

3.4 Нормална форма на Чомски за безконтекстни граматика

Нека $\Gamma = \langle V, W, S, P \rangle$ е безконтекстна граматика и $L(\Gamma) \neq \emptyset$. Тогава за всяко $\omega \in L(\Gamma)$ може да се построи дърво на извод в Γ . Под дълбочина на това дърво ще разбираме дължината на най-дългия път по дървото от корена към листата.

Определение 3.4.1 *Една безконтекстна граматика е приведена във формата на Чомски, ако правилата ѝ са от вида $A \rightarrow BC$ или $A \rightarrow a$, като $A, B, C \in W, a \in V$.*

Теорема 3.4.1 *Всеки ε -свободен безконтекстен език може да бъде породен от безконтекстна граматика във формата на Чомски, т. е. граматика с правила от вида $A \rightarrow BC, A, B, C \in W$ или $A \rightarrow a, A \in W, a \in V$.*

Доказателство: Нека L е ε -свободен език, породен от безконтекстната граматика

$$\Gamma = \langle V, W, S, P \rangle.$$

Да построим еквивалентна граматика Γ' на Γ с

$$\Gamma' = \langle V, W', S, P' \rangle,$$

която е във формата на Чомски. Това можем да извършим в следната последователност.

Най-напред отделяме всички правила от вида $A \rightarrow B$, за които $A \in W$ и $B \in W$ в P и ги наричаме *първични*. За всеки нетерминален символ A да разгледаме множествата

$$U(A) = \{A \rightarrow B \in P \mid B \in W\}$$

$$\text{и} \quad N(A) = \{A \rightarrow \omega \in P \mid A \rightarrow \omega \notin U(A)\}.$$

Заменяме за всяко $A \in W$, множеството $U(A)$ с множеството от правила

$$\Delta(A) = \{A \rightarrow \alpha \mid A \stackrel{\Gamma}{\models} B \text{ и } B \rightarrow \alpha \in N(B)\}$$

По-нататък всички правила, които в дясната си страна имат поне по две букви, от които поне една е терминален символ, да наречем *вторични*.

На всеки терминален символ a , появяващ се в дясната страна на вторични правила, съпоставяме нов нетерминален символ A_a и добавяме ново правило в P' от вида $A_a \rightarrow a$.

В резултат от това съпоставяне можем да заменим всяко правило от вида $A \rightarrow \omega$, $\omega = X_1 X_2 \dots X_m \in (V \cup W)^*$ с правилото $A \rightarrow Y_1 Y_2 \dots Y_m$, където

$$Y_i = \begin{cases} X_i & \text{ако } X_i \in W \\ A_a & \text{ако } X_i = a \in V. \end{cases}$$

и новите правила $A_a \rightarrow a$, ако $a \in V$ е буква на ω . Да означим с W' множеството от символи в W и новите символи A_a .

Да наречем сега всички правила в P' *третични*, ако в дясната им страна има поне три нетерминални символи. Всяко третично правило има вида

$$A \rightarrow B_1 B_2 \dots B_m, \quad m > 2, \quad B_i \in W'.$$

Заменяме всички такива правила с правила от вида

$$A \rightarrow B_1 C_1, \quad C_1 \rightarrow B_2 C_2, \quad \dots, \quad C_{m-3} \rightarrow B_{m-2} C_{m-2}, \quad C_{m-2} \rightarrow B_{m-1} B_m$$

и добавяме в W' новите символи C_1, C_2, \dots, C_{m-2} .

Начина по който образувахме P' и W' , превръща проверката, че $L(\Gamma) = L(\Gamma')$ в лесно самостоятелно упражнение. \square

Пример 3.4.1 Нека Γ е граматика, породена от правилата

$$S \rightarrow A|ABA, \quad A \rightarrow aA|a|B, \quad B \rightarrow bB|b.$$

Тогава $U(S) = \{S \rightarrow A\}$, $U(A) = \{A \rightarrow B\}$, $U(B) = \emptyset$. Заменяме $U(S)$ с $S \rightarrow aA|a|bB|b$, а $U(A)$ заменяме с $A \rightarrow bB|b$. Така след първата стъпка имаме следните правила

$$S \rightarrow ABA|aA|a|bB|b, \quad A \rightarrow aA|a|bB|b, \quad B \rightarrow bB|b.$$

Добавяме правилата $A_a \rightarrow a$ и $A_b \rightarrow b$ заедно с правилата

$$S \rightarrow ABA|A_aA|a|A_bB|b, \quad A \rightarrow A_aA|a|A_bB|b, \quad B \rightarrow A_bB|b.$$

На третата стъпка заменяме правилото $S \rightarrow ABA$ с правилата $S \rightarrow AC_1$, $C_1 \rightarrow BA$ и така получаваме окончателно за P'

$$S \rightarrow AC_1|A_aA|a|A_bB|b, \quad A \rightarrow A_aA|a|A_bB|b, \quad B \rightarrow A_bB|b$$

$$A_a \rightarrow a, \quad A_b \rightarrow b, \quad C_1 \rightarrow BA$$

и

$$W' = \{S, A, B, C, A_a, A_b, C_1\}.$$

Лема 3.4.1 Ако Γ е безконтекстна граматика във формата на Чомски, то максималната дължина на дума в Γ , която може да се изведе от дърво с дълбочина p е 2^{p-1} . \square

Теорема 3.4.2 Съществува ефективен алгоритъм за установяване дали дадена безконтекстна граматика $\Gamma = \langle V, W, S, P \rangle$ поражда празен език, т.е. дали $L(\Gamma) = \emptyset$ или не.

Доказателство: Най-напред, ако в P има правило от вида $S \rightarrow \varepsilon$, то тогава очевидно $L(\Gamma) \neq \emptyset$.

В случая когато $S \rightarrow \varepsilon \notin P$, се построяват всички дървета на извод с дълбочина равна на $|W|$. Ако нито едно такова дърво не поражда дума от терминални символи (т.е. има поне един лист, който е нетерминален символ), то $L(\Gamma) = \emptyset$, а в противен случай $L(\Gamma) \neq \emptyset$. \square

Доказателството на лемата следва непосредствено от това, че Γ е във формата на Чомски.

Теорема 3.4.3 Ако L е безконтекстен език то съществуват естествени числа l_1 и l_2 така, че всяка дума $\omega \in L$, за която $|\omega| > l_1$, може да се представи във вида $\omega = uvzxy$ като:

- (i) $|vzx| \leq l_2$;
- (ii) $|vx| \geq 1$;

(iii) за всяко цяло i , $uv^i zx^i y \in L$.

Доказателство: Нека $L \setminus \{\varepsilon\}$ е език, породен от безконтекстната граматика $\Gamma = \langle V, W, S, P \rangle$, която е във формата на Чомски. Да положим $n = |W|$ и $l_1 = 2^{n-1}$, $l_2 = 2^n$ и нека $\omega \in L$ с $|\omega| > 2^{n-1}$.

Да означим с Π максималния, по дължина, път в едно дърво на извод T на ω . Тъй като Γ е във формата на Чомски и $|\omega| > l_1$ следва, че пътят Π има повече от $n + 1$ възли, от които точно един е терминален символ. Но тогава Π ще минава през два върха N_1 и N_2 , означени с един и същи нетерминален символ A . Да приемем, че N_1 е по-близо от N_2 до корена на дървото, а ако има повече такива двойки възли, избираме тази за която N_1 е най-далече от корена.

Понеже Π е максимален, то тази негова част, която е в дървото с корен N_1 ще е максимален подпът. Ето защо дълбочината на това поддърво T_1 на T ще е не по-голяма от $n + 1$.

Следователно по указаното поддърво от A може да се изведе дума в Γ с дължина не повече от 2^n . Нека например ω_1 е поддума, която се извежда от N_1 т.е. буквите и са листа на нашето поддърво и следователно $|\omega_1| \leq 2^n$ и $A \models \omega_1$. Нека T_2 е поддърво на T_1 и T с корен N_2 и z е дума от терминални символи, изведена от N_2 , чрез T_2 . Тогава може да напишем $\omega_1 = vzx$, като v и x не могат едновременно да са ε тъй като първото правило в извода $A \models \omega_1$ трябва да е от вида $A \rightarrow BC$, следователно $|vx| > 1$. Понеже ω_1 е поддума на ω следва, че $\omega = u\omega_1 y = uvzx y$. Следователно съществува изводът

$$S \models \omega = uvzx y, \text{ за който}$$

$$S \models uAy \models uvAxy \models \omega, \text{ където } |vzx| \leq l_2.$$

От последната схема на извод следва, че $A \models vAx \models vzx$. Следователно за всяко $i \geq 1$ ще имаме

$$A \models v^i A x^i \models v^i z x^i \quad \text{т.е.} \quad uv^i z x^i y \in L.$$

□

Теорема 3.4.3 може да се илюстрира с помоща на следния чертеж.

Фиг. 3.4.1

Теорема 3.4.4 *Съществуват контекстни езици, които не са безконтекстни.*

Доказателство: Да разгледаме езика $L = \{a^n b^n c^n | n \geq 1\}$.

В началото на тази глава се убедихме, че той е контекстен език (Виж $L(\Gamma_2)$ в Пример 3.1.1). Да се убедим, че той не е безконтекстен.

Тъй като в L има думи с произволно голяма дължина, ако допуснем, че L е безконтекстен, то от Теорема 3.4.3 следва, че съществуват числа l_1 и l_2 така, че всяка дума ω с $|\omega| \geq l_1$, се представя във вида $\omega = uvzxu$ и за всяко i , $uv^i zx^i u \in L$. Следователно

$$uv^i zx^i u = a^t b^t c^t$$

От това равенство следва, че буквите във v и x не могат да се различават по между си. От $|vx| \geq 1$ следва, че може да се избере i така, че горното равенство да се наруши. \square

Теорема 3.4.5 *Съществува алгоритъм за установяване на това дали даден безконтекстен език е краен или не.*

Доказателство: От Теорема 3.4.3 следва, че $L = L(\Gamma)$ е безкраен точно тогава, когато съществува дума $\omega \in L$, за която $|\omega| > l_1 = 2^{n-1}$, където $n = |W|$. Нека $l_2 = 2^n$ и ω_0 е най-кратката дума в L , за която $|\omega_0| > l_1$. Ще докажем, че $|\omega_0| \leq l_1 + l_2$. Да допуснем, че $|\omega_0| > l_1 + l_2$ и следователно L няма думи с дължина между l_1 и l_2 . От Теорема 3.4.3 следва, че $\omega_0 = uvzxy$ и $uzy \in L$. Но w, x, u и z могат да се изберат така, че дължината на vzx да не е по-голяма от l_2 . (Защо?) Следователно

$$|uzy| > |\omega_0| - l_2 > l_1,$$

което противоречи на избора на ω_0 . От тук следва, че $l_1 < |\omega_0| \leq l_1 + l_2$.

Сега алгоритъмът за решаване на въпроса за това дали L е краен, се състои в пораждаването на всички думи в Γ с дължина между l_1 и $l_1 + l_2$ и в зависимост от това дали има такива думи или не се дава отговор и на въпроса дали L е краен или не. \square

З а д а ч и

1. Да се приведе във формата на Чомски граматиката, която поражда езика:

а) $L_1 = \{a^n b^n c^m \mid n, m \geq 1\}$;

б) $L_2 = \{a^n b^m c^m \mid n, m \geq 1\}$.

2. Да се приведе във форма на Чомски, граматиката с правила:

$$S \rightarrow A|ABC, \quad A \rightarrow aA|a|C, \quad B \rightarrow bAb|ab, \quad C \rightarrow abcC|c.$$

3. Да се докаже, че езикът L не е безконтекстен:

а) $L = \{a^n b^m c^p \mid n \geq m \geq p \geq 1\}$;

б) $L = \{ww'w \mid w \in \{a, b\}^*, w' \text{ е } w, \text{ написана в обратен ред.}\}$

4. Да се докаже, че ако L_1 и L_2 са безконтекстни езици, то безконтекстни са и езиците $L_1 L_2$ и $L_1 \cup L_2$.

5. Нека L е безконтекстен език и $\varphi : V^* \rightarrow V^*$ е хомоморфизъм. Да се докаже, че $\varphi(L)$ е безконтекстен език.