ reláció reflexív tranzitív lezártját jelöli.

A $\Longrightarrow$ reláció tranzítív lezártját $\Longrightarrow^+$-val jelöljük.

A kezdőszimbólumból levezethető sztringeket **mondatformának** nevezzük.

A generált nyelv - Definíció

Legyen $G = (N, T, P, S)$ egy tetszőleges generatív grammatika. A G **grammatika által generált** $L(G)$ **nyelv** alatt az

$$L(G) = \{w | S \overset{*}{\underset{G}{\Longrightarrow}} w, w \in T^*\}$$

szavakból álló halmazt értjük.

Azaz, a G grammatika által generált nyelv a T^* halmaz azon elemeiből áll, amelyek levezethetők a G grammatika S kezdőszimbólumából.

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$, $T = \{a, b\}$ és $P = \{S \to aSb, S \to ab, S \to ba\}$.

Akkor $L(G) = \{a^n abb^n, a^n bab^n | n \geq 0\}$.

Példa egy levezetésre:

$$S \Longrightarrow_G aSb \Longrightarrow_G aaSbb \Longrightarrow_G aababb.$$

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S, X, Y\}$, $T = \{a, b, c\}$. Legyen

$$
\begin{aligned}
P = \{S \to abc, &\quad S \to aXbc, \\
Xb \to bX, &\quad Xc \to Ybcc, \\
bY \to Yb, &\quad aY \to aaX, \quad aY \to aa\}.
\end{aligned}
$$

Akkor $L(G) = \{a^n b^n c^n | n \geq 1\}$.

Példa egy levezetésre:

$$S \Longrightarrow_G aXbc \Longrightarrow_G abXc \Longrightarrow_G abYbcc \Longrightarrow_G aYbbcc \Longrightarrow_G aabbcc.$$

Ekvivalens Grammatikák és Nyelvek

Két generatív grammatikát (gyengén) **ekvivalensnek** nevezünk, ha ugyanazt a nyelvet generálják.

Két nyelvet **gyengén ekvivalensnek** mondunk, ha legfeljebb az üres szóban különböznek.

A Chomsky-féle hierarchia

A $G = (N, T, P, S)$ **generatív grammatikát** i**-típusúnak mondjuk**, $i = 0, 1, 2, 3$, ha P szabályhalmazára teljesülnek a következők:

- $i = 0$: Nincs korlátozás.
- $i = 1$: P minden szabálya $u_1Au_2 \to u_1vu_2$ alakú, ahol $u_1, u_2, v \in (N \cup T)^*$, $A \in N$, és $v \neq \varepsilon$, kivéve az $S \to \varepsilon$ alakú szabályt, feltéve, hogy P-ben ilyen szabály létezik. Ha P tartalmazza az $S \to \varepsilon$ szabályt, akkor S nem fordul elő P egyetlen szabályának jobboldalán sem.
- $i = 2$: P minden szabálya $A \to v$ alakú, ahol $A \in N$ és $v \in (N \cup T)^*$.
- $i = 3$: P minden szabálya vagy $A \to uB$ vagy $A \to u$, alakú, ahol $A, B \in N$ és $u \in T^*$.

Példa

Legyen $G = (N, T, P, S)$ egy generatív grammatika, ahol $N = \{S\}$, $T = \{a, b\}$ és $P = \{S \to aSb, S \to ab, S \to ba\}$.

Ez a grammatika 2-típusú (környezetfüggetlen).

A Chomsky-féle hierarchia - folytatás

Legyen $i = 0, 1, 2, 3$. Egy L nyelvet i**-típusúnak mondunk**, ha i-típusú grammatikával generálható.

Az i-típusú nyelvek osztályát $\mathcal{L}_i$-vel jelöljük.

A 0-típusú grammatikát **mondatszerkezetű grammatikának**, az 1-típusú grammatikát **környezetfüggő grammatikának**, a 2-típusú grammatikát **környezetfüggetlen grammatikának** is nevezzük. A 3-típusú grammatikát **reguláris** vagy **véges állapotú** grammatikának is mondjuk.

A 0,1,2,3-típusú nyelvek osztályait rendre **rekurzíven felsorolható**, **környezetfüggő**, **környezetfüggetlen**, valamint **reguláris nyelvosztálynak** is mondjuk.

A Chomsky-féle hierarchia - folytatás

Nyilvánvaló, hogy $\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_0$ és $\mathcal{L}_1 \subseteq \mathcal{L}_0$.

A későbbiekben megmutatjuk, hogy

$$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0.$$

Megjegyzés: Könnyen észrevehetjük, hogy a $\mathcal{L}_2$ és a $\mathcal{L}_1$ nyelvosztályok közötti, a tartalmazásra vonatkozó reláció nem azonnal látható a megfelelő grammatikák definíciójából.

Generatív grammatikák egy normálformája

Tétel:

Minden $G = (N, T, P, S)$ generatív grammatikához meg tudunk konstruálni egy vele ekvivalens és azonos típusú $G' = (N', T, P', S)$ generatív grammatikát úgy, hogy P' egyetlen szabályának baloldalán sem fordul elő terminális szimbólum.

A bizonyítás vázlata:

- 2- és 3-típusú grammatikák esetében az állítás azonnal adódik a definíciókból.

- Legyen $G = (N, T, P, S)$ 0-típusú vagy 1-típusú grammatika.

  Megkonstruáljuk a $G' = (N', T, P', S)$ grammatikát.

  Tekintsük az $N' = N \cup \bar{T}$ halmazt, ahol $\bar{T} = \{\bar{a} \mid a \in T\}$. Képezzük P'-t a P szabályhalmazból úgy, hogy minden $a \in T$ szimbólumot $\bar{a}$ szimbólumra cserélünk minden egyes olyan szabály mindkét oldalán P-ben, ahol a előfordul, továbbá az így kapott szabályhalmazhoz adjuk hozzá minden $a \in T$ szimbólumra a $\bar{a} \to a$ szabályt.

  Álljon P' az így kapott szabályokból.

Bizonyításvázlat - folytatás

(1) Megmutatjuk, hogy $L(G) \subseteq L(G')$.

Azonnal látható, hogyha $u = a_1 \dots a_n \in L(G)$, ahol $a_i \in T$, $1 \leq i \leq n$, akkor $v = \bar{a}_1 \dots \bar{a}_n$ levezethető G'-ben. Ekkor a $\bar{a}_i \to a_i$ szabályok alkalmazásával u is levezethető G'-ben.

Az üres szó esetében nyilvánvaló, hogyha $\varepsilon \in L(G)$, akkor $\varepsilon \in L(G')$ is teljesül.

Bizonyításvázlat - folytatás

(2) Megmutatjuk, hogy $L(G') \subseteq L(G)$.

Definiáljuk a h homomorfizmust úgy, hogy $h(\bar{a}) = a$ minden $\bar{a} \in \bar{T}$ szimbólumra és $h(x) = x$ minden $x \in (N \cup T)$ szimbólumra.

Ha $u \Longrightarrow_{G'} v$, akkor fennáll $h(u) \Longrightarrow^*_G h(v)$ is. Ha a v szó levezethető az u szóból valamely $\bar{a} \to a$ szabály alkalmazásával, akkor $h(u) = h(v)$. Egyébként az $u \Longrightarrow_{G'} v$ levezetés P valamely szabályának alkalmazását kívánja meg, és így $h(u) \Longrightarrow_G h(v)$ szintén fennáll. Vagyis, $u \Longrightarrow^*_{G'} v$ teljesülése maga után vonja $h(u) \Longrightarrow^*_G h(v)$ teljesülését. Azaz, ha $S \Longrightarrow^*_{G'} w$, ahol $w \in T^*$, akkor $S = h(S) \Longrightarrow^*_G h(w) = w$.

Nyelvosztályok zártsági tulajdonságai

Az unió, a konkatenáció, valamint a lezárás (a $*$) műveleteket együttesen **reguláris** műveleteknek nevezzük.

Tétel:

Az $\mathcal{L}_i$, $i = 0, 1, 2, 3$ nyelvosztályok mindegyike zárt a reguláris műveletekre nézve.

Tétel:

Az $\mathcal{L}_i$, $i = 0, 1, 2, 3$ nyelvosztályok mindegyike zárt a reguláris műveletekre nézve.

Bizonyításvázlat:

Legyen L és L' két i-típusú nyelv, ahol $i = 0, 1, 2, 3$. Tegyük fel, hogy L és L' rendre generálhatók az i-típusú $G = (N, T, P, S)$ és $G' = (N', T', P', S')$ grammatikákkal. Az általánosság megszorítása nélkül feltehetjük, hogy G és G' a korábbiakban ismertetett normálformában adott (a szabályok baloldalán nincs terminális szimbólum), valamint hogy $N \cap N' = \emptyset$.

Bizonyításvázlat - folytatás

Unió:

(1) $i = 0, 2, 3$ esetében legyen $S_0 \notin (N \cup N')$ és legyen

$$G_u = (N \cup N' \cup \{S_0\}, T \cup T', P \cup P' \cup \{S_0 \to S, S_0 \to S'\}, S_0).$$

(2) Nyilvánvaló, hogy G_u egyazon típusú, mint G és G'.

(3) Az is azonnal látható, hogy $L(G) \cup L(G') \subseteq L(G_u)$.

(4) $L(G_u) \subseteq L(G) \cup L(G')$ szintén fennáll, mivel N és N' diszjunktak és az $S_0 \to S$, $S_0 \to S'$ szabályok garantálják, hogy $L(G_u)$ egyetlen elemének levezetésekor sem használhatunk szabályt mind a P és mind a P' szabályhalmazból.

Bizonyításvázlat - folytatás

Az $i = 1$ és $\varepsilon \notin (L \cup L')$ esetben megkonstruálunk egy G_u grammatikát az előbbi módon.

Ha $i = 1$ és $\varepsilon \in (L \cup L')$, akkor először tekintjük az $L_1 = L - \{\varepsilon\}$ és az $L_2 = L' - \{\varepsilon\}$ nyelveket. Tegyük fel, hogy a G_1 és a G_2 grammatikák 1-típusúak, valamint rendre generálják az L_1 és az L_2 nyelveket. Ezután az előbbieknek megfelelően megkonstruálunk egy G_u grammatikát, amely az $(L_1 \cup L_2)$ nyelvet generálja. Majd bevezetünk egy új S_1 kezdőszimbólumot és a G_u szabályhalmazához hozzáadjuk az $S_1 \to S_0$ és $S_1 \to \varepsilon$ szabályokat. Az így nyert grammatika az $(L \cup L')$ nyelvet generálja.

Bizonyításvázlat - folytatás

Konkatenáció:

Tekintsük először az $i = 0, 2$ eseteket. Legyen $S_0 \notin (N \cup N')$ és

$$G_c = (N \cup N' \cup \{S_0\}, T \cup T', P \cup P' \cup \{S_0 \to SS'\}, S_0).$$

(2) Nyilvánvaló, hogy G_c egyazon típusú, mint G és G'.

(3) Az is azonnal látható, hogy $L(G)L(G') \subseteq L(G_c)$.

Konkatenáció - folytatás

(4) Megmutatjuk, hogy $L(G_c) \subseteq L(G)L(G')$. Tekintsük az

$$S_0 \Longrightarrow u_1 \Longrightarrow u_2 \Longrightarrow \ldots \Longrightarrow u_m = u,\ m \geq 1$$

G_c-beli levezetést, ahol $u \in (T \cup T')^*$.

j-szerinti indukcióval megmutatjuk, hogy

$$u_j = v_j v_j'$$

alakú valamely v_j és v_j'-re, ahol $S \Longrightarrow_G^* v_j$ és $S' \Longrightarrow_{G'}^* v_j'$ teljesül. A $j = 1$ esetben az állítás triviális, hiszen $u_1 = SS'$ kell, hogy legyen. Tegyük fel, hogy az állítás igaz u_j-re. Akkor viszont igaz u_{j+1}-re is, mivel N és N' diszjunktak, terminális szimbólum nem fordul elő a baloldalon, és ahhoz, hogy az u_{j+1} szót megkapjuk, vagy a v_j, vagy a v_j' mondatformát át kell írnunk. Ez alapján az $L(G_c)$ minden eleme egyben eleme az $L(G)L(G')$ nyelvnek is.

Konkatenáció - folytatás

(5) Legyen $i = 1$.

(a) Ha $\varepsilon \notin L$ és $\varepsilon \notin L'$ akkor G_c-t az előzőeknek megfelelően konstruáljuk meg.

(b) Ha $\varepsilon \in LL'$, akkor először vegyük az $L_1 = L - \{\varepsilon\}$ és $L_2 = L' - \{\varepsilon\}$ nyelveket, és konstruáljuk meg G_c-t a fenti módon. Az LL' nyelv megegyezik a következő nyelvek valamelyikével:

$L_1L_2 \cup L_2$, $L_1L_2 \cup L_1$, $L_1L_2 \cup L_1 \cup L_2 \cup \{\varepsilon\}$,

attól függően, hogy $\varepsilon \in L$ és $\varepsilon \notin L'$, vagy fordítva, vagy ε mindkét nyelv eleme.

Mindegyik esetben $LL' \in \mathcal{L}_1$ következik abból, hogy $L_1L_2 \in \mathcal{L}_1$ és $\mathcal{L}_1$ zárt az unió műveletére nézve.

Konkatenáció - folytatás

Legyen $i = 3$.

A P szabályhalmazból megkonstruálunk egy P_1 szabályhalmazt úgy, hogy minden $A \to u$ alakú szabályt, ahol $A \in N$ és $u \in T^*$ felcserélünk egy $A \to uS'$ alakú szabályra ($S' \notin (N \cup T)$) és a többi szabályt változatlanul hagyjuk. A

$$G_c = (N \cup N', T \cup T', P_1 \cup P', S)$$

grammatika nyilvánvalóan 3-típusú és generálja az $L(G)L(G')$ nyelvet.

Megmutatjuk, hogy $L(G_c) \subseteq L(G)L(G')$. A G_c grammatikában minden terminális szóhoz vezető levezetés $S \Longrightarrow^*_{G_c} wS' \Longrightarrow^*_{G_c} ww'$ alakú, ahol ahhoz, hogy a w szót előállítsuk P-beli szabályokat, ahhoz, hogy a w' szó elemeit előállítsuk, P'-beli szabályokat kell használnunk. Azaz, $L(G_c) \subseteq L(G)L(G')$. A fordított irányú tartalmazás könnyen látható.

Bizonyításvázlat - folytatás

A lezárás:

(1) Legyen $i = 2$ és legyen $S_0 \notin N$. Akkor

$$G_* = (N \cup \{S_0\}, T, P \cup \{S_0 \to \varepsilon, S_0 \to SS_0\}, S_0)$$

generálja az L^* nyelvet.

(2) Legyen $i = 3$. Definiáljuk a P_* szabályhalmazt úgy, hogy $A \to uS$ eleme P_*-nak minden $A \to u$ P-beli szabályra, ahol $u \in T^*$. Akkor

$$G_* = (N \cup \{S_0\}, T, P_* \cup P \cup \{S_0 \to \varepsilon, S_0 \to S\}, S_0)$$

grammatika generálja az L^* nyelvet.

A lezárás - folytatás

(3) Legyen $i = 0, 1$ és $\varepsilon \notin L$. Tegyük fel, hogy $S_0, S_1 \notin N$.

Legyen

$G_* = (N \cup \{S_0, S_1\}, T, P \cup \{S_0 \to \varepsilon, S_0 \to S, S_0 \to S_1 S\} \cup \{S_1 a \to S_1 S a \mid a \in T\} \cup \{S_1 a \to S a \mid a \in T\}, S_0)$ grammatika.

Legyen $L_* = L(G_*)$. Könnyen látható, hogy $L^* \subseteq L_*$. Megmutatjuk a fordított irányú tartalmazást.

A lezárás - folytatás

$G_* = (N \cup \{S_0, S_1\}, T, P \cup \{S_0 \to \varepsilon, S_0 \to S, S_0 \to S_1S\} \cup \{S_1a \to S_1Sa \mid a \in T\} \cup \{S_1a \to Sa \mid a \in T\}, S_0)$

Megmutatjuk, hogy $L(G_*) \subseteq L^*$.

Tekintsük az

$$S_0 \Longrightarrow_{G_*} u_1 \Longrightarrow_{G_*} u_2 \Longrightarrow_{G_*} \ldots \Longrightarrow_{G_*} u_m = u,\ m \geq 1$$

levezetést, ahol $u \in T^*$.

Ha az első lépésben az $S_0 \to \varepsilon$ szabályt használjuk, akkor $m = 1$ és $u_m = \varepsilon \in L^*$.

Ha $u_1 = S$, akkor $u \in L^*$, egyébként $u = S_1S$ és minden j-re, $1 \leq j \leq m$ indukcióval j szerint megmutatható, hogy u_j alakja a következő két alak közül valamelyik:

(a) $S_1v_1 \ldots v_k$, $k \geq 1$, ahol $S \Longrightarrow_G^* v_l$, $l = 1, \ldots, k$ és a $v_2, \ldots, v_k$ szavak mindegyike terminális szimbólummal kezdődik; vagy

(b) $v_1 \ldots v_k$, $k \geq 0$, ahol $S \Longrightarrow_G^* v_l$, $l = 0, \ldots, k$ és a $v_2, \ldots, v_k$ szavak mindegyike terminális szimbólummal kezdődik.

Az $u = S_1S$ eset az (a) esetnek felel meg. A G_* szabályait megvizsgálva láthatjuk, hogy u_{j+1} vagy (a), vagy (b) formájú, ha u_j rendelkezik ezzel a tulajdonsággal. azaz (a) vagy (b) formájú. Azaz, $L(G_*) \subseteq L^*$.

A lezárás - folytatás

Ha $i = 0, 1$ és $\varepsilon \in L$, akkor először veszünk egy G_1 grammatikát, ahol $L(G_1) = L - \{\varepsilon\}$.

(a) $i = 1$ esetében elhagyjuk az $S \to \varepsilon$ szabályt.

(b) $i = 0$ esetben a következőképpen járunk el: Jelölje P_ε azon $u \to \varepsilon$ alakú szabályok halmazát, amelyek elemei P-nek.

Akkor legyen $P_1 = (P - P_\varepsilon) \bigcup_u P_u$, ahol $P_u = \{uX \to X, Xu \to X \mid X \in (N \cup T), u \to \varepsilon \in P\}$.

G_1 típusa ugyanaz marad, mint G típusa és $(L - \{\varepsilon\})^* = L^*$.

Korollárium

Ha az L nyelv i-típusú, $i = 0, 1, 2, 3$, akkor L^+ is az.

Néhány további tulajdonság

(1) A $\mathcal{L}_i$, ahol $i = 0, 1, 2$ is zárt a megfordítás (tükrözés) műveletére nézve.

(2) $\mathcal{L}_i$, ahol $i = 0, 2$ zárt a homomorfizmus és $\mathcal{L}_1$ zárt a ε-mentes homomorfizmus műveletére nézve.

(3) Minden véges nyelv eleme $\mathcal{L}_3$-nak.