

## NEOHRANIČENÁ SLOVA A (LOKÁLNÍ) PERIODY

Řekneme, že slovo  $w$  je *ohraničené*, pokud existuje neprázdné slovo  $u \neq w$ , které je současně prefixem a sufixem  $w$ . Takové slovo  $u$  nazveme *hranicí* slova  $w$ . Je-li slovo  $u$  samo ohraničené a má hranici  $z$ , je  $z$  také hranice  $w$ . Z toho snadno odvodíme, že nejkratší hranice ohraničeného slova není ohraničená. Je-li hranice  $u$  slova  $w$  delší než  $|w|/2$ , pak se  $u$  překrývá samo se sebou a je tedy ohraničené. Z toho vidíme, že každé ohraničené slovo je tvaru  $uvu$ ,  $u \neq \varepsilon$ . Dále zřejmě platí, že  $|w| - |u|$  je perioda slova  $w$ . Označíme-li nejkratší periodu slova  $w$  symbolem  $\pi(w)$ , dostáváme, že pro ohraničené slovo  $w$  platí  $\pi(w) = |w| - |u|$ , kde  $u$  je jeho nejdelší hranice. Slovo je tedy *neohraničené* (tedy není ohraničené), právě když  $\pi(w)$  je maximální možné, totiž  $|w|$ .

**Lyndonova slova.** Primitivní slovo je takové, které není celočíselnou mocninou žádného kratšího slova. Každé ohraničené slovo je však racionální mocninou nějakého kratšího slova, např.  $aabaaba = (aab)^{\frac{7}{3}}$ . Ukážeme ale, že každé primitivní slovo je konjugované s nějakým neohraničeným slovem. Takových slov může být více, ale jedním z nich je vždy tzv. *Lyndonovo slovo*, které definujeme jako primitivní slovo, které je minimální ve své konjugační třídě (tj. ekvivalenční třídě vzájemně konjugovaných slov) vzhledem k nějakému lexikografickému uspořádání  $\triangleleft$ .

Připomeňme, že je-li slovo primitivní, jsou primitivní i všechna slova s ním konjugovaná. Z primitivity plyne, že Lyndonovo slovo odpovídá jediné konjugaci, stejný výsledek dvou různých konjugací by znamenal, že slovo není primitivní (je-li  $uv$  primitivní, pak  $uv \neq vu$ ).

Poznamenejme několik vlastností lexikografických uspořádání. Každé takové uspořádání je určeno nějakým lineárním uspořádáním písmen. Jsou-li  $\triangleleft$  a  $\blacktriangleleft$  lexikografická uspořádání daná opačnými uspořádáními písmen (tj.  $a \triangleleft b \Leftrightarrow b \blacktriangleleft a$ ), pak pro dvě prefixově nesrovnatelná slova  $u$  a  $v$  platí rovněž  $u \triangleleft v \Leftrightarrow v \blacktriangleleft u$ . Pokud je ovšem  $u$  prefix  $v$ , chovají se uspořádání stejně:  $u \triangleleft v$  i  $u \blacktriangleleft v$ .

Dále platí, že pro každé  $z$  je  $u \triangleleft v \Leftrightarrow zu \triangleleft zv$ . Pokud jsou  $u$  a  $v$  prefixově nesrovnatelná (speciálně pokud jsou stejně dlouhá a různá), máme také  $u \triangleleft v \Leftrightarrow uz \triangleleft vz'$  pro každé  $z$  a  $z'$ .

*Věta.* Každé Lyndonovo slovo je neohraničené.

*Důkaz.* Sporem. Nechť je Lyndonovo slovo  $w$  tvaru  $uvu$ . Z minimality plyne  $uvu \triangleleft vuv$ , tedy  $uv \triangleleft vu$ , a proto také  $uuv \triangleleft uvu$ , což je spor s minimalitou  $w$ .  $\square$

Lyndonova slova lze charakterizovat jako „samominimální“:

*Věta.* Slovo  $w$  je Lyndonovo, právě když je nejmenší ze svých sufixů (tj. je vzhledem k  $\triangleleft$  striktně menší než jakýkoli vlastní sufix).

*Důkaz.* Nechť je  $w$  Lyndonovo slovo a nechť  $p$  je prefix  $w$  a  $s$  je sufix  $w$  tak, že  $|p| = |s|$ . Protože je  $w$  Lyndonovo (a tedy neohraničené), platí  $p \neq s$  a  $p \triangleleft s$ , a tedy také  $w \triangleleft s$ .

Nechť naopak  $w$  Lyndonovo není. Pokud  $w$  není primitivní, je lexikograficky větší než vlastní primitivní kořen. Pokud je  $w = uv$  primitivní a  $vu$  je Lyndonovo, platí  $vu \triangleleft uv$ , a tedy také  $v \triangleleft uv$ .  $\square$

*Věta.* Nechť  $u$  i  $v$  jsou Lyndonova slova (vzhledem k  $\triangleleft$ ). Pokud  $u \triangleleft v$  a  $u \neq v$ , pak je i  $uv$  Lyndonovo slovo.

*Důkaz.* Nejprve ukažme, že  $uv \triangleleft v$ . Pokud je  $v = uv'$ , pak dokazované  $uv \triangleleft uv' = v$  plyne z  $v \triangleleft v'$ . Pokud  $u$  není prefix  $v$ , pak  $uv \triangleleft v$  plyne z  $u \triangleleft v$ .

Nechť je nyní  $z \neq v$  vlastní sufix  $uv$ . Pokud  $uv = uv'z$ , pak  $uv \triangleleft v \triangleleft z$ . Pokud  $z = z'v$ , pak  $u \triangleleft z'$ , a tedy  $uv \triangleleft z'v$ .

Slovo  $uv$  je tedy menší než všechny jeho vlastní sufixy a důkaz je hotov.  $\square$

Následující věta mluví o existenci a jednoznačnosti tzv. *Lyndonovy faktorizace*.

*Věta.* Nechť  $\triangleleft$  je lexikografické uspořádání. Každé slovo  $w$  lze jednoznačně napsat jako součin nerostoucí posloupnosti Lyndonových slov (vzhledem k  $\triangleleft$ ).

*Důkaz.* Uvažujme faktorizaci  $w = u_1u_2 \cdots u_k$  definovanou tak, že pro každé  $j = 1, 2, 3, \dots, k$  je  $u_j$  nejdelší sufix  $(u_1u_2 \cdots u_j)$ , který je Lyndonovým slovem. Jinak řečeno, ze slova  $w$  postupně odebíráme jeho nejdelší Lyndonovy sufixy.

Faktorizace je dobře definovaná, protože každé slovo délky jedna je Lyndonovo. Všechna slova ve faktorizaci jsou Lyndonova a pokud  $u_i \triangleleft u_{i+1}$  a  $u_i \neq u_{i+1}$ , pak z předchozí věty dostáváme spor s maximalitou  $u_{i+1}$ . Posloupnost je tedy nerostoucí.

Nechť nyní  $w = v_1v_2 \cdots v_m = z_1z_2 \cdots z_\ell$  jsou dvě faktorizace splňující předpoklady věty. Bez újmy na obecnosti předpokládáme, že  $v_1$  je prefix  $z_1$  a nechť  $r$  je neprázdný prefix  $v_j$ , který je sufixem  $z_1$  (kde  $j$  je nejmenší index takový, že  $z_1$  je prefixem  $v_1v_2 \cdots v_j$ ). Protože Lyndonovo slovo je menší než jakýkoli jeho sufix, dostáváme

$$z_1 \triangleleft r \triangleleft v_j \triangleleft v_1 \triangleleft z_1,$$

a tedy  $j = 1$  a  $z_1 = v_1$ . Důkaz je dokončen indukcí.  $\square$

### Lokální perioda a kritická faktorizace.

*Definice.* Lokální periodou slova  $w$  na pozici  $k \in \{0, 1, \dots, |w|\}$  rozumíme délku nejkratšího slova  $x$ , které je sufixově srovnatelné s  $u$  a prefixově srovnatelné s  $v$ , kde  $w = uv$  a  $k = |u|$ .

Jinak řečeno, lokální perioda je délka  $x$  u nejkratšího čtverce  $x^2$ , který je centrován na pozici  $k$ . Všimněme si, že takové  $x$  je vždy neohrazené: pokud by totiž  $u$  byla hranice  $x$ , byl by  $u$  kratší čtverec, rovněž centrován na pozici  $k$ .

Je snadné si všimnout, že lokální perioda je na každé pozici rovna nejvýše  $\pi(w)$ . Je-li lokální perioda na pozici  $|u|$  rovna  $\pi(w)$ , nazýváme ji *kritickou pozicí* a faktorizaci  $w = uv$ , či přesněji dvojici  $(u, v)$ , *kritickou faktorizací* slova  $w$ .

*Věta* (Věta o kritické faktorizaci – Critical Factorization Theorem). Každé slovo  $w$  připouští kritickou faktorizaci.

*Důkaz.* Zvolme dvě opačná lexikografická uspořádání  $\triangleleft$  a  $\blacktriangleleft$ . Nechť  $\alpha$  je maximální sufix  $w$  vzhledem k  $\triangleleft$  a  $\beta$  maximální sufix  $w$  vzhledem k  $\blacktriangleleft$ . Pokud je  $w$  mocninou jednoho písmene, je tvrzení zřejmé, v opačném případě je jistě  $\alpha \neq \beta$ . Bez újmy na obecnosti předpokládáme  $|\alpha| < |\beta|$ .

Ukážeme, že  $k = |w| - |\alpha|$  je kritický bod. Nechť je  $x^2$  nejmenší čtverec centrován na pozici  $k$ . Uvažme postupně čtyři možnosti vztahu  $x$  ke slovům  $\alpha$  a  $w\alpha^{-1}$ .

1. Je-li  $x$  prefix  $\alpha$  a sufix  $w\alpha^{-1}$ , je  $x\alpha$  sufixem  $w$ , a tedy  $x\alpha \triangleleft \alpha$ . Odtud dostáváme  $\alpha \triangleleft x^{-1}\alpha$ , což je spor s maximalitou  $\alpha$ .

2. Je-li  $\alpha$  prefix  $x$  a  $x$  je sufix  $w\alpha^{-1}$ , je  $\alpha$  prefixem  $x\alpha$ , a tedy  $\alpha \triangleleft x\alpha$ , což je opět spor.

3. Nechť je  $x$  prefixem  $\alpha$  a  $w\alpha^{-1}$  sufixem  $x$ . Budiž  $x^i x'$  nejdelší prefix  $\alpha$ , který má periodu  $|x|$ . Pak existují písmena  $c \neq d$  taková, že  $x^i c$  a  $x^i x' d$  jsou prefixy  $\alpha$ .

Z maximality  $\alpha$  plyne  $d \triangleleft c$ . Označme  $v = \beta\alpha^{-1}$ . Protože  $v$  je sufixem  $x$ , je  $vx'd$  faktorem slova  $w$ . To je ovšem spor s maximalitou  $\beta$ , protože  $vx'c$  je prefixem  $\beta$  a  $vx'c \triangleleft vx'd$ .

Zbývá tedy možnost, že  $\alpha$  je prefixem a  $w\alpha^{-1}$  sufixem  $x$ , což znamená, že  $|x|$  je periodou  $w$ , což jsme chtěli ukázat.  $\square$

\*

**Nejdelší neohraničené faktory.** Označme  $\tau(w)$  délku nejdelšího neohraničeného faktoru slova  $w$ . Je-li  $v$  faktor slova  $w$  delší než  $\pi(w)$ , pak je jistě ohraničený, protože má rovněž periodu  $\pi(w)$ . Platí tedy  $\tau(w) \leq \pi(w)$ . Je-li  $|w| \geq 2\pi(w)$ , pak nutně platí rovnost, protože slovo  $w$  obsahuje nějaké Lyndonovo slovo délky  $\pi(w)$ ; obsahuje totiž všechny prvky konjugační třídy svého periodického základu.

Neohraničené faktory délky  $\pi(w)$  můžeme také hledat pomocí kritických pozic. Je-li totiž  $x^2$  minimální čtverec centrovaný na nějaké kritické pozici slova  $w$  a alespoň jedno  $x$  leží celé ve  $w$ , je  $\tau(w) = \pi(w)$ , protože  $x$  je neohraničené, a z definice kritické pozice máme  $|x| = \pi(w)$ . To umožňuje výše uvedený odhad mírně zlepšit.

*Věta.* Je-li  $|w| \geq 2\pi(w) - 3$ , pak  $\tau(w) = \pi(w)$ .

*Důkaz.* Nechť  $w = vc_1c_2c_3v$ , kde  $|vc_1c_2c_3| = \pi(w)$  (případy, kdy  $\pi(w) < 3$  jsou snadné). Existuje-li kritická pozice  $k \leq \pi(w) - 3$ , pak  $x$ , pro které je  $x^2$  centrován ve  $k$ , leží celé ve slově  $w$  a jsme hotovi.

Uvažujme kritickou pozici  $\ell$  popsanou v důkazu Věty o kritické faktorizaci. Předpokládejme nejprve, že je rovna  $\pi(w) - 1$ . Pak je  $c_3$  maximální sufix  $vc_1c_2c_3$  vzhledem k nějakému lexikografickému uspořádání. Písmeno  $c_3$  se tedy nevyskytuje ve slově  $vc_1c_2$ , a  $vc_1c_2c_3$  je neohraničené a délky  $\pi(w)$ .

Obsahuje-li slovo  $w$  tři různá písmena, můžeme zvolit  $\triangleleft$  tak, že  $c_2$  není ani nejmenší, ani největší písmeno, a tudíž příslušný kritický bod leží mimo  $\pi(w) - 2$  a jsme hotovi.

Zbývá situace, kdy  $\ell = \pi(w) - 2$  a  $w$  obsahuje právě dvě písmena. Je-li  $c_2 \neq c_3$ , zvolíme uspořádání  $\triangleleft$ , ve kterém je  $c_3$  minimální a  $c_2$  maximální. Opět není těžké ověřit, že  $c_2$  se nevyskytuje ve  $v$ . Pak je ale  $vc_1 = c_3v$  a  $w$  má v rozporu s předpokladem periodu  $|vc_1c_2|$ . Nechť tedy  $c_2 = c_3 = b$  a  $c_1 = a$ . Protože  $bb$  je lexikograficky maximální sufix  $vabb$ , nemá  $bb$  žádný výskyt ve  $v$ . Pokud je  $vabb$  neohraničené, jsme hotovi. V opačném případě je  $ba$  prefix  $v$  a není těžké ověřit, že  $b^{-1}vabbb$  je neohraničený faktor  $w$ .  $\square$

Mez uvedená ve větě je optimální, jak ukazují slova  $w = (aba)^k abba \cdot (aba)^k$ , která jsou délky  $2\pi(w) - 4$  a nejdelší neohraničené faktory jsou  $(aba)^k abb$  a  $bba(aba)^k$  délky  $\pi(w) - 1$ .

Souvislost rovnosti  $\pi(w) = \tau(w)$  s délkou  $w$  vyjádřenou pomocí  $\pi(w)$  je tedy přesně popsána. Podobně se můžeme ptát na souvislost s délkou  $w$  vyjádřenou pomocí  $\tau(w)$ . Jinak řečeno, je-li  $v$  neohraničené slovo, jak moc ho můžeme prodloužit do slova  $w$ , aby  $|v| = \tau(w) < \pi(w)$ . V tomto případě platí, že  $|w| \geq \frac{7}{3}\tau(w)$  implikuje  $\pi(w) = \tau(w)$ . Tato mez je asymptoticky optimální, jak ukazuje následující slovo:

$$w = a^n ba^{n+1} ba^n ba^{n+2} ba^n ba^{n+1} ba^n$$

kde  $n \geq 0$ ,  $\tau(w) = 3n + 6$  (dvě nejdelší neohraničená slova jsou  $ba^{n+1}ba^nba^{n+2}$  a  $a^{n+2}ba^nba^{n+1}b$ ),  $\pi(w) = 4n + 7$  a  $|w| = 7n + 10$ . Tedy  $\tau(w) < \pi(w)$  a  $|w| = \frac{7}{3}\pi(w) - 4$ .

Důkaz je obtížný. Dokážeme ale tvrzení týkající se speciálního případu, kdy dovolujeme neohraničené slovo prodlužovat jen jedním směrem (např. doprava). Řekněme, že  $vu$  je *Duvalovo prodloužení* neohraničeného slova  $v$ , pokud  $|v| = \tau(vu) < \pi(vu)$ . Jak ukazují slova  $v = a^i ba^j bb$  a  $u = a^j ba^i$ , kde  $1 \leq i < j$ , je možné Duvalovo prodloužení o  $|v| - 2$ . To je optimální hranice. Ukážeme ale jen následující slabší tvrzení (případ  $|v| - 1$  je jen o něco techničtější).

*Věta.* Nechť  $v$  je neohraničené slovo. Jeho Duvalovo prodloužení délky  $|v|$  není možné.

*Důkaz.* Nechť  $u \neq v$  je slovo délky  $|v|$  takové, že  $\pi(vu) > |v|$ , tedy  $u \neq v$ . Najdeme neohraničený faktor  $vu$  delší než  $v$ .

Nechť  $\triangleleft$  a  $\blacktriangleleft$  jsou dvě opačná lexikografická uspořádání taková, že  $\alpha$  je  $\triangleleft$ -maximální sufix  $v$ ,  $\beta$  je  $\blacktriangleleft$ -maximální sufix  $v$  a  $|\alpha| < |\beta|$ . Z maximality  $\alpha$  plyne, že má jen jeden výskyt ve  $v$ . Nechť má nějaký další výskyt ve  $vu$  a nechť je např.  $v'\alpha$  prefix  $vu$ , kde  $|v'\alpha| > |v|$ . Tvrdíme, že  $v'\alpha$  je neohraničené. Případná nejkratší hranice  $r$  je totiž jistě kratší než  $v$ . Pokud by měla délku alespoň  $|\alpha|$ , bylo by  $\alpha$  sufixem  $r$ , a protože  $r$  je prefixem  $v$ , dostáváme výskyt  $\alpha$  uvnitř  $v$ , což je spor. Pokud by bylo naopak  $|r| < |\alpha|$ , bylo by  $r$  hranicí  $v$ , což je opět spor.

Nechť má tedy  $\alpha$  jen jeden výskyt ve  $vu$ . Nechť  $u = u'u''$ , kde  $|u''| = |\alpha|$ . Označme  $p$  nejkratší periodu slova  $\alpha u'$ . Slovo  $uv$  má nejkratší periodu ostře větší než  $p$ , jinak by se  $\alpha$  v  $\alpha u'$  opakovalo, což jsme vyloučili. Nechť je  $\alpha u'zc$  nejkratší prefix slova  $\alpha u$  s periodou ostře větší než  $p$ , kde  $c$  je písmeno a  $z \in \Sigma^*$ . Nechť  $\alpha u'z = ws$ , kde  $|w| = p$ ; neboli  $w$  je periodický základ slov  $\alpha u'$  a  $\alpha u'z$ , zatímco přidáním písmene  $c$  je perioda  $p$  porušena. Všimněme si, že  $s \neq \alpha$  je prefix  $\alpha$ , zatímco  $sc$  nikoli. Nechť je tedy  $sd$  prefix  $\alpha$ , kde  $d \neq c$  je písmeno. Pokud je  $\alpha u'zc$  neohraničené, jsme hotovi. Nechť je tedy ohraničené a nechť je  $rc$  jeho nejdelší hranice. Z definice  $s$  plyne, že  $r$  je kratší než  $s$ , a je tedy hranicí  $s$ . Pak je  $rc$  prefix  $\alpha$  a  $rd$  je faktorem  $\alpha$ . Z maximality  $\alpha$  dostáváme  $d \triangleleft c$ .

Uvažujme nyní slovo  $\beta wsc$  a předpokládejme, že  $tc$  je jeho nejdelší hranice. Je-li  $t$  sufix  $s$ , je  $tc$  prefix  $\beta$  a  $td$  faktor  $\alpha$ , což je vzhledem k  $tc \blacktriangleleft td$  spor s maximalitou  $\beta$ . Je-li naopak  $s$  sufix  $t$ , je  $sc$  (jakožto sufix  $tc$ ) faktorem  $u$ , což je spor s maximalitou  $\alpha$ , protože  $sd$  je prefix  $\alpha$  a  $sd \triangleleft sc$ .

Slovo  $\beta wsc$  je tedy neohraničené a jsme hotovi.  $\square$