Formális nyelvek, nyelvtanok, gépek

Számítógépes nyelvészet – 2018 tavasz
5. óra

Simon Eszter – Mittelholcz Iván

MTA Nyelvtudományi Intézet

Tartalom

A Chomsky-féle nyelvhierarchia

Definíció (nyelv)

Egy tetszőleges véges A halmazból alkotott $A*$ halmaz tetszőleges részhalmazát (vagyis az A fölötti füzérekből összegyűjtött tetszőleges halmazt) nyelvnek nevezzük, az A halmazt pedig e nyelv ábécéjének.

A nyelvtan formális meghatározása

Definíció (nyelvtan)

A $G =< V_T, V_N, S, R >$ négyest formális nyelvtannak nevezzük, ahol
V_T: a terminális elemek ábécéje
V_N: a nem-terminális elemek ábécéje
S: a nyelvtan kezdőszimbóluma
R: a nyelvtan szabályainak a halmaza

A Chomsky-féle nyelv(tan)hierarchia

Definíció (Chomsky-féle nyelv(tan)hierarchia)

Egy $G =< V_T, V_N, S, R >$ nyelvtan i típusú ($i \in \{0, 1, 2, 3\}$) a Chomsky-féle nyelvtanhierarchia szerint, amennyiben az R szabályhalmaz minden elemére teljesül az adott típusban előírt, a szabály felépítésére vonatkozó előírás.

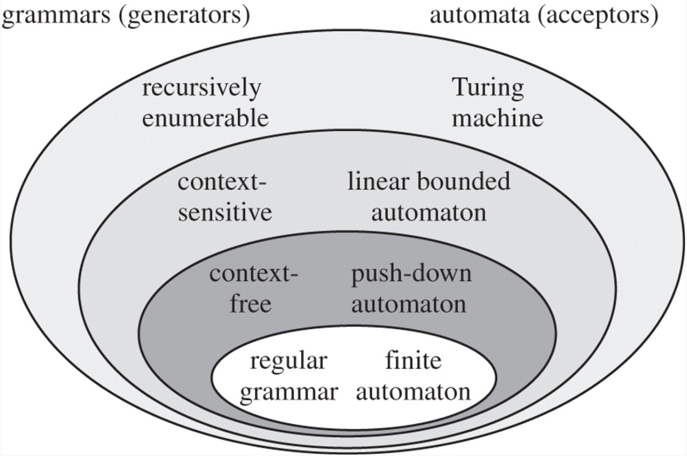
Jelölések: $\alpha, \beta, \gamma \in (V_T \cup V_N)*$

$A, B \in V_N$

$x \in V_T*$

0. típus (megszorítatlan újraíró rendszer)	$\alpha \to \beta$, ahol $\alpha \neq \varepsilon$
1. típus (környezetfüggő)	$\alpha A \beta \to \alpha\gamma\beta$, ahol $\gamma \neq \varepsilon$
2. típus (környezetfüggetlen)	$A \to \gamma$
3. típus (reguláris vagy jobblineáris)	$A \to xB$ vagy $A \to x$

Nyelvtanok és gépek



Megszorítatlan (0. típusú) nyelvek

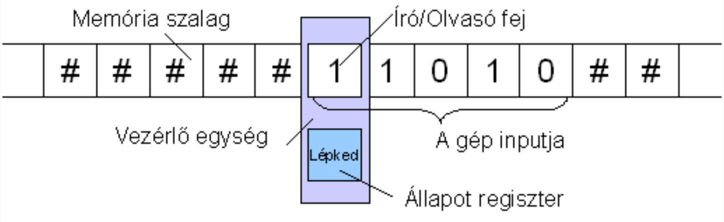
$\alpha \to \beta$, ahol $\alpha \neq \varepsilon$

- megszorítatlan újraíró rendszer
- a leghasznosabb új eszköz a szimbólumcsere
- a 0. típusú nyelveket Turing-géppel lehet elfogadtatni $\to$ a legáltalánosabb nyelvtantípusnak a létező és elképzelhető legáltalánosabb absztrakt gép felel meg
- „a Turing-géppel minden kiszámolható, ami egyáltalán kiszámolható, és minden meghatározható, ami egyáltalán meghatározható emberi elménk számára”

Turing-gép

- író-olvasó fej, egy kockákra osztott végtelennek tekintett szalag, amelynek minden kockáján egy szimbólum áll
- a fej jobbra és balra is mozoghat
- az író-olvasó fejnek különböző állapotai vannak, amelyek megszabják, hogy az éppen leolvasott szimbólumot átírja-e, vagy lépjen tovább
- a végtelennek tekintett szalag egy véges részén van csak információ, a többi # jelet tartalmazó üres hely: $...\#a_1a_2...a_n\#...$
- a számítás kezdetén a fej a bal szélső nem-üres szimbólumon áll, utána: $< q_i, a, q_j, X >$
- ha $X \in \{J, B\}$, akkor jobbra vagy balra lép egyet, különben marad és átír
- a Turing-gép akkor fogad el egy füzért, ha a számítás véges számú lépés után leáll

Turing-gép



Memória szalag
Író/Olvasó fej
1 1 0 1 0 #
Vezérlő egység
A gép inputja
Lépked
Állapot regiszter

És így született meg a Turing-gép...

A környezetfüggő (1. típusú) nyelvek

$\alpha A\beta \rightarrow \alpha\gamma\beta$, ahol $\gamma \neq \varepsilon$

- a környezetfüggő szabály egyetlen nem-terminális szimbólum átalakításáról gondoskodik
- a bemeneti oldalon a szimbólum kétoldali környezetére is hivatkozhatunk
- a nem-terminális szimbólum megsemmisítése nem megengedett $\rightarrow$ hosszúságot nem csökkentő szabályok
- a környezetfüggő nyelvek absztrakt számítógépes jellemzésére a lineárisan korlátozott Turing-gépek szolgálnak $\rightarrow$ az író-olvasó fej nem léphet ki egy kezdő- és egy végszimbólumokkal jelölt területen kívülre

Környezetfüggő szabályok

Noam Chomsky and Morris Halle: The Sound Pattern of English (1968)

$$A \rightarrow B/[precontext\_\_postcontext]$$

tundrai nyenyec cirill–IPA konverzió

ю $\rightarrow ju/\#\_\_$

ю $\rightarrow ju/\_\_$ [аеёиоуыэюя]

ю $\rightarrow^{j} u/$ [бвгджзйклмнңпрстфхцчшщ] $\_\_$

else:

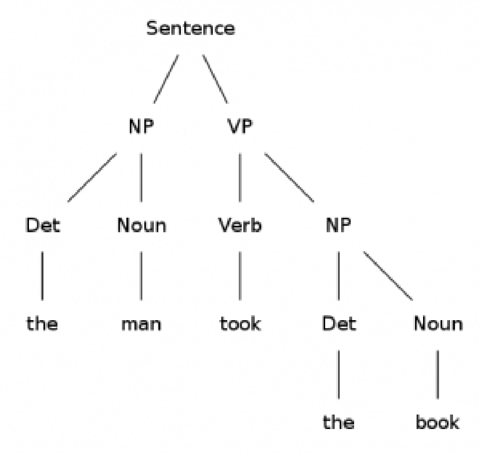
ю $\rightarrow ju/\_\_$

A környezetfüggetlen (2. típusú) nyelvek

$A \to \gamma$

- csak a szabály bemeneti oldala van korlátozva: egyetlen nem-terminális szimbólum állhat ott
- a kimeneti oldalon megengedett az üres sztring is
- Chomsky-féle normálalak: $A \to a$ vagy $A \to BC$
- elfogadó gép: veremautomata $\to$ mintha egy véges automata meg lenne toldva egy veremszerű memóriával
- verem: „last in, first out” (LIFO)

Környezetfüggetlen szabályok



Az első környezetfüggetlen elemzési fa (Chomsky, 1956: Three models for the description of language. *IRI Transactions on Information Theory, 2(3)*, 113–124.).

A reguláris (3. típusú) nyelvek

$A \rightarrow xB$ vagy $A \rightarrow x$

- a reguláris nyelveket leíró reguláris nyelvtanok reguláris kifejezésekkel ekvivalensek
- a reguláris nyelvtanok lehetnek jobblineárisak vagy ballineárisak
- egy jobblineáris szabály bemeneti oldalán egyetlen nem-terminális szimbólum állhat, és maximum egy nem-terminális állhat a kimeneti oldalán, és ez utóbbinak a legutolsónak kell lennie
- elfogadó gép: véges állapotú automata (finite state automaton, FSA)

Hol vannak a természetes nyelvek?

- valahol a környezetfüggetlen és a környezetfüggő között → enyhén környezetfüggő nyelvtanok
- környezetfüggetlen szabályokkal a természetes nyelvi jelenségek nagyon nagy része leírható, de nem minden → ellenpélda egy svájci német dialektusból

Shieber (1985): keresztező függőség az igék és tárgyuk között

Jan säit das...

...mer	em Hans	es huus	hälfed	aastriche
mi	Hans.DAT	a ház.ACC	segített	fest

'Jan azt mondta, hogy segítettünk Hansnak festeni a házat.'

két ugyanolyan füzér konkatenációjával előálló ismétléses füzérek környezetfüggő nyelve: $\{xx \mid x \in \{a, b\}*\}$

Automaták és transzducerek

Véges állapotú automata

Definíció (véges állapotú automata)

Az $M =< K, \Sigma, d, q_0, F >$ ötös egy véges állapotú automata, ahol
K: az automata állapotainak véges halmaza
Σ: az ábécé
q_0: a kezdőállapot
F: a végállapotok halmaza
d: az átmenetek halmaza.
Az M automata egy szituációja: $< x, q, y >$, ahol $q \in K$ (egy létező állapot) és $x, y \in \Sigma*$ (az ábécé fölötti füzérek).

Determinisztikus bégetőautomata

$K = \{q_0, q_1, q_2, q_3, q_4\}$
$\Sigma = \{b, e, !\}$
$F = \{q_4\}$
$d = \{< q_0, b, q_1 >, < q_1, e, q_2 >, < q_2, e, q_3 >, < q_3, e, q_3 >,$
$< q_3, !, q_4 >\}$

Nem-determinisztikus bégetőautomaták

$K = \{q_0, q_1, q_2, q_3, q_4\}$
$\Sigma = \{b, e, !\}$
$F = \{q_4\}$
$d = \{< q_0, b, q_1 >, < q_1, e, q_2 >, < q_2, e, q_2 >, < q_2, e, q_3 >,$
$< q_3, !, q_4 >\}$

VAGY

$K = \{q_0, q_1, q_2, q_3, q_4\}$
$\Sigma = \{b, e, !, \varepsilon\}$
$F = \{q_4\}$
$d = \{< q_0, b, q_1 >, < q_1, e, q_2 >, < q_2, e, q_3 >, < q_3, \varepsilon, q_2 >,$
$< q_3, !, q_4 >\}$

Állapot-átmenet tábla

	bemenet		
állapot	b	e	!
0	1	-	-
1	-	2	-
2	-	3	-
3	-	3	4
4:	-	-	-

Példaautomata

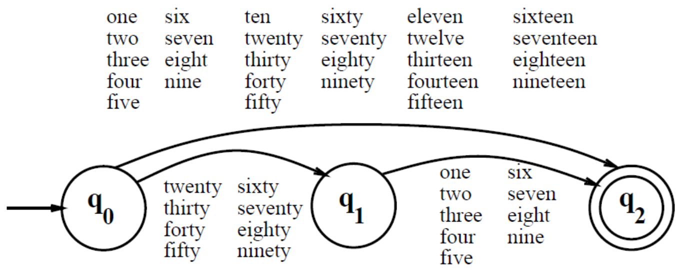


Figure 2.16 An FSA for the words for English numbers 1 – 99.

Levezetés

Ajánlott irodalom

- Alberti Gábor: *Matematika a természetes nyelvek leírásában*. Segédkönyvek a nyelvészet tanulmányozásához 52. Tinta Könyvkiadó, Budapest, 2006.
- Daniel Jurafsky & James H. Martin: *Speech and Language Processing*. 1st edition. Chapter 2 & 13. Prentice Hall, 2000.

Házi feladat