23. fejezet

Formális nyelvek

Formális nyelvtanok és Chomsky-féle nyelvosztályok. A reguláris nyelvek alapvető tulajdonságai és alkalmazásai. Környezetfüggetlen nyelvek és elemzésük. Matematikai gépek – véges automata és veremautomata.

23.1. Formális nyelvtanok

23.1.1. Alapfogalmak

A formális nyelvek alapelemei

23.1.1. Definíció (Ábécé). *Ábécének nevezünk egy tetszőleges véges szimbólumhalmazt. Az ábécé elemeit betűknek hívjuk.*

23.1.2. Definíció (Szó). *Az X ábécé betűinek egy véges (esetleg üres) sorozatát az X ábécé feletti szónak nevezzük.*

Jelölések. Az ábécé feletti összes szó halmazát jelölje X^*! Azt a legszűkebb ábécét, mely felett u jelsorozat még szó, jelölje $X(u)$! Az ε legyen az üres szó jele!

23.1.3. Definíció (Nyelv). *Az X ábécé feletti nyelvnek nevezzük az X^* egy részhalmazát, azaz az X feletti szavak egy tetszőleges halmazát.*

Jelölések. Legyen $X(L)$ a legszűkebb olyan ábécé, amely felett az L halmaz nyelv!

Műveletek

23.1.4. Definíció (Szavak konkatenációja). *Legyenek u és v szavak egy X ábécé felett. Ekkor a két szó konkatenációjának nevezzük azt a jelsorozatot, amelyet a két szó jeleinek egymás után fűzésével kapunk. Jelölése: uv.*

Megjegyzés. A konkatenáció *asszociatív*, de *nem kommutatív* művelet, melynek *egységeleme* ε. Így X^* a konkatenációval mint művelettel és ε-nal *egységelemes félcsoport*ot alkot.

23.1.5. Definíció (Megfordítás). *Az u szó megfordítása legyen jeleinek fordított sorrendben vett sorozata, és jelölje ezt u^{-1}!*

23.1.6. Definíció (Szó hossza). *Adott u szó hossza a benne szereplő jelek száma. Jelölése $l(u)$ vagy $|u|$.*

Jelölések. $X := X(u)$. $y \in X$ esetén jelölje $l_y(u)$ az u szóban az y jel előfordulásainak számát, továbbá ha $H \subseteq X$, akkor $l_H(u)$ az u-ban a H halmazbeli jelek összes előfordulásainak számát.

23.1.7. Definíció (Homomorfizmus). *Homomorfizmusnak nevezünk két ábécé közötti konkatenációtartó leképezést. Egy $h : X^* \to Y^*$ leképezés konkatenációtartó, ha tetszőleges $u, v \in X^* : h(uv) = h(u)h(v)$.*

Megjegyzések. Ha h homomorfizmus, akkor $h(\varepsilon) = \varepsilon$. A homomorfizmus nyelvekre vonatkozó kiterjesztése: $h(L) = \bigcup_{u \in L} h(u)$.

23.1.8. Tétel (Véges sok nyelv uniója nyelv). *Véges sok nyelv (halmaz) uniójához található olyan ábécé, amely felett az unió is nyelv. Ez az ábécé: $\bigcup_{i=1}^{k} X(L_i)$.*

23.1.9. Definíció (Nyelvek konkatenációja). *L_1 és L_2 nyelvek konkatenációja az $L_1L_2 = \{uv | u \in L_1 \wedge v \in L_2\}$.*

Megjegyzések. Nyelvek konkatenációja asszociatív. A konkatenáció az unióra nézve mindkét oldalról disztributív.

23.1.10. Definíció (Maradéknyelv). *Ha L nyelv és $u \in L$ szó, akkor L-nek az u-ra vonatkozó maradéknyelve $L_u = \{uv | uv \in L\}$.*

23.1.11. Definíció (Helyettesítés). *Nyelvek közötti, unió-, konkatenáció-, egységelem- és zéruselemtartó leképezés.*

23.1.12. Definíció (Lezárás). *L nyelv lezárásán a következő nyelvet értjük: $L^* = \bigcup_{i=0}^{\infty} L^i$, ahol L^i az L i-szeres, önmagával vett konkatenációját jelöli.*

Pozitív lezárásnak nevezzük és L^+-szal jelöljük a lezárásnak azt a változatát, ahol az üres nyelvet ($L^0 = \{\varepsilon\}$) elhagyjuk az unióképzésből.

23.1.2. Nyelvek megadása

A formális nyelvek megadási módszereitől elvárjuk, hogy a leírás véges legyen.

1. Véges nyelvek esetén felsorolhatjuk a nyelv elemeit.
2. A nyelvet megadhatjuk logikai formulával.
3. Megadható nyelv strukturális rekurzív leírás segítségével.
4. Algoritmus segítségével mely felsorolhatja a nyelv összes elemét, illetve egy szóról megmondhatja, hogy eleme-e a nyelvnek.

   23.1.13. Definíció (Rekurzívan felsorolhat nyelv). *Egy nyelv rekurzívan felsorolható, ha létezik olyan algoritmus, mely felsorolja minden elemét (nem feltétlenül véges időn belül).*

   23.1.14. Definíció (Rekurzív nyelv). *Egy nyelvet rekurzívnak nevezünk, ha létezik eldöntő algoritmus, ami inputként egy szót kapva mindig terminál, és megmondja a szóról, hogy eleme-e a nyelvnek.*

   23.1.15. Definíció (Parciálisan rekurzív nyelv). *Egy nyelv parciálisan rekurzív, ha létezik olyan algoritmus, amely mindig terminál* igen *válasszal, ha az inputként kapott szó eleme a nyelvnek; és nem terminál vagy* nem *válasszal terminál, ha nem eleme.*

5. Matematikai géppel is megadható egy nyelv.
6. Nyelv definiálható produkciós rendszerrel (axiómák és következtetési szabályok segítségével).

Produkciós rendszerek

23.1.16. Definíció (Produkciós rendszer). *Produkciós rendszer alatt a* $\Pi = (X, P, A_x)$ *hármast értjük, ahol* X *ábécé,* $P \subseteq X^* \times X^*$ *a produkciós szabályok véges halmaza, és* $A_x \subseteq X^*$ *véges axiómahalmaz.*

Levezetések. Egy szót *közvetlenül levezethetőnek* mondunk egy Π produkciós rendszerben, ha az egyik szóból a másik előállítható valamely résszónak a P szabályok valamelyike segítségével történő helyettesítéssel.

Egy szó *közvetett módon levezethető* a produkciós rendszerben, ha valahány (véges) levezetési lépés segítségével előállítható a kiinduló szóból a célszó.

Generatív nyelvtanok

A generatív nyelvtanok speciális produkciós rendszerek.

23.1.17. Definíció (Formális nyelvtan). *Olyan* $G = (T, N, \mathcal{P}, S)$ *négyes, ahol* T *a terminális jelek ábécéje,* N *a nyelvtani jelek ábécéje,* $\mathcal{P}$ *véges szabályhalmaz, melyben bármely szabály bal oldalán szerepel legalább egy nyelvtani jel, és* $S \in N$ *a kezdőszimbólum.*

23.1.18. Definíció (Mondatforma). *Mondatformának nevezzük terminális és nemterminális szimbólumok véges sorozatát.*

23.1.19. Definíció (Generált nyelv). *A* G *formális nyelvtan által generált nyelv szavai a kezdőszimbólumból levezethető szavak.*

23.1.20. Tétel. *Nem minden nyelv írható le nyelvtannal. A bizonyítás az* $\mathcal{L}_0^T \subset 2^{T^*}$ *összefüggés belátásával történik.*

Nyelvtanok osztályozása

Jelölések.

- $A, B, C \in N$ nyelvtani jelek,
- S a kezdőszimbólum,
- p egy szabály bal oldala,
- q tetszőleges szabály jobb oldala,

- $\alpha_{1,2} \in (T \cup N)^*$ mondatforma,
- KES a korlátozott epszilon-szabályt jelöli, azaz *ha a nyelvtanban szerepel az $S \to \varepsilon$ szabály, akkor S nem szerepelhet szabály jobb oldalán.*

O	Alapnyelv	Megszorított	Normálforma
0	$p \to q$	$p \to q$ $(q \neq \varepsilon)$, $S \to \varepsilon$ KES	$AB \to A$, $BA \to A$ + az 1-es és 2-es normálforma szabályai.
1	$p \to q$, $l(p) \leq l(q)$, $S \to \varepsilon$ KES	$\alpha_1 A \alpha_2 \to \alpha_1 q \alpha_2$ $(q \neq \varepsilon)$, $S \to \varepsilon$ KES	$AB \to AC$, $BA \to CA$ + 2-es normálforma szabályai.
2	$A \to q$	$A \to q$ $(q \neq \varepsilon)$, $S \to \varepsilon$ KES	$A \to BC$, $A \to t$, $S \to \varepsilon$ KES
3	$A \to nB$, $A \to n$	$A \to tb$, $A \to t$, $S \to \varepsilon$ KES	$A \to tB$, $A \to \varepsilon$

Megjegyzés. Az 1. típusú nyelvek rekurzívak, azaz létezik hozzájuk eldöntő algoritmus (általában $\mathcal{O}(e^x)$ futásidővel).

Chomsky-féle hierarchia. $\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_1 \subseteq \mathcal{L}_0$.

23.1.21. Tétel (A nyelvosztályok változatainak kapcsolata). *Az alapformák, megszorított formák, illetve normálformák által meghatározott nyelvosztályok mind a négy osztály esetén azonosak, azaz $\mathcal{L}_i = \mathcal{L}_{msz_i} = \mathcal{L}_{nf_i}$.*

23.2. A 3. típusú nyelvek

23.2.1. Tétel (Kis Bar-Hillel lemma). *Tetszőleges $L \in \mathcal{L}_3$ nyelvhez $\exists n = n(L) > 0$ egész szám, hogy $\forall u \in L, l(u) \geq n : u = xyz$ a következő tulajdonságokkal:*

- $y \neq \varepsilon$,
- $l(xy) \leq n$,
- $\forall i = 0, 1, \ldots : xy^i z \in L$.

23.2.2. Tétel (Zártság). *$\mathcal{L}_3$ zárt az unió, konkatenáció, lezárás, komplementerképzés, metszet, különbség és szimmetrikus differencia műveletekre nézve.*

23.2.3. Definíció (Reguláris nyelv). *A reguláris nyelvek definícióját strukturális indukcióval adjuk meg:*

- *Reguláris nyelvek az elemi nyelvek:* $\{\}, \{\varepsilon\}, \{t\}$, *ahol* $t \in \mathcal{U}$.
- *Reguláris műveletek az* unió, konkatenáció *és a* lezárás.
- *Reguláris nyelvek az elemi nyelvekből a reguláris műveletek véges sokszori alkalmazásával előállítható nyelvek.*

23.2.4. Definíció (Reguláris kifejezés). *A reguláris kifejezések a reguláris nyelvek egyszerűsített leírását jelentik, ahol:*

- $\{t\}$ *helyett* t *szerepel,*
- *az unió jele* $\cup$ *helyett* $+$.

23.2.5. Tétel (Kleene-tétel). *A reguláris nyelvek osztálya megegyezik a véges determinisztikus automatával előállítható nyelvek osztályával.*

23.2.6. Tétel. *A véges determinisztikus automatával előállítható nyelvek osztálya megegyezik a véges, nemdeterminisztikus automatával előállítható nyelvek osztályával.*

23.2.7. Tétel. *A véges determinisztikus automatával előállítható nyelvek osztálya megegyezik a 3. nyelvosztállyal* ($\mathcal{L}_3$).

23.2.1. Reguláris nyelvek felhasználási területei

Bár a reguláris nyelvtanok az összes nyelvek viszonylag szűk osztályát jelölik ki, könnyű kezelhetőségük miatt gyakorlati alkalmazásuk rendszeres.

Lexikális elemző (scanner): A fordítóprogramok legalsó szintjén, a lexikális elemek felderítésében nagy szerepet játszanak a 3. típusú nyelvtanok: általában reguláris kifejezésekkel azonosítják a nyelv lexikális elemeit.

Mintaillesztés: Mintaillesztési feladatokban (pl. adatok között/szövegben történő keresés) reguláris kifejezéssel írható le az illesztendő minta. Ebből általában véges determinisztikus automatát (például *Knuth-Morris-Pratt automatát*) generálnak.

23.3. Automaták

Az automaták a matematikai gépek egy csoportját alkotják. Ebben a fejezetben a véges (determinisztikus és nemdeterminisztikus) automatákkal, illetve veremautomatákkal foglalkozunk.

23.3.1. Véges determinisztikus automaták

23.3.1. Definíció (Véges determinisztikus automata). *Egy véges determinisztikus automatát az $\mathcal{A} = (Q, T, \delta, q_0, F)$ ötös ír le, ahol:*

- *Q az állapotok véges halmaza,*
- *T a bementi értékek ábécéje (így véges),*
- *$\delta : Q \times T \to Q$ az átmeneti függvény,*
- *$q_0 \in Q$ a kezdő állapot,*
- *$F \subseteq Q$ a végállapotok halmaza.*

Megjegyzés. δ az automata egy lépésben végzett működését írja le. Kiterjeszthető $\hat{\delta} : Q \times T^* \to Q$ alakra, ahol:

- $\hat{\delta}(q, \varepsilon) = q$,
- $\hat{\delta}(q, ut) = \delta(\hat{\delta}(q, u), t)$.

23.3.2. Definíció (Automata által felismert nyelv). *Az automata által felismert nyelv azon szavak halmaza, amelyekre a kezdő állapotból valamelyik végállapotba jut, azaz $L(\mathcal{A}) = \{u \in T^* | \delta(q_0, u) \in F\}$.*

Az automata megadási módszerei

Átmeneti gráf: Olyan irányított gráf, melynek pontja az állapotok, és A állapotból t-vel címkézett él fut B állapotba, ha $\delta(A, t) = B$.

Táblázat: egyik tengelyén az állapotok, a másikon a terminálisok találhatók – lényegében az átmeneti gráf mátrixreprezentációja.

Képlet: a δ függvény megadása valamilyen zárt számítási képlettel, formulával.

Automata és a 3. típusú nyelvek kapcsolata

23.3.3. Tétel. $\mathcal{L}_{DA} = \mathcal{L}_3$.

3. típusú nyelvtan konstrukciója automatából. A konstrukció alapötlete a következő:

1. hozzuk normálformára a nyelvtant,
2. a nyelvtani jeleket feleltessük meg az automata állapotainak,
3. q_0-t rendeljük az S startszimbólumhoz,
4. $q \to tq' \in \mathcal{P} \Leftrightarrow \delta(q,t) = q'$,
5. $q \to \varepsilon \in \mathcal{P} \Leftrightarrow q \in F$.

Automata konstrukciója 3. típusú nyelvből. A fentihez hasonló, fordított konstrukcióval készíthető 3. típusú nyelvtan alapján automata, ez azonban nem mindig lesz determinisztikus – ha nemdeterminisztikus állapotátmeneti függvényt kapunk, az automatát még determinisztikussá kell tennünk (ld. az automaták determinisztikussá tételéről szóló tételt).

23.3.2. Véges nemdeterminisztikus automata

23.3.4. Definíció (Véges nemdeterminisztikus automata). *Véges nemdeterminisztikus automata felépítésében azonos a determinisztikus automatával,* de $\delta : Q \times T \to 2^Q$, *illetve* $\hat{\delta}(q, ut) = \bigcup_{q' \in \hat{\delta}(q,u)} \delta(q', t)$.

23.3.5. Tétel (Automata determinisztikussá tétele). *Véges nemdeterminisztikus automatához konstruálható véges determinisztikus automata, amely ugyanazt a nyelvet ismeri fel.*

1. $\mathcal{A}' = (2^Q, T, \delta', q_0, \mathcal{F})$,
2. $\delta'(\{q_1, \ldots, q_s\}, t) = \bigcup_{i=1}^{s} \delta(q_i, t)$,
3. $\mathcal{F} = \{E \subset Q | E \cap F \neq \{\}\}$.

23.3.3. Veremautomaták

23.3.6. Definíció (1-verem automata). *Az egy veremmel rendelkező (1-verem) automatát a* $(V) = (Q, T, \Sigma, \delta, q_0, \sigma_0, F)$ *hetessel azonosítjuk, ahol*

- Q *az állapotok véges halmaza,*
- T *a terminális ábécé,*
- Σ *a veremábécé,*
- $\delta : Q \times T \cup \{\varepsilon\} \times \Sigma \to 2^{Q \times \Sigma^*}$ *az átmeneti függvény, hogy* $|\delta(q, t, \sigma)| < \infty$,
- q_0 *a kezdőállapot,*
- σ_0 *a verem kezdőjele,*
- F *a végállapotok halmaza.*

Megjegyzés. A több veremmel rendelkező automaták esetén a definíció érteemszerűen kibővül vermenként további veremábécékkel, illetve kezdőjelekkel. Az átmeneti függvény, illetve a konfiguráció definíciója is bővül.

Megjegyzés. A veremautomaták a működés definíciójától függően kétféle módon fogadhatnak el szót: *elfogadó állapottal*, vagy *üres veremmel*. Utóbbi esetben a definícióban az elfogadó állapotokat nem szokás jelölni. A kétféle automata által felismert nyelvosztályok azonosak.

23.3.7. Definíció (Konfiguráció). *Veremautomata egy konfigurációján egy* $[q, v, \alpha] \in Q \times T^* \times \Sigma^*$ *hármast értünk, melyben* q *az aktuális állapot,* v *a hátralévő inputszöveg és* α *a verem tartalma.*

Megjegyzések. A veremautomaták általában nemdeterminisztikusak, mivel az átmeneti függvényben a terminális szimbólum helyén megengedik az üres szót (ε) is.

A veremautomaták és a 2. típusú nyelvek kapcsolata

23.3.8. Tétel. *Az 1-verem automatákkal felismerhető nyelvek osztálya megegyezik a 2. típusú grammatikák által generált nyelvek osztályával, azaz* $\mathcal{L}_\mathcal{V} = \mathcal{L}_2$.