# Pumping lemma pro bezkontextové jazyky

V kapitole **9.5 CFG v Chomského normální formě** jsme naznačili, že CNF je kanonický tvar, který významně použijeme v dokazování pomocí Pumping lemmatu pro bezkontextové jazyky. Ještě než si uvedeme znění lemmatu, vysvětlíme, proč oba zmíněné pojmy spolu souvisejí.

## Věta 1.

Nechť $G = (N, \Sigma, P, S)$ je bezkontextová gramatika v Chomského normální formě, $w \in L(G)$ je libovolné slovo. Označme derivační strom pro slovo $w$ symbolem $T$. Pak platí tvrzení: má-li každá cesta ve stromu $T$ délku maximálně $n$ ($n \in \mathbb{N}$), pak $T$ obsahuje nejvýše $2^{n-1}$ listů.

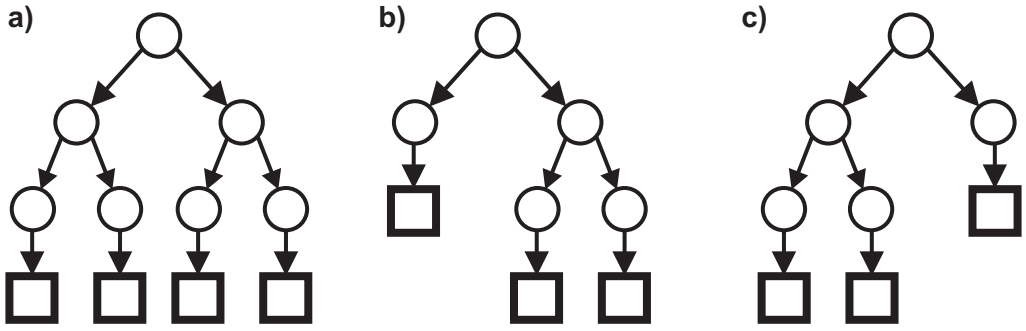
## Poznámka.

Ještě než provedeme důkaz, ukážeme si několik konkrétních případů derivačního stromu pro libovolné slovo $w \in L(G)$, kde $G$ je CFG v Chomského normální formě.

1. Začneme stromem, kde libovolná cesta má délku maximálně 1 (tj. $n = 1$). Uvědomte si, že v tom případě vypadá strom tak, že obsahuje pouze kořen, z něhož vede jediná cesta k listu. Počet listů je tedy $1 = 2^{n-1} = 2^{1-1} = 2^0$.

2. Je-li $n$ maximálně 2, tj. libovolná cesta ve stromu je délky nejvýše 2 ($n \leq 2$), pak máme opět jedinou možnost, jak strom může vypadat (nepočítáme-li situaci z 1. případu). Z kořene $A$ vedou dvě cesty do vnitřních uzlů $B, C \in N$ (dle pravidla $A \to BC$) a z každého z nich potom jedna cesta do listu (dle pravidel $B \to x$, $C \to y$, kde $x, y$ jsou libovolné terminály). Počet listů je 2, což opět odpovídá podmínce věty, že má být nejvýše $2^{n-1} = 2^{2-1} = 2$.

3. Je-li $n$ maximálně 3, pak libovolná cesta ve stromu má délku nejvýše 3 ($n \leq 3$). V tu chvíli už máme tři možné situace, jak strom může vypadat. Pro lepší přehlednost označíme listy čtvercem a vnitřní uzly kružnicí:




   Všimněte si, že nejvyšší počet listů je v případě a), a to 4. Týká se to situace, kdy každá cesta ve stromu má délku právě 3. Je tedy zřejmé, že největší počet listů ve stromu, jehož libovolná cesta má délku max. 3, je v případě, že každá cesta má délku 3. V ostatních případech (b, c) je počet listů menší než 4. I zde tedy platí závěr věty: počet listů je max. $4 = 2^{n-1} = 2^{3-1} = 2^2$.

Takto bychom mohli pokračovat dále. Místo toho však podáme obecný důkaz věty pro libovolné $n \in \mathbb{N}$ znamenající max. možnou délku cesty ve stromu.

**Důkaz.** Nechť platí předpoklady věty, tj. mějme bezkontextovou gramatiku $G = (N, \Sigma, P, S)$ v Chomského normální formě. Zvolme slovo $w \in L(G)$ libovolně a označte symbolem $T$ jeho derivační strom. Dále předpokládejme, že každá cesta ve stromu $T$ má délku maximálně $n$.

Uvažujme podstrom $T_1$ stromu $T$, který vznikne tak, že ze stromu $T$ odstraníme listy a cesty do nich. Pro takový strom platí, že délka jeho libovolné cesty je nejvýše $n - 1$.

Předpokládejme navíc striktně, že

$$(*) \text{ každá cesta ve stromu } T_1 \text{ má délku právě } n - 1.$$

(Je to případ, kdy strom $T_1$ může mít nejvíce listů.) Protože $T_1$ vznikl z derivačního stromu $T$ vyjmutím jeho listů, platí, že každý vnitřní uzel $T_1$ má právě dva následníky (mohli jsme použít pouze pravidla $A \to BC$). To znamená, že kořen má 2 přímé následníky (1. díl cesty), 4 následníky ve 2. úrovni „pod“ kořenem (2. díl cesty), …, $2^{n-1}$ následníků v $(n-1)$-té úrovni „pod“ kořenem [$(n-1)$-tý díl cesty]. Z toho vyplývá, že počet listů stromů $T_1$ je právě $2^{n-1}$. Připojíme-li ke stromu $T_1$ původní listy stromu $T$, zjišťujeme, že jejich počet je také $2^{n-1}$.

Protože jsme v předpokladu (*) počítali s tím, že délka každé cesty ve stromu $T_1$ je právě $n - 1$, úpravou (*)

$$(**) \text{ každá cesta ve stromu } T_1 \text{ má maximálně } n - 1$$

zajistíme tvrzení věty, tj. že počet listů stromu $T_1$, potažmo $T$ je nejvýše $2^{n-1}$.

## Lemma (Věta o vkládání – Pumping lemma)

Buď $L$ bezkontextový jazyk.
Pak existují $p, q \in \mathbb{N}$ takové, že každé slovo $w \in L,\ |w| > p$, lze psát ve tvaru $w = uvwxy$ tak, že:

a) alespoň jedno ze slov $v, x$ je neprázdné ($v, x \neq \varepsilon$),

b) $|vwx| \leq q$,

c) pro každé $i \in \mathbb{N}_0$ platí $uv^i wx^i y \in L$.

## Poznámka.

1. Konstanta $p$ se může rovnat $q$.

2. Tvrzení je ve tvaru $P \Rightarrow Q$, kde

   - $P$ je předpoklad, že $L$ je bezkontextový jazyk, a
   - $Q$ je zbytek tvrzení, tj. že existují konstanty $p, q \in \mathbb{N}_0$ takové, že…

3. V případě, že $Q$ platí, nelze říci, jestli platí či neplatí $P$ (tj. zda $L$ je či není bezkontextový jazyk, protože implikace $0 \Rightarrow 1$ nebo $1 \Rightarrow 1$ jsou obě pravdivé).

   Pokud ale předpokládáme, že $P$ platí (tj. $L$ je bezkontextový) a dokážeme $\neg Q$, pak je platnost $P \wedge \neg Q$ ve sporu s Pumping lemmatem, protože jsme ukázali pravdivost $\neg(P \Rightarrow Q) = P \wedge \neg Q$. Náš předpoklad, že $P$ platí, není správně, a tudíž $P$ neplatí (tedy jazyk $L$ není bezkontextový).

4. Z předchozího bodu vyplývá, že Pumping lemma lze použít pouze k důkazu toho, že daný jazyk $L$ **není** bezkontextový.

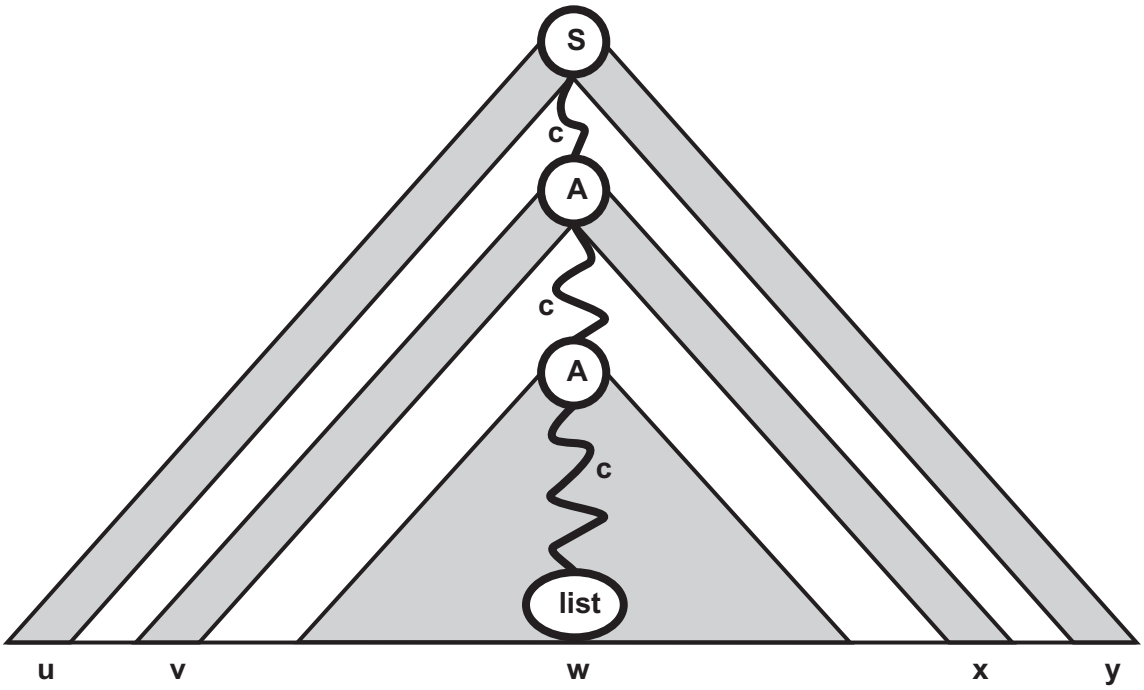
**Důkaz Pumping lemmatu.** Předpokládejme, že $L$ je bezkontextový jazyk. Potom existuje bezkontextová gramatika $G = (N, \Sigma, P, S)$, která jazyk $L$ generuje, tj. $L = L(G)$. Bez újmy na obecnosti předpokládejme, že $G$ je v Chomského normální formě. Položme $k$ rovno počtu neterminálů, $p = 2^{k-1},\ q = 2^k$.

Zvolme slovo $z \in L$ takové, že $|z| > p = 2^{k-1}$. Tedy derivační strom pro $z$ má více než $2^{k-1}$ listů. Potom (dle věty 1 na začátku tohoto dokumentu) existuje v derivačním stromu pro $z$ cesta delší než $k$. Počet vnitřních uzlů cesty je tedy větší než $k$, což znamená, že minimálně jeden neterminál se v derivační cestě musí opakovat.

Zvolme pevně derivační strom $T$ pro slovo $z$ a v něm cestu $c$ delší než $k$, která zároveň bude nejdelší možná. [Předpoklad maximální možné délky cesty $c$ je pro nás v další fázi důkazu důležitý!]

Na cestě $c$ lze zvolit tři uzly $u_1, u_2, u_3$ s těmito vlastnostmi:

1. uzly $u_1, u_2$ jsou označeny tím samým neterminálem $A$,
2. $u_1$ je blíže ke kořeni než $u_2$,
3. uzel $u_3$ je list,
4. cesta od $u_1$ do $u_3$ má délku nejvýše $k + 1$.




Uzel $u_1$ označený neterminálem $A$ určuje strom $T_1$ ($u_1$ je kořenem $T_1$). Protože $T_1$ je podstromem $T$, musí existovat slova $u, y$ a derivace

$$(1) \quad S \Rightarrow^* uAy$$

Uzel $u_2$ označený taktéž neterminálem $A$ určuje strom $T_2$, přičemž platí, že $T_2$ je podstromem $T_1$. Z toho opět vyplývá, že existují řetězce $v, x$ a derivace

$$(2) \quad A \Rightarrow^* vAx$$

Výsledkem podstromu $T_1$ je nějaké slovo $z_1$, které je podslovem slova $z$. Dle vlastnosti 4 má podstrom $T_1$ nejvýše $2^k$ listů (opět to vyplývá z Věty 1). To znamená, že i délka slova $|z_1| \leq 2^k = q$. Pokud by existovala cesta z uzlu $u_1$ k nějakému jinému listu $u_4$, která by byla delší než $k+1$, pak by to byl spor s předpokladem, že naše cesta $c$ je nejdelší.

Výsledkem podstromu $T_2$ je nějaké slovo $w$, pro které platí, že je podslovem slova $z_1$. Zároveň jistě existuje derivace tvaru

$$(3) \quad A \Rightarrow^* w$$

Nyní použijme jedenkrát derivaci (1), $i$-krát derivaci (2) a nakonec jedenkrát derivaci (3):

$$(4) \quad S \Rightarrow^* uAy \Rightarrow^* uv^iAx^iy \Rightarrow^* uv^iwx^iy$$

Z derivace (4) je patrné, že slovo $uv^iwx^iy \in L(G)$ (bod c je dokázán).

Navíc v předchozím jsme tvrdili, že pokud cestu z uzlu $u_1$ do listu $u_3$ uskutečníme jednou, bude mít podstrom $T_1$ příslušný uzlu $u_1$ maximálně $2^k = q$ listů. Výsledkem stromu $T_1$ je slovo $z_1$, potažmo $vwx$, jehož délka je menší nebo rovna $q = 2^k$ (bod b dokázán).

Dále si uvědomme, že uzel $u_1$ s návěštím $A$ je vnitřní uzel, tj. musí existovat derivace $A \Rightarrow BC$. Gramatika $G$ je v Chomského normální formě, tj. bez $\varepsilon$-pravidel, musí tedy platit, že alespoň jeden z neterminálů $B, C$ se odvodí na neprázdné slovo. Díky tomu můžeme psát $vx \neq \varepsilon$ (bod a je dokázán).

Tím je důkaz hotov.

## Poznámka.

Raději si uvedeme přesné znění negace tvrzení $Q$: pro všechna $p, q \in \mathbb{N}_0$ existuje slovo $w \in L$, $|w| > p$, tak, že pro každé možné rozdělení slova $z = uvwxy$, kde

a) $vx \neq \varepsilon$ a zároveň

b) $|vwx| \leq q$,

platí, že existuje $i \in \mathbb{N}_0$ tak, že $uv^iwx^iy \notin L$.

## Poznámka.

Mějme libovolný jazyk $L$, u kterého chceme ukázat, že není bezkontextový. Nejjednodušší strategie důkazu pomocí Pumping lemmatu spočívá v tom, že volíme slovo $z \in L(G)$ závislé pouze na jediné obecně zadané konstantě $n \in \mathbb{N}_0$. Tj. platí $n = p = q$. Nemůžeme však zvolit nějaké konkrétní slovo, protože potom bychom nesplnili počáteční podmínku tvrzení $\neg Q$: pro všechna $p, q \in \mathbb{N}_0 \ldots$, respektive pro všechna $n \in \mathbb{N}_0 \ldots$ Zároveň by u naší volby slova $z$ mělo platit, že $|z| > n$.

Poté hledáme všechna možná rozdělení slova $z$ na pět částí $u, v, w, x, y$, přičemž klademe důležitou podmínku: $|vwx| \leq n$. Pro každé rozdělíme hledáme konkrétní konstantu $i \in \mathbb{N}_0$ tak, abychom zajistili $uv^iwx^iy \notin L$. Podaří-li se nám to pro všechna rozdělení, pak jsme ukázali platnost $\neg Q$, což spolu s předpokladem $P$ (jazyk $L$ je bezkontextový) zajišťuje spor s Pumping lemmatem. Toto lemma je však dokázáno, je tedy chyba v našem předpokladu $P$, a proto platí jeho opak. Jazyk $L$ není bezkontextový.