

THUEJEV IZREK O BREZKVADRATNIH BESEDAH IN NEIZOGIBNI VZORCI

LUKA LJEVAR

Fakulteta za matematiko in fiziko
Univerza v Ljubljani

Predstavljena je kombinatorika besed in vzorci v besedah. Najprej so podane osnovne definicije. Pomembni pojmi so beseda, faktor in vzorec. Povedano je, kaj je vzorec, kdaj ga beseda vsebuje in kdaj se mu izogne. Posebej zanimiv je vzorec, ki se imenuje kvadrat. Formuliran in dokazan je Thuejev izrek, ki pove, da nad abecedo iz treh črk, obstajajo poljubno dolge besede, ki ne vsebujejo kvadrata. Nato je prikazana konstrukcija neskončne brezkvadratne besede nad abecedo iz treh črk. Nato je podana definicija brezprekrivnih besed. Primer takšne besede je Thue-Morseova beseda. Nato je predstavljen pojem (ne)izogibnosti vzorcev v besedah. Predstavljene so besede Zimina, ki določajo neizogibne vzorce. S pomočjo Ziminovega algoritma in pojma ireducibilnosti so karakterizirani neizogibni vzorci. Za neizogibne vzorce se poraja vprašanje, kako dolga mora biti beseda, da se vzorec gotovo pojavi. Na to vprašanje delno odgovorijo stolpične ocene, ki so obravnavane na koncu.

THUE'S THEOREM ABOUT SQUARE-FREE WORDS AND UNAVOIDABLE PATTERNS

Combinatorics of words and patterns in words are briefly presented. Basic definitions are given at the beginning. Important terms are word, factor, and pattern. The definition of patterns is given next. It is explained when a word either contains or avoids a pattern. A pattern called a square is of particular interest. Thue's theorem, which says that there exist words of arbitrary length containing only three letters that do not contain a square, is formulated and proven. A construction of an infinite square-free word over a three-element alphabet follows next. Overlap-free words are defined. An example of such a word is the Thue-Morse word. Then the (un)avoidability of patterns in words is explained. Zimin words are presented and it is demonstrated that Zimin words determine unavoidable patterns. With the help of the Zimin algorithm and a notion of irreducibility, unavoidable patterns are characterised. For unavoidable patterns, one can ask, what is the minimum length of the word such that a given pattern certainly appears in a given word. A partial answer to this question is given by tower-type bounds, which are considered at the end.

1. Uvod

Ukvarjali se bomo s kombinatoriko besed in z vzorci v besedah. Zato začnimo z intuitivnim premislekom o tem, kaj so besede in vzorci. Besede, kot jih poznamo iz vsakdanjega življenja, so zaporedja črk oziroma simbolov iz abecede. To idejo posplošimo tako, da za abecedo vzamemo poljubno končno množico. Za nas bo beseda poljubno končno ali neskončno zaporedje simbolov iz abecede. Ko besedo imamo, jo lahko na različne načine „razstavimo“. Na primer, besedo „matematika“ lahko razstavimo na dele „mate“, „mati“ in „ka“. Tem kosom besede bomo rekli faktorji. Faktor je lahko tudi ena sama črka ali pa cela beseda, lahko pa je tudi prazna beseda.

V nadaljevanju bodo imeli pomembno vlogo vzorci v besedah. Zato za začetek zgradimo nekaj intuicije. Zamislimo si, da imamo na papirju zapisano besedo, ki smo jo na poljuben način razkosali na faktorje. Fiksirajmo enega izmed faktorjev, recimo mu f_0 in ga na poljuben način razkosajmo na manjše faktorje. Nato vsakega izmed manjših faktorjev faktorja f_0 pobarvajmo z barvnimi flomastri, pri čemer smemo z enako barvo pobarvati le enake faktorje. Na primer v besedi „matematika“ si izberemo faktor „matema“, ki ga nadaljnje razkosamo na faktorje „ma“, „te“ in „ma“. Z enako barvo smemo pobarvati faktorja „ma“ in „ma“, ne smemo pa z enako barvo pobarvati faktorjev „ma“ in „te“. Kot vzorec si predstavljamo zaporedje barv, ki smo ga pri tem dobili. Če lahko v besedi najdemo dva zaporedna neprazna enaka faktorja, pravimo, da beseda vsebuje kvadrat.

Slika 1. Primer možne faktorizacije besede „matematika“.

Na začetku se bomo ukvarjali s kvadrati in dokazali Thuejev izrek, ki pove, da za vsako naravno število n obstaja beseda nad abecedo iz treh črk dolžine vsaj n , ki kvadrata ne vsebuje. Nato se bomo ukvarjali s splošnejšimi vzorci in njihovo vsebovanostjo v besedah nad različnimi abecedami.

2. Osnovne definicije

Za začetek povejmo, kaj je beseda.

Definicija 1. **Beseda** je končno ali neskončno zaporedje simbolov iz končne množice Σ , ki ji pravimo **abeceda**. S Σ^* označimo množico vseh besed nad abecedo Σ . Z ε označimo prazno besedo.

Sedaj bomo besedo razkosali na dele, ki jim rečemo faktorji.

Definicija 2. **Faktor besede** w je poljubno podzaporedje sosednjih simbolov iz besede w .

Povejmo, kaj je vzorec in kdaj ga neka beseda vsebuje oziroma se mu izogne.

Definicija 3. Naj bo Ω končna množica.¹ Imejmo preslikavo $h : \{\omega \in \Omega \mid \omega \text{ simbol v } p\} \rightarrow \Sigma^*$. **Morfizem** ali **substitucija** besedi $p = a_0 a_1 \dots a_n$ priredi besedo $h(p) = h(a_0)h(a_1) \dots h(a_n)$, kjer so $a_0, \dots, a_n$ črke besede p .

Opomba 4. Preslikava h je v splošnem lahko neinjektivna, na primer beseda $'aa'$ vsebuje vzorec xy , vzamemo lahko namreč $h(x) = h(y) = a$.

Definicija 5. Pravimo, da beseda $w \in \Sigma^*$ vsebuje **vzorec** $p \in \Omega^*$, če obstaja taka substitucija h , da je $h(p)$ faktor besede w . Beseda w se izogne vzorcu p , če ga ne vsebuje.

Primer vzorca, ki bo za nas še posebej zanimiv, je kvadrat. Beseda vsebuje kvadrat, če lahko v njej najdemo dva zaporedna neprazna enaka faktorja.

Definicija 6. Pravimo, da je beseda w **brezkvadratna**, če ne vsebuje vzorca xx . Vzorcju xx bomo rekli tudi **kvadrat**.

Zgornjo definicijo ponazorimo na konkretnih primerih.

Primer 7. Beseda „homomorfizem“ ni brezkvadratna, saj vsebuje kvadrata „momo“ in „omom“. Beseda „epimorfizem“ je brezkvadratna.

Smiselno se je vprašati, ali bodo vse dovolj dolge besede nad neko abecedo vsebovale nek vzorec.

Primer 8. Vse dovolj dolge besede nad dvoelementno abecedo $\{0, 1\}$ vsebujejo kvadrat. S preprosto obravnavo primerov (na sliki 2 spodaj) vidimo, da že poljubna beseda dolžine štiri iz 0 in 1 vsebuje kvadrat. Z rdečo in modro barvo smo na spodnji sliki označili zaporedna neprazna enaka faktorja, torej faktorja, ki tvorita kvadrat.

<u>0000</u>	01 <u>00</u>	10 <u>00</u>	11 <u>00</u>
<u>0001</u>	01 <u>01</u>	10 <u>01</u>	11 <u>01</u>
<u>0010</u>	01 <u>10</u>	10 <u>10</u>	11 <u>10</u>
<u>0011</u>	01 <u>11</u>	10 <u>11</u>	11 <u>11</u>

Slika 2. Poljubna beseda iz $\{0, 1\}$ dolžine štiri vsebuje kvadrat.

¹Neka druga abeceda.

Ker že vse besede dolžine štiri nad abecedo $\{0, 1\}$ vsebujejo kvadrat, ga bodo seveda vsebovale tudi vse daljše besede nad $\{0, 1\}$. Torej poljubna beseda nad $\{0, 1\}$ dolžine vsaj štiri vsebuje kvadrat.

3. Thuejev izrek

Ravnokar smo videli, da vsaka dovolj dolga beseda nad dvoelementno abecedo vsebuje kvadrat. Naravno vprašanje je, ali to velja tudi nad troelementno abecedo. Odgovor je ne. Leta 1906 je norveški matematik Axel Thue (1863 - 1922) dokazal, da nad troelementno abecedo obstaja beseda poljubne dolžine, ki kvadrata ne vsebuje.

Izrek 9. (Thue, 1906) Za vsak $n \in \mathbb{N}$ obstaja brezkvadratna beseda dolžine vsaj n nad abecedo $\Sigma_3 = \{a, b, c\}$, ki se izogne faktorjema aca in $bc b$.

Opomba 10. Thuejev izrek nam torej da več kot zgolj brezkvadratno besedo. Za skonstruirano besedo bo veljalo celo, da se izogne faktorjema aca in $bc b$. Ta dodatek bo koristil v kasnejših konstrukcijah.

Dokaz. Podajmo skico dokaza. Podroben dokaz se nahaja v [3].

Naj bo $u \in \Sigma_3^*$ poljubna brezkvadratna beseda, ki se izogne faktorjema aca in $bc b$ in se ne začne s črko c . Konstruirali bomo daljšo besedo w , ki zadošča istima lastnostima.

Izberimo novo tričrkovno abecedo $\{\alpha, \beta, \gamma\}$, ki je disjunktna z $\{a, b, c\}$. Najprej vsako pojavitev ac v besedi u zamenjamo z $a\beta\alpha$ in vsako pojavitev bc v u zamenjamo z $b\alpha\beta$. Tako tvorimo novo besedo u' . Opazimo, da je prirejanje $u \mapsto u'$ injektivno, saj najdemo inverzni predpis. Če iz u odstranimo vse simbole β in α preimenujemo v c , dobimo nazaj besedo u .

1. Če bi u' vsebovala vzorec xx , bi v x nastopal eden izmed simbolov a, b ali c . Iz načina inverznega prirejanja izhaja, da bi kvadrat vsebovala tudi u , torej u ne bi bila brezkvadratna. To je protislovje.
2. Beseda u' ne vsebuje vzorcev $x\alpha x$ in $x\beta x$. Predpostavimo, da u' vsebuje $x\alpha x$. Po konstrukciji je simbol β neposredno pred simbolom α ali pa je neposredno za njim. Torej velja bodisi $x\alpha x = y\beta\alpha y\beta$ bodisi $x\alpha x = \beta y\alpha\beta y$. Ko bi odstranili vse simbole α in zamenjali simbole β s simbolom c , bi bodisi dobili faktor oblike $y c y c$ bodisi faktor oblike $c y c y$ v u , za katero pa smo predpostavili, da je brezkvadratna. Dobili smo protislovje. Podobno se prepričamo, da u' ne vsebuje vzorca $x\beta x$.

Beseda u' je torej brezkvadratna in se izogne vzorcema $x\alpha x$ ter $x\beta x$.

Za vsakim simbolom v besedi u' postavimo simbol γ in tako konstruiramo besedo u'' . Če je na primer $u' = a\beta ab$, potem je $u'' = a\gamma\beta\gamma a\gamma b\gamma$. Beseda u'' je brezkvadratna, sicer bi iz besede u'' odvezli vse simbole γ in tako rekonstruirali besedo u' , ki bi vsebovala kvadrat. To je protislovje, saj je u' brezkvadratna.

Sedaj zamenjamo vsak a v u'' z $\alpha\beta\alpha$ in zamenjamo vsak b z $\beta\alpha\beta$. Tako konstruiramo besedo w . Če je na primer $u'' = a\gamma\beta\gamma a\gamma b\gamma$, potem je $w = \alpha\beta\alpha\gamma\beta\gamma\alpha\gamma\beta\alpha\beta\gamma$.

Dobili smo besedo w nad abecedo $\Sigma'_3 = \{\alpha, \beta, \gamma\}$. Iz konstrukcije je jasno, da je beseda w strogo daljša od besede u in se ne začne s simbolom γ . Pokažimo, da je tudi w brezkvadratna beseda, ki se izogne faktorjema $\alpha\gamma\alpha$ in $\beta\gamma\beta$.

1. V besedi u po simbolu a ne sme slediti simbol a , sicer bi imeli vzorec xx v u . Torej lahko po a sledita zgolj b in c . Če simbolu a sledi c , v besedi u' , dobimo faktor $a\beta\alpha$. To potrjuje smiselnost zamenjav v zgornjih primerih. Torej po a lahko v besedi u' sledita le β ali b . Na povsem enak način razmislimo, da lahko po b v besedi u' sledita le α in a . Ker je beseda u' brezkvadratna, simbolu α ne sme slediti α . Pokažimo še, da po α ne sme slediti a . Če bi v u' imeli faktor αa , bi glede na konstrukcijo v u' imeli celo $a\beta\alpha a$, kar pa bi pomenilo, da beseda u vsebuje faktor aca , kar je v protislovju s predpostavkami izreka. Torej lahko po α sledita le b in β . Na podoben način razmislimo, da lahko po β sledita le a in α .

Iz zgornjega premisleka sledi, da se v u' izmenično pojavljata simbol iz množice $\{a, \alpha\}$ in simbol iz množice $\{b, \beta\}$. V u'' vstavimo γ za vsakim simbolom, nato pa v w zamenjamo vsak a z $\alpha\beta\alpha$ in zamenjamo vsak b z $\beta\alpha\beta$. Zato so vsi tri-simbolni faktorji v w , ki imajo simbol γ na sredini bodisi oblike $\alpha\gamma\beta$ bodisi oblike $\beta\gamma\alpha$. Torej w ne vsebuje faktorjev $\alpha\gamma\alpha$ in $\beta\gamma\beta$.

2. Pokažimo še, da je w brezkvadratna. Denimo, da beseda w vsebuje vzorec xx . Iz konstrukcije w sledi, da med poljubnima dvema sledečima si γ lahko imamo le štiri možne faktorje: α , β , $\alpha\beta\alpha$ in $\beta\alpha\beta$. Ker noben od teh štirih faktorjev ne vsebuje kvadrata, mora x vsebovati vsaj en γ .

Predpostavimo, da x vsebuje le en γ in izpeljimo protislovje. Če x vsebuje le en γ , potem xx mora vsebovati $\gamma\alpha\gamma\alpha$, $\gamma\beta\gamma\beta$, $\gamma\alpha\beta\alpha\gamma\alpha$ ali $\gamma\beta\alpha\gamma\beta$. Torej w vsebuje $\alpha\gamma\alpha$ ali $\beta\gamma\beta$, kar je protislovje s točko 1. Torej mora x vsebovati vsaj dve ponovitvi γ . Označimo: $X = \{\alpha, \beta, \alpha\beta\alpha, \beta\alpha\beta\}$. Potem je x oblike:

$$x = p\gamma x_1\gamma x_2\gamma \dots \gamma x_m\gamma q$$

za $m \geq 1$ in $x_1, x_2, \dots, x_m \in X$ ter $qp \in X$. Če je $q = \varepsilon$, potem $p\gamma x_1\gamma x_2\gamma \dots \gamma x_m\gamma$ povzroči kvadrat² v u'' in enako za $p = \varepsilon$. Torej je $qp = \alpha\beta\alpha$ ali $qp = \beta\alpha\beta$. Brez škode za splošnost predpostavimo, da je $(q, p) = (\alpha, \beta\alpha)$. Glede na našo konstrukcijo, se w ne more začeti z $p\gamma = \beta\alpha\gamma$, torej w vsebuje faktor $qp\gamma x_1\gamma x_2\gamma \dots \gamma x_m\gamma$. V tem primeru u'' ponovno vsebuje kvadrat. To je protislovje, saj je u'' brezkvadratna. Torej w mora biti brezkvadratna.

Pokazali smo torej, da lahko za poljubno brezkvadratno besedo $u \in \Sigma_3^*$, ki se ogne aca in $bc b$, konstruiramo daljšo besedo $w \in \Sigma_3'^*$, ki se ogne $\alpha\gamma\alpha$ ter $\beta\gamma\beta$

■

S pomočjo ravnokar dokazanega bomo skonstruirali neskončno brezkvadratno besedo nad troelementno abecedo. Faktorju na začetku besede bomo dali posebno ime.

Definicija 11. Če se faktor nahaja na začetku besede, mu pravimo **predpona** ali **koren**.

Primer 12. Faktor „**kor**“ je koren besede „**kor**en“. Faktor „**en**“ ni koren besede „**kor**en“, saj se ne nahaja na začetku besede..

²Za podrobnosti glejte [3].

Izrek 13. (Thue, 1906) Obstaja neskončna brezkvadratna beseda nad trielementno abecedo. Torej obstaja takšno zaporedje besed $\{w_n\}_{n \geq 0}$, da je w_n koren w_{n+1} in je w_n brezkvadratna za vsak $n \geq 0$.

Dokaz. Naj bo u poljubna brezkvadratna beseda nad abecedo $\{a, b, c\}$, ki se ogne faktorjema aca in $bc b$ ter se začne s simboloma a ali b . Konstruirajmo novo besedo $\sigma(u)$ s pomočjo morfizma σ , definirane s predpisi:

$$\begin{aligned}\sigma(a) &= abac, \\ \sigma(b) &= babc, \\ \sigma(c) &= \begin{cases} bcac, & \text{če pred } c \text{ stoji } a, \\ acbc, & \text{če pred } c \text{ stoji } b. \end{cases}\end{aligned}$$

Če vzamemo na primer $u = abc$, dobimo $\sigma(u) = abac\ babc\ acbc$. Opazimo, da $\sigma(u)$ sovпада z besedo w iz dokaza prejšnjega izreka, kjer α, β, γ zamenjamo z a, b, c . Torej s pomočjo σ konstruiramo daljšo brezkvadratno besedo. Oglejmo si primer, ko je $w_0 = a$ in je za $n \geq 1$: $w_{n+1} = \sigma(w_n)$.

$$\begin{aligned}w_0 &= a \\ w_1 &= abac \\ w_2 &= abac\ babc\ abac\ bcac \\ w_3 &= abac\ babc\ abac\ bcac \dots babc\ acbc\ abac\ bcac \\ &\vdots\end{aligned}$$

Tako s predpisom $w_n = \sigma^n(a)$ dobimo daljšo brezkvadratno besedo, kjer je w_n koren od w_{n+1} ³ za vsak $n \geq 1$ in je a koren od w_0 . ■

4. Brezprekrivne besede in Thue-Morseove besede

Sedaj si oglejmo brezprekrivne besede.

Definicija 14. Beseda w je **brezprekrivna**, če se izogne vzorcu $xyxyx$.

Primer 15. Primer besede nad $\{a, b, c, d, e, f\}$, ki ni brezprekrivna je:

$$cccabcde\ fabcde\ fabcaaa.$$

Opazimo namreč:

$$ccc \underbrace{abc}_x \underbrace{def}_y \underbrace{abc}_x \underbrace{def}_y \underbrace{abc}_x aaa.$$

Primer brezprekrivnih besed nad dvoelementno abecedo so Thue-Morseove besede.

³Dejstvo, da je beseda w_n koren besede w_{n+1} je natančno utemeljeno v [3].

Definicija 16. (Thue-Morseova beseda)

Naj bo $A = \{a, b\}$ dvosimbolna abeceda. Definirajmo zaporedji besed $\{u_n\}_{n \geq 0}$ in $\{v_n\}_{n \geq 0}$ rekurzivno kot $u_0 = a$, $v_0 = b$ in $u_{n+1} = u_n v_n$, $v_{n+1} = v_n u_n$ za vsak $n \geq 0$:

$$\begin{array}{ll} u_0 = a & v_0 = b \\ u_1 = ab & v_1 = ba \\ u_2 = abba & v_2 = baab \\ u_3 = abbabaab & v_3 = baababba \\ \vdots & \vdots \end{array}$$

Besede u_n za $n \in \mathbb{N}$ imenujemo **Thue-Morseove besede**.

Opomba 17. Na enak način, kot smo konstruirali neskončno brezkvadratno besedo, lahko tu definiramo neskončno Thue-Morseovo besedo. Po definiciji je za vsak $n \geq 0$ namreč beseda u_n koren besede u_{n+1} .

V [3] je Thue dokazal, da so besede u_n brezprekrivne, poleg tega pa je pokazal tudi povezavo med brezprekrivnimi in brezkvadratnimi besedami.

Izrek 18. (Thue, 1912) Naj bo w neskončna brezprekrivna beseda nad abecedo $\{a, b\}$ in g substitucija na besedah nad trielementno abecedo $\Sigma_3 = \{a, b, c\}$ dana s predpisi:

$$g(a) = abb, \quad g(b) = ab, \quad g(c) = a.$$

Potem je w v sliki g in je vsaka beseda iz praslike $g^*(w)$ neskončna brezkvadratna beseda nad abecedo Σ_3 .

Torej je $g(w') = w$, kjer je w' brezkvadratna in w brezprekrivna. Imejmo besedo $w = abba baab baab abba baab abba abba baab$. Uporabimo izrek 18, da dobimo brezkvadratno besedo $w' = abcacbabcbac...$

5. Izogibnost in neizogibnost vzorcev

Uvedimo nekaj terminologije.

Definicija 19. Vzorec p je **q-izogiben**, če za vsak $n \in \mathbb{N}$ obstaja beseda dolžine več kot n nad q -elementno abecedo, ki se izogne vzorcu p . Vzorec je **q-neizogiben**, če vse dovolj dolge besede nad q simboli vsebujejo vzorec p . Vzorec p je **neizogiben**, če je q -neizogiben za vsak $q \in \mathbb{N}$ in je **izogiben**, če obstaja $q \in \mathbb{N}$, da je p q -izogiben.

Zgornjo definicijo ponazori naslednji primer.

Primer 20. V primeru 8 smo razmislili, da že vse besede dolžine štiri nad abecedo $\{0, 1\}$ vsebujejo kvadrat. Torej tudi vse daljše besede nad dvoelementno abecedo vsebujejo kvadrat. Za trielementno abecedo to po Thuejevem izreku ne drži. Zato je vzorec xx 2-neizogiben, je pa 3-izogiben. Torej xx ni neizogiben, je pa izogiben.

6. Besede Zimina in ireducibilnost

6.1 Besede Zimina

Spoznali bomo besede Zimina, ki so primeri neizogibnih vzorcev.

Definicija 21. Besede Zimina so besede nad abecedo $\{x_1, x_2, \dots, x_N\}$, definirane rekurzivno kot:

$$\begin{aligned} Z_1 &= x_1 \\ Z_2 &= Z_1 x_2 Z_1 = x_1 x_2 x_1 \\ &\vdots \\ Z_n &= Z_{n-1} x_n Z_{n-1} \end{aligned}$$

Dokazali bomo, da besede Zimina podajajo neizogibne vzorce.

Lema 22. Vzorec $Z_2 = xyx$ je neizogiben.

Dokaz. Imejmo poljubno abecedo Σ nad k simboli. Naj bo $w \in \Sigma^*$ beseda dolžine $2k + 1$. Potem mora biti (po posplošenem Dirichletovem načelu) nek simbol ponovljen vsaj trikrat in torej lahko besedo w faktoriziramo kot:

$$w = w_1 a w_2 a w_3 a w_4.$$

Vzorec xyx se pojavi s substitucijo $h(x) = a$ in $h(y) = w_2 a w_3$

■

Lema 23. Če je p neizogiben v abecedi Σ in je $\alpha \in \Sigma$ simbol, ki se ne pojavi v p , potem je tudi vzorec $p\alpha p$ neizogiben v Σ .

Dokaz. Naj bo $|\Sigma| = k$. Ker je p neizogiben nad Σ , obstaja $l \in \mathbb{N}$, da vsaka beseda $u \in \Sigma^*$ dolžine vsaj l vsebuje vzorec p . Sedaj si oglejmo besedo w dolžine $l(k^l + 1) + k^l$ nad abecedo Σ .

$$w = w_1 s_1 w_2 s_2 \dots w_{k^l} s_{k^l} w_{k^l+1},$$

kjer je $|w_i| = l$ za vse $i \in \{1, 2, \dots, k^l + 1\}$ in s_i so posamezni simboli za vse $i \in \{1, 2, \dots, k^l\}$. Obstaja natanko k^l paroma različnih besed dolžine l , torej morata obstajati $i \neq j$, da je $w_i = w_j$. Ker sta w_i in w_j ločena z vsaj enim simbolom, w vsebuje vzorec $p\alpha p$.

■

Izrek 24. (Zimin, 1984) Vse besede Zimina Z_n , $n \geq 1$ so neizogibne.

Dokaz. Vzorec $Z_1 = x$ je trivialno neizogiben. Lema 22 nam pove, da je vzorec $Z_2 = xyx$ neizogiben. Denimo, da je vzorec Z_{n-1} neizogiben. Po lemi 23 je tudi vzorec $Z_{n-1}xZ_{n-1}$ neizogiben. Torej je vzorec Z_n neizogiben.

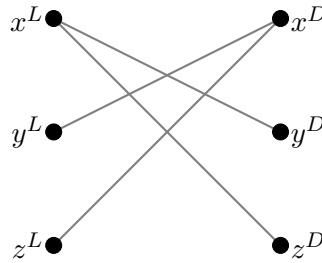
■

6.2 Ziminov redukcijski algoritem

Sedaj bomo s pomočjo **Ziminovega algoritma** definirali ireducibilnost vzorcev. S tem pojmom si bomo pomagali pri karakterizaciji neizogibnih vzorcev.

Naj bo $p \in \Sigma^*$ vzorec. Konstruirajmo **graf sosednjosti vzorca p** , $G(p)$ na naslednji način:

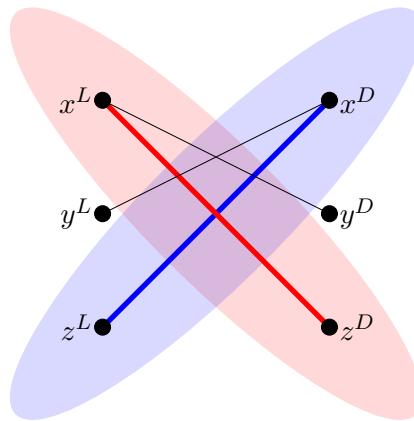
- Za vsak simbol $x \in \Sigma$ definiramo vozlišči x^L in x^D . Naj bo Σ^L leva množica vozlišč in naj bo Σ^D desna množica vozlišč. Z drugimi besedami, elementi Σ^L so natanko x^L in elementi Σ^D so natanko x^D , ko x preteče Σ . Množici Σ^L in Σ^D tvorita dvodelni razcep množice vozlišč grafa $G(p)$.
- Za simbola $x, y \in \Sigma$, vozlišči x^L in y^D povežemo natanko tedaj, ko je xy faktor v p .



Slika 3. Graf sosednjosti $G(xyxzx)$ vzorca $xyxzx$ nad abecedo $\Sigma = \{x, y, z\}$.

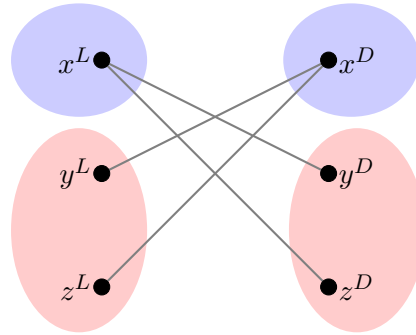
Definicija 25. Neprazna podmnožica $F \subset \Sigma$ je **prosta**, če za $x, y \in F$, ne obstaja pot iz x^L v y^D .

Primer 26. Podmnožica $\{x, z\}$ ni prosta podmnožica $\{x, y, z\}$, saj med x^L in z^D ter med z^L in x^D potekata povezavi v $G(xyxzx)$.



Slika 4. Podmnožica $\{x, z\}$ ni prosta podmnožica $\{x, y, z\}$, saj med x^L in z^D ter med z^L in x^D potekata povezavi.

Primer 27. Podmnožici $\{x\}$ in $\{y, z\}$ sta prosti glede na $G(xyxzx)$, saj, kot vidimo na sliki 5, ne obstaja pot iz x^L v x^D . Prav tako ne obstaja pot iz y^L v y^D ali z^D ter ne obstaja pot iz z^L v y^D ali z^D .



Slika 5. Med x^L in x^D ter med $\{y^L, z^L\}$ in $\{y^D, z^D\}$ ni poti. Zato sta $\{x\}$ in $\{y, z\}$ prosti podmnožici $\{x, y, z\}$.

S pomočjo vzorca p in proste podmnožice F ustvarimo novi vzorec p' tako, da iz p izbrišemo vse simbole iz F . Pravimo, da p v **enem koraku reduciramo** do p' . Če obstaja zaporedje enokoračnih redukcij iz p v vzorec q , pravimo, da lahko p **reduciramo** do q .

Definicija 28. Vzorec p je **reducibilen**, če ga lahko reduciramo do ε in je **ireducibilen**, če ni reducibilen.

Izrek 29. (Zimin, 1984) Vzorec je izogiben natanko tedaj, ko je ireducibilen.

Dokaz tega izreka se nahaja v [4].

S pomočjo tega izreka zlahka vidimo, da je kvadrat izogiben vzorec. Zadošča namreč videti, da je vzorec xx ireducibilen. Zato si narišimo graf sosednosti za vzorec xx . Graf sosednosti $G(xx)$ bo imel dve vozlišči, x^L in x^D . Ker je xx faktor v xx (trivialno), sta x^L in x^D povezana v $G(xx)$. Zato $\{x\}$ ni prosta podmnožica, torej vzorca xx ne moremo reducirati do prazne besede ε .



Slika 6. Podmnožica $\{x\}$ ni prosta podmnožica $\{x\}$, saj med x^L in x^D poteka povezava v $G(xx)$.

Da dokažemo ireducibilnost nekega vzorca, se moramo prepričati, da nobeno zaporedje enokoračnih redukcij ne vodi do praznega vzorca ε . Ta rekurzivni algoritem bomo imenovali **Ziminov algoritem**.

Opomba 30. Preprosta vaja za bralca je pokazati, da je končna redukcija vedno enaka besedi, ki jo dobimo, če iz vzorca p pobrišemo vse simbole iz unije vseh prostih množic. Po vsaki enokoračni redukciji tudi vsem prostim množicam odstranimo elemente, ki smo jih odstranili v redukciji.

Primer 31. Vzorec $xyxzx$ je reducibilen. Ker je $\{x\}$ prosta podmnožica $\{x, y, z\}$, lahko vzorec najprej reduciramo do yz . Ker je tudi $\{y, z\}$ prosta podmnožica $\{x, y, z\}$, lahko nato yz reduciramo do praznega vzorca ε .

7. Stolpične ocene

Izrek 32. (Zimin, 1984) Vzorec je neizogiben natanko tedaj, ko je vsebovan v besedi Zimina.

Dokaz tega izreka se nahaja v [4].

Oglejmo si kvantitativne ocene za vzorce, vsebovane v besedah Zimina. Zaradi gnezdenih eksponentov, ki se v teh ocenah pojavljajo, jih bomo imenovali stolpične ocene.

Definicija 33. Za fiksno besedo Zimina Z_n definiramo $f(n, q)$ kot najmanjše takšno število, da vsaka beseda dolžine $f(n, q)$ nad q simboli vsebuje vzorec Z_n .

Lema 34. $f(n+1, q) \leq (f(n, q) + 1) \cdot (q^{f(n, q)} + 1) - 1$

Dokaz. Oglejmo si naslednjo besedo:

$$w = \underbrace{ab \dots cx}_{f(n, q)} \underbrace{de \dots fy}_{f(n, q)} \underbrace{zs \dots t}_{f(n, q)}.$$

$\underbrace{\hspace{15em}}_{q^{f(n, q)} + 1}$

Beseda w je sestavljena iz $q^{f(n, q)} + 1$ besed dolžine $f(n, q)$, ki so med sabo ločene z natanko enim simbolom. Beseda w je dolžine $(f(n, q) + 1) \cdot (q^{f(n, q)} + 1) - 1$. Vsaka od manjših podbesed mora po definiciji vsebovati vzorec Z_n . Ker obstaja $q^{f(n, q)}$ besed dolžine $f(n, q)$ nad q simboli, po Dirichletovem načelu w vsebuje dve kopiji iste besede, ločeni z vsaj enim simbolom. Torej w mora vsebovati vzorec Z_{n+1} . ■

Izrek 35. (Cooper, Rorabaugh, 2014, [7]) Za vsak $n \geq 1$ in $q \geq 1$ ter $Q := 2q + 1$ velja:

$$f(n, q) \leq Q^{Q^{Q^{\dots^Q}}},$$

kjer je eksponentov nad Q natanko $n - 1$.

Dokaz. Dokažimo z indukcijo po n . Bazni primer je $n = 1$. Vsaka neprazna beseda očitno vsebuje vzorec $Z_1 = x$, torej je $f(1, q) = 1 \leq 2q + 1 = Q$.

Predpostavimo sedaj, da trditev drži za n in označimo $T = f(n, q)$. Lema 34 pove, da je $f(n+1, q) \leq (T+1)(q^T + 1) - 1$. Razmisliti še moramo, da je $(T+1)(q^T + 1) - 1 \leq Q^T = (2q+1)^T$.

Desno stran neenačbe razvijemo po binomskem izreku: $(2q+1)^T = \sum_{k=0}^T \binom{T}{k} (2q)^k = (2q)^T + \sum_{k=0}^{T-1} \binom{T}{k} (2q)^k$. Ocenimo: $\sum_{k=0}^{T-1} \binom{T}{k} (2q)^k \geq \sum_{k=0}^{T-1} \binom{T}{k} = 2^T - 1$. Torej je

$$(2q+1)^T \geq 2^T q^T + 2^T - 1.$$

Sedaj ocenimo še levo stran neenačbe. Za naravno število T je $T+1 \leq 2^{T/4}$. Zato je

$$(T+1)(q^T + 1) - 1 \leq 2^T (q^T + 1) - 1 = 2^T q^T + 2^T - 1.$$

Torej je $(T+1)(q^T + 1) - 1 \leq (2q+1)^T = Q^T$, kar smo želeli dokazati. ■

Lema 36. (Ritter, Shur, 2014, [6]) $f(3, q) \leq 2^{q+1} \cdot (q+1)!$

Definicija 37. Pravimo, da je beseda w **2-minimalna**, če vsebuje Z_2 , vendar se vsi njeni pravi faktorji⁵ ognejo Z_2 .

⁴To se dokaže s preprosto indukcijo. Prepuščamo bralcu za zabavo.

⁵Pravi faktorji so faktorji, ki niso enaki celotni besedi.

Dokaz. Ugotovimo, da mora biti 2-minimalna beseda nujno oblike aaa ali $ab_1^{j_1}b_2^{j_2}\dots b_r^{j_r}a$, kjer so $b_i \neq a$ paroma različni in je za vsak $i : j_i \in \{1, 2\}$. Sledi, da je število 2-minimalnih besed nad q -simbolno abecedo končno. Naj bo $t(2, q)$ število 2-minimalnih besed nad q -simbolno abecedo. Obstaja q besed prve oblike in za nek dan r je $q \cdot (q-1) \cdot (q-2) \cdot \dots \cdot (q-r) \cdot 2^r$ besed druge oblike. Faktor 2^r se pojavi, ker imamo za vsak $i \in \{1, \dots, r\}$ dve izbiri za eksponent i .

Sledi:

$$t(2, q) = q + \sum_{r=1}^{q-1} q \cdot (q-1) \cdot (q-2) \cdot \dots \cdot (q-r) \cdot 2^r \leq q! \sum_{r=0}^{q-1} 2^r = q! \cdot (2^q - 1).$$

Sedaj si oglejmo besedo:

$$w = \underbrace{ab \dots cx}_{f(2, q)} \underbrace{de \dots fy}_{f(2, q)} \underbrace{zs \dots t}_{f(2, q)}.$$

$t(2, q) + 1$

Vsaka beseda dolžine $f(2, q)$ mora vsebovati vsaj eno 2-minimalno besedo. Po Dirichletovem načelu morata biti dve izmed teh $t(2, q) + 1$ besed enaki. Torej beseda w vsebuje vzorec $Z_3 = xyx$, kjer ponovljeni besedi gradita x in simboli med njima gradijo y .

Iz dokaza izreka 35 se spomnimo, da je $f(2, q) \leq 2q + 1$, od koder dobimo:

$$f(3, q) \leq (f(2, q) + 1) \cdot (t(2, q) + 1) - 1 \leq (2q + 2)2^q q! = 2^{q+1}(q + 1)!$$

■

Izrek 38. (Conlon, Fox, Sudakov, 2017, [5]) Za vsaka $n \geq 3$, $q \geq 35$:

$$f(n, q) \leq q^{q^{q^{\dots^{q-o(q)}}}},$$

kjer je eksponentov nad q natanko $n - 1$ in je člen $o(q)$ neodvisen od n ter $\lim_{q \rightarrow \infty} o(q) = 0$.

Dokaz tega izreka se nahaja v [5].

Vsi ti rezultati nam dajo oceno minimalne dolžine besede, ki gotovo vsebuje vzorec Z_n . Vendar to ni indikativno za večino besed. Naslednji izrek, katerega dokaz se nahaja v [1] na strani 8 oziroma v [5], določa prag, nad katerim se bo Z_n pojavil v večini besed.

Izrek 39. (Conlon, Fox, Sudakov, 2017, [5])

Pri fiksnem $n \in \mathbb{N}$ in $q \rightarrow \infty$ je prag dolžine za pojavitev Z_n v naključni besedi nad q simboli $q^{2^{n-1}-(n+1)/2}$. To pomeni, da ko $q \rightarrow \infty$, se verjetnost pojavitve vzorca Z_n v besedi dolžine $q^{2^{n-1}-(n+1)/2}$ približuje 1.

Opomba 40. Nad slovensko abecedo, kjer je $q = 25$, bomo po izreku 39 z veliko verjetnostjo našli vzorec Z_3 v besedi dolžine $25^{2^{(3-1)}-(3+1)/2} = 625$, ampak ocena za minimalno dolžino besede⁶, ki gotovo vsebuje Z_3 , je približno 10^{34} , kar je neverjetnih 31 velikostnih redov razlike.

⁶Po lemi 36.

LITERATURA

- [1] Wu, Menghua. *Survey of Combinatorics on Words and Patterns*. 2018.
- [2] Lothaire, M. *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications, vol. 90, 2002.
- [3] Thue, Axel. *Über unendliche Zeichenreihen*. Kongelige Norske Videnskabers Selskabs Skrifter. 1912.
- [4] Zimin, A. I. *Blocking sets of terms*. Mathematics of the USSR-Sbornik. 1984. Vol. 47. Strani 353-364.
- [5] Conlon, David. Fox, Jacob. Sudakov, Benny. *Tower-type bounds for unavoidable patterns in words*. Transactions of the American Mathematical Society. 2017.
- [6] Ritter, Wojciech in Shur, Arseny M. *On searching zimin patterns*. CoRR abs/1409.8235. 2014
- [7] Cooper, Joshua in Rorabaugh, Danny. *Bounds on Zimin Word Avoidance*. ArXiv e-prints. 2014.