Cvičenie č. 3
Bezkontextové gramatiky (1. časť)

Peter Kostolányi

5. októbra 2022

1 Bezkontextové a regulárne gramatiky

Bezkontextové gramatiky sú frázové gramatiky, v ktorých ľavá strana každého prepisovacieho pravidla pozostáva z jediného neterminálu. Definície najdôležitejších pojmov súvisiacich s frázovými gramatikami boli zavedené na prednáške. V rámci ich rekapitulácie sa obmedzíme výhradne na špeciálny prípad bezkontextových gramatík.

Definícia 1. *Bezkontextová gramatika* je štvorica $G = (N, T, P, \sigma)$, kde N je abeceda[1] neterminálnych symbolov, T je abeceda terminálnych symbolov, $N \cap T = \emptyset$, $P \subseteq N \times (N \cup T)^*$ je *konečná* množina prepisovacích pravidiel a $\sigma \in N$ je počiatočný neterminál.

Prepisovacie pravidlá $(\xi, x) \in P$ obyčajne zapisujeme ako $\xi \to x$. Ide iba o špeciálny spôsob zápisu usporiadanej dvojice používaný v tomto kontexte a symbol $\to$ sám o sebe nemá definovaný žiaden význam. Zápis $\xi \to x_1, \xi \to x_2, \ldots, \xi \to x_k$ skracujeme aj ako $\xi \to x_1 \mid x_2 \mid \ldots \mid x_k$.

Definícia 2. Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika. *Krok odvodenia* v gramatike G je binárna relácia $\Rightarrow_G$ na $(N \cup T)^*$ taká, že

$$\begin{aligned}&\forall u, v \in (N \cup T)^* : u \Rightarrow_G v \text{ práve vtedy, keď} \\ &\qquad \exists u_1, u_2, x \in (N \cup T)^* \ \exists \xi \in N : u = u_1 \xi u_2 \ \wedge \ v = u_1 x u_2 \ \wedge \ \xi \to x \in P.\end{aligned}$$

V prípade, že je gramatika G zrejmá z kontextu, píšeme namiesto $\Rightarrow_G$ iba $\Rightarrow$.

Keďže je krok odvodenia $\Rightarrow$ definovaný ako binárna relácia na $(N \cup T)^*$, dajú sa naň aplikovať všetky bežné operácie na takýchto reláciách. Pre $k \in \mathbb{N}$ tak napríklad môžeme hovoriť o k-tej mocnine relácie $\Rightarrow$, označovanej symbolom $\Rightarrow^k$. V relácii $\Rightarrow^k$ sú potom všetky dvojice slov u, v také, že slovo v sa dá odvodiť zo slova u na práve k krokov odvodenia. Podobne možno definovať reflexívno-tranzitívny uzáver $\Rightarrow^*$ a tranzitívny uzáver $\Rightarrow^+$ relácie $\Rightarrow$. V týchto reláciách sú všetky dvojice slov u, v také, že slovo v sa dá odvodiť zo slova u na ľubovoľný resp. ľubovoľný nenulový počet krokov. Relácia $\Rightarrow^0$ je identita, podobne ako nultá mocnina ľubovoľnej inej relácie.

Zneužívajúc notáciu často hovoríme o „odvodení $\sigma \Rightarrow w_1 \Rightarrow \ldots \Rightarrow w_n$“, o „odvodení $\sigma \Rightarrow^n w_n$ (dĺžky n)“, alebo o „odvodení $\sigma \Rightarrow^* w_n$“, aj keď každý z týchto zápisov vyjadruje iba reláciu medzi slovami a ťažko tak hovoriť o „entite odvodenia“. Žiaden z uvedených zápisov navyše neposkytuje kompletnú informáciu, ktorú od konceptu odvodenia intuitívne očakávame. Problém formálneho uchopenia pojmu odvodenia sa však ukazuje ako pomerne delikátny a v literatúre sa často možno stretnúť s chybnými definíciami neskôr kompenzovanými intuíciou. My sa k týmto záležitostiam vrátime na nasledujúcom cvičení, kde sa pokúsime aj o formálne presnú definíciu. Väčšinou sa ale budeme pridŕžať zaužívaných zvyklostí z úvodu tohto odstavca.

Definícia 3. Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika. *Jazyk* generovaný gramatikou G je daný ako $L(G) = \{w \in T^* \mid \sigma \Rightarrow^* w\}$.

O slovách z jazyka $L(G)$ tiež hovoríme, že sú *generované* gramatikou G. Z definície vyplýva, že takéto slová pozostávajú výhradne z terminálnych symbolov. Slovo nad abecedou $N \cup T$, pre ktoré existuje v gramatike G odvodenie, nazývame *vetnou formou* v gramatike G. Pre jazyk všetkých vetných foriem v gramatike G budeme *pre účely týchto cvičení* používať označenie $F(G)$.

[1] Čiže neprázdna konečná množina.

Definícia 4. Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika. *Vetná forma* v gramatike G je slovo $x \in (N \cup T)^*$ také, že $\sigma \Rightarrow^* x$. Pre jazyk všetkých vetných foriem v gramatike G píšeme $F(G) = \{x \in (N \cup T)^* \mid \sigma \Rightarrow^* x\}$.

Jazyk nazveme *bezkontextovým*, ak existuje bezkontextová gramatika, ktorá ho generuje.

Definícia 5. *Bezkontextový jazyk* je jazyk L taký, že $L = L(G)$ pre nejakú bezkontextovú gramatiku G. *Triedu* všetkých *bezkontextových jazykov* označujeme $\mathscr{L}_{CF}$.

Skutočnosť, že hovoríme o *triede* bezkontextových jazykov a nie o *množine* bezkontextových jazykov, nie je náhodná. Pokiaľ sa totiž neobmedzíme na nejakú konkrétnu množinu symbolov, ktoré môžu byť prvkami abecied, množina všetkých bezkontextových jazykov neexistuje. Trieda sa od množiny líši predovšetkým tým, že sama nie je množinou, a teda existuje napríklad aj trieda všetkých množín a podobne.

Regulárne gramatiky sú bezkontextové gramatiky, v ktorých sa môže každý neterminál prepísať iba na niekoľko (aj nula) terminálov nasledovaných najviac jedným neterminálom.

Definícia 6. *Regulárna gramatika* je bezkontextová gramatika $G = (N, T, P, \sigma)$, pre ktorú platí $P \subseteq N \times (T^* \cup T^* N)$.[2]

Keďže sme regulárne gramatiky definovali ako špeciálne bezkontextové gramatiky, vzťahuje sa väčšina definícií a označení zavedených vyššie aj na regulárne gramatiky. V nasledujúcom sa budeme zaoberať metódami dokazovania, že daná gramatika skutočne generuje zamýšľaný jazyk; aj keď budeme hovoriť o bezkontextových gramatikách vo všeobecnosti, rovnaké metódy bude možné použiť aj pre regulárne gramatiky, ktoré sú ich špeciálnym prípadom.

Poznámka 1. *Lineárna gramatika* je bezkontextová gramatika $G = (N, T, P, \sigma)$ taká, že množina pravidiel P je konečnou podmnožinou množiny $N \times (T^* \cup T^* N T^*)$ – na pravej strane každého pravidla sa teda môže vyskytovať najviac jeden neterminál. Regulárne gramatiky sú očividne špeciálnym prípadom lineárnych gramatík a niekedy sa preto nazývajú aj *sprava lineárne gramatiky.*

Poznámka 2. Regulárne gramatiky sú ďalším z radu objektov, ktorých definícia nie je úplne ustálená. Kým niektorí autori napríklad týmto pojmom označujú sprava lineárne gramatiky, ktorých pravidlá nemôžu na pravej strane obsahovať viac ako jeden terminál, inde sa zas pod regulárnou gramatikou rozumie bezkontextová gramatika, ktorá je lineárna sprava *alebo* zľava – množina prepisovacích pravidiel *zľava lineárnej gramatiky* $G = (N, T, P, \sigma)$ je konečnou podmnožinou množiny $N \times (T^* \cup N T^*)$. Neskôr pochopíme, že tieto drobné variácie v skutočnosti nie sú nijak podstatné.

2 Dokazovanie správnosti konštrukcie gramatiky

V úlohách nasledujúcich za týmto oddielom je väčšinou cieľom *zostrojiť* bezkontextovú gramatiku G generujúcu daný jazyk L a *dokázať správnosť* konštrukcie, čo možno vyjadriť rovnosťou $L(G) = L$.

Po nájdení samotnej gramatiky, ktoré zvyčajne býva jedinou skutočne tvorivou časťou celej úlohy, je teda ešte potrebné dokázať rovnosť jazyka $L(G)$ generovaného zostrojenou gramatikou G a jazyka L zo zadania úlohy.

Dokázať rovnosť $L(G) = L$ znamená overiť platnosť obidvoch inklúzií $L(G) \subseteq L$ a $L(G) \supseteq L$.

Pri dôkaze inklúzie $L(G) \subseteq L$ pre bezkontextovú gramatiku $G = (N, T, P, \sigma)$ uvažujeme ľubovoľné slovo $w \in L(G)$ a ukazujeme, že toto slovo patrí do jazyka L. Predpoklad $w \in L(G)$ znamená, že $w \in T^*$ a súčasne $\sigma \Rightarrow^* w$, v dôsledku čoho $\sigma \Rightarrow^n w$ pre nejaké $n \in \mathbb{N}$. O slove w v tomto momente nevieme nič ďalšie a potrebujeme ukázať, že $w \in L$. Ako dôkazová metóda sa teda ponúka matematická indukcia vzhľadom na dĺžku n uvažovaného odvodenia slova w.

[2] Keďže má ísť o špeciálny prípad bezkontextových gramatík, musí byť množina pravidiel P konečná.

V rámci indukčného kroku takéhoto dôkazu predpokladáme platnosť tvrdenia pre $n = k$ a dokazujeme jeho platnosť pre $n = k + 1$. Ak teda $\sigma \Rightarrow^{k+1} w$, môžeme sa pokúsiť prepísať toto odvodenie napríklad ako

$$\sigma \Rightarrow^k y \Rightarrow w$$

a aplikovať indukčný predpoklad na odvodenie $\sigma \Rightarrow^k y$. Tu však narážame: indukčný predpoklad nám o slove y nehovorí nič, pretože slovo y nepatrí do jazyka $L(G)$; ide iba o *vetnú formu* v gramatike G, ktorá nutne obsahuje aspoň jeden neterminál. Oplatí sa teda dokazované tvrdenie *zosilniť* tak, aby sme do úvah zahrnuli všetky vetné formy. Napríklad môžeme nájsť jazyk F všetkých vetných foriem v gramatike G a namiesto inklúzie $L(G) \subseteq L$ dokazovať inklúziu $F(G) \subseteq F$.

Pri dôkaze inklúzie $L(G) \supseteq L$ naopak uvažujeme ľubovoľné slovo $w \in L$ a ukazujeme, že pre toto slovo v gramatike G existuje odvodenie. Obvykle tak robíme indukciou vzhľadom na nejakú štrukturálnu vlastnosť slova w, ktorú volíme na základe jazyka L a skonštruovanej gramatiky G pre tento jazyk. Ak napríklad $L = \{a, b\}^*$, je prirodzené tvrdenie dokazovať indukciou na dĺžku slova; pri jazyku $L = \{a^n b^n \mid n \in \mathbb{N}\}$ sa ponúka dôkaz indukciou vzhľadom na n a napríklad pri jazyku $L = \{abu \mid u \in \{a, b\}^*\}$ je prirodzeným prístupom indukcia vzhľadom na dĺžku slova u. Často je potrebných aj viacero indukcií, ako čoskoro uvidíme na príklade.

Nech ale túto štrukturálnu vlastnosť zvolíme akokoľvek, je často žiadúce vedieť na odvoditeľnosť slova v gramatike usúdiť z indukčného predpokladu o odvoditeľnosti nejakého iného slova. Preto môže byť aj tu užitočné dokazované tvrdenie *zosilniť*. Opäť môžeme napríklad identifikovať jazyk F všetkých vetných foriem v G a namiesto inklúzie $L(G) \supseteq L$ dokázať inklúziu $F(G) \supseteq F$.

Rovnosť $L(G) = L$ pre bezkontextovú gramatiku $G = (N, T, P, \sigma)$ tak typicky môžeme dokázať nasledovne.

1. Identifikujeme jazyk F všetkých vetných foriem v gramatike G; preň by, samozrejme, vždy malo platiť $F \cap T^* = L$.

2. Namiesto inklúzie $L(G) \subseteq L$ dokážeme silnejšie tvrdenie $F(G) \subseteq F$ – čiže „ak $x \in (N \cup T)^*$ a $\sigma \Rightarrow^* x$, tak $x \in F$“. Obyčajne tak robíme indukciou vzhľadom na dĺžku uvažovaného odvodenia slova x.

3. Namiesto inklúzie $L(G) \supseteq L$ dokážeme silnejšie tvrdenie $F(G) \supseteq F$ – čiže „ak $x \in F$, tak $\sigma \Rightarrow^* x$“. Obyčajne tak robíme indukciou vzhľadom na určité štrukturálne vlastnosti slova x.

Uvedený postup je pri konštrukcii gramatík pre „učebnicové“ bezkontextové jazyky použiteľný skoro univerzálne, ale občas môže byť zbytočne prácny. V rámci nasledujúcich riešených úloh si preto okrem príkladov použitia tohto postupu ukážeme aj ďalšie možnosti dokazovania správnosti konštrukcie bezkontextových gramatík.

3 Riešené úlohy

Pri riešení prvých dvoch spomedzi nasledujúcich úloh sa budeme verne pridŕžať postupu opísaného v predchádzajúcom oddiele.

Úloha 1. Zostrojte bezkontextovú gramatiku generujúcu jazyk $L = \{w \in \{a, b\}^* \mid w = w^R\}$ a dokážte správnosť svojej konštrukcie.

Riešenie. Ukážeme, že $L = L(G)$ pre gramatiku $G = (N, T, P, \sigma)$, kde $N = \{\sigma\}$, $T = \{a, b\}$ a

$$P = \{\sigma \to a\sigma a \mid b\sigma b \mid a \mid b \mid \varepsilon\}.$$

Dokážeme, že $F(G) = F$ pre jazyk $F = \{usu^R \mid u \in T^*;\ s \in \{\varepsilon, a, b, \sigma\}\}$. Rovnosť $L(G) = L$ z tohto vyplynie ako bezprostredný dôsledok, keďže evidentne $F \cap T^* = L$.

$\subseteq$: Uvažujme ľubovoľné slovo $x \in F(G)$. To je vetnou formou gramatiky G – existuje teda $n \in \mathbb{N}$ také, že $\sigma \Rightarrow^n x$. Indukciou vzhľadom na n dokážeme, že $x \in F$.

Pre $n = 0$ musí byť $x = \sigma \in F$.[3]

Predpokladajme teda platnosť tvrdenia pre $n = k$ a dokážme ho pre $n = k + 1$. Ak $\sigma \Rightarrow^{k+1} x$, pre nejaké slovo y nutne $\sigma \Rightarrow^{k} y \Rightarrow x$. Na odvodenie $\sigma \Rightarrow^{k} y$ sa vzťahuje indukčný predpoklad, a teda $y \in F$. To znamená, že $y = vtv^R$ pre nejaké $v \in T^*$ a $t \in \{\varepsilon, a, b, \sigma\}$.

Keďže $y \Rightarrow x$, musí slovo y obsahovať aspoň jeden neterminál, z čoho $t = \sigma$ – čiže $y = v\sigma v^R$ pre nejaké $v \in T^*$. Slovo x musí zo slova y vzniknúť prepísaním neterminálu σ podľa niektorého z pravidiel v množine P. V prípade, že ide o niektoré z pravidiel $\sigma \to a\sigma a$ alebo $\sigma \to b\sigma b$, dostávame $x = va\sigma av^R = (va)\sigma(va)^R$ resp. $x = vb\sigma bv^R = (vb)\sigma(vb)^R$; v oboch prípadoch je $x \in F$.[4] Pre pravidlá $\sigma \to a$, $\sigma \to b$ a $\sigma \to \varepsilon$ postupne dostávame $x = vav^R$, $x = vbv^R$ a $x = vv^R$; vo všetkých prípadoch opäť $x \in F$.[5]

$\supseteq$: Nech $x \in F$ – čiže $x = usu^R$ pre nejaké $u \in T^*$ a $s \in \{\varepsilon, a, b, \sigma\}$. Indukciou vzhľadom na $|u|$ dokážeme, že $u \in F(G)$ – čiže $\sigma \Rightarrow^* u$.

Pre $|u| = 0$ je $x \in \{\varepsilon, a, b, \sigma\}$. Avšak $\sigma \Rightarrow \varepsilon$, $\sigma \Rightarrow a$, $\sigma \Rightarrow b$ a $\sigma \Rightarrow^0 \sigma$ – teda iste $\sigma \Rightarrow^* x$.

Nech tvrdenie platí pre $|u| = k$; dokážme ho pre $|u| = k+1$. Nech $x = usu^R$ pre nejaké $u \in T^{k+1}$ a $s \in \{\varepsilon, a, b, \sigma\}$. Položme $u = vc$ pre $v \in T^k$ a $c \in T$ – čiže $x = vcscv^R$.

Z indukčného predpokladu je $\sigma \Rightarrow^* v\sigma v^R$. Použitím prepisovacieho pravidla $\sigma \to c\sigma c \in P$ dostávame $\sigma \Rightarrow^* v\sigma v^R \Rightarrow vc\sigma cv^R$. To je pre $s = \sigma$ už postačujúce. Pre zvyšné s obsahuje množina P pravidlo $\sigma \to s$, použitím ktorého dostávame $\sigma \Rightarrow^* v\sigma v^R \Rightarrow vc\sigma cv^R \Rightarrow vcscv^R$. $\square$

Úloha 2. Zostrojte bezkontextovú gramatiku generujúcu jazyk $L = \{a^i b^j c^j d^i \mid i, j \in \mathbb{N}\}$ a dokážte správnosť svojej konštrukcie.

Riešenie. Dokážeme, že jazyk L je generovaný bezkontextovou gramatikou $G = (N, T, P, \sigma)$, kde $N = \{\sigma, \alpha\}$, $T = \{a, b, c, d\}$ a

$$P = \{\sigma \to a\sigma d \mid \alpha \\ \alpha \to b\alpha c \mid \varepsilon\}.$$

Namiesto rovnosti $L(G) = L$ dokážeme silnejšie tvrdenie $F(G) = F$, kde F je jazyk obsahujúci práve

(*i*) všetky slová $a^i\sigma d^i$, kde $i \in \mathbb{N}$;

(*ii*) a všetky slová $a^i b^j s c^j d^i$, kde $i, j \in \mathbb{N}$ a $s \in \{\alpha, \varepsilon\}$.

Zrejme pritom $F \cap T^* = L$, takže z rovnosti $F(G) = F$ naozaj vyplynie aj $L(G) = L$.

$\subseteq$: Nech $x \in F(G)$, čiže $\sigma \Rightarrow^n x$ pre nejaké $n \in \mathbb{N}$. Indukciou vzhľadom na n dokážeme, že $x \in F$.

Pre $n = 0$ je $x = \sigma$, čo je slovo tvaru (*i*). Nech teraz tvrdenie platí pre $n = k$; uvažujme $n = k + 1$. Odvodenie $\sigma \Rightarrow^{k+1} x$ vieme v takom prípade prepísať ako $\sigma \Rightarrow^k y \Rightarrow x$, kde na odvodenie $\sigma \Rightarrow^k y$ sa vzťahuje indukčný predpoklad. To znamená, že slovo y je jedného z tvarov (*i*) alebo (*ii*).

Ak $y = a^i\sigma d^i$ pre nejaké $i \in \mathbb{N}$, musí slovo x vzniknúť zo slova y použitím niektorého z pravidiel $\sigma \to a\sigma d$ alebo $\sigma \to \alpha$ na jediný výskyt neterminálu σ v slove y. V prvom prípade dostávame $x = a^{i+1}\sigma d^{i+1}$, čo je slovo tvaru (*i*). V zostávajúcom prípade zas $x = a^i\alpha d^i$, čo je slovo tvaru (*ii*). V oboch prípadoch je teda $x \in F$.

Ak $y = a^i b^j s c^j d^i$ pre nejaké $i, j \in \mathbb{N}$ a $s \in \{\alpha, \varepsilon\}$, musí slovo y vďaka vzťahu $y \Rightarrow x$ obsahovať aspoň jeden neterminál – z čoho dostávame $s = \alpha$ a $y = a^i b^j \alpha c^j d^i$ pre nejaké $i, j \in \mathbb{N}$. Slovo x teraz musí vzniknúť zo slova y použitím niektorého z pravidiel $\alpha \to b\alpha c$ alebo $\alpha \to \varepsilon$ na neterminál α. V prvom prípade dostávame $x = a^i b^{j+1}\alpha c^{j+1} d^i$ a v druhom $x = a^i b^j c^j d^i$. V oboch prípadoch ide o slovo tvaru (*ii*), a teda $x \in F$.

[3] V množinovom zápise jazyka F stačí vziať $u = \varepsilon$ a $s = \sigma$.
[4] Stačí totiž vziať $s = \sigma$ a $u = va$ resp. $u = vb$.
[5] Stačí vziať $u = v$ a $s = a$, $s = b$, resp. $s = \varepsilon$.

$\supseteq$: Dokážeme, že všetky slová v jazyku F – čiže všetky slová tvaru (i) alebo (ii) sú vetnými formami v gramatike G.

Urobme tak najprv pre slová tvaru (i) a indukciou vzhľadom na i dokážme, že pre všetky $i \in \mathbb{N}$ je $\sigma \Rightarrow^* a^i \sigma d^i$. Pre $i = 0$ je skutočne $\sigma \Rightarrow^0 \sigma$. Nech teraz tvrdenie platí pre $i = k$ a uvažujme $i = k+1$. Z indukčného predpokladu potom $\sigma \Rightarrow^* a^k \sigma d^k \Rightarrow a^{k+1} \sigma d^{k+1}$, kde v poslednom kroku bolo použité pravidlo $\sigma \to a\sigma d$.

Pre slová tvaru (ii) teraz najprv indukciou vzhľadom na j dokážme, že pre všetky $i, j \in \mathbb{N}$ je $\sigma \Rightarrow^* a^i b^j \alpha c^j d^i$. Pre $j = 0$ z tvrdenia dokázaného v predchádzajúcom odstavci pre všetky $i \in \mathbb{N}$ dostávame $\sigma \Rightarrow^* a^i \sigma d^i \Rightarrow a^i \alpha d^i = a^i b^0 \alpha c^0 d^i$, kde v poslednom kroku bolo použité pravidlo $\sigma \to \alpha$. Nech ďalej tvrdenie platí pre $j = \ell$ a uvažujme $j = \ell + 1$. Z indukčného predpokladu potom pre všetky $i \in \mathbb{N}$ vyplýva $\sigma \Rightarrow^* a^i b^\ell \alpha c^\ell d^i \Rightarrow a^i b^{\ell+1} \alpha c^{\ell+1} d^i$, kde v poslednom kroku odvodenia bolo použité pravidlo $\alpha \to b\alpha c$.

Zostáva dokázať, že pre všetky $i, j \in \mathbb{N}$ je $\sigma \Rightarrow^* a^i b^j c^j d^i$. Z dokázaného ale $\sigma \Rightarrow^* a^i b^j \alpha c^j d^i$ a použitím pravidla $\alpha \to \varepsilon$ dostávame $\sigma \Rightarrow^* a^i b^j \alpha c^j d^i \Rightarrow a^i b^j c^j d^i$. $\square$

Dokazovanie správnosti konštrukcie bezkontextovej gramatiky G možno často zjednodušiť tým, že namiesto tvrdenia $F(G) = F$ dokážeme predsa len o niečo slabšie tvrdenie – ktoré je však, samozrejme, stále dostatočne silné na to, aby z neho vyplynula rovnosť $L(G) = L$.

Inklúzia $L(G) \subseteq L$ totiž vyplynie z príslušnosti všetkých vetných foriem v G *do nejakého* jazyka $X \subseteq (N \cup T)^*$ takého, že $X \cap T^* = L$. Podobne inklúziu $L(G) \supseteq L$ získame dôkazom odvoditeľnosti všetkých slov *z nejakého* jazyka $X \subseteq (N \cup T)^*$ takého, že $X \cap T^* = L$.

Nie je teda vždy nutné charakterizovať všetky vetné formy v gramatike G. Na dôkaz inklúzie $L(G) \subseteq L$ stačí o vetných formách v gramatike G dokázať, že patria do nejakého jazyka, ktorého terminálne slová patria do jazyka L. Pokojne pritom môže ísť aj o nadjazyk jazyka $F(G)$. Pri dôkaze inklúzie $L(G) \supseteq L$ sa naopak stačí presvedčiť o tom, že sú v gramatike G odvoditeľné všetky slová z nejakého nadjazyka jazyka L, ktorý zvolíme tak, aby sa dobre robil dôkaz indukciou. Tento nadjazyk jazyka L ale nemusí nutne obsahovať všetky vetné formy v gramatike G.

Pri riešení nasledujúcej úlohy aplikujeme obidve tieto pozorovania. Úlohu je však stále možné – v princípe celkom jednoducho – vyriešiť aj pomocou postupu z oddielu 2.

Úloha 3. Zostrojte bezkontextovú gramatiku generujúcu jazyk $L = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$ a dokážte správnosť svojej konštrukcie.

Riešenie. Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika taká, že $N = \{\sigma\}$, $T = \{a, b\}$ a

$$P = \{\sigma \to \sigma a \sigma b \sigma \mid \sigma b \sigma a \sigma \mid \varepsilon\}.$$

Dokážeme, že $L(G) = L$.

$\subseteq$: Namiesto samotnej inklúzie dokážeme nasledujúce silnejšie tvrdenie: pre všetky vetné formy x v gramatike G je $\#_a(x) = \#_b(x)$. Keďže každé $w \in L(G)$ je zároveň aj vetnou formou v G, vyplynie z uvedeného tvrdenia aj dokazovaná inklúzia $L(G) \subseteq L$.[6]

Uvažujme preto ľubovoľnú vetnú formu x v gramatike G. Iste existuje $n \in \mathbb{N}$ také, že $\sigma \Rightarrow^n x$. Indukciou vzhľadom na n dokážeme, že $\#_a(x) = \#_b(x)$.

Tvrdenie platí pre $n = 0$ – v takom prípade totiž $x = \sigma$. Predpokladajme teda, že tvrdenie platí pre $n = k$ a uvažujme $n = k + 1$. Ak $\sigma \Rightarrow^{k+1} x$, tak pre nejaké $y \in (N \cup T)^*$ je $\sigma \Rightarrow^k y \Rightarrow x$. Na odvodenie $\sigma \Rightarrow^k y$ sa vzťahuje indukčný predpoklad, takže $\#_a(y) = \#_b(y)$. Slovo x vznikne zo slova y použitím niektorého z pravidiel v P. V prípade použitia pravidla $\sigma \to \sigma a \sigma b \sigma$ alebo $\sigma \to \sigma b \sigma a \sigma$ je zjavne $\#_a(x) = \#_a(y) + 1 = \#_b(y) + 1 = \#_b(x)$; v prípade použitia pravidla $\sigma \to \varepsilon$ je $\#_a(x) = \#_a(y) = \#_b(y) = \#_b(x)$.

[6]Nejde ale o dôkaz inklúzie $F(G) \subseteq F$ z oddielu 2, pretože nie každé slovo $x \in (N \cup T)^*$ spĺňajúce $\#_a(x) = \#_b(x)$ je vetnou formou v G. To je ale pre účely tejto inklúzie vedľajšie.

$\supseteq$: Indukciou vzhľadom na n najprv dokážme, že pre všetky $n \in \mathbb{N}$ sú v gramatike G odvoditeľné všetky slová $\sigma a_1 \sigma a_2 \sigma \ldots \sigma a_{2n} \sigma$ také, že $a_1, \ldots, a_{2n} \in \{a, b\}$ a $\#_a(a_1 \ldots a_{2n}) = \#_b(a_1 \ldots a_{2n})$. Pre $n = 0$ je jediným takýmto slovom slovo σ, pričom $\sigma \Rightarrow^0 \sigma$. Predpokladajme teda, že tvrdenie platí pre $n = k$ a uvažujme $n = k + 1$. Nech $a_1, \ldots, a_{2k+2} \in \{a, b\}$ sú také, že $\#_a(a_1 \ldots a_{2k+2}) = \#_b(a_1 \ldots a_{2k+2})$. Potom nutne existuje[7] $i \in [2k + 1]$ také $a_i = a$ a $a_{i+1} = b$ alebo $a_i = b$ a $a_{i+1} = a$. Z indukčného predpokladu tak dostávame

$$\sigma \Rightarrow^* \sigma a_1 \sigma a_2 \sigma \ldots \sigma a_{i-1} \underline{\sigma} a_{i+2} \sigma \ldots \sigma a_{2k+2} \sigma \Rightarrow a_1 \sigma a_2 \sigma \ldots \sigma a_{i-1} \sigma a_i \sigma a_{i+1} \sigma a_{i+2} \sigma \ldots \sigma a_{2k+2},$$

kde v poslednom kroku odvodenia aplikujeme niektoré z pravidiel $\sigma \to \sigma a \sigma b \sigma$ alebo $\sigma \to \sigma b \sigma a \sigma$ na podčiarknutý výskyt neterminálu σ.

Indukciou vzhľadom na k teraz môžeme dokázať, že sú v gramatike G pre všetky $n \in \mathbb{N}$ a všetky $a_1, \ldots, a_{2n} \in \{a, b\}$ také, že $\#_a(a_1 \ldots a_{2n}) = \#_b(a_1 \ldots a_{2n})$ odvoditeľné aj slová $a_1 \ldots a_k \sigma(a_{k+1}\sigma)(a_{k+2}\sigma) \ldots (a_{2n}\sigma)$ pre $k = 0, \ldots, 2n$. Pre $k = 0$ totiž z vyššie dokázaného vyplýva $\sigma \Rightarrow^* \sigma(a_1\sigma) \ldots (a_{2n}\sigma)$ a z platnosti dokazovaného tvrdenia pre $k = s < 2n$ dostávame aj $\sigma \Rightarrow^* a_1 \ldots a_s \sigma(a_{s+1}\sigma)(a_{s+2}\sigma) \ldots (a_{2n}\sigma) \Rightarrow a_1 \ldots a_s a_{s+1} \sigma(a_{s+2}\sigma) \ldots (a_{2n}\sigma)$, kde v poslednom kroku odvodenia sme použili pravidlo $\sigma \to \varepsilon$ na najľavejší výskyt neterminálu σ; to dokazuje platnosť tvrdenia pre $k = s + 1$.

Zisťujeme teraz, že v gramatike G sú odvoditeľné aj všetky slová z jazyka L – pre $w \in L$ totiž $w = a_1 \ldots a_{2n}$ pre nejaké $n \in \mathbb{N}$ a $a_1, \ldots, a_{2n} \in \{a, b\}$ také, že $\#_a(a_1 \ldots a_{2n}) = \#_b(a_1 \ldots a_{2n})$. Z dokázaného pritom dostávame

$$\sigma \Rightarrow^* a_1 a_2 \ldots a_{2n} \sigma \Rightarrow a_1 a_2 \ldots a_{2n} = w,$$

kde v poslednom kroku aplikujeme na jediný výskyt neterminálu σ pravidlo $\sigma \to \varepsilon$. □

Pre niektoré gramatiky G môže byť jazyk vetných foriem $F(G)$ pomerne komplikovaný a dôkaz tvrdenia $L(G) = L$ môže byť prirodzenejší s využitím „rekurzívnej štruktúry" jazyka L. Kým doteraz zohrával pri induktívnych dôkazoch oboch inklúzií $L(G) \subseteq L$ a $L(G) \supseteq L$ kľúčovú úlohu posledný krok odvodenia, môže byť niekedy výhodnejšie zamerať sa na prvý krok odvodenia a aplikovať nasledujúce – intuitívne pomerne zrejmé – tvrdenie, ktorého dôkaz je prenechaný ako jedna z úloh na najbližšie cvičenia.

Tvrdenie 1. *Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika, $\xi_1, \ldots, \xi_k \in N$, $u_0, \ldots, u_k \in T^*$ a $n \in \mathbb{N}$. Pre slovo $w \in (N \cup T)^*$ je potom $u_0 \xi_1 u_1 \xi_2 u_2 \ldots u_{k-1} \xi_k u_k \Rightarrow^n w$ práve vtedy, keď existujú $x_1, \ldots, x_k \in (N \cup T)^*$ a $n_1, \ldots, n_k \in \mathbb{N}$ také, že $w = u_0 x_1 u_1 x_2 u_2 \ldots u_{k-1} x_k u_k$, $n_1 + \ldots + n_k = n$ a pre $i = 1, \ldots, k$ je $\xi_i \Rightarrow^{n_i} x_i$.*

V dôsledku toho $u_0 \xi_1 u_1 \xi_2 u_2 \ldots u_{k-1} \xi_k u_k \Rightarrow^* w$ práve vtedy, keď existujú $x_1, \ldots, x_k \in (N \cup T)^*$ také, že $w = u_0 x_1 u_1 x_2 u_2 \ldots u_{k-1} x_k u_k$ a pre $i = 1, \ldots, k$ je $\xi_i \Rightarrow^* x_i$.

Opísaný prístup teraz v rámci riešenia nasledujúcej úlohy demonštrujeme na ďalšej gramatike G generujúcej jazyk $L = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$. Inklúziu $L(G) \subseteq L$ by sme aj pre túto gramatiku mohli dokázať prakticky rovnako ako v úlohe 3; výhodnosť alternatívneho prístupu sa však naplno ukáže pri dôkaze opačnej inklúzie.

Úloha 4. Nech $G = (N, T, P, \sigma)$ je bezkontextová gramatika taká, že $N = \{\sigma\}$, $T = \{a, b\}$ a

$$P = \{\sigma \to a\sigma b\sigma \mid b\sigma a\sigma \mid \varepsilon\}.$$

Dokážte, že $L(G) = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$.

Riešenie. Položme $L := \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$. Dokážme najprv, že

$$\text{ak } w \in L, \text{ tak } w = \varepsilon, \text{ alebo existujú } x, y \in L \text{ také, že } w = axby \text{ alebo } w = bxay. \qquad (*)$$

[7] V opačnom prípade by bolo buď $a_1 = \ldots = a_{2k+2} = a$, alebo $a_1 = \ldots = a_{2k+2} = b$, pričom v ani jednom prípade zjavne nemôže byť $\#_a(a_1 \ldots a_{2k+2}) = \#_b(a_1 \ldots a_{2k+2})$.

Pre $u \in \{a,b\}^*$ označme $\Delta(u) = \#_a(u) - \#_b(u)$; je pritom jasné, že pre všetky $v \in \{a,b\}^*$ a $c \in \{a,b\}$ je $\Delta(vc) = \Delta(v) \pm 1$, pričom $v \in L$ práve vtedy, keď $\Delta(v) = 0$. Pre ľubovoľné *neprázdne* slovo $w \in L$ uvažujme jeho *najkratší neprázdny* prefix u taký, že $u \in L$. Nech $u = a_1 \ldots a_k$ pre $a_1, \ldots, a_k \in \{a,b\}$. Ak $a_1 = a$, z uvedeného vyplýva, že pre $i = 1, \ldots, k-1$ je $\Delta(a_1 \ldots a_i) > 0$; keďže ale $\Delta(u) = 0$, nutne $a_k = b$, a teda $w = axby$ pre $x = a_2 \ldots a_{k-1}$ a nejaké slovo $y \in \{a,b\}^*$. Z príslušnosti slova $u = axb$ do L potom dostávame $x \in L$. Keďže ďalej $w = uy$ a $u, w \in L$, nutne aj $y \in L$. Podobne by sme pre $a_1 = b$ ukázali, že $w = bxay$ pre nejaké $x, y \in L$.

Môžeme teraz dokázať samotnú rovnosť $L(G) = L$.

$\subseteq$: Indukciou vzhľadom na $n \in \mathbb{N}$ dokážeme, že $w \in L$ kedykoľvek $w \in T^*$ a $\sigma \Rightarrow^n w$.

Ak $n = 0$, neexistuje žiadne $w \in T^*$ také, že $\sigma \Rightarrow^n w$; tvrdenie teda platí. Predpokladajme teraz platnosť tvrdenia pre $n = 0, \ldots, k$ a uvažujme $n = k+1$. Odvodenie $\sigma \Rightarrow^{k+1} w$ potom môžeme prepísať ako $\sigma \Rightarrow u \Rightarrow^k w$ pre nejaké $u \in (N \cup T)^*$. Keďže $\sigma \Rightarrow u$, je $\sigma \to u \in P$, a teda $u \in \{\varepsilon, a\sigma b\sigma, b\sigma a\sigma\}$. Ak $u = \varepsilon$, nutne $k = 0$ a $w = \varepsilon \in L$; tvrdenie teda platí. Pre $u = a\sigma b\sigma$ z tvrdenia 1 vyplýva, že $w = axby$ pre nejaké $x, y \in T^*$ také, že pre nejaké prirodzené $k_1, k_2 \leq k$ je $\sigma \Rightarrow^{k_1} x$ a $\sigma \Rightarrow^{k_2} y$. Z indukčného predpokladu teda $x, y \in L$, a teda aj $w = axby \in L$. Pre $u = b\sigma a\sigma$ možno argumentovať analogicky.

$\supseteq$: Evidentne $\sigma \Rightarrow \varepsilon$, a teda $\varepsilon \in L(G)$. Ak teraz $x, y \in L$ sú také, že $x, y \in L(G)$, tak $\sigma \Rightarrow^* x$ a $\sigma \Rightarrow^* y$, v dôsledku čoho vďaka tvrdeniu 1 aj $\sigma \Rightarrow a\sigma b\sigma \Rightarrow^* axby$ a $\sigma \Rightarrow b\sigma a\sigma \Rightarrow^* bxay$. Z tvrdenia $(*)$ teda vyplýva, že musia byť v $L(G)$ všetky slová z jazyka L. $\square$

Pri riešení predchádzajúcej úlohy sme do veľkej miery využili skutočnosť, že jediným neterminálom gramatiky G bol počiatočný neterminál σ. V prípade, že gramatika G obsahuje viacero užitočných neterminálov, môžu sa aj niektoré z nich objaviť vo vetnej forme odvoditeľnej z počiatočného neterminálu na jeden krok. Pre takúto gramatiku by sme teda opäť museli dokazovať silnejšie tvrdenie, ktoré by tentokrát charakterizovalo jazyky slov odvoditeľných z jednotlivých neterminálov gramatiky (a nie iba z počiatočného neterminálu σ).