

Automaták és formális nyelvek

Rövid összefoglaló

A félév során formális nyelvekről és automatákról tanultunk. A formális nyelvek, vagy röviden „nyelvek” számunkra szavak általában végtelen halmazai, a szavak pedig véges szimbólum halmazok (ábécék) elemeiből (betűk) képzett szimbólum sorozatok (azaz sztringek). Például, $\Sigma = \{a, b, c, d\}$ egy véges szimbólumhalmaz, vagyis egy ábécé. Az ab , $bbcd$, d , $cccc$, $dadcbb$ ennek az ábécének a betűiből képzett szavak, más szóval a Σ ábécé feletti szavak. Egy Σ ábécé feletti nyelv (vagyis a Σ betűiből képzett szavakból álló nyelv) a Σ betűiből képzett összes lehetséges szavak halmazának valamilyen részhalmaza.

A nyelvek halmazok, tehát azokat a műveletek, amik általában a halmazokon értelmesek, azokat nyelvekre is értelmezzük (például unió, metszet, komplementer, különbség). A szavakon értelmezett speciális művelet a konkatenáció, amit kiterjesztünk szóhalmazokra, azaz nyelvekre is, illetve a konkatenáció alapján a nyelveken értelmezzük a „hatványozást” és a konkatenáció lezárását is („csillag”).

A formális nyelvek általában végtelen sok szóból állnak, ezért őket az elemeik felsorolása helyett más módon kell megadni. Két módszer a generatív grammatikákkal (nyelvtanokkal) illetve az automatákkal történő megadás. Egy grammatika szimbólum-átírási szabályok halmaza, amelyek segítségével egy kezdőszimbólumból szavakat generálunk: amely szavak generálhatók, azokat tartalmazza a grammatika által megadott nyelv, amelyek nem generálhatók, azokat nem tartalmazza. Egy automata valamilyen absztrakt matematikai „gép”, azaz egy eljárás, amivel egy elképzelt bemeneti szalagon adott bemeneti szó vizsgálatát tudjuk elvégezni: ha a „vizsgálat” elfogadja a szót, akkor az benne van az automata által megadott (leírt, definiált) nyelvben, ha nem fogadja el, akkor nincs.

A formális nyelvek osztályozhatók (jellemezhetők, csoportosíthatók, kategorizálhatók, stb.) annak alapján, hogy milyen típusú automaták illetve grammatikák szükségesek a megadásukhoz. Az a típus amivel először foglalkoztunk, a reguláris nyelvek osztálya. Ezeket úgy definiáltuk, hogy a véges automatának nevezett automata-típus segítségével adhatók meg. Megmutattuk azt is, hogy ebből a szempontból nem számít az, hogy a véges automata determinisztikus, vagy nem-determinisztikus: bár a determinisztikus véges automaták a nemdeterminisztikus véges automaták speciális esetei, mégis teljesül, hogy bármilyen nyelv, ami megadható nemdeterminisztikus véges automatával, az megadható determinisztikussal is. (Ezért létezik olyan eljárás, amit a véges automaták „determinizálásának” neveztünk, és aminek segítségével egy tetszőleges nem-determinisztikus véges automatát determinisztikussá tudunk alakítani, úgy, hogy az általa leírt (megadott, definiált, stb.) nyelv ne változzon meg.)

A reguláris nyelvekkel illetve a véges automatákkal kapcsolatban fontos még, hogy minden reguláris nyelvhez tartozik egy minimális állapotszám, amit a nyelvet leíró automatának mindenképpen szükséges tartalmaznia ahhoz, hogy az adott nyelvet le tudja írni. Ez a minimális állapotszám a nyelv tulajdonsága, vagyis minden egyes reguláris nyelvhez létezik egy minimális állapotszámú determinisztikus automata (ennél kevesebb állapottal bíró automatával a nyelv nem adható meg), ráadásul ez az automata egyértelmű, ami azt jelenti, hogy (az állapotok átnevezésétől eltekintve) nem tudunk két különböző minimális állapotszámú automatát konstruálni egy nyelvhez, csak egyet. (A minimális állapotszám a nyelv szavainak megkülönböztethetősége

alapján definiált reláció ekvivalencia osztályainak számával van kapcsolatban, ezekről szól néhány nem túl bonyolult tétel.)

A reguláris nyelveket megadhatjuk grammatikával is, ehhez olyan grammatikára van szükség, amelyben az átírási szabályok alakja $A \rightarrow aB$ vagy $A \rightarrow a$ vagy $A \rightarrow \lambda$, ahol A, B valamilyen nemterminális szimbólum, a valamilyen terminális szimbólum, λ pedig az úgynevezett üres szó. Az ilyen szabályokat tartalmazó grammatikát reguláris grammatikának nevezzük.

A reguláris nyelveket az automatákon és a grammatikákon kívül megadhatjuk ún. reguláris kifejezésekkel is, ezek bizonyos alapkifejezésekből jól definiált módon felépülő összetett kifejezések, amelyek szintén jól definiált módon egyszerű halmazoknak az unió, konkatenáció és „csillag” műveletek (azaz az ún. reguláris műveletek) segítségével előálló kombinációjával keletkező szóhalmazokat, vagyis nyelveket írnak le.

Összefoglalva: A reguláris grammatikák, a nemdeterminisztikus és determinisztikus véges automaták és a reguláris kifejezések ekvivalensek a nyelvleíró „erő” szempontjából: mindannyian a reguláris nyelveket írják le, azaz, amelyik nyelv megadható az egyik formalizmus segítségével, az megadható bármelyik másik segítségével is (és minden reguláris nyelv ilyen).

Nem minden formális nyelv reguláris azonban. Ezzel kapcsolatban foglalkoztunk a reguláris nyelvekre vonatkozó pumpálási lemmával, ami nagyjából azt állítja, hogy ha egy nyelv reguláris, akkor az általa tartalmazott elég hosszú szavakra teljesül egy bizonyos pumpálási tulajdonság. Ennek alapján meg tudtuk mutatni bizonyos nyelvekről (pl. az $\{a^n b^n \mid n \geq 1\}$ nyelvről), hogy nem lehetnek regulárisak, hiszen akármilyen hosszú szót is veszünk belőlük, azokra nem teljesül a pumpálási tulajdonság.

A nem reguláris (tehát a véges automatával és a vele a fentiek szerint ekvivalens formalizmusokkal nem megadható) nyelvek közül először a környezetfüggetlen nyelvekkel foglalkoztunk. Ezeket úgy definiáltuk, mint az olyan generatív grammatikával megadható (generálható) nyelveket, melyek $A \rightarrow \alpha$ alakú átírási szabályokat tartalmaznak, ahol A valamilyen nemterminális szimbólum, α pedig valamilyen nemterminálisokból és terminálisokból álló szó (lehet az üres szó is). (Látszik, hogy ezek a szabályok általánosabbak mint a reguláris grammatikák szabályai, ott nem lehetett pl. olyan szabály, ami a jobb oldalon egynél több nemterminálist tartalmaz, itt lehet olyan.) A környezetfüggetlen grammatikák kapcsán beszéltünk a környezetfüggetlen levezetéseket reprezentáló fákról, a levezetési fákról, és definiáltuk az egyértelmű nyelvtan fogalmát: azt a környezetfüggetlen nyelvtant nevezzük egyértelműnek, amelyben minden egyes levezethető (generálható) szónak csupán olyan levezetései vannak, amikhez egyetlen levezetési fa tartozik (azaz az egyes szavak lehetséges levezetései nem különböznek lényegesen, csak a szabályok alkalmazásának sorrendjében). Általában egy nyelv sok különböző grammatikával generálható, ezek között lehetnek egyértelműek és nem egyértelműek is. Említettük azonban, hogy vannak olyan nyelvek, amelyek csak nem egyértelmű grammatikával generálhatók

Láttuk azt is, hogy a környezetfüggetlen grammatikák szabályhalmazai átalakíthatók úgy (a generált nyelv megváltoztatása nélkül), hogy ne tartalmazzanak törlő szabályokat és láncszabályokat sem. Sőt, átalakíthatók úgy is, hogy csak $A \rightarrow BC$, $A \rightarrow a$, $A \rightarrow \lambda$ alakú szabályokat tartalmazzanak, ahol A, B, C valamilyen nemterminális szimbólum, a valamilyen terminális szimbólum, λ pedig az üres szó. (Azokról a grammatikákról, amik csak ilyen szabályokat tartalmaznak, azt mondjuk, hogy Chomsky féle normálformában vannak. Itt beszéltünk még a Cocke-Younger-Kasami algoritmusról is, amelyikkel egyszerűen el tudjuk dönteni egy adott szóról, hogy egy szintén adott, Chomsky normálformában lévő grammatikával generálható-e.)

Ha a környezetfüggetlen nyelveket automatával akarjuk megadni, akkor az ún. veremautomatákra van szükségünk: nemdeterminisztikus veremautomatával tetszőleges környezetfüggetlen (azaz környezetfüggetlen grammatikával generálható) nyelv megadható, és viszont: minden nemdeterminisztikus veremautomata által elfogadott nyelvhez konstruálható a nyelvet generáló környezetfüggetlen grammatika. (E két állítás közül viszonylag könnyű megmutatni, hogy minden

környezetfüggetlen grammatika által generált nyelvhez lehet a nyelvet elfogadó veremautomatát konstruálni, ezt meg is mutattuk.) A veremautomaták esetében is definiáltuk a nemdeterminisztikus mellé a determinisztikus változatot, ezek azonban (a véges automatáktól eltérően) nem ugyanazoknak a nyelveknek az elfogadására képesek: van olyan környezetfüggetlen nyelv, ami determinisztikus veremautomatával nem adható meg (fogadható el), nemdeterminisztikussal viszont igen.

Láttuk azt is, hogy vannak olyan nyelvek, amelyek egyáltalán nem generálhatók semmilyen környezetfüggetlen grammatikával. Ebben a a környezetfüggetlen nyelvekre vonatkozó pumpálási lemma volt segítségünkre: Hasonlóan a reguláris nyelvekre vonatkozó változathoz, a lemma arról szól, hogy ha egy környezetfüggetlen grammatika által generált szó elég hosszú, akkor egy bizonyos pumpálási tulajdonságnak teljesülnie kell rá (ez a tulajdonság nem ugyanolyan, mint a reguláris nyelvek esetén). Ezek szerint, ha egy nyelvről sikerül megmutatni, hogy akármilyen hosszú szót is veszünk belőle, a pumpálási tulajdonság nem teljesülhet, akkor az a nyelv nem lehet generálható környezetfüggetlen grammatikával, azaz nem lehet környezetfüggetlen. (Ilyen nyelv volt pl. az $\{a^n b^n c^n \mid n \geq 1\}$.)

A következő automata típus, amiről szó volt, az ún. Turing gép. Ez az automata az olvasás mellett írhat is a képeletbeli (potenciálisan végtelen hosszúságú) szalagra, így elég bonyolult ahhoz, hogy a „számítás” fogalmának matematikai modellje legyen (ezzel a céllal vezette be Alan Turing a XX. század első felében). Erről szól az ún. Church-Turing tézis, miszerint a Turing gép minden „automatizálható” számítást, vagy algoritmust végre tud hajtani: amit egy Turing géppel nem tudunk megoldani (kiszámítani) annak a megoldására (kiszámítására) egyáltalán nem tudunk algoritmust adni, azaz annak a megoldására (kiszámítására) semmilyen más számítógép sem lehet képes. Ide kapcsolódik az ún. univerzális Turing gép, amelyiknek a bemeneti szó mellé egy tetszőleges másik Turing gépet leíró „programot” is írhatunk a szalagjára, aminek hatására a bemenetet a programban leírt Turing géppel megegyező módon fogja feldolgozni.

A Turing gép segítségével definiáltunk két új nyelvosztályt, a rekurzív és a rekurzívan felsorolható nyelvek osztályát: Egy nyelvet rekurzívnak nevezünk akkor, ha konstruálható hozzá olyan Turing gép, mely a nyelv szavait kapva bemenetül elfogadó állapotba kerül a nyelvhez nem tartozó szavak hatására pedig egy ún. elutasító állapotba. Egy nyelvet rekurzívan felsorolhatónak nevezünk, ha konstruálható hozzá olyan Turing gép, mely a nyelv szavait kapva bemenetül elfogadó állapotba kerül (a nyelvhez nem tartozó szavak hatása nem feltétlen definiált, elképzelhető az is, hogy ezekre a szavakra a gép nem áll meg). (Az előbbiekkel ekvivalens, ha azt mondjuk, hogy a rekurzív nyelvek karakterisztikus függvénye Turing géppel kiszámítható, illetve a rekurzívan felsorolható nyelvek szavait egy Turing gép fel tudja sorolni, vagyis a kezdő állapotában elindítva, olyan (végtelen hosszú) működést produkál, melyben a végtelen szalagjára a nyelv szavait sorban felírja.)

A Turing gépekkel kapcsolatos vizsgálódások (mivel a Turing gép a számítás illetve az algoritmus fogalmának matematikai modellje) sok érdekes eredményt adnak problémák algoritmikus megoldhatóságáról illetve megoldhatatlanságáról, függvények kiszámíthatóságáról illetve kiszámíthatatlanságáról, de a formális nyelvekkel kapcsolatban talán nekünk most érdekesebb, hogy a Turing gépekkel elfogadható nyelvek (a rekurzívan felsorolható nyelvek) ugyanazok, amelyek általános alakú szabályokat tartalmazó generatív grammatikákkal generálhatók. Általános alakú egy szabály, ha $\alpha \rightarrow \beta$ alakú, ahol α, β nemterminálisokból és terminálisokból álló tetszőleges szavak (bár α tartalmaz legalább egy nemterminálist).

Mivel az összes eddigi szabálytípus az általános alakú szabály speciális esete, az összes eddigi nyelvosztály nyelvei generálhatóak általános alakú szabályokat tartalmazó grammatikával, azaz az összes eddigi nyelvosztály nyelvei rekurzívan felsorolhatóak.

A generatív grammatikákban használatos szabályoknak még egy típusáról beszéltünk, az ún. környezetfüggő szabályokról, ezeket úgy definiáltuk, hogy $\alpha \rightarrow \beta$ alakúak, ahol az β szó

hosszúsága nem lehet rövidebb, mint az α hossza. Az ilyen szabályokat tartalmazó grammatika segítségével megadható nyelveket környezetfüggő nyelveknek nevezzük. Ha a törlő szabályoktól eltekintünk (eltekinthetünk, mert kiküszöbölhetőek), akkor a környezetfüggetlen grammatikák is környezetfüggőek, tehát a környezetfüggetlen nyelvek is környezetfüggőek. Tudjuk azt is, hogy van olyan nyelv, amelyik nem környezetfüggetlen, de környezetfüggő. Ezek szerint a reguláris, a környezetfüggetlen, a környezetfüggő és a rekurzívan felsorolható nyelvek egy „szép” hierarchiát alkotnak, miszerint egyik sem egyezik meg a másikkal, de sorban tartalmazzák egymást, a legszűkebb reguláris nyelvektől kezdve sorban a legtágabb rekurzívan felsorolható nyelvek osztályáig. (Ezt nevezzük a nyelvosztályok Chomsky féle hierarchiájának.) Egy kérdés marad még: Hol helyezkednek el ebben a hierarchiában a rekurzív nyelvek. A környezetfüggő nyelvekről megmutattuk, hogy rekurzívak, tehát a rekurzív nyelvek osztálya tartalmazza a reguláris, a környezetfüggetlen és a környezetfüggő nyelvek osztályát is, azt pedig szintén láttuk, hogy vannak rekurzívan felsorolható, de nem rekurzív nyelvek.

Irodalom: (Lásd a Neptun Meet Street-en, vagy a Google Drive-on, utóbbihoz a linket megkapták emailben.)

- Demetrovics J., J. Denev, R. Pavlov: *A számítástudomány matematikai alapjai*, Nemzeti Tankönyvkiadó, 1985.
- Dömösi P., Falucskai J., Horváth G., Mecsei Z. és Nagy B.: *Formális nyelvek és automaták*, jegyzet 2011.
http://www.tankonyvtar.hu/en/tartalom/tamop425/0046_formalis_nyelvek_es_automatak/0046_formalis_nyelvek_es_automatak.pdf
- J. Martin: *Introduction to Languages, and the Theory of Computation, 4th edition*, McGraw-Hill, New York, NY, 2011.
- M. Sipser *Introduction to the Theory of Computation*, Thomson, 2006.