PRO GRADU -TUTKIELMA

Tuomas Neulaniemi

Bar-Hillelin ja Ogdenin lemmoista

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tietojenkäsittelytieteiden tutkinto-ohjelma
Syyskuu 2015

Tampereen yliopisto
Informaatiotieteiden yksikkö
NEULANIEMI, TUOMAS: Bar-Hillelin ja Ogdenin lemmoista
Pro gradu -tutkielma, 59 sivua
Syyskuu 2015

Tiivistelmä

Tässä pro gradu -tutkielmassa tutkimme formaalien kielten teoriaa, erityisesti Bar-Hillelin ja Ogdenin lemmoja, joita kutsutaan pumppauslemmoiksi. Tutkimme lisäksi lineaarisille ja ei-lineaarisille kontekstittomille kielille tarkoitettua lemmaa. Ogdenin todistusta seuraa luontaisen moniselitteisyyden käsittely. Tutkimme vaihtolemmaa ja Parikhin lausetta, joita seuraa vahvojen iterointilemmojen esittely. Lopuksi katsastamme lyhyesti lemmojen tietojenkäsittelytieteellisiä sovelluksia.

Avainsanat ja -sanonnat: kontekstittomat kielet, pumppauslemma, Bar-Hillelin lemma, Ogdenin lemma, luontainen moniselitteisyys.

Sisältö

1 Johdanto **1**

2 Määritelmiä ja merkintöjä **2**

3 Pumppauslemma säännöllisille kielille **3**
3.1 Esimerkkejä 4
3.2 Ehrenfeuchtin, Parikhin ja Rozenbergin lemma 5

4 Bar-Hillelin lemma kontekstittomille kielille **13**
4.1 Esimerkkejä 15
4.2 Bar-Hillelin lemman toteuttavista kielistä 19

5 Lemmat lineaarisille ja ei-lineaarisille kontekstittomille kielille **22**

6 Ogdenin lemma **25**
6.1 Luontainen moniselitteisyys 28
6.2 Ogdenin lemman toteuttavista kielistä 33
6.3 Ogdenin lemman yleistäminen 34

7 Vaihtolemma **38**

8 Parikhin lause **43**

9 Baderin ja Mouran lemman yleistäminen vahvoilla iterointilemmoilla **47**

10 Tietojenkäsittelytieteellisiä sovelluksia **53**

11 Yhteenveto **56**

Viiteluettelo **58**

1 Johdanto

Kielitieteilijä Noam Chomsky tutki 1950-luvulla luonnollisten kielten syntaksia ja löysi kuvailevassa kielentutkimuksessa ongelmia, sillä luonnollisessa kielessä ei ollut riittävän selkeitä kielioppeja. Chomsky totesi, että on löydettävä lingvistisen rakenteen yleinen teoria ja kielen säännöt. Chomsky formalisoi generatiiviset kieliopit ja niistä hän loi neljä kategoriaa käsittävän kielihierarkian (Chomskyn kielihierarkia), jossa ovat seuraavat kielikategoriat: Luokka $\mathscr{L}_0$ on rekursiivisesti numeroituvat kielet, joihin sisältyvät kontekstiset ($\mathscr{L}_1$) kielet, joihin sisältyvät kontekstittomat ($\mathscr{L}_2$) kielet ja joihin sisältyvät säännölliset kielet, jotka muodostavat Chomskyn hierarkian luokan $\mathscr{L}_3$. Chomskyn kielihierarkia on otettu käyttöön tietojenkäsittelyteoriassa. Tietojenkäsittelytieteellisissä sovelluksissa tarvitaan yleensä kontekstittomia ja säännöllisiä kielioppeja ja kieliä.

Bar-Hillelin lemmaa (vuonna 1961) kontekstittomille kielille edelsi (vuonna 1959) vastaava lemma säännöllisille kielille, jolla oli osoitettavissa jokin formaali kieli eisäännölliseksi. Säännöllisille kielille tarkoitetulla lemmalla pumpataan merkkijonoja yhdestä kohdasta. Bar-Hillelin lemmassa merkkijonojen pumppautuvuus laajeni yhdestä kohdasta kahteen kohtaan. Tämän jälkeen oli mahdollista todistaa kontekstittomien kielten perheen aito sisältyminen kontekstisten kielten perheeseen. Kontekstiton tarkoittaa, että kontekstittomassa kieliopissa säännöt noudattavat muotoa $A \to \alpha$, missä A on apumerkki ja α on merkkijono. Sääntöjen muodosta seuraa, että niitä sovelletaan aina yhteen apumerkkiin kerrallaan. Johdot ovat siis "kontekstittomia".

Ogdenin lemma esitettiin vuonna 1968 tarkoituksena vahvistaa Bar-Hillelin jälkimmäistä lemmaa, jonka todettiin olevan riittämätön osoittamaan joitakin kieliä kontekstittomiksi. Ogdenin lemma mahdollisti lisäksi joidenkin kontekstittomien kielten todistamisen luontaisesti moniselitteiseksi, mikä tarkoittaa, että kieltä ei voida tuottaa yksiselitteisellä kontekstittomalla kieliopilla.

Esittelemme aluksi säännöllisille joukoille tarkoitetun lemman (luku 3). Seuraavaksi (luku 4) siirrymme kontekstittomiin kieliin tutkimalla sekä Bar-Hillelin että Ogdenin lemmoja. Tutkimme lisäksi, millä tavoin Bar-Hillelin lemmaa on muotoiltu (luku 4) vahvemmaksi. Lisäksi käsittelemme (luku 4) tapausta, jossa Bar-Hillelin lemma ei ole riittävä osoittamaan, että kieli ei ole kontekstiton. Mainitut lemmat antavat välttämättömiä ehtoja tiettyihin kieliperheisiin kuulumiselle. Tarkastelemme lisäksi (luku 5) lineaarisille ja ei-lineaarisille kontekstittomille kielille tarkoitettua lemmaa. Ogdenin todistusta seuraa (luku 6) luontaisen moniselitteisyyden käsittely. Tarkastelemme vaihtolemmaa (luku 7) ja Parikhin lausetta (luku 8). Esittelemme myös vahvoja iterointilemmoja (luku 9). Lopuksi (luku 10) katsastamme lemmojen tietojenkäsittelytieteellisiä sovelluksia.

2 Määritelmiä ja merkintöjä

Tässä luvussa esittelemme käyttämiämme määritelmiä ja merkintöjä.

Sanaa merkitään w, sanan pituutta |w|. Tyhjää lausetta merkitään λ. Luonnollisten lukujen joukkoja merkitään $\mathbb{N}$ ja alkulujen joukkoa $\mathbb{P}$. Eulerin funktiota merkitään ϕ.

Tarkastellaan aakkostoa Σ. Aakkoston Σ kaikkien mahdollisten merkkijonojen joukkoa merkitään Σ^*. Aakkoston Σ ei-tyhjien merkkijonojen joukkoa merkitään Σ^+.

Kontekstiton kielioppi on järjestelmä $G = (N,T,P,S)$, missä N on apumerkkien aakkosto, T on perusmerkkien aakkosto, P on aakkostossa $N \cup T$ määritelty äärellinen joukko sääntöjä, joista jokainen on muotoa $A \rightarrow \alpha$, missä A on apumerkki ja α on joukon $(N \cup T)^*$ merkkijono, ja S tarkoittaa joukon N erillistä alkiota, aloitusmerkkiä.

Jos $A \rightarrow \beta$ on sääntöjoukon P sääntö ja sekä α että γ ovat joukon $(N \cup T)^*$ merkkijonoja, niin kieliopissa G on voimassa $\alpha \mathrm{A} \gamma \Rightarrow \alpha\beta\gamma$. Sääntöä $A \rightarrow \beta$ sovelletaan merkkijonoon $\alpha A\gamma$ merkkijonon $\alpha\beta\gamma$ tuottamiseksi; sanotaan myös että $\alpha A\gamma$ johtaa suoraan merkkijonon $\alpha\beta\gamma$ kieliopissa G.

Oletetaan, että $\alpha_1, \alpha_2, \ldots, \alpha_m$ ovat merkkijonoja $\in (N \cup T)^*$, $m \geq 1$, ja $\alpha_1 \Rightarrow \alpha_2$, $\alpha_2 \Rightarrow \alpha_3, \ldots, \alpha_m - 1 \Rightarrow \alpha_m$. Tällöin merkitsemme $\alpha_1 \Rightarrow^* \alpha_m$ ja sanomme, että α_1 tuottaa merkkijonon α_m kieliopissa G, tai että α_m voidaan tuottaa merkkijonosta α_1, jolloin derivaatiopolku siis kulkee merkkijonojen $\alpha_2, \ldots, \alpha_{m-1}$ kautta. Merkintä $\Rightarrow$ tarkoittaa siis derivaatiota, eli johtaa-relaatiota, $\Rightarrow^+$ merkitsee relaation $\Rightarrow$ transitiivista sulkeumaa ja $\Rightarrow^*$ sen transitiivista refleksiivistä sulkeumaa.

Jos on voimassa $S \Rightarrow^* \delta$, niin δ on kieliopin lausejohdos. Kieliopin G kaikkien lausejohdosten joukosta käytetään merkintää $SF(G)$. Kontekstittoman kieliopin tuottama kieli määritellään seuraavasti: Olkoon $G = (N,T,P,S)$ kontekstiton kielioppi. Kieliopin G tuottama kieli on $L(G) = \{w \in T^* \mid S \Rightarrow_G^* w\}$. Kieli L on kontekstiton, jos on olemassa sellainen kontekstiton kielioppi G, että $L = L(G)$. Kun relaatio $\Rightarrow$ pätee erityisesti kieliopissa G, merkitään se muodossa $\Rightarrow_G$.

Kieli on säännöllinen, jos se on tuotettavissa säännöllisellä kieliopilla. Säännöllisen kieliopin säännöt noudattavat seuraavia muotoja: i) $A \rightarrow$ a, ii)$A \rightarrow$ aB,
iii) $A \rightarrow \lambda$, missä $A, B \in N$ ja a $\in T$.

Kieli on lineaarinen, jos on olemassa sellainen kielioppi, jossa jokainen sääntö on muotoa $A \rightarrow$ v tai $A \rightarrow vBw$, missä $A, B \in N$ ja v,w $\in T$*, ja joka tuottaa kyseisen kielen. Tällaista kielioppia sanotaan lineaariseksi. Jos kontekstitonta kieltä ei voida tuottaa lineaarisella kieliopilla, niin kieli on ei-lineaarinen.

3 Pumppauslemma säännöllisille kielille

Tässä luvussa esittelemme pumppauslemman, äärellisen automaatin ja kaksi esimerkkiä pumppauslemman käytöstä, joista ensimmäisessä alkuluvut esitetään binäärisinä ja toisessa merkkijonon pituus esitetään täydellisenä neliönä. Luvun lopuksi esittelemme Ehrenfeuchtin, Parikhin ja Rozenbergin lemman, joka karakterisoi säännölliset kielet eräänlaisen pumppauksen avulla.

Pumppauslemma on tehokas apuväline joidenkin kielten ei-säännölliseksi tai eikontekstittomaksi todistamisessa. Lisäksi se on tehokas apuväline äärellisiin automaatteihin liittyvissä päätösongelmissa, kuten onko annetun automaatin hyväksymä kieli äärellinen vai ääretön. [Hopcroft and Ullman, 1979]

Äärellinen automaatti on järjestelmä $A = (Q, \Sigma, \delta, q_o, F)$, missä
Q on äärellinen tilojen joukko,
Σ on äärellinen syötemerkkien aakkosto,
δ on siirtymärelaatio ($\delta : Q \times \Sigma \to Q$),
q_o on alkutila, ($q_o \in Q$), ja
F on lopputilojen joukko. ($F \subseteq Q$).

Mikäli kieli on säännöllinen, niin se on edellä kuvatun kaltaisen automaatin hyväksymä. Äärelliset automaatit siis hyväksyvät saman kieliperheen, jonka säännölliset kieliopit tuottavat.

Tarkastellaan n-tilaista äärellistä automaattia ja sen syötemerkkijonoa $a_1 a_2 \dots a_m$, m > n, jolla pätee $\delta(q_o, a_1 a_2 \dots a_i) = q_i$, kun i = 1,2,... m. Kun automaatin tilojen lukumäärä on vain n, niin on sellaiset kokonaisluvut j ja k, $0 \leq j < k \leq n$, että $q_i = q_k$. Koska j < k, merkkijono $a_{j+1} \dots a_k$ on pituudeltaan vähintään 1, ja koska k ≤ n, sen pituus on korkeintaan n. Tällöin myös $a_1 a_2 \dots a_j a_{k+1} a_{k+2} \dots a_m \in L$, koska on olemassa polku $q_o \to q_m$, joka kulkee tilan q_j kautta, mutta ei silmukan ympäri, johon liittyy merkkijono $a_{j+1} \dots a_k$. [Hopcroft and Ullman, 1979]. Vastaavasti silmukka voidaan kiertää mielivaltaisen monta kertaa, joten automaatin hyväksymään kieleen kuuluvat kaikki lauseet $a_1 \dots a_j (a_{j+1} \dots a_k)^n a_{k+1} \dots a_m$, n ≥ 0. Näin ollaan todistettu seuraava lause.

Lause 3.0.1 (Pumppauslemma säännöllisille kielille, pumping lemma).
Olkoon L säännöllinen kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k \in \mathbb{N}$, että kun $|z| > k$, niin sana z voidaan kirjoittaa muodossa $z = uvw$ siten, että
1. $|v| \geq 1$.
2. $|uv| \leq k$.
3. $uv^i w \in L, i \geq 0$.

Määritelmä 3.0.2 Olkoot $L \subseteq \Sigma^*$, x $\in \Sigma^*$ ja x = uvw. Silloin v on pumppaus osasanalle x suhteessa kieleen L jos ja vain jos, kaikille $i \geq 0$, $uv^iw \in L$.

3.1 Esimerkkejä

Seuraavaksi osoitamme käyttämällä lausetta 3.0.1, että seuraavissa esimerkeissä olevat kielet eivät ole säännöllisiä.

Esimerkki 3.1.1. [Shallit, 2008]
Olkoon kieli $L = \{1, 11, 101, 111, 1011, 1101, 10001, 10011, 11001, \dots\}$, missä alkuluvut esitetään binäärisinä. Kieli on esitettävissä myös muodossa $L = \{w \in \mathbb{P} | w \in \{0,1\}^*\}$, missä binääriesitykset alkavat 1-bitillä. Alkuluvuista esimerkiksi 71 on binäärisenä 1000111. Kieleen kuulumisen ehtona luvun on siis oltava alkuluku, joten esimerkiksi 11110100 $\notin L$. Osoitamme, että kieli L ei ole säännöllinen.

Tarvitsemme kaksi tulosta lukuteorian alueelta, joista ensimmäisen mukaan alkulukujen määrä on ääretön. Toisen tuloksen eli Fermatin pienen lauseen mukaan a^{p-1} ≡ 1 mod p, missä a on luonnollinen luku, p on alkuluku ja luku a ei muodostu alkuvun p monikerroista. Voimme esimerkiksi valita a=2 ja p=17, jolla saamme $2^{17-1} \equiv$ 1 mod 17. Havaitsemme pätevän, että 2^{p-1}(p - 1)! ≡ (p - 1)! mod p. [Rosen, 2010]

Oletamme, että kieli L on säännöllinen. Olkoon k lauseessa 3.0.1 annettu kokonaisluku. Olkoon p sellainen alkuluku, että p > 2^k. Olkoon z binääriesitys alkuluvulle p. Lauseen 3.0.1 nojalla z voidaan kirjoittaa muodossa z=uvw, missä |v| ≥ 1 ja uv^iw on alkuluvun binääriesitys kaikille i. Olkoot n_u, n_v ja n_w osasanojen u, v ja w arvoja, jotka merkitsevät binäärilukuja. Koska sanan z binääriesitys alkaa 1-bitillä, niin n_u ≥ 1. Jos u = w = λ, niin $n_u = n_w = 0$. Valitsemme i = p, jolloin saamme uv^pw - uvw ≡ 0 mod p. [Shallit, 2008]. Nyt uv^pw on alkuluvun q binääriesitys. Numeerinen arvo luvulle q on $n_u 2^{|w|+p|v|} + n_v\, 2^{|w|}(1+2^{|v|}+ \dots + 2^{(p-3)|v|} + 2^{(p-2)|v|} + 2^{(p-1)|v|}) + n_w$.

Binääriesityksen alkamisesta 1-bitillä vastaa siis n_u, mutta sitä vastoin n_v ja n_w alkavat 0- tai 1-bitillä. Pumppauksen kohteena on n_v. [Hopcroft and Ullman, 1979].

Geometrisen sarjan summan kautta lausekkeen $n_u 2^{|w|+p|v|} + n_v\, 2^{|w|}(1 + 2^{|v|}+ \dots$ + $2^{(p-3)|v|} + 2^{(p-2)|v|} + 2^{(p-1)|v|}) + n_w$ keskiosa $(2^{|v|}+ \dots + 2^{(p-3)|v|} + 2^{(p-2)|v|} +$ $2^{(p-1)|v|}) = 2^{|w|}\ \frac{2^{p|v|}-1}{2^{|v|}-1}$, missä p ja v eivät ole vakioita. Valitsemalla p=5 saamme $\frac{2^{5|v|}-1}{2^{|v|}-1} \equiv 1$ mod 3. Yhtälöstä on todettavissa, että $\frac{2^{p|v|}-1}{2^{|v|}-1} \not\equiv 1$ mod p. [Shallit, 2008]

Sarjan summa on siis ∞. Meidän on selvitettävä, päteekö q>p>1. Mikäli pätee, niin tiedämme, että q ei voi olla alkuluku. Fermatin pienen lauseen nojalla $2^{p-1} \equiv 1$ mod p. Korottaessamme molemmat puolet potenssiin |v|, saamme $2^{(p-1)|v|} \equiv 1$ mod p. Täten $2^{p|v|} = 2^{(p-1)|v|}\ 2^{|v|} \equiv 2^{|v|}$ mod p. [Hopcroft and Ullman, 1979]. Pätee myös, että $2^{(p!-1)} + (2^{|v|} / 2) = 2^{(p+|v|)-1} + 2 \equiv 1$ mod p. Olkoon s = $1+ 2^{|v|}+ \dots + 2^{(p-1)|v|}$, jonka voimme merkitä $\sum_{p=1}^{v} 2^{p-1} = \frac{1-2^{v+1}}{1-2} = 2^{v+1} - 1$.

Nyt $(2^{|v|})$s = $2^{p|v|}$ -1, joten $(2^{|v|}$- 1)(s-1) on jaettavissa luvulla p, eli p ≡ 0 mod ($2^{|v|}$ - 1)(s - 1). Mutta 1 ≤ |v| ≤ n, joten 2 ≤ $2^{|v|}$ ≤ 2^k < p, jonka seurauksena p ei voi jakaa lukua $2^{|v|}$ - 1, ja koska $2^{|v|}$ - 1 > p, niin $2^{|v|}$ - 1 jakaa lukua s -1. Siis s ≡ 1 mod p = p ≡ 0 (mod s - 1). Myöskään p ei voi jakaa lukua $2^{|v|!}$ - 1. [Shallit, 2008]

Koska s ei jaa lukua p, niin $2^s - 1$ ei jaa lukua $2^p - 1$ [Rosen, 2010]. Mutta kuitenkin uv^pw = q = $n_u 2^{|w|+p|v|}$ + n_v $2^{|w|}$ s + n_w, joten q ≡ $n_u 2^{|w|+|v|}$ + n_v $2^{|w|}$ + n_w mod p. On voimassa, että p! ≢ 1 mod p. Koska myös pätee, että p^2 ≡ 0 mod p, niin neliönjäännöksiä ei voi syntyä. Päädymme siihen, että q ≡ p mod p. Koska sanassa z ei ole 0-alkuisia binääriesityksiä, niin $uv^pw > uvw$ = p. [Shallit, 2008]

Yrityksemme oli todeta pumppauslemmaa käyttämällä kaikki binääriesitettävät luvut alkuluvuiksi. Koska q > p > 1, niin q ei voi olla alkuluku, mikä merkitsee ristiriitaa. [Hopcroft and Ullman, 1979]. Samoin ristiriita ilmenee siten, että $uv^pw \notin L$. Täten kieli L ei ole säännöllinen. [Shallit, 2008]

Esimerkki 3.1.2. [Hopcroft et al., 2005]
Olkoon kieli $L = \{0^j \mid$ j on täydellinen neliö$\}$. Oletetaan, että kieli on säännöllinen. Valitaan $z = 0^{k^2}$. Esimerkiksi luvut 49 ja 121 ovat täydellisiä neliöitä. Kirjoitettaessa z=uvw, tiedämme, että osasana v muodostuu 0-merkeistä, joiden lukumäärä on välillä 1 - j. Täten merkkijonon uvvw pituus on välillä $k^2 + 1$ ja $k^2 + j$. Jos j = 16, niin erotus seuraavaan neliöön on (2k + 1), jolloin seuraava neliö siis on 25. Koska seuraava täydellinen neliö merkkijonon k^2 jälkeen on $(j + 1)^2 = k^2 + 2j + 1$, tiedämme merkkijonon uvvw pituuden sijaitsevan kahden peräkkäisen täydellisen neliön k^2 ja $(j + 1)^2$ välillä, siis $k^2 <$ |uvvw| $< (j + 1)^2$.

Jos i=2, niin saamme $k^2 = |z| < |uv^2w| \leq k^2 + j(j + 1)^2$. Valinnalla i=4 saamme k^2 = $|0^{k^2}| <$ |uv^{4}w| ≤ k^2 + j = j(j + 1) < $(j + 1)^2 <$ j! + $\sqrt{k^2}$. Seuraavaan neliöön oleva väli kasvaa lukujen myötä ja suurten lukujen kohdalla neliöiden väli on suuri, ja tälle välille sijoittuu arvoja i+ψ, missä ψ tarkoittaa mitä tahansa suurta lukua. Tavoite on, että jokainen suuren luvun ψ arvo toteuttaa täydellisen neliön. Päättelemme, että |uvvw| ei voi olla täydellinen neliö, koska $k^2 < k^2$ + i < $(j + 1)^2$, jolloin $\sqrt{|uvvw|} \neq k^2$. Mikäli kieli olisi säännöllinen, niin uvvw, kuten uv^{4}w, kuuluisi kieleen L, mikä muodostaisi ristiriidan oletuksen kanssa, missä merkeistä 0, joiden pituus on täydellinen neliö, koostuva kieli on säännöllinen. [Hopcroft et al., 2005]

3.2 Ehrenfeuchtin, Parikhin ja Rozenbergin lemma

Käytämme seuraavia merkintöjä: olkoot a, b ja σ merkkejä, olkoot x, y ja z merkkijonoja ja i, j, k, l, m, n ja p lukuja.

Määritelmä 3.2.1.
Pinoautomaatti (pushdown automata) on järjestelmä $M = (Q, \Sigma, \Gamma, \delta, q_o, F)$, missä
Q on äärellinen tilajoukko,
Σ on äärellinen syötemerkkien aakkosto,
Γ äärellinen pinoaakkosto,
δ on siirtymärelaatio δ: Q x (Σ ∪ {λ}) x Γ → $2^{Qx\Gamma}$* (merkintä $2^{Qx\Gamma}$* tarkoittaa joukon (Q x Γ)* osajoukkojen kokoelmaa),
q_o on alkutila,
Z_o on pinon alkumerkki ja
F on lopputilojen joukko.
[Hopcroft and Ullman, 1979]

Lause 3.2.2 On olemassa numeroituvasti ääretön määrä kieliä, jotka toteuttavat säännöllisille kielille tarkoitetun pumppauslemman. Jotkut näistä kielistä ovat kontekstittomia, mutta eivät säännöllisiä.

Todistus. Olkoon aakkosto $\Sigma_1 = \{a, b\}$ ja X ⊆ Σ_1^*. Valitsemme aakkoston Σ, jossa on 16 kirjainta ja koodataan X aakkoston Σ* kieleksi $L(X)$ siten, että $L(X)$ toteuttaa pumppauslemman ja kuvaus X → $L(X)$ on injektiivinen, eli 1 - 1. Koska X on aakkoston Σ_1^* mielivaltainen alijoukko, niin joukolle X vaihtoehtojen määrä on numeroituvasti ääretön ja samoin alijoukolle $L(X)$. Tämä tulee todistamaan lauseen ensimmäisen osan. Osoitamme lisäksi, että jos X = {$a^n b^n | n \geq 0$ }, niin $L(X)$ on kontekstiton, mutta ei säännöllinen.

Olkoon $\Sigma = \{a_{i,j} | 0 \leq i, j \leq 3\}$. Määritämme kaksi kuvausta f_a, f_b aakkostosta Σ: $f_a(a_{i,j}) = a_{i+1,j}$ (mod 4), $f_b(a_{i,j}) = a_{i,j+1}$ (mod 4). Funktiot f_a ja f_b ovat aakkoston Σ permutaatioita ja lisäksi niillä on ominaisuus, että kahden funktion soveltamisessa ei voi koskaan olla samaa vaikutusta kuin yhden funktion soveltamisessa. Sillä esimerkiksi kaikille merkeille $\sigma \in \Sigma$ pätee, että $f_b(f_a(\sigma)) \neq f_a(\sigma) \neq f_a(f_a(\sigma))$. Tämä johtuu siitä, että kaksi funktiota lisää molempia alaindeksejä i, j yhdellä (mod 4), tai yhtä alaindeksiä kahdella (mod 4) ja yksi funktion sovellus ei voi koskaan saavuttaa tätä.

Olkoon laillinen merkkijono mikä tahansa merkkijono x = $(\sigma_1)^{n_1}(\sigma_2)^{n_2} \ldots (\sigma_m)^{n_m}$, missä m ≥ 1, σ_1 on $a_{0,0}$ ja kaikille i < m, σ_{i+1} on joko $f_a(\sigma_i)$ tai $f_b(\sigma_i)$. Potenssit n_i ovat kaikki positiivisia. Jos ajattelemme siirtoa merkiltä σ_1 merkkiin σ_{i+1}, minkä olisi aiheuttanut merkki a tai b, riippuen siitä onko $\sigma_{i+1} = f_a(\sigma_i)$ tai $\sigma_{i+1} = f_b(\sigma_i)$, silloin merkkijonossa x on m - 1 siirtoa ja vastaavasti m - 1 merkkiä muodostavat merkkijonon y aakkostossa Σ*. Koska jokainen n_i on positiivinen, x on yksikäsitteisesti määrittänyt merkkijonon y. Näin ollen x koodaa merkkijonon y.

Täten merkkijono x = $a_{0,0}a_{1,0}a_{1,0}a_{1,1}$ on laillinen, $n_1 = 1$, $n_2 = 2$, ja $n_3 = 1$. Koodattu merkkijono y on ab. Mutta y ei määritä potenssien n_i tarkkoja arvoja ja täten sillä on monta koodia x. Minkä tahansa osan $(\sigma_i)^{n_i}$ todistaminen edellä mainitusta laillisesta

merkkijonosta johtaa tilanteeseen, joka ei vastaa merkkiä a tai b, ja täten saatavan merkkijonon on oltava laiton.

Merkkijonon parillisuus on kaikkien alaindeksien i, j (mod 2) summa. Täten edellä tarkastellun merkkijonon y parillisuus on 0.

Olkoon $L(X) = \{x|x$ on laillinen ja x koodaa merkin y $\in X\} \cup \{x|x$ on laiton ja sillä on parillisuus 0$\}$. Väitämme, että kuvaus L on 1-1. Jos siis X $\neq$ X', niin $L(X) \neq L(X')$. Kuulukoon x joukkoon X - X'. Silloin mikä tahansa laillinen merkkijono y, joka koodaa merkkijonon x, kuuluu kieleen $L(X)$ - $L(X')$, ja täten $L(X) \neq L(X')$.

Osoitamme seuraavaksi, että $L(X)$ toteuttaa aina pumppausehdon. Olkoon k=5. Olkoot xyz $\in \Sigma^*$ ja |y| $\geq$ 5. Tarkastelemme erikseen tapauksia, joissa xyz on laillinen ja laiton.

Oletetaan aluksi, että xyz on laillinen ja y sisältää duplikaatin $\sigma\sigma$. Olkoon y = uσw, missä myös viimeinen osasanan u merkki on σ ja olkoon $v = \sigma$. Silloin kaikille arvoille i pätee, että xu$(\sigma)^i$wz on laillinen ja koodaa saman merkkijonon, jonka xyz koodaa. Täten xu$(\sigma)^i$wz $\in L(X)$, jos ja vain jos xyz $\in L(X)$.

Oletetaan seuraavaksi, että xyz on laillinen, mutta y ei sisällä duplikaatteja. On tarkasteltava parillisuuksia. Olkoon esimerkiksi xyz $\in L(X)$ ja sillä parillisuus 1. Nyt yhdellä merkkijonon y kahdesta viimeisestä merkistä on parillisuus 1, koska peräkkäisten merkkien parillisuudet vuorottelevat. Olkoon tuo merkki v ja ilmaistaan y niin, että y=*uvw*. Silloin kaikille i $\geq$ 1 pätee, että xu(v)^{i}wz koodaa saman merkkijonon kuin xyz ja on laillinen, joten xu(v)^{i}wz $\in L(X)$. Kun i=0, xu(v)^{i}wz = xuwz, ja sillä on parillisuus 0 ja se on laiton, joten edelleen xu(v)^{i}wz $\in L(X)$. Tapaukset, joissa merkkijonon xyz parillisuus on 0 tai xyz $\notin L(X)$ ovat vastaavia.

Oletetaan nyt, että xyz on laiton. Laittomuus saattaa johtua sitä, että aloitusmerkki on muu kuin $a_{0,0}$ tai siirto on epäonnistunut. Joka tapauksessa xyz sisältää sellaisen osamerkkijonon y', jonka pituus on $\leq$ 2, että kyseisen merkkijonon säilyttäminen säilyttää laittomuuden. Joten koska |y| $\geq$ 5, voimme löytää sellaisen merkkijonon y osamerkkijonon v', jonka pituus on 2, että v' $\cap$ y'. (Tämä olisi automaattista, jos y' olisi merkkijonossa x tai z ja voi myös tulla tavoitetuksi, jos y' lomittuu merkkijonon y kanssa.) Olkoon v ei-tyhjä merkkijonon v' osamerkkijono, jolla on parillisuus 0. Täytyy olla sellainen v, jolla on yksi tai kaksi merkkiä. Jos on merkki, jolla on parillisuus 0, niin v voi olla tuo merkki. Muutoin millä tahansa merkkijonolla v, jonka pituus on 2, on parillisuus 0. Olkoon y = *uvw*. Silloin kaikille i $\geq$ 0, merkkijonoilla xu(v)^{i}wz on sama parillisuus kuin merkkijonolla xyz ja merkkijono on laiton. Joten xu(v)^{i}wz kuuluu kieleen L(X), jos ja vain jos xyz kuuluu. Täten lauseen ensimmäinen osa on todistettu.

Lauseen toisen osan todistamiseksi tarkastellaan tilannetta, jossa joku pinoautomaatti tunnistaa kielen X, jos ja vain jos joku toinen pinoautomaatti tunnistaa kielen

$L(X)$. Täten jos X on kontekstiton, mutta ei säännöllinen, niin kielellä $L(X)$ on sama ominaisuus.

Tarkastellaan merkkijonoja x = $a^n b^n$, missä n on jaettavissa luvulla 4, ja merkkijonoja y, jotka esittävät näitä merkkijonoja x. Tarkastellaan kontekstittoman kieliopin sääntöjä
$S \rightarrow A_{0,0}A_{1,0}A_{2,0}A_{3,0}SA_{0,1}A_{0,2}A_{0,3}A_{0,0}$,
$S \rightarrow A_{0,0}$,
$A_{i,j} \rightarrow a_{i,j}A_{i,j}$,
$A_{i,j} \rightarrow a_{i,j}$.

Tuotettu kieli on kontekstiton ja se on niiden merkkijonojen y joukko, jotka esittävät joitakin muotoa $a^n b^n$, missä n = 0 mod 4, olevia merkkijonoja. (Tapaukset n = i mod 4, kun i = 1, 2, 3, ovat vastaavia). Täten $L(X)$ ∩ Laiton on kontekstiton, ja se on neljän kontekstittoman joukon yhdiste. Mutta silloin $L(X)$ = ($L(X)$ ∩ Laiton) ∪ (Laiton ∩ 0-parillisuus), ja Laiton, 0-parillisuus ovat säännöllisiä. Täten $L(X)$ on kontekstiton, ja se on kahden kontekstittoman joukon yhdiste. Mutta $L(X)$ ei voi olla säännöllinen, sillä se sisältää sellaisia merkkijonoja y_i, että y_i = $(a_{0,0}a_{1,0}a_{2,0}a_{3,0})^{i/4}a_{0,0}$ ja sellaisia z_i, että z_i = $(a_{0,1}a_{0,2}a_{0,3}a_{0,0})^{i/4}$ ja i on jaettavissa luvulla 4. Nyt kaikille i, j, pätee, että jos i ≠ j niin $y_i z_i \in L(X)$ ja $y_j z_i \notin L(X)$. Neroden lauseen (ks. [Nerode, 1958]) nojalla kieli $L(X)$ ei ole säännöllinen. [Ehrenfeucht et al., 1981] □

Kieli $L(X)$ ei siis ole säännöllinen, koska $y_i z_i$ kuuluu kieleen $L(X)$, joka on kontekstiton, mutta $y_j z_i$ ei kuulu kieleen $L(X)$ - mille välttämätön ja riittävä ehto on kirjainten i ja j erisuuruus.

Esitämme seuraavaksi Jaffen ehdon, koska on tutkittu, onko sellaista pumppausta, joka olisi riittävä ja välttämätön ehto säännöllisyydelle.

Lause 3.2.3 (Jaffe) Kieli on säännöllinen, jos ja vain jos on olemassa sellainen k, että kaikille x ∈ Σ^*, jos |x| ≥ k, niin on olemassa sellaiset u,v,w, että x = uvw, v ≠ λ ja kaikille z, v on pumppaus merkkijonolle xz suhteessa kieleen L, toisin sanoen kaikille i ≥ 0, kaikille $z \in \Sigma^*$, u(v^i)w$z \in L$, jos ja vain jos x$z \in L$.

Todistus. Sivuutetaan (ks. [Jaffe, 1978]). □

Mutta Jaffen ehto ei ole paikallinen, sillä annetulla osasanan x arvolla vaaditaan pumppaus, joka ei toimi ainoastaan osasanaan x, vaan lisäksi yhdenmukaisesti kaikille xz ∈ L. Kysymyksenä on, että voidaanko löytää paikallinen pumppausehto, joka on ekvivalentti säännöllisyyden kanssa. Lause 3.2.7 antaa kysymykseen positiivisen vastauksen. [Ehrenfeucht et al., 1981]

Seuraavaksi käyttöön ottamamme sana lohko tarkoittaa osasanaa.

Määritelmä 3.2.4 Kielellä $L \subseteq \Sigma^*$ on lohkopumppausominaisuus, jos on sellainen k, että kaikille x, w, $y_1, \ldots, y_k$, w' $\in \Sigma^*$, jos x = w$y_1 \ldots y_k$w', niin on olemassa sellaiset m, j , $1 \leq m < j \leq k$, että $y_{m+1} \ldots y_j$ on pumppaus merkkijonolle x suhteessa kieleen L.

Määritelmässä 3.2.4 w, $y_1, \ldots, y_k$ ja w' ovat siis lohkoja.

Määritelmä 3.2.5 Kielellä $L \subseteq \Sigma^*$ on lohkoperuutusominaisuus, jos on sellainen k, että kaikille x,w, $y_1, \ldots, y_k$, w' $\in \Sigma^*$, jos x = w$y_1 \ldots y_k$w', niin on olemassa sellaiset m, j , $1 \leq m < j \leq k$, että w$y_1 \ldots y_m y_{j+1} \ldots y_k$w' $\in L$ jos ja vain jos x $\in L$.

Havaitsemme, että lohkoperuutusominaisuus on lohkopumppausominaisuuden erikoistapaus.

Tarkastellaan lausetta 3.2.3. Jos pumppaisimme arvolla i=0, päätyisimme lohkoperuutukseen.

Määritelmä 3.2.6 Jos kielellä L on lohkoperuutusominaisuus tietylle luvulle k, niin merkitsemme, että $L \in \varphi_k$.

Lause 3.2.7 Säännöllisyys, lohkopumppausominaisuus ja lohkoperuutusominaisuus ovat ekvivalentteja.

Todistus. Todettakoon ensin, että säännöllisyys implikoi lohkopumppausominaisuuden. Oletetaan, että kieli L on säännöllinen. Olkoon M kielen L tunnistava automaatti ja olkoon k automaatin M tilojen lukumäärä. Oletetaan, että x $\in \Sigma^*$ ja x = w$y_1 \ldots y_k$w'. Olkoon s^j tila, jonka automaatti M on saavuttanut, kun se on lukenut merkkijonon y_j. Silloin $s^0, s^1, \ldots s^k$ ovat tilojen k + 1 esiintymää ja on oltava sellaiset m, j, että m < j, mutta $s^m = s^j$. Silloin $v = y_{m+1} \ldots y_j$ on vaadittu pumppaus merkkijonolle x suhteessa kieleen L. Jos kieli L toteuttaa lohkopumppausominaisuuden, silloin se triviaalisti toteutaa lohkoperuutusominaisuuden. Täten lause palautuu seuraavaan apulauseeseen. [Ehrenfeucht et al., 1981] □

Seuraavaksi esitämme version Ramseyn lauseesta. Kun X on joukko, niin X[2] merkitsee joukon X kaikkia kahden alkion alijoukkoja. Mikäli joukolla X on n alkiota, niin joukolla X[2] on n(n - 1) / 2 alkiota.

Lause 3.2.8 (Ramseyn lause). Jokaisella luonnollisella luvulla k on olemassa sellainen luku r(k), että jos joukolla X on r(k) alkiota tai enemmän ja X[2]=Z $\cup$ Z', niin on sellainen Y $\subseteq$ X, että joukolla Y on vähintään k + 1 alkiota ja Y[2]=Z $\cup$ Z tai Y[2]=Z $\cup$ Z'.

Todistus. Sivuutetaan (ks. [Ramsey, 1954]). □

Apulause 3.2.9 Lohkoperuutusominaisuus implikoi säännöllisyyden.

Apulauseen todistus palautuu kolmeen seuraavaan apulauseeseen:

Apulause 3.2.10 On olemassa vain äärellisen monta kieltä joukossa φ_k (k on vakio).

(Olkoon annettu kieli L ja merkkijono x, ja olkoon edelleen L_x joukko $\{z|xz \in L\}$).

Todistus. On riittävää osoittaa, että jos L, L' ovat joukossa φ_k ja kaikille merkkijonoille x, missä |x| < r(k), x ∈ L, jos ja vain jos x ∈ L', niin $L = L$'. Jos aakkostolla Σ on n alkiota, n > 1, niin on korkeintaan m = $n^{r(k)}$ merkkijonoa, joiden pituus on < r(k) ja täten korkeintaan 2^m kieltä joukossa φ_k.
Väite: Osoitamme induktiolla luvun n suhteen, että jos |x| = n, silloin x ∈ L, jos ja vain jos ∈ L'. Oletetaan väitteen koskevan kaikkia arvoja p < n. Oletetaan, että |x| = n ja n ≥ r(k). Soveltaaksemme Ramseyn lausetta määrittelemme joukot Z ja Z' seuraavasti: Kirjoitetaan x = wy_1 ... $y_{r(k)}$w', missä kaikki merkit y_j ovat eityhjiä. Olkoon X = $\{0,\ldots,r(k)\}$ ja luvuille m, j ∈ X, m < j, olkoon $\{m,j\}$ ∈ Z jos wy_1 ... $y_m y_{j+1}$w' ... $y_{r(k)}$w' ∈ L ja muutoin olkoon $\{m,j\}$ ∈ Z'. Silloin X[2]=Z ∪ Z' ja Ramseyn lauseen nojalla, on olemassa Y, jossa on sellaiset k + 1 alkiota, että Y[2]⊆ Z tai Y[2]⊆ Z'. Kummassakin tapauksessa joukon Y alkiot jakavat merkkijonon x k + 2 merkkijonoon, joista merkkijono u on ennen joukon Y ensimmäistä alkiota, u' viimeisen alkion jälkeen, ja välissä ovat merkkijonot $z_1, \ldots, z_k$. Jokainen z z_i on yhdiste vähintään yhdestä peräkkäisestä merkkijonosta y, ja x = uz_1 ... z_ku'.

Saamme kaksi tapausta sen suhteen, sisältyykö Y[2] joukkoon Z vai joukkoon Z':

i) Y[2]⊆ Z. Nyt merkkijonosta x peräkkäisten lohkojen z' poistaminen vastaa jotain joukkoa $\{m,j\}$ joukossa Y[2] ja lyhennetty x' on aina kielessä L. Kuitenkin, peruutusehdon nojalla on jokin merkkijonon z' peräkkäinen lohko, jonka poistaminen johtaa sellaiseen merkkijonoon x', että x' ∈ L, jos ja vain jos x ∈ L. Täten x ∈ L.
ii) Y[2]⊆ Z'. Vastaava argumentti osoittaa, että x ∉ L.

Tapausten i) ja ii) nojalla x ∈ L, jos ja vain jos on sellainen joukko Y, että sillä on sellaiset k + 1 alkiota, että Y[2]⊆ Z. Samat tosiasiat pätevät myös kielelle L', jossa on samat Z ja Z'. Näin on, koska L ja L' vastaavat toisiaan kaikille merkkijonoille, joiden pituus on vähemmän kuin n. Täten merkkijonolle x saamme, että x ∈ L, jos ja vain jos x ∈ L'. Täten väite ja apulause on todistettu. [Ehrenfeucht et al., 1981] □

Apulause 3.2.11 Jos kieli L sisältyy joukkoon φ_k, niin myös L_σ kaikilla $\sigma \in L$ (k on vakio) kuuluu joukkoon φ_k.

Todistus. Oletetaan, että $z \in \Sigma^*$ ja z = w$y_1 \ldots y_k$w'. Tarkastellaan merkkijonoa σ z = w"$y_1 \ldots y_k$w', missä w" = σw. Koska $L \in \varphi_k$, on olemassa sellaiset m, j, 1 ≤ m < j ≤ k, että w"$y_1 \ldots y_m y_{j+1} \ldots y_k$w' ∈ L, jos ja vain jos σz ∈ L. Mutta silloin w$y_1 \ldots y_m y_{j+1} \ldots y_k$w' ∈ L_σ jos ja vain jos $z \in L_\sigma$. Täten kieli L_σ sisältyy myös joukkoon φ_k. [Ehrenfeucht et al., 1981] □

Apulause 3.2.12 Olkoon J sellainen kielten ominaisuus, että (i) on olemassa vain äärellisen monta kieltä, joilla on J; (ii) kaikille $\sigma \in \Sigma$, jos kielellä L on J, silloin kielellä L_σ on J. Silloin J implikoi säännöllisyyden.

Todistus. Oletetaan, että kielellä L_0 on ominaisuus J. Määritämme automaatin M seuraavasti:
K = kaikki kielet L, joilla on ominaisuus J,
s_0 = aloitustila = L_0,
M(L, σ) = seuraava kieli (tila), kun σ on luettu = L_σ,
F = tunnistavien tilojen joukko = $\{L | \lambda \in L\}$.

Osoitamme induktiolla merkkijonon x suhteen, että $M(L, \text{x})$ on L_x kaikille x aakkostossa Σ^*. Kun $L_\lambda = L$, ja kaikille σ aakkostossa Σ, niin

$$L_{x\sigma} = (L_x)_\sigma = M(L, \text{x})_\sigma = M(M(L, \text{x}), \sigma) = M(L, \text{x}\sigma).$$

Täten M tunnistaa merkkijonon x jos ja vain jos $M(L_{10}, \text{x}) \in$ F
jos ja vain jos $\lambda \in M(L_{10}, \text{x})$
jos ja vain jos $\lambda \in (L_{10})_x$
jos ja vain jos x $\lambda \in L_{10}$
jos ja vain jos x $\in L_{10}$. [Ehrenfeucht et al., 1981] □

Ehrenfeucht ja muut [1981] tarjoavat välttämättömän ja riittävän ehdon sekä toteavat, ettei säännöllisille kielille tarkoitettu lause 3.0.1 implikoi säännöllisyyttä. Lauseen 3.2.2 todistuksessa mainitaan, että lauseen ensimmäinen osa on todistettu, mikä johtuu siitä, että xu(v)^{i}wz kuuluu kieleen $L(X)$, jos ja vain jos xyz kuuluu. On merkittävää, että lauseen 3.2.2 todistuksen aakkostossa on ainoastaan kaksi kirjainta, a ja b. Todistuksessa käytetään äärellistä alijoukkoa sekä tutkitaan, vastaavatko kuvaukset f_a ja f_b toisiaan. Myös Jaffen ehto kohdistuu säännöllisiin kieliin. Ehrenfeucht ja muut ovat vahvistaneet kyseistä lemmaa lohkopumppausominaisuudella, joka on yhtäpitävä säännöllisyyden kanssa. Kieli on säännöllinen, jos se toteuttaa lohkopumppausominaisuuden. He tutkivat annetun lemman pumppausominaisuuden implikoituvuutta säännöllisyyteen. Mainittu ominaisuus koostuu lohkoperuutuksesta ja Ramseyn teoreeman äärellisen version soveltamisesta. Lohkoperuutus,

jossa peruutetaan yksittäinen lohko, on riittävä ehto lohkopumppausominaisuudelle, siis lohkopumppausominaisuus on välttämätön ehto lohkoperuutusominaisuudelle - mutta luonnollisesti tämä ei päde toisinpäin. Kielellä on siis myös peruutuksellisia ominaisuuksia.

Ehrenfeucht ja muut todistavat, että mikäli kieli L toteuttaa lohkopumppausominaisuuden, niin kieli on säännöllinen. Todistuksessa Ramseyn lauseella on ratkaiseva osa. Pumppaus kohdistuu yhteen osamerkkijonoon. Määritelmässä 3.2.3 luvut m ja j liittyvät pumppaukseen ja pumppaus kohdistuu lohkoon $y_{m+1} \ldots y_j$. Huomaamme, että vakio k saa arvoksi vähintään 2. Määritelmässä 3.2.4 luvut m ja j liittyvät lohkoperuutukseen, jossa peruutetaan y-merkkejä, mutta ei niistä kaikkia. Sillä käytettäessä lohkoperuutusta sanaan w$y_1 \ldots y_m y_{j+1} \ldots y_k$w', jota merkitsemme z, saadaan sana x,w, $y_1, \ldots, y_k$, w jota merkitsemme z'. Sana z' kuuluu kieleen L. Peruutettava lohko on tällöin $y_m \ldots y_{j+1}$.

4 Bar-Hillelin lemma kontekstittomille kielille

Tässä luvussa esittelemme Bar-Hillelin lemman ja esimerkkejä lemman käytöstä. Lopuksi esittelemme Bar-Hillelin lemman toteuttavia kieliä.

Lause 4.0.1 (Bar-Hillelin lemma, pumping lemma)
Olkoon L kontekstiton kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k \in \mathbb{N}$, että kun $|z| > k$, niin sana z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että
1. $|vx| \geq 1$.
2. $|vwx| \leq k$.
3. $uv^iwx^iy \in L, i \geq 0$.

Todistus. Olkoon $G = (N,T,P,S)$ kielen L tuottava Chomskyn normaalimuodossa oleva kielioppi. Olkoon kieliopin G apumerkkien lukumäärä q. Valitaan $k = 2^q$. Jos Chomskyn normaalimuodossa olevan kieliopin derivaatiopuussa pisin polku juuresta lehtisolmuun on pituudeltaan j, niin derivaatiopuuhun liittyvän sanan pituus on korkeintaan 2^{j-1}. Tarkastellaan sellaista kielen L sanaa z, jonka pituus on suurempi kuin k. Sanan z derivaatiopuussa on oltava sellainen polku juuresta lehteen, että sen pituus on suurempi kuin q. Tällaisella polulla on oltava solmut n_1 ja n_2, jotka täyttävät seuraavat kolme ehtoa:

i) solmuihin liittyy sama apumerkki A.
ii) solmu n_1 on lähempänä juurta kuin solmu n_2.
iii) etäisyys solmusta n_1 lehtisolmuun on korkeintaan q + 1.

Olkoon T_1 osapuu, jonka juuri on n_1, ja T_2 se osapuu, jonka juuri on n_2. Osapuussa T_1 ei ole lukua (q + 1) pidempää polkua, joten osapuuhun T_1 liittyvän osasanan z_1 pituus on korkeintaan 2^q. Jos osapuuhun T_2 liittyvästä osasanasta käytetään merkintää z_2, niin z_1 voidaan kirjoittaa muodossa $z_1 = z_3z_2z_4$. Koska solmusta n_1 edettäessä käytetään muotoa $A \rightarrow BC$ olevaa sääntöä, niin $|z_3z_4| > 0$. On siis $A \Rightarrow^* z_3Az_4 \Rightarrow^* z_3z_2z_4$, missä $|z_3z_2z_4| \leq k$, ja kaikilla arvoilla i = 0,1,2, . . . , on olemassa johto $A \Rightarrow^* (z_3)^iz_2(z_4)^i$. Koko sana z voidaan kirjoittaa muodossa $z = uz_3z_2z_4y$. Merkitsemällä z_3 = v, z_2 = w ja z_4 = x saadaan lauseen väite. [Hopcroft and Ullman, 1979] □

Tarkastelemme lauseen 4.0.1 ehtoja. Ehto 1 tarkoittaa, että osasanat v ja x eivät voi molemmat olla tyhjiä. Mutta joko v tai x voi olla tyhjä. Silloin pumppaus tapahtuu yhdessä paikassa. Ehto mahdollistaa sen, että derivaatiopuussa syntyy yksi toistettava osapuu. [Hopcroft and Ullman, 1979]

Ehto 2 tarkoittaa, että osasanojen v ja x välissä olevan osasanan w pituutta ei ole erikseen määritelty, mutta se on pienempi kuin k. [Hopcroft and Ullman, 1979] Huomaamme, että osasanojen u ja y pituutta ei ole rajoitettu.

Ehto 3 mahdollistaa mielivaltaisen pumppauskertojen i toistamisen osasanoihin v ja x. Kun i > 1, sanan pituus kasvaa, sillä osasanoista v ja x koostuva lohko pumpataan eli toistetaan i kertaa. Kun i=0, niin osasanat v ja x poistetaan, jolloin sanan z pituus pienenee ja kyseinen sana supistuu kolmeen osasanaan. Mikäli $z \in L$, niin siihen kuuluu myös mielivaltaisen suuri määrä merkkijonoja, jolloin kielessä L on siis äärettömän monta merkkijonoa. Pumppauksen lopputuloksena on siis uusia sanoja, jotka kuuluvat kieleen L. [Hopcroft and Ullman, 1979]

Osasana w ei voi olla tyhjä, sillä sen yllä olevan apumerkin on johdettava perusmerkkiin. Pumppaus voi sisältää useita eri merkkejä, sillä se riippuu täsmälleen siitä, mitä merkkejä v ja x sisältävät. [Hopcroft and Ullman, 1979]

Teemme seuraavaksi huomioita todistuksesta. Ehdossa i) apumerkin A avulla muodostetaan äärettömän pitkiä merkkijonoja. Ehdossa ii) solmu n_1 on edeltäjä solmulle n_2. Solmu n_1 liittyy osasanoihin u ja y. Solmu n_2 liittyy osasanoihin v ja x. On voimassa $|z_3 z_4| > 0$ ja $|vx| \geq 1$. Kummankin osapuun T_1 ja T_2 korkeus on korkeintaan q. Osajohto $A \Rightarrow^*$ w suoritetaan osajohdon $A \Rightarrow^*$ vAx jälkeen. Jos i=0, niin vAx ei siis tule toistetuksi. Derivaatiopuussa jokaisen, lukuunottamatta viimeistä, apumerkin A alla on osapuu, johon liittyy apumerkki A, kunnes lopulta päädytään osasanaan w. Osapuiden lukumäärä ei voi lisääntyä toistamatta apumerkkiä A, joka on ensimmäinen kahdesti toistuva apumerkki. Toistettaessa apumerkkiä A ja päädyttäessä osasanaan w tiedämme, että kyseisessä polussa on vähintään kaksi apumerkkiä. Tällöin ehto ii) toteutuu, sillä toinen osasanoista v ja x on ei-tyhjä. Derivaatiopuun korkeus riippuu pisimmän polun pituudesta. Kun siis i > 0 ja mitä suurempi i on, niin sitä erottuvammin puuhun on muodostunut pisin polku, jonka pidentyminen on tapahtunut askeleittain pumppausten myötä.

Tarkastelkaamme lemman ehdot täyttävää sanaa z = haqlkrfbmceo $\in L$. Olettakaamme osasanan v koostuvan merkeistä lkrf ja osasanan x koostuvan merkeistä ce. Nyt uv^0wx^0y = haqbmo $\in L$, uv^2wx^2y = haqlkrflkrfbmceceo $\in L$ ja uv^4wx^4y = haqlkrflkrflkrflkrfbmcecececeo $\in L$. Osasanoihin v ja x kohdistuu aina täsmälleen sama määrä pumppaustoistoja, ja ne suoritetaan samanaikaisesti.

Kielioppi G tuottaa siis kaikki muotoa uv^iwx^iy, $i \geq 0$, olevat merkkijonot. Sanan pituus ei ole rajoitettu, jolloin kieli ei ole äärellinen. [Hopcroft and Ullman, 1979]

Toistettaessa apumerkkiä A johdot ovat muotoa: $S \Rightarrow^*$ uAy, $A \Rightarrow^*$ vAx, $A \Rightarrow^*$w joillekin osamerkkijonoille u, v, w, x, y. Toistettaessa saamme $S \Rightarrow^*$ uAy $\Rightarrow^*$ uv$A$xy $\Rightarrow^*$ uvv$A$xxy $\Rightarrow^* \ldots \Rightarrow^* uv^iAx^iy \Rightarrow^* uv^iwx^iy$, jota olemme toistaneet i kertaa. Täten $uv^iwx^iy \in L$. [Shallit, 2008]

4.1 Esimerkkejä

Käyttäessämme seuraavaksi lemmaa on tavoitteenamme osoittaa, että tarkasteltava kieli L ei ole kontekstiton. Osoitamme sen seuraavilla vaiheilla:

Oletamme aluksi, että kieli L on kontekstiton. Seuraavaksi etsimme sopivan sanan z, joka kuuluu kieleen L. Osoitamme, että uv^iwx^iy ei kuulu kieleen L jollain arvolla i. [Hopcroft and Ullman, 1979]

Osoitamme käyttämällä Bar-Hillelin lemmaa, että seuraavat esimerkeissä olevat kielet eivät ole kontekstittomia.

Esimerkki 4.1.1 Olkoon kieli $L = \{\alpha^\psi \mid \psi$ on alkuluku$\}$. Kieleen kuuluu siis ainoastaan merkkijonoja, joiden pituus on alkuluku. Oletetaan, että kieli L on kontekstiton. Nyt voimme kirjoittaa $\alpha^\psi = uvwxy$, missä $|vx| \geq 1$ ja $|vwx| \leq k$, jolloin $uv^iwx^iy \in L$, kaikille $i \geq 0$. Olkoot u=α^n, v=α^m, w=α^q, x=α^r, y=α^t. Tällöin α^ψ = $\alpha^n\ \alpha^m\ \alpha^q\ \alpha^r\ \alpha^t$, tai ψ = n+m+q+r+t, missä m+r > 0, m+q+r ≤ k. Täten ψ_2 = (n+q+t) + (m+r)i, kaikille i > 0, missä ψ_2 on alkuluku tai $\psi_2 = \xi + \Upsilon$i kaikille i > 0, ψ_2 on alkuluku. [Simon, 1999] Edellä siis ositimme alussa valitsemamme alkuluvun, olkoon se esimerkiksi 7901, jolle pätee, että $\phi(7901) = 7900$, jonka saamme kaavalla ϕ(n) = n $\prod_{p|n}$ (1 - $\frac{1}{p}$). [Rosen, 2010] Pumpattavat lohkot sijaitsevat "$osassa$" (m+r). Voimme kirjoittaa $\alpha^n\alpha^{m(i)}\alpha^q\alpha^{r(i)}\alpha^t$. Valitsemme i=$\xi$+$\Upsilon$+1, missä ξ olkoon esimerkiksi 4196 ja Υ olkoon esimerkiksi 3705, ja jolloin ψ_2=ξ+Υ(ξ+Υ+1) = 29281106 ei ole alkuluku. [Simon, 1999]

Käyttäessämme lemmaa ja edellä valitsemiamme lukuja, ja tietämättä osasanojen v ja x sijainteja, valitkaamme kuitenkin ξ=uwy=4196 ja Υ=vx=3705, jolloin saamme arvolla i=2 merkkijonon pituudeksi 11606, joka on alkuluvun sijaan komposiitti ja jolloin $uv^2wx^2y \notin L$. [Simon, 1999] Mikäli merkkijonon pituus on pariton, niin voimme tutkia, onko se alkuluku soveltamalla esimerkiksi Wilsonin lausetta, jonka mukaan n on alkuluku, jos n = 2, että (n - 1)! = -1 mod n. Jos saamme tulokseksi (n - 1)! = 0 mod n, niin luku on komposiitti, josta voimme käyttää merkintää ζ. Tällöin voimme merkitä $\zeta^{(\xi-1)(\Upsilon-1)} \equiv 1 \bmod \xi$. [Rosen, 2010] Luvuista ξ ja Υ toinen tai molemmat voivat olla merkkijonoja, joiden pituus on alkuluku, mutta toisen ei välttämättä tarvitse olla. Oletuksemme on osoittautunut vääräksi. Havaitsemamme ristiriidan perusteella kieli L ei ole kontekstiton. [Simon, 1999]

Esimerkki 4.1.2. Olkoon kieli $L = \{a^lb^m | l = m^2\}$. Kieleen kuuluminen liittyy tässä tapauksessa merkkien a ja b lukumääriin. Tehdään vastaoletus, että kieli L on kontekstiton. Tarkastelemme mitä tahansa sanaa z, $|z| > k$. Mikäli sanassa z on a-merkkejä n kappaletta, niin merkkejä b on n^2 kappaletta. Pumpattavissa osamerkkijonoissa v ja x voi olla molemmissa vain yhtä merkkiä, sillä muutoin merkkien a ja b järjestys menisi sekaisin pumpattaessa. Joten $|vwx|$ voi sisältää ainoastaan yhtä merkkiä. Olkoot a- ja b-merkkien lukumäärät vastaavasti t ja s. Tällöin arvolla

i=2 saadaan $a^{n+t}b^{n^2+s}$. Muutoin jos sekä v että x sisältäisivät a-merkkiä, niin saataisiin $a^{n+ts}b^{n^2}$ tai jos v ja x sisältäisivät myös b-merkkiä, niin saataisiin $a^{n+ts}b^{n^2+ts}$. [Automaatit-kurssi]

Kun v ja x sisältävät yhtä merkkiä, niin n sisältää arvon t ja n^2 arvon s. Ratkaisemalla s saadaan s = 2nt + t^2. Kun i=2, on b-merkkien lukumäärän n^2 + s oltava sama kuin $(n + t)^2$. Kun i=3, on b-merkkien lukumäärän n^2 + 2s oltava sama kuin $(n + 2t)^2$ eli n^2 + 4nt + $2t^2$ = n^2 + 4nt + $4t^2$. Yhtäsuuruus n^2 + 4nt + $2t^2$ = n^2 + 4nt + $4t^2$ on mahdollista ainoastaan siinä tapauksessa, kun t=0, jolloin s=0, sillä esimerkiksi arvoilla i=2, n=4 ja t=3 a- ja b-merkkien määrä ei toteuta kielen ehtoa. Tämä on ristiriita, joten vastaoletus osoittautuu vääräksi, jolloin kieli L ei ole kontekstiton. [Automaatit-kurssi]

Esimerkki 4.1.3. Olkoon kieli $L = \{ww|w \in 0,1^*\}$. Oletetaan, että kieli L on kontekstiton. L koostuu toistuvista merkkijonoista, kuten 1010, 001001, 00100010, 110110 ja 111000111000. Tarkastellaan merkkijonoa $z = 0^k1^k0^k1^k$, jonka pituus $4k > k$. Osoitamme, että uwy ei kuulu kieleen L. Koska $|vwx| \leq k$, niin $|uwy| \geq 3k$. Mikäli uwy on jokin toistuva merkkijono, esimerkiksi tt, niin t on pituudeltaan vähintään $3k/2$. Tarkastelemme missä kaikkialla osasana vwx voi sijaita sanassa z. [Hopcroft et al., 2005]

1. Osasana vwx on ensimmäisen 0-lohkon sisällä. Tällöin myös u kuuluu samaan 0-lohkoon. Nyt $1^n0^n1^n$ sisältää osasanan y, joten tarkastelemme osasanaa $0^r1^n0^n1^n$. Koostukoon vx n, n $\geq$ 0, kappaleesta 0-merkkejä. Nyt uwy alkaa merkkijonolla $0^{k-n}1^k$. Koska $|uwy| = 4k$ - n, ja jos uwy = tt, niin |t| = $2k$ - n/2. Täten t loppuu vasta ensimmäisen 1-merkkejä sisältävän lohkon jälkeen, toisin kuin uwy, joka loppuu 1-merkkeihin, jolloin uwy ei olla muotoa tt. Pumppaukset tapahtuvat ensimmäisessä lohkossa. [Hopcroft et al., 2005]

2. Osasana vwx lomittuu ensimmäiseen 0-lohkoon ja ensimmäiseen 1-lohkoon. Jos x = λ, niin vx voi muodostua ainoastaan 0-merkeistä. Tässä tapauksessa merkkijono on muotoa $0^r1^s0^n1^n$. Nyt argumentti sen puolesta, että uwy ei ole muotoa tt, vastaa tapausta 1. Jos osasanassa vx on vähintään yksi 1, silloin osamerkkijonon t, joka on pituudeltaan vähintään $3k/2$, täytyy päättyä osasanaan 1^n, koska uwy päättyy 1^n toisin kuin vx. On yksi 1-merkeistä koostuva lohko, jonka pituus on k ja joka on lohkoista viimeinen. Täten t ei voi toistua osasanassa uwy. [Hopcroft et al., 2005]

3. Osasana vwx sijaitsee ensimmäisessä 1-lohkossa, jossa pumppaukset nyt tapahtuvat. Osasana on muotoa $0^n1^s0^n1^n$. Argumentti osasanan uwy kuulumattomuudesta kieleen L on vastaava kuin tapauksen 2 jälkimmäinen osa. Tässä tapauksessa 1^n sisältää osasanan y. [Hopcroft et al., 2005]

4. Osasana vwx lomittuu ensimmäiseen 1-lohkoon ja toiseen 0-lohkoon, eli 10 sisältää osasanan vwx, jolloin $0^n1^s0^r1^n$. Tässä tapauksessa $uvwxy$ on $0^n1^{n-k}0^{n-k}1^n$

= 4n - $2k$. Nyt 0^n sisältää osasanan u ja 1^n sisältää osasanan y. Jos vx ei sisällä 0-merkkejä, argumentti on sama kuin jos vwx sisältyisi ensimmäiseen 1-lohkoon. Jos osasanaan vx sisältyy vähintään yksi 0, niin uwy alkaa 0-lohkolla, samoin t, jos uwy=tt. Nyt osasanassa uwy ei ole toista 0-lohkoa toiselle kopiolle t-merkkiä. Tällöin uwy ei kuulu kieleen L. [Hopcroft et al., 2005]

5. Vielä on käsittelemättä tapaukset, joissa vwx sijaitsee sanan z jälkimmäisessä lohkossa kahdesta lohkosta koostuvassa lohko-osassa. Tällöin z on joko muotoa $0^n1^n0^r1^n$ tai muotoa $0^n1^n0^n1^s$. Argumentti on symmetrinen niiden tapausten kanssa, joissa vwx sisältyy sanan z ensimmäiseen kahdesta lohkosta koostuvaan lohko-osaan. On siis yhteensä neljä vaihtoehtoa, joissa vwx sijaitsee yhdessä lohkossa. [Hopcroft et al., 2005]

Koska lomittuminen ei voi koskea samanaikaisesti kolmea tai neljää lohkoa, niin esimerkiksi merkkijonoja $0^r1^s0^r1^s$, $0^r1^s0^r1^n$ tai $0^n1^s0^r1^s$ ei voi olla olemassa. Lomittuminen voi kahden lohkon osalta toteutua peräkkäisiin lohkoihin. Havaitsemme, että missään mahdollisessa tapauksessa uwy ei kuulu kieleen L. Täten kieli L ei ole kontekstiton. [Hopcroft et al., 2005]

Esimerkki 4.1.4. Olkoon kieli $L = \{a^e b^f c^g | e < f < g\}$. Oletetaan, että kieli L on kontekstiton. Tarkastellaan kieleen L kuuluvaa lausetta $a^k b^{k+1} c^{k+2}$, jolloin eroavuutta olisi mahdollisimman vähän eli ainoastaan yhden esiintymän suuruisesti. Kieleen kuuluminen liittyy tässä tapauksessa järjestykseen ja sitä kautta myös merkkien lukumääriin. Merkkien a, b ja c lukumäärät eivät voi olla samat. Tutkikaamme ensin tapausta, jossa v koostuu a-merkeistä. Tällöin arvolla i=2 saadaan e ≥ f tai e ≥ g, joista kumpikaan ei ole mahdollista kielessä L. Seuraavaksi tutkimme tilannetta, jossa e=0. Saamme tarkasteltavaksi kolme eri tapausta, jotka ovat seuraavat: [Automaatit-kurssi]

a) v ja x sisältää b- ja c-merkkejä. Nyt a-merkkien lukumäärä on 0, mikä toteuttaa kieleen kuulumisen ehdon. Täten $uv^iwx^iy \in L$.
b) v ja x sisältää b-merkkejä. Tässä tapauksessa tiedämme, että b-merkkejä on vähintään kolme kappaletta. Saamme $uv^0wx^0y \in L$, mutta $uv^2wx^2y \notin L$.
c) v ja x sisältää c-merkkejä. Tässä tapauksessa b-merkkien lukumäärä säilyy samana. Tiedämme, että c-merkkejä on vähintään kolme kappaletta. Mikäli pumpataan arvolla i=0, niin b-merkkejä on saman verran tai enemmän kuin c-merkkejä, jolloin $uv^0wx^0y \notin L$. Mutta arvolla i=2 saadaan b^kc^{k+1}.
Osoittamamme ristiriidan nojalla kieli L ei ole kontekstiton. [Automaatit-kurssi]

Kun e=0, niin kieli on muotoa $L = \{b^f c^g | f < g\}$, joka on kontekstiton. Pinoautomaatilla tunnistettavuuden lisäksi kieli on tuotettavissa seuraavalla kieliopilla:
$S \rightarrow HQ$
$H \rightarrow bHc \mid \lambda$
$Q \rightarrow cQ|c$.

Esimerkki 4.1.5. Olkoon kieli $L = \{a^m b^n a^m b^n | m, n \geq 1\}$. Oletetaan, että kieli L on kontekstiton. Jokainen riittävän pitkä sana $a^m b^n a^m b^n$ on esitettävissä muodossa $uvwxy$ siten, että $uv^i wx^i y$ kaikilla i ≥ 0. Kielen L sanoilla on ominaisuus, että niillä kaikilla on kolme rajaa seuraavasti: aa . . . a | bb . . . b | aa . . . a | bb . . . b. Jokainen v ja x on yksittäisen kirjaimen potenssi. Sana z on valittavissa siten, että $z = a^p b^p a^p b^p$, jonka pituus on 4p. Mikäli v on a-merkin potenssi ja x on b-merkin potenssi, niin osasanan v on oltava a-merkkien lohko merkkijonossa $a^m b^n a^m b^n$. Nyt kahdesta a-merkeistä koostuvasta lohkosta toinen on pidempi kuin toinen. Täten $uv^2 wx^2 y \notin L$. On siis oltava $v = a^p$ ja x = a^q, tai $v = b^p$ ja x = b^q, jolloin on riittävää tarkastella edellistä tapausta. [Howie, 1991]

On oltava p = q, jotta a-merkkien lohkot lisääntyvät yhtäläisesti, kun i kasvaa merkkijonossa $uv^i wx^i y$. Tällöin saamme $a^m b^n a^m b^n = a^r a^p a^s b^n a^t a^p a^l b^n$, joka tarkoittaa että u = a^r, $v = a^p$, w = $a^s b^n a^t$, x = a^p, $y = a^l b^n$, missä r, s, t, l $\geq$ 0, p > 0, r + p + s = t + p + l = m. Voidaan todeta, että t + l = r +s. Tällöin $uv^i wx^i y = a^{m+(i-1)p} b^n a^{m+(i-1)p} b^n \in L$, kaikilla i ≥ 0 (tai vastaavasti $uv^i wx^i y = a^m b^{n+(i-1)p} a^m b^{n+(i-1)p}$). Olemme olettaneet, että edellä mainitut osamerkkijonoihin jaot ovat ainoita tapoja takaamaan jokaisen $uv^i wx^i y$ kuulumisen kieleen L. Sanoissa $a^m b^n a^m b^n$ pätee 2m + 2n > k. Mutta olettaessamme, että n > k, niin $|vwx|$ = 2p + s + t + n > k sekä $|uwy|$ = 2n + r + s + t + l, jolloin osamerkkijonoihin jako ei toteuta lemman ehtoja. Koska lemma edellyttää osamerkkijonoihin jaon olevan mahdollinen kaikille riittävän pitkille sanoille, oletuksemme on väärä, eikä kieli L ole kontekstiton. [Howie, 1991]

Esimerkki 4.1.6. Tarkastellaan tämän kohdan lopuksi tapausta, jossa Bar-Hillelin lemma ei ole riittävä osoittamaan, että kieli ei ole kontekstiton. Olkoon kieli $L = \{a^q b^r c^s d^t | q = 0 \vee r = s = t\}$. Oletetaan, että kieli L on kontekstiton. Mikäli valitsemme $z = b^r c^s d^t$ ja kirjoitamme $z = uvwxy$, niin on aina mahdollista valita u, v, w, x ja y siten, että $uv^i wx^i y$ kuuluu kieleen L, kaikilla i > 0. Valitkaamme osasana vwx sisältämään ainoastaan merkkejä b. Mikäli valitsemme $z = a^q b^r c^r d^r = uvwxy$, niin v ja x saattavat sisältää ainoastaan merkkejä a, jolloin $uv^i wx^i y$ edelleen kuuluu kieleen L, kaikilla i ≥ 0. Jokainen pumpatuksi tuleva merkkijono kuuluu kieleen L. [Hopcroft and Ullman, 1979]
Tapaus 1: q=0, jolloin merkkien b, c ja d lukumäärälle ei ole rajoituksia. Mutta määrät eivät ole samat. Tällöin sanan z ensimmäinen merkki ei ole a. Tässä tapauksessa valinnat v=b, u=w=x=λ pätevät.
Tapaus 2: q > 0, jolloin tiedämme, että r=s=t ja valinnat v=a, u=w=x=λ pätevät. [Martin, 2003]

Koska emme pysty osoittamaan Bar-Hillelin lemmalla, että kieli ei ole kontekstiton, niin tarvitsemme vahvemman lemman. Tutkimme kyseistä kieltä luvussa 6 esimerkissä 6.0.2, jossa kielen todistaminen ei-kontekstittomaksi onnistuu käyttämällä Ogdenin lemmaa.

Havaitsemme, että sekä säännöllisten että kontekstittomien kielten lemmalla merkkijonoja voidaan pumpata yhdestä kohdasta, kun jälkimmäisessä tarkasteltavan kie-

len sanassa z joko v tai x on tyhjä.

4.2 Bar-Hillelin lemman toteuttavista kielistä

Sandor Horvath [1978] kutsuu Bar-Hillelin lemmaa lauseessa 4.0.1 esitetyssä muodossa täydeksi muodoksi. Hän esittelee kieliä, jotka toteuttavat Bar-Hillelin lemman, jolloin hän määrittelee kyseiset kielet "täyteen BH" -perheeseen kuuluviksi. $\mathscr{B}_0$ tarkoittaa täyttä BH-perhettä, $\mathscr{B}_1$ perhettä, joka täyttää heikomman ehdon uv^iwx^iy, kun i $\geq$ 1. $\mathscr{BR}_0$ ja $\mathscr{BR}_1$ tarkoittavat vastaavasti BH-perheitä, joissa |vw| = 0, jolloin merkkijonosta uv^iwx^iy merkkejä sisältää osamerkkijono ux^iy. $\mathscr{L}_3$ merkitsee säännöllisiä kieliä, $\mathscr{L}_0$ rekursiivisesti numeroituvia kieliä, L_{re} rekursiivisia kieliä ja L_L kielten luokkaa, jossa on äärellinen $\Sigma_1 \subseteq \Sigma$, missä $L \subseteq \Sigma^*$. Tunnetusti seuraavat sisältymiset pätevät:

$\mathscr{L}_3 \subseteq \mathscr{L}_2 \subseteq \mathscr{L}_1 \subseteq L_{re} \subseteq \mathscr{L}_0 \subseteq L_L$.

Esittelemme pian tuloksen, jonka avulla mainitut neljä BH-perhettä liittyvät toisiinsa joukko-opillisen sisältymisen näkökulmasta. [Horvath, 1978]

Määritelmä 4.2.1.
Lineaarinen rajoitettu (linear bounded automata) automaatti on järjestelmä $M = (Q, \Sigma, \Gamma, \delta, q_o, c, \$, F)$, missä
Q on äärellinen tilajoukko,
Σ on äärellinen nauhamerkkien aakkosto,
Γ nauha-aakkosto,
δ on siirtymärelaatio δ: $Q \times (\Sigma \cup \{\lambda\}) \times \Gamma \rightarrow Q \times \Gamma \times$ (L,R,S),
q_o on alkutila,
c on nauhan vasen loppumerkki,
\$ on nauhan oikea loppumerkki,
F on lopputilojen joukko.
[Hopcroft and Ullman, 1979]

Apulause 4.2.2. Kieliperheiden $\mathscr{B}_0$, $\mathscr{B}_1$, $\mathscr{BR}_0$ ja $\mathscr{BR}_1$ keskuudessa pätevät seuraavat ominaisuudet:
$\mathscr{B}_0 \subseteq \mathscr{B}_1$, $\mathscr{BR}_1 \subseteq \mathscr{B}_1$, $\mathscr{B}_0$ - $\mathscr{BR}_1 \neq \emptyset$, $\mathscr{BR}_1$ - $\mathscr{B}_0 \neq \emptyset$, ja $\mathscr{BR}_0 \subseteq \mathscr{B}_0 \cap \mathscr{BR}_1$.
[Horvath, 1978]

Todistus. Olkoot $L_{11} = \{a^m b^n c^n | 0 \leq m, \leq n\}$, $L_{12} = \{a^m b^m | m \geq 0\}$,
$L_{13} = \{a^{m^2} b^n | m \geq 0, n \geq 1\}$ ja $L_{14} = \{a^m b^m c^n | m \geq 0, n \geq 1\}$.
Tällöin pätee $L_{11} \in \mathscr{B}_1$ - $\mathscr{B}_0$, $L_{11} \in \mathscr{B}_1$ - $\mathscr{BR}_1$, $L_{12} \in \mathscr{B}_0$ - $\mathscr{BR}_1$, $L_{13} \in \mathscr{BR}_1$ - $\mathscr{B}_0$ ja $L_{14} \in (\mathscr{B}_0 \cap \mathscr{BR}_1)$ - $\mathscr{BR}_0$. [Horvath, 1978] □

Kielistä L_{11}, L_{13} ja L_{14} ovat kontekstisia ja ne ovat tunnistettavissa lineaarisesti rajoitetulla automaatilla. L_{12} on kontekstiton ja se on tunnistettavissa pinoautomaatil-

la. Horvath todistaa täyden BH-ominaisuuden olevan ainoastaan riittävä ehto kielen olemiselle kontekstiton.

Käsitteiden säilyttävä katenaatiosulkeutuminen, säilyttävä morfismi, käänteinen morfismi osalta ks. [Salomaa, 1973].

Määritelmä 4.2.3. AFL on abstrakti kieliperhe jos ja vain jos se sisältää ei-tyhjän kielen ja se on suljettu seuraavien operaatioiden suhteen: yhdiste, säilyttävä katenaatiosulkeutuminen, säilyttävä morfismi, käänteinen morfismi, ja säännöllisten kielten leikkaus. AFL on täysi, jos ja vain jos se on suljettu mielivaltaisen morfismin suhteen. [Salomaa, 1973]

Apulause 4.2.4. Kieliperheet $\mathscr{B}_0$, $\mathscr{B}_1$, $\mathscr{BR}_0$ ja $\mathscr{BR}_1$ toteuttavat kaikki muut AFL-aksioomat paitsi sulkeutuvuus käänteisen morfismin ja säännöllisten joukkojen leikkauksen suhteen. [Horvath, 1978]

Todistus. Todistamme kaksi väitettä koskien ei-sulkeutuvuutta. Olkoon $L_{15} = L_{13} \cup a^*$ ja olkoot h: a $\mapsto$ a, b $\mapsto$ ab morfismi ja $L_{16} = a^*$b ($\in L_{13}$). Täten saamme $L_{15} \in \mathscr{BR}_0$, kun taas $h^{-1}(L_{15}) = \{a^{m^2-1}b | m \geq 1\} \cup a^* \notin \mathscr{B}_1$ ja $L_{15} \cap L_{16}$ $\{a^{m^2}b | m \geq 0\} \notin \mathscr{B}_1$. [Horvath, 1978] □

Todistuksessa h^{-1} merkitsee käänteistä morfismia ja silloin kieliluokka ei muutu. Havaitsemme, että operaation myötä n katoaa ja parametrin m arvo muuttuu niin, että m ≥ 1. h^{-1} (L_{15}) koostuu kielen L_{15} merkkijonoista.

Lause 4.2.5. $\mathscr{B}_0 \cap (\mathscr{L}_1 - \mathscr{L}_2) \neq \emptyset$. [Horvath, 1978]

Todistus. Muodostamme kieliperheen $\mathscr{B}_0 \cap (\mathscr{L}_1 - \mathscr{L}_2)$ kielen L. Tarkastellaan kieltä $L \subseteq \{a, b, c\}$*, joka muodostuu sanoista, jotka saadaan kielen L' $= \{r^j s^l t^m | j, m \geq l \geq 0\}$ sanoista a^+ b^+ korvaamalla mielivaltainen alkio kullakin kirjaimella r ja t, ja merkkijonon a^+ c^+ korvaamalla mielivaltainen alkio kullakin kirjaimella s. Kutsumme korvattuja sanoja r-, s- tai t-osasanoiksi sen mukaan, mitä alkion w kirjainta ne edustavat. On todistettava, että kieli L ei ole kontekstiton. Tehdään vastaoletus: olkoon L jonkin kontekstittoman kieliopin tuottama, missä sääntöjen oikeiden puolien maksimaalinen pituus on d. Olkoot z_1, z_2, ... kielen L alkioiden sellainen ääretön sarja, että sanan z_i s-osasanojen lukumäärä l_1 lähestyy ääretöntä, jos i $\to \infty$. Kullekin alkiolle i olkoon osasanan T_i sanan z_i derivaatiopuu ja T_i' pienin derivaatiopuun T_i sellainen osapuu, että sen perusmerkkijonot sisältävät kaikki sanan z_i s-osasanat. Osapuun T_i' välittömien osapuiden joukossa on yksi, olkoon sen juuri A_i, jonka perusmerkkijono sisältää vähintään $(p_i + 1 - d)/d$ s-osasanaa, eikä sisällä sekä r- tai t-osasanaa kerrallaan. On oltava sellainen apumerkki D, joka esiintyy sarjassa A_i mielivaltaisen usein. Jos A_{i_1} ja A_{i_2} ovat apumerkin D kaksi sellaista esiintymää, että i_2 - i_1 on riittävän suuri, sen jälkeen korvaamalla osapuun T_{i_2} A_{i_2}- osapuu A_{i_1}-

osapuulle T_{i_1} saamme kielen L alkion, missä s-osasanojen lukumäärä ylittää joko r-osasanojen tai t-osasanojen lukumäärän, jolloin tuloksena on ristiriita. [Horvath, 1978] □

Seuraavaksi esitetään toinen todistus lauseelle 4.2.5.

Todistus. Olkoot a, b, c $\in \Sigma_1$, h $\subseteq \Sigma_1^*$, h $\in \mathscr{L}_1$ - $\mathscr{L}_2$, ja $L = (\{a^n bc^n | n \geq 1\}$ h) $\cup$ (b Σ_1^*) ($\in \mathscr{L}_1$). Oletetaan, että $L \in \mathscr{L}_2$, joten se on jonkin pinoautomaatin M hyväksymä. Pinoautomaatista M on rakennettavissa toinen sellainen M_1, että mikä tahansa sana w $\in \Sigma_1^*$ on pinoautomaatin M_1 hyväksymä, jos ja vain jos abcw on pinoautomaatin M hyväksymä. Toisin sanoen h on pinoautomaatin M_1 hyväksymä, mikä merkitsee ristiriitaa. [Horvath, 1978] □

Lause 4.2.6. $\mathscr{B}_0 \cap (L_{re}$ - $\mathscr{L}_1) \neq \emptyset$. [Horvath, 1978]

Todistus. Otetaan kielen (L_{re} - $\mathscr{L}_1$) lause h ja määritetään L täsmälleen kuten lauseen 4.2.5 todistuksessa. On todistettava, että L ei ole kontekstinen. Olkoon L jonkin lineaarisen rajoitetun automaatin M hyväksymä, sen jälkeen toisen lineaarisen rajoitetun automaatin M_1 hyväksymä, joka ensin liittää merkkijonon abc sen syötteen w alkuun ja sen jälkeen toimii samoin kuten M toimisi syötteellä abcw, hyväksyy lauseen h, mikä merkitsee ristiriitaa. [Horvath, 1978] □

Lause 4.2.7. $\mathscr{B}_0 \cap (\mathscr{L}_0$ - $L_{re}) \neq \emptyset$ ja $\mathscr{B}_0 \cap (L_L$ - $\mathscr{L}_0) \neq \emptyset$. [Horvath, 1978]

Todistus. Sama argumentti kuin lauseen 4.2.6 todistuksessa, paitsi että nyt h $\in \mathscr{L}_0$ - L_{re}, tai h $\in L_L$ - $\mathscr{L}_0$ vastaavasti, ja M sekä M_1 ovat lineaarisen rajoitetun automaatin sijaan Turingin (ks. [Hopcroft and Ullman, 1979]) koneita. [Horvath, 1978] □

5 Lemmat lineaarisille ja ei-lineaarisille kontekstittomille kielille

Tässä luvussa esittelemme pumppauslemmat lineaarisille ja ei-lineaarisille kontekstittomille kielille.

Lause 5.1. (Horvathin lemma lineaarisille kontekstittomille kielille)
Olkoon L lineaarinen kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku k, että kun $|z| \geq k$, niin sana z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että
1. $|uvxy| \neq 0$.
2. $|vwx| \leq k$.
3. $uv^iwx^iy \in L, i \geq 0$.

Todistus. Sivuutetaan (ks. [Mateescu and Salomaa, 1997]). □

Lauseella 5.1 on osoitettavissa, että esimerkiksi kieli $L = \{a^ib^ic^jd^j | i, j \geq 0\}$ ei ole lineaarinen. Lemma perustuu Bar-Hillelin lemmaan, josta tämän ja lauseen 5.2 yhteydessä puhumme. Eroavuudet ovat oletusten lisäksi lauseiden 4.0.1 ja 5.1 ehdossa 1. Horvath on laajentanut ehtoa 1 lisäämällä siihen osasanat u ja y. Ehdossa 1 on mahdollista, että ainoastaan yksi osasanoista ei ole tyhjä, ja kahden osasanan tyhjyys luonnollisesti selittyy ehdossa 3 pumpattaessa merkkijonoja arvolla i=0. Täten joko u tai y on ei-tyhjä.

Horvath esittelee myös uuden lemman, joka kaikista aiemmista lemmoista poiketen tarjoaa välttämättömän ehdon kielten kuulumiselle ei-lineaaristen kontekstittomien kielten luokkaan. Hän osoittaa myös, että lemma ei ole verrattavissa Bar-Hillelin lemmaan.

Lause 5.2. (Uusi pumppauslemma)
Olkoon L ei-lineaarinen kontekstiton kieli. Silloin on olemassa äärettömän monta sellaista sanaa $z \in L$, että ne voidaan kirjoittaa muodossa $z = uvwxy$ siten, että
1. $|su| \neq 1$.
2. $|wy| \neq 1$.
3. $rs^itu^ivw^jxy^jz \in L, i, j \geq 0$.

Todistus. Olkoon $G = (N,T,P,S)$ sellainen kontekstiton kielioppi, että $L(\mathrm{G}) = L$ ja olkoon $G_A = (N,T,P,A)$ kaikille $A \in N$. Koska kieli L on ei-lineaarinen, on olemassa sellaiset apumerkit $A, B \in N$ ja sellaiset merkkijonot $\alpha, \beta, \gamma \in T^*$, että S

$\Rightarrow^* \alpha A \beta B \gamma$, missä molemmat kielistä $L(G_A)$ ja $L(G_B)$ ovat äärettömiä. Silloin sanat $\alpha L(G_A) \beta L(G_B) \gamma \in L$. Soveltamalla Bar-Hillelin lemmaa kieliin $L(G_A)$ ja $L(G_B)$ saamme $\alpha u_1 v_1^i w_1 x_1^i y_1 \beta u_2 v_2^j w_2 x_2^j y_2 \gamma \in L$. Nyt voidaan valita r = αu_1, s = v_1, t = w_1, u = x_1, $v = y_1 \beta u_2$, w = v_2, x = w_2, y = x_2, $z = y_2 \gamma$, mikä on eksponenteitta konkatenoituna yllä. [Horvath, 2006] □

Lauseessa 5.1 tarkastellaan lineaarista kieltä, kun taas lauseessa 5.2 tarkastellaan ei-lineaarista kontekstitonta kieltä. Kaikissa edellisissä lemmoissa annettu kokonaisluku on poistettu lauseesta 5.2. Lauseen 5.2 ehdossa 1 osasana vx on korvattu osasanalla su, jonka pituus ei voi olla 1. Ehdossa 2 osasanat v ja x on nyt korvattu osasanalla y, ja osasanan wy pituus ei voi olla 1. Horvath tulee itse asiassa sijoittaneeksi Bar-Hillelin lemmaan kuuluvan sanan $uvwxy$ ehdon 3 sisälle, mutta eksponentteja on jopa neljä. Ensimmäinen ja toinen eksponentti ovat erisuuret kuin kolmas ja neljäs eksponentti. Ehdossa 3 osasanojen määrä lisääntyy neljällä. Lisäksi pumpattavia lohkoja on siis neljä. Ehtoon 3 lähes mahtuukin kaksi Bar-Hillelin lemman sanaa $uvwxy$, joista jälkimmäisessä pumpattava lohko alkaisi poikkeuksellisesti ensimmäisessä osasanassa, jossa (kuten jälkimmäisen merkkijonon kolmannessa merkissä) olisi eri eksponentti kuin ensimmäisessä merkkijonossa. Täten ajateltaessa merkkijonoja erillisinä, jälkimmäisessä pumpattavat lohkot on siirretty yksi askel vasemmalle ja lisäksi osasanoja on viiden sijaan neljä. Lauseessa 5.2 ei siis aseteta vastaavia ehtoja ei-pumpattaville merkkijonoille kuin kaikissa muissa esittelemissämme lemmoissa. Lisäksi Horvath [2006] yhtenäistää ehtoja 1 ja 2 niin, että lemmassa annettu k ei liity kumpaankaan, vaan hän siis poistaa kokonaisluvun k ja määrittää molemmissa ehdoissa osasanan pituuden olevan erisuuri kuin 1. On siis mahdollista, että kaikki pumpattavat merkkijonot ovat tyhjiä, jolloin ei ole merkitystä, millä arvolla pumpataan.

Lisäksi kummankin ehdon kaikki osasanat tulevat pumppauksen kohteeksi, toisin kuin Bar-Hillelin lemmassa, jossa ehdon 2 osasana w jää aina pumppauksen ulkopuolelle. Lisäksi Horvathin ehdoissa 1 ja 2 kaikki neljä osasanaa ovat eri osasanoja, kun taas Bar-Hillelilla osasanoja on viisi ja niistä v ja x esiintyvät kahteen kertaan, mutta w ainoastaan kerran. Voimme ajatella Horvathin ehdon 1 korvaavan Bar-Hillelin ehdon 1, koska osasana su on ensimmäisessä kaksi samaa eksponenttia käsittävässä osasanassa ja nyt Horvathin ja Bar-Hillelin sanat olisivat pituudeltaan samat. Mutta jos Horvathin ehto 2 korvaa Bar-Hillelin ehdon 1, niin viisiosainen sana korvautuu neljällä ja pumppaus tapahtuukin ensimmäisessä ja kolmannessa lohkossa. Lemmojen välisenä eroavuutena havaitsemme myös sen, että Horvath rajaa pumppauksen ulkopuolelle viisi osasanaa Bar-Hillelin rajatessa kolme. Riippuen tarkasteltavasta kielestä Horvathin lemma mahdollistaa suuremman merkkien sekaisin menemisen kuin Bar-Hillelin lemma.

Se, kumpaa lemmaa kannattaa soveltaa, on kieliriippuvaista. Nimittäin toisinaan Horvathin lemman ehdot eivät toteudu jossain kielessä ja toisinaan Bar-Hillelin lemman ehdot eivät toteudu jossain muussa kielessä. Joissakin tapauksissa meidän on

kokeiltava molempia lemmoja. [Horvath, 2006]

Uusi lemma pätee, toisin kuin Bar-Hillelin lemma, seuraavassa tapauksessa: Olkoon $\psi \subseteq \{1^2, 2^2, 3^2, \dots\}$ ja kieli $L_\psi = \{a^l b^l a^m b^m | l, m \geq 1; l \in \psi \vee l \in \psi\} \cup \{a^n b^n | n \geq 1\}$. Joukko ψ koostuu siis täydellisistä neliöistä, jollaisia olemme tarkastelleet luvussa 3. Mutta Bar-Hillelin lemma on vahvempi kuin uusi lemma tarkasteltaessa esimerkiksi kieltä $L = \{a^i b^i c^j d^j | i, j \geq 0\} \cup \{d^l e^l f^l | l \geq 0\}$. [Horvath, 2006]

Uudella lemmalla on sovellus, jossa uuden lemman ehtoja toteuttamaton kontekstiton kieli on lineaarinen. Esimerkkinä katsomme seuraavaa: Olkoon kieli $L = \{a^i b^{2i} | i \geq 0\}$ joka voidaan tuottaa kieliopilla $G = (\{S, B\}, \{a, b\}, S, \{S \to aSB, S \to \lambda, B \to bb\})$. Koska L ei toteuta uuden lemman ehtoja, niin kieli on lineaarinen. [Horvath, 2006]

6 Ogdenin lemma

Tässä luvussa esittelemme Ogdenin lemman, joka on vahvempi kuin Bar-Hillelin lemma, ja esimerkkejä Ogdenin lemman käytöstä. Sen jälkeen tutkimme luontaista moniselitteisyyttä, josta tarkastelemme myös esimerkkiä, ja esittelemme Ogdenin lemman toteuttavia kieliä. Lopuksi tarkastelemme Ogdenin lemman yleistettyä versiota.

Lause 6.0.1 Jokaiselle kontekstittomalle kieliopille $G = (N,T,P,S)$ on olemassa sellainen kokonaisluku k, että kaikille sanoille $z \in L(G)$, jos k tai enemmän erillistä merkkiä sanassa z on merkitty, niin on olemassa sellainen $A \in N$ ja sellaiset osamerkkijonot u, v, w, x ja $y \in T^*$, että

1. $S \Rightarrow^* uAy \Rightarrow^* uvAxy \Rightarrow^* uvwxy = z$.
2. w sisältää vähintään yhden merkityn merkin.
3. vx sisältää vähintään yhden merkityn merkin.
4. vwx sisältää korkeintaan k merkittyä merkkiä.

Todistus. Tarkastellaan sanaa z vastaavaa derivaatiopuuta. Kutsumme solmua s Θ-solmuksi, jos solmulla s on sellaiset välittömät jälkeläiset t_1 ja t_2, että sekä solmulla t_1 että t_2 on jälkeläisiä, joilla on vähintään yksi merkitty merkki sanassa z. Olkoon d minkä tahansa säännön oikean puolen maksimaalinen pituus sääntöjoukossa P. Jos jokainen polku sanan z derivaatiopuussa sisältää korkeintaan i Θ-solmua, silloin z sisältää korkeintaan d^i merkittyä merkkiä. Nyt niistä poluista, jotka sisältävät mahdollisimman paljon Θ-solmuja ja jotka päättyvät solmuissa, jotka ovat merkittyjä merkkejä sanassa z, valitsemme jonkin polun $s_0, \ldots, s_{n-1}$. Olkoon ξ on merkkien lukumäärä aakkostossa N ja olkoon $k = d^{2\xi+3}$. Nyt polku $s_0, \ldots, s_{n-1}$ sisältää vähintään $2\xi + 3$ Θ-solmua. Valitsemme polun alaosasta osapolun $s_m, s_{m+1}, \ldots, s_{n-1}$, joka sisältää täsmälleen $2\xi + 3$ Θ-solmua.

Jaamme Θ-solmut tässä osapolussa kahteen joukkoon J_Q ja J_R niin, että joko joukon J_Q tai J_R täytyy sisältää vähintään ξ+2 Θ-solmua. Θ-solmu s_j, missä m ≤ j < n, on joukossa J_Q (vastaavasti J_R), jos sillä on välitön jälkeläinen t, jolla ei ole jälkeläisenä s_{n-1}, mutta jolla on merkitty jälkeläinen, joka on solmun s_{n-1} vasemmalla (vastaavasti oikealla) puolella.

Tarkastelemme tapausta, jossa joukolla J_Q on vähintään $\xi + 2$ alkiota. Löydetään pienin sellainen kokonaisluku h, että $s_h \in J_Q$. Koska on ainoastaan ξ erilaista solmutunnistetta, täytyy olla sellaiset kokonaisluvut i ja j, missä h < i < j, että solmut s_i ja s_j ovat joukossa J_Q ja molemmilla on sama tunniste A. Saamme toivotun johdon $S \Rightarrow^*$ uAy $\Rightarrow^*$ uv$A$xy $\Rightarrow^* uvwxy = z$ asettamalla w perusmerkkien merkkijonoksi, joka periytyy solmusta s_j, jne.

Nyt tarkastelemme muita vaadittuja ominaisuuksia. Osamerkkijono w sisältää vähintään yhden merkityn merkin, koska s_{n-1} on solmun s_j jälkeläinen. Osamerkkijono v sisältää merkityn merkin, koska s_i on joukossa J_Q. Solmu s_i on solmun s_h jälkeläinen, joten koska s_h on joukossa J_Q, myös u sisältää merkityn merkin. Kun luku i > m, niin polku $s_i, \ldots, s_{n-1}$ sisältää korkeintaan $2\xi + 3$ Θ-solmua. Koska polku $s_0, \ldots, s_{n-1}$ sisältää maksimaalisen määrän Θ-solmuja, kukin solmun s_i sisältävä polku osapuussa sisältää korkeintaan $2\xi + 3$ Θ-solmua. Täten vwx sisältää korkeintaan $d^{2\xi+3}$ = k merkittyä merkkiä. Tapaus, jossa ainoastaan joukossa J_R on vähintään $\xi + 2$ alkiota, käsitellään vastaavasti. [Ogden, 1968] □

Teemme ensin havaintoja lauseesta 6.0.1. Ehto 3 voidaan kirjoittaa myös seuraavasti: Joko osamerkkijonoista u ja v molemmat sisältävät merkittyjä merkkejä, tai x ja y molemmat sisältävät merkittyjä merkkejä. [Ogden, 1968] Merkitsemme vähintään k merkkiä sanassa z valintamme mukaisesti. Mutta kaikkien merkkien merkitseminen ei vastaisi Ogdenin lemman tarkoitusta. Mikäli osamerkkijonot v ja x (i voi siis olla 0) tulevat pumpatuiksi, niin vähintään toinen niistä, joita kutsumme pumpattaviksi lohkoiksi, sisältää merkityn merkin. Merkkijono vwx sisältää vähintään kaksi merkittyä merkkiä. Merkittyjen merkkien kuuluessa kolmeen osamerkkijonoon sisältyy w aina johonkin niistä. Ogdenin tarkoituksena on kohdistaa pumppaus tiettyihin merkittyihin osamerkkijonoihin. Pumppaus voi kohdistua myös joihinkin merkitsemättömiin osamerkkijonoihin. Mikäli poistamme kaikki merkitsemiset Ogdenin lemmasta, niin lemma on ”redusoitavissa” Bar-Hillelin lemmaksi.

Seuraavaksi esitämme havaintoja lauseen 6.0.1 todistuksesta. Aluksi valittu jokin polku johtaa siis perusmerkkijonoihin. Toistettavat apumerkit sijaitsevat kulloisenkin polun varrella. Jälkeläisten t_1 ja t_2 sisältämien merkittyjen merkkien lukumäärä ei ole välttämättä sama. Merkittyjen merkkien lukumäärälle on voimassa $d^i > d^{\xi+1}$ ja edelleen $d^{2\xi+3} > d^{2\xi} + 3$. Θ-solmut ovat juuren ja alimpina olevien jälkeläisten, joilla ei siis ole jälkeläisiä, välissä. Lisäksi Θ-solmut ovat keskeisessä roolissa, koska niillä on sekä vasemmalla että oikealla puolella jälkeläisiä, joilla on vähintään yksi merkitty jälkeläinen. On olemassa muitakin solmuja, joiden jälkeläisillä ei ole merkittyjä merkkejä, jolloin kyseiset solmut eivät ole olennaisia. vx sisältää vähintään yhden merkityn merkin, koska derivaatiopuussa osamerkkijonoon v ja/tai x sijoittuu merkittyjä jälkeläisiä, jotka ovat lähtöisin Θ-solmusta. Derivaatiopuussa on siis solmu s_i, johon liittyy apumerkki A ja tämän solmun alla on solmu s_j, johon liittyy sama apumerkki A. Solmuun s_h ei liity apumerkkiä A. Solmu s_{n-1} sijaitsee pumpattaessa solmun s_j alapuolella. Solmu s_j kohdistuu osasanan w ylläolevan osapuun vasemmalle ja oikealle puolelle. Jälkeläiset t_1 ja t_2 sijaitsevat derivaatiopuussa esimerkiksi osasanan w yläpuolella. J_Q tarkoittaa vasenta joukkoa. Osasanat u ja y liittyvät solmuun s_i sekä v ja x solmuun s_j. Solmu s_i kohdistuu derivaatiopuussa vasempaan ja oikeaan reunaan. Pienin Θ-solmujen lukumäärä polulla on $2\xi + 3$, koska valittiin $k = d^{2\xi+3}$. Koska polku $s_0, \ldots, s_{n-1}$ sisältää vähintään $2\xi + 3$ Θ-solmua, niin kyseinen polku sisältää solmuja t_1 ja t_2 vähintään $2\xi + 3$ kappaletta. Solmuun s_j

voi liittyä enemmän merkittyjä merkkejä kuin solmuun s_i. Solmujen s_h ja s_m välissä on jälkeläisiä. Jokainen polku, joka on solmussa s_i, voi enimmillään sisältää yhtä paljon Θ-solmuja kuin polku $s_i, \ldots, s_{n-1}$, kun i > m. Kokonaisuudessaan voimme kuvata polkua seuraavasti: $s_0 < s_h < s_i < s_j < s_m < s_{m+1} < s_{n-1} < s_n$. Polku $s_j, \ldots, s_{n-1}$ siis sisältää vähemmän Θ-solmuja kuin polku $s_0, \ldots, s_{n-1}$. Havaitsemme, että $\xi + 2$ sisältyy tapauksessamme joukkoon J_Q. Mikäli joukossa J_R on $\xi + 1$ Θ-solmua, niin mainitut joukot sisältävät yhteensä vähintään $2\xi + 3$ Θ-solmua. Merkkijonojen $uvwxy$ ja uv^5wx^5y välillä edetään solmusta s_j kohti solmua s_{n-1}.

Esimerkki 6.0.2. Jatkamme esimerkin 4.1.6 käsittelyä. Olkoon kieli $L = \{a^q b^r c^s d^t | q = 0 \vee r = s = t\}$. Oletetaan, että kieli L on kontekstiton. Olkoon k kokonaisluku Ogdenin lemman mukaan ja olkoon $z = ab^k c^k d^k$. Nyt merkitsemme kaikki paitsi sanan z ensimmäisen merkin, jolloin $b^k c^k d^k$ merkitään. Oletetaan, että u, v, w, x ja y toteuttavat lemman ehdot. Nyt osasanan vx täytyy sisältää jokin merkeistä b, c tai d, siis vähintään yksi merkitty merkki. Osasanassa vx ei voi olla b, c ja d, sillä muutoin pumppaus kohdistuisi jokaiseen merkeistä b, c ja d, mikä ei ole mahdollista. Osasanan vwx täytyy sisältää korkeintaan k merkkiä, jotka on merkitty. Täten merkkijonossa uv^2wx^2y on yksi a, eikä merkkijonossa ole samaa määrää merkkejä b, c ja d. Myöskään merkkijono uv^0wx^0y ei kuulu kieleen L, sillä merkkijonossa on yksi a, eikä merkkien b, c ja d lukumäärä ole sama. Tällöin ainoastaan uv^iwx^iy arvolla s=1 $\in L$. Täten olemme osoittaneet, että kieli L ei ole kontekstiton. [Martin, 2003; Hopcroft and Ullman, 1979]

Esimerkki 6.0.3. Olkoon kieli $L = \{a^q b^r c^s | q \neq r \neq s \neq q\}$. Nyt merkkien a,b ja c lukumäärällä on merkitystä, mutta lukumäärien ei ole välttämätöntä esiintyä nousevassa tai laskevassa järjestyksessä. Lukumääristä q, r ja s yksi voi olla 0. Tehdään vastaoletus, että kieli L on kontekstiton. Olkoon k kokonaisluku Ogdenin lemman mukaan. Tarkastellaan sanaa $z{=}a^k b^{k+k!} c^{k+2k!}$. Olkoot a-merkit merkittyjä ja olkoon $z{=}uvwxy$. Jos joko v tai x sisältää kaksi erillistä merkkiä ja i=3 ja jos esimerkiksi a^+b^+ sisältää osasanan v, ja siis i=3, jolloin v=$a^q b^r a^q b^r a^q b^r$, niin sanassa uv^3wx^3y jotkut b-merkit esiintyvät ennen a-merkkejä, minkä seurauksena $uv^3wx^3y \notin L$. Nyt osasanoista v ja x vähintään toisen täytyy sisältää a-merkkejä, koska ainoastaan ne ovat merkittyjä. Täten mikäli b^*c^* sisältää osasanan x, niin a^+ sisältää osasanan v. Jos a^+ sisältää osasanan x, niin a^* sisältää osasanan v, jotta a esiintyisi ennen b- tai c-merkkiä. Jos r=0, niin pumpattaessa arvolla i=0 saamme tulokseksi q=0 ja r=0, joten saatava sana ei voi kuulua kieleen L. Tässä tapauksessa c-merkin lukumäärällä ei ole merkitystä. Mutta muutoin on oltava niin, että joko a^+ tai b^+ sisältää osasanan v tai että b^+ tai c^+ sisältää osasanan x. [Hopcroft and Ullman, 1979]

Mahdollisia ovat tapaukset, joissa a^+ sisältää osasanat u, v, w ja x. Silloin b^*c^+ tai b^+c^* sisältää osasanan y. Tarkastelemme tapausta, jossa b^* sisältää osasanan x ja a^+ sisältää osasanan v, jolloin osasana w saattaa sisältää b-merkkejä. Olkoon $\varsigma = |v|$. Täten $1 \leq \varsigma \leq$ k. Voidaan todeta, että k! $\equiv 0$ mod k. Olkoon ψ sellainen kokonaisluku, että $\varsigma\psi =$ k!. Silloin z' = $uv^{2\psi+1}wx^{2\psi+1}y \in$ L. Mutta $v^{2\psi+1} = a^{2\varsigma\psi+\varsigma} = a^{2k!+\varsigma}$, ja pätee myös $|a^{2\varsigma\psi+\varsigma}| = 2\varsigma\psi + \varsigma \equiv 0$ mod k. [Hopcroft and Ullman, 1979]

Koska osasana uwy sisältää täsmälleen $(k - \varsigma)$ kappaletta a-merkkejä, jolloin a-merkkien lukumäärä voi olla 0, sanassa z' on (2k! + k) kappaletta a-merkkejä. Mutta koska osasanoissa v ja x ei ole c-merkkejä, niin uwy ei voi sisältää niitä, sillä u sijaitsee ennen osasanoja v ja x sekä w sijaitsee ennen osasanaa x, joten c-merkit sijaitsevat osasanassa y. Tiedämme, että sanassa z' on (2k! + k) tai vastaavasti $2\varsigma\psi + \varsigma$ kappaletta c-merkkejä, minkä seurauksena $z' \notin L$. Saamme tulokseksi vastaavan ristiriidan, jos a^+ tai c^* sisältää osasanan x, joten kieli L ei ole kontekstiton. [Hopcroft and Ullman, 1979]

6.1 Luontainen moniselitteisyys

Jos kaikki kontekstittoman kielen tuottavat kieliopit ovat moniselitteisiä, niin kieli on luontaisesti moniselitteinen (inherently ambiguous). Kielioppi on moniselitteinen, kun jollakin sen tuottaman kielen sanalla on vähintään kaksi vasenta johtoa. [Hopcroft and Ullman, 1979]

Lause 6.1.1. Olkoot M = $\{a^i ba^{i+1} | i \geq 0\}$, $L_{10} = abM^*$ ja $L_{11} = M^*\{a\}^* b$. Kieli $L_{10} \cup L_{11}$ on luontaisesti moniselitteinen. [Ogden, 1968]

Todistus. Oletetaan, että kontekstiton kielioppi G tuottaa kielen $L_{10} \cup L_{11}$ ja että k_0 on lauseen 6.0.1 määräämä kokonaisluku. Olkoon p = k_0!. Tavoitteemme on osoittaa, että sanalla $z = aba^2ba^3b \ldots ba^{4p}ba^{4p+1}b$ on kaksi eri vasenta johtoa. Ensin tarkastelemme sanaa $z_0 = aba^2ba^3b \ldots ba^{4p-1}ba^pa^pba^{2p+1}b$. Oletamme sanan z viimeistä edellisen a-lohkon ensimmäiset p merkkiä merkityiksi.

Koska p > k_0, saadaan lauseen 6.0.1 nojalla A_0 ja u_0, v_0, w_0, x_0 ja y_0, joilla kaikilla on ominaisuudet 1-4. Koska $z_0 \in L_{10}$, ja $z_0 \notin L_{11}$, havaitsemme, että ominaisuudet 1-3 pakottavat osasanan v_0 olemaan sanan z_0 merkityissä merkeissä ja osasanan x_0 olemaan sanan z_0 viimeisessä a-lohkossa. Osasana $v_0w_0x_0$ ei voi kokonaan sisältyä sanan z_0 viimeistä edelliseen a-lohkoon, koska $S \Rightarrow^* u_0w_0y_0$. Osasana $v_0w_0x_0$ ei voi sisältää sanan z_0 kolmanneksi viimeistä b-merkkiä, koska osasanojen v_0 ja x_0 täytyy yhdessä sisältää parillinen lukumäärä b-merkkejä ja $S \Rightarrow^* u_0{v_0}^2w_0{x_0}^2y_0$. Olkoon j_0 = $|v_0|$ = $|x_0|$, ja käyttämällä ominaisuutta 4, saamme $k_0 \geq |v_0| = j_0 > 0$. Nyt mille tahansa q ≥ 0, saamme ominaisuudesta 1, että

$S \Rightarrow^* u_0A_0y_0 \Rightarrow^* \ldots \Rightarrow^* u_0{v_0}^qA_0{x_0}^qy_0 \Rightarrow^* u_0{v_0}^qw_0{x_0}^qy_0$.

Valitsemalla $q_0 = 2k_0!/j_0 + 1$, saamme johdon

$S \Rightarrow^* u_0{v_0}^{q0}w_0{x_0}^{q0}y_0 = aba^2b \ldots ba^{4p-1}ba^{\,p+(q_0-1)j_0}\mathrm{a}^p\ \mathrm{ba}^{2p+(q_0-1)j_0+1}\ \mathrm{b}$.

Johdettu sana on z, koska (q_0 - 1)j_0 = 2p.

Tarkastellaan sanaa $z_1 = aba^2ba^3b \ldots\ ba^{4p-2}ba^{2p-1}ba^pa^pba^{4p+1}b$. Tässä tapauksessa oletamme viimeiset p merkkiä viimeistä edelliseen a-lohkoon merkityiksi. Koska edelleen p > k_0, saamme A_1 ja u_1, v_1, w_1, x_1 ja y_1, joilla on ominaisuudet 1-4. Kuten tapauksen z_0 kohdalla, saamme sellaiset j_1 ja q_1, että
$S \Rightarrow^* u_1{v_1}^{q1}w_1{x_1}^{q1}y_1 = aba^2b \ldots\ ba^{4p-2}ba^{2p+(q_1-1)j_1-1}ba^pa^{p+(q_1-1)j_1}ba^{4p+1}b = z$.

Havaitsemme, että kaksi sanan z johtoa ovat erilliset, sillä ensimmäisessä osasana ${v_0}^{q0}w_0{x_0}^{q0}$ on johdettu apumerkistä A_0, toisessa osasana ${v_1}^{q1}w_1{x_1}^{q1}$ on johdettu apumerkistä A_1. Valitsimme merkityt merkit kahdessa tapauksessa siten, että nämä kaksi osasanaa lomittuvat, mutta kumpikaan ei ole toiseensa sisältyvä. Täten olemme osoittaneet kieliopin G moniselitteisyyden. [Ogden, 1968] □

Seuraavaksi esitämme huomioita todistuksesta. Sanassa z_0 on siis merkitty toiseksi viimeisen osamerkkijonon a^p sisältämät ensimmäiset p merkkiä ja vastaavasti sanassa z_1 toiseksi viimeisen osamerkkijonon a^p sisältämät viimeiset p merkit. Jos osasanojen v_0 ja x_0 yhdessä sisältämien a-merkkien parillinen lukumäärä on ${\Psi_a}^{v_0x_0}$, niin pätee, että ${\Psi_a}^{v_0x_0} > j_0$. Vastaavasti b-merkkien osalta voidaan todeta, että ${\Psi_b}^{v_0x_0} < {\Psi_b}^{w_0y_0}$. Jos a-merkkien lukumäärä osasanassa u_0 on ζ_a, niin voidaan todeta, että ${\zeta_a}^{u_0} > {\zeta_a}^{y_0}$. Edelleen todetaan, että ${\Psi_a}^{v_0x_0} < {\zeta_b}^{u_0w_0y_0}$. On voimassa $k_0! > k_0 \geq |x_0| = j_0 = |v_0| \geq 1$ ja $2p > k_0 \geq |v_1| = |x_1| = j_1 \geq 1$. Voidaan todeta, että $|x_0| < 2j_0 \leq 2k_0 < k_0! + 3p - v_0$. Kun q ≥ 0, niin pumppaus arvolla i = 0 on mahdollista. Havaitsemme, että $ba^pa^pba^{2p+1}b = ba^{4p+1}bb = ba^{4k_0!+1}bb$. Valinnassa q_0 saadaan siis $(q_0 - 1)j_0 + (q_0 - 1)j_0 + 1 = 2p + 2p + 1 = 4p + 1$. Olisi virheellistä tehdä valinta $q_0 = 3k_0!/j_0 + 1$, jolla saataisiin $aba^2b \ldots\ ba^{4p-1}ba^{p+(q_0-1)j_0}\mathrm{a}^{p+(q_0-1)j_0}\mathrm{ba}^{2p+(q_0-1)j_0+1}\mathrm{b}$, sillä silloin pumpattaisiin kolmesta paikasta. Voidaan merkitä myös $3p / |x_0| + 1 = 3k_0! / |v_0| + 1$. Johdetussa sanassa, joka on z, a^pb kuuluu osasanaan w_0 ja viimeinen b kuuluu osasanaan y_0. Osasana $v_0w_0x_0$ ei voi sisältää sanan z_0 myöskään neljänneksi viimeistä tai viimeistä b-merkkiä. Osasana x_0 ei voi sisältää viimeistä b-merkkiä, eikä $\mathrm{a}^{2k_0!+1}$ voi sisältyä osasanaan $u_0v_0w_0$. Osasana y_0 ei voi sisältää a-merkkiä, eikä w_0y_0 voi sisältää parillista lukumäärää b-merkkejä. Sanoissa z_0 ja z_1 viimeisen osasanan merkki on sama. Selvästi j_0 ei voi sisältyä osasanaan $u_0w_0y_0$ eikä j_1 osasanaan $u_1w_1y_1$. Lomittuminen ei kohdistu osasanoihin u_0, y_0, u_1 ja y_1. Valittaessa lomittuminen kohdistumaan osasanoihin u_0, w_0 ja y_0 pumppaus ei olisi mahdollista lomittuviin osasanoihin. Kukin v_0, x_0, v_1 ja x_1 sisältää yhtä merkkiä. Sanassa z_1 toiseksi viimeisen osamerkkijonon a^p sisältämät viimeiset p merkit, jotka ovat merkittyjä, tulevat pumpatuiksi, mutta sanassa z_0 toiseksi viimeisen osamerkkijonon a^p sisältämät ensimmäiset p merkkiä, jotka ovat merkittyjä, eivät tule pumpatuiksi.

Seuraavaksi esitämme vaihtoehtoisen todistuksen lauseelle 6.1.1.

Todistus. Kieli $L = \{a^rb^sc^t | r = s \vee s = t\}$ on kontekstiton, kuten on $\{a^rb^r | i \geq 0\}\mathrm{c}^*$ ∪ a^* $\{b^rc^r | i \geq 0\}$, joista molemmat osat ovat kontekstittomia (kahden kontekstittoman kielen unioni on aina kontekstiton). Olkoon k Ogdenin lemmassa annettu kokonaisluku. Tarkastellaan sanaa $z = a^mb^mc^{m+m!}$, missä m=max(k,3), ja merkitään

a-merkit. Ogdenin lemman mukaan on olemassa osamerkkijonoihin jako $z = uvwxy$ ja sellainen apumerkki A, että $S \Rightarrow^*$ uAy, $A \Rightarrow^*$ vAx, ja $A \Rightarrow^*$ w. Täten uv^iwx^iy $\in L$, kaikille i ≥ 0.

Nyt vx sisältää korkeintaan m kappaletta b-merkkiä, koska niiden lukumäärä määritellään sanassa z. Koska m ≥ 3, niin m < m! ja $|\alpha|_b \leq 2$m < m + m! $\leq |\alpha|_c$. Täten $|\alpha|_a$ $= |\alpha|_b$. Nyt vx sisältää saman verran a- ja b-merkkejä, ja a-merkkejä on vähintään yksi, sillä vx sisältää vähintään yhden merkityn merkin. Joten jos v ja x kumpikin sisältäisi useamman kuin yhdentyyppisiä merkkejä, niin $\alpha \notin L$. Täten $v = a^s$, x = b^s, kun 1$\leq$ s $\leq$ m.

Olkoon i = $\frac{m!}{s}$ +1, jolla saamme johdon $\beta = uv^iwx^iy = a^{m+m!}b^{m+m!}c^{m+m!}$ käyttämällä johtoa $A \Rightarrow^*$ vAx i kertaa. Seuraavaksi menettelemme samoin sanalle z' = $a^{m+m!}b^mc^m$, paitsi, että nyt merkitsemmekin c-merkit. Saamme osamerkkijonoihin jaon $z' = u'v'w'x'y'$ ja sellaisen apumerkin A', että $S \Rightarrow^*$ u'Ay', $A' \Rightarrow^*$ v'A'x' ja $A' \Rightarrow^*$ w'. Vastaavasti saamme v' = $b^{s'}$, x' = $c^{s'}$ jollekin s', 1$\leq$ s' $\leq$ m. Olkoon i' = $\frac{m!}{s'}$ +1, jolla saamme johdon β = u'$v'^{i'}$ w'$x'^{i'}$ y' = $a^{m+m!}b^{m+m!}c^{m+m!}$. Johdot eivät ole samanlaiset ja tällöin kieli L on luontaisesti moniselitteinen. [Shallit, 2008] □

Seuvaaksi tarkastelemme todistusta. Käyttämällä johtoa $A \Rightarrow^*$ vAx i kertaa voisimme esimerkiksi edetä $S \Rightarrow^*$ uAy $\Rightarrow^*$ uv$A$xy $\Rightarrow^*$ uvv$A$xxy ..., jolloin v ja x esiintyvät mielivaltaisen monta kertaa. Käyttämällä johtoa $A \Rightarrow^*$ vAx i kertaa, siirrymme derivaatiopuussa ylhäältä alaspäin. Mikäli merkittyjä c-merkkejä olisi vähintään δ kappaletta, ja tutkisimme, päteekö s=t, niin merkitsisimme äskeisen sijaan i = $\frac{m!}{t}$ + δ. Mikäli siis $v = a^s \wedge$ x = b^s, m=5 ja i=25, niin saamme c-merkkien arvoksi luonnollisesti Δ!=Δ*(Δ-1)*(Δ-2)*(Δ-3)*(Δ-4) + Δ, missä m=Δ, ja jolloin $a^{m+m!}b^{m+m!}c^{m+m!}$. Mutta tästä ei vielä seuraa, että r=s tai s=t johtavat samaan lopputulokseen. Sanassa z' voisimme c-merkin sijaan merkitä kaikki b-merkit, mutta silloin emme voisi merkitä kaikkia a-merkkejä. Pumppauksen kohteena ovat siis a- ja b-merkit, eikä niiden järjestys voi mennä sekaisin. Ratkaisevaa on a- ja b-merkkien lukumääräinen samuus. Mutta tapauksissa $v = a^rb^r$ ja x = a^rb^r, tai $v = a^r$ ja x = a^rb^r, tai $v = a^rb^r$ ja x = a^r, tai $v = a^sb^t$ ja x = a^sb^t ja missä s $\neq$ t, kielen L mukainen ehto ei toteutuisi. Apumerkit eivät voi olla samat ja siksi merkitsimme niiden olevan A ja A'. Tarvitsimme vähintään kaksi apumerkkiä, joiden derivaatiopolkujen on oltava toisistaan poikkeavat. Kielessä $L = \{a^rb^sc^t | r = s\}$ on esimerkiksi sana $a^{11}b^{11}c^{13}$ ja kielessä $L = \{a^rb^sc^t | s = t\}$ on esimerkiksi sana $a^{17}b^{19}c^{19}$. Osasanat vwx ja $v'w'x'$ eivät sisälly toisiinsa.

Apulause 6.1.2.
Olkoot (N_i, M_i), 1 $\leq$ i $\leq$ r, kokonaislukujoukkojen pareja. (Joukot voivat olla äärellisiä tai äärettömiä). Olkoon $S_i = \{(n,m) | $n $\in N_i$, m $\in M_i\}$ ja olkoon S = $S_1 \cup S_2 \cup$ ... $\cup S_r$. Jos pätee, että kukin kokonaislukujoukkojen pari (n,m) $\in$ S kaikille n ja m, missä n $\neq$ m, niin (n, n) $\in$ S äärettömän monelle arvolle n. [Hopcroft and Ullman, 1979]

Todistus. Sivuutetaan (ks. [Hopcroft and Ullman, 1979]). □

Apulause 6.1.3.
Olkoon G yksiselitteinen kontekstiton kielioppi. Nyt on muodostettavissa sellainen yksiselitteinen kontekstiton kielioppi G', joka tuottaa saman kielen kuin G, että kieliopilla G' ei ole turhia merkkejä tai sääntöjä, ja kaikille apumerkeille A lukuunottamatta mahdollisesti kieliopin G' lähtömerkkiä, on olemassa johto $A \Rightarrow_G{}^* x_1\, Ax_2$, missä $x_1 \wedge x_2 \neq \lambda$. [Hopcroft and Ullman, 1979]

Todistus. Sivuutetaan (ks. [Hopcroft and Ullman, 1979]). □

Esimerkki 6.1.4. Kontekstiton kieli $L = \{a^n b^n c^m d^m | n \wedge m \geq 1\} \cup$
$\{a^n b^m c^m d^n | n \wedge m \geq 1\}$ on luontaisesti moniselitteinen.

Todistus. Oletetaan, että on yksiselitteinen kielioppi G, joka tuottaa kielen L. Voimme muodostaa apulauseen 6.1.3 nojalla yksiselitteisen kieliopin $G = (N,T,P,S)$ tuottamaan kielen L, jossa ei ole tarpeettomia merkkejä, ja kullekin $A \in N$ - $\{S\}$ pätee, että $A \Rightarrow^* x_1 A x_2$ jollekin x_1 ja $x_2 \in T^*$, ja molemmat eivät ole tyhjiä merkkijonoja.

Kieliopilla G on seuraavat ominaisuudet [Hopcroft and Ullman, 1979]:

1) Jos $A \Rightarrow^* x_1 A x_2$, silloin x_1 ja x_2 molemmat muodostuvat ainoastaan yhdentyyppisistä merkeistä (a,b,c tai d), sillä muutoin olisi johto

$$S \Rightarrow^* w_1 A w_3 \Rightarrow^* w_1 x_1 x_1 A x_2 x_2 w_3 \Rightarrow^* w_1 x_1 x_1 w_2 x_2 x_2 w_3,$$

jollekin w_1, w_2 ja w_3. Saatu perusmerkkijono ei kuulu kieleen L .
2) Jos $A \Rightarrow^* x_1 A x_2$, silloin x_1 ja x_2 koostuvat eri merkeistä. Muutoin, apumerkin A sisältävässä johdossa voisimme lisätä kielen L lauseen yhden merkin lukumäärää lisäämättä muiden merkkien lukumäärää, jonka seurauksena lause ei kuuluisi kieleen L.
3) Jos $A \Rightarrow^* x_1 A x_2$, niin $|x_1| = |x_2|$. Muutoin kielen L sanoissa olisi yhtä merkkiä enemmän kuin mitään muuta merkkiä.
4) Jos $A \Rightarrow^* x_1 A x_2$, ja $A \Rightarrow^* x_3 A x_4$, niin x_1 ja x_3 koostuvat samasta merkistä, samoin x_2 ja x_4. Muutoin ominaisuus 1) ei toteutuisi, sillä esimerkiksi x_2 voisi sisältää b-merkkejä ja d-merkkejä.
5) Jos $A \Rightarrow^* x_1 A x_2$, niin joko
a) x_1 koostuu ainoastaan a-merkeistä ja x_2 ainoastaan b- tai d-merkeistä,
b) x_1 koostuu ainoastaan b-merkeistä ja x_2 ainoastaan c-merkeistä, tai
c) x_1 koostuu ainoastaan c-merkeistä ja x_2 ainoastaan d-merkeistä.

Muissa tapauksissa on helppoa johtaa kieleen L kuulumaton merkkijono. Täten muut apumerkit kuin S ovat jaettavissa neljään joukkoon C_{ab}, C_{ad}, C_{bc} ja C_{cd}. Joukko C_{ab} on sellainen apumerkkien $A \in N$ joukko, että $A \Rightarrow^* x_1 \mathrm{A} x_2$, missä $x_1 \in$ a* ja $x_2 \in$ b*; joukot C_{ad}, C_{bc} ja C_{cd} määritellään vastaavasti.
6) Jos johdossa on joukon C_{ab} tai C_{cd} apumerkkejä, niin siinä ei ole joukon C_{ad} ja C_{bc} apumerkkejä tai päinvastoin. Muutoin voisimme lisätä kielen L lauseen kolmen eri merkin lukumäärää lisäämättä neljännen eri merkin lukumäärää. Siinä tapauksessa kielessä L olisi lause, jolle yhden merkin esiintymien lukumäärä on pienempi kuin toisen merkin.

Huomaamme, että jos johto sisältää joukon C_{ab} tai C_{cd} apumerkin, niin täytyy päteä, että tuotettu perusmerkkijono kuuluu kieleen $\{a^n b^n c^m d^m | n \wedge m \geq 1\}$. Oletetaan, että joukon C_{ab} apumerkki A esiintyy sellaisen lauseen johdossa, joka ei kuulu kieleen $\{a^n b^n c^m d^m | n \wedge m \geq 1\}$. Silloin osasanan x täytyy olla muotoa $a^n b^m c^m d^n$, $m \neq n$. Koska A on joukossa C_{ab}, lause $a^{n+p} b^{m+p} c^m d^n$, $m \neq n$, jollekin p > 0, voi tulla tuotetuksi. Sellainen lause ei kuulu kieleen L. Vastaava argumentti pätee joukon C_{cd} apumerkille A. Vastaavan päättelyn tuloksena on, että jos johto sisältää joukon C_{ad} tai C_{bc} apumerkin A, niin tuotetun lauseen täytyy olla kielessä $\{a^n b^m c^m d^n, | n \wedge m \geq 1\}$. Jaamme kieliopin G kahteen osaan:

$G_1 = (\{S\} \cup C_{ab} \cup C_{cd}, T, P_1, S)$ ja

$G_2 = (\{S\} \cup C_{ad} \cup C_{bc}, T, P_2, S)$,

missä P_1 sisältää kaikki ne sääntöjoukon P säännöt, joissa apumerkit ovat joukoista C_{ab} tai C_{cd}, joko oikealla tai vasemmalla, ja P_2 sisältää kaikki sääntöjoukon P säännöt, joissa apumerkki on joukosta C_{ad} tai C_{bc}, joko oikealla tai vasemmalla. Lisäksi P_1 sisältää kaikki sääntöjoukon P muotoa $S \rightarrow a^n b^n c^m d^m$, n ≠ m, ja P_2 sisältää kaikki sääntöjoukon P muotoa $a^n b^m c^m d^n$, n ≠ m, olevat säännöt. Sääntöjoukon P muotoa $S \rightarrow a^n b^n c^n d^n$ olevat säännöt eivät ole sääntöjoukossa P_1 tai sääntöjoukossa P_2.

Koska G tuottaa kielen $\{a^n b^n c^m d^m | n \wedge m \geq 1\} \cup \{a^n b^m c^m d^n | n \wedge m \geq 1\}$, kieliopin G_1 täytyy tuottaa kaikki lauseet kielessä $\{a^n b^n c^m d^m | n \wedge m \geq 1\}$ sekä mahdollisesti joitakin lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$, ja kieliopin G_2 täytyy tuottaa kaikki lauseet kielessä $\{a^n b^m c^m d^n | n \wedge m \geq 1\}$ sekä mahdollisesti joitakin lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$.

Osoitamme seuraavaksi, että tällainen tapaus pätee ainoastaan silloin, jos G_1 ja G_2 molemmat tuottavat kaikki paitsi äärellisen määrän lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$. Täten kaikki paitsi äärellinen määrä lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$ ovat molempien G_1 ja G_2 tuottamia ja siis kieliopin G kahden erillisen vasemman johdon tuottamia. Tuloksena on ristiriita, sillä G ei ole yksiselitteinen.

Todetaksemme, että G_1 ja G_2 molemmat tuottavat kaikki paitsi äärellisen määrän lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$, numeroimme sääntöjoukon P_1 säännöt muotoa $S \rightarrow \alpha$ numeroilla 1, ..., r. Kun 1 ≤ i ≤ r, ja jos $S \rightarrow \alpha$ on i. sääntö, olkoon N_i kaikkien niiden arvojen n joukko, että on olemassa johto $S \Rightarrow_{G1} \alpha \Rightarrow^*_{G1} a^n b^n c^m d^m$

jollekin arvolle m, ja olkoon M_i kaikkien niiden arvojen m joukko, että on olemassa johto $S \Rightarrow_{G1} \alpha \Rightarrow^*_{G1} a^n b^n c^m d^m$ jollekin arvolle n. Vastaava argumentti pätee kieliopille G_2. Huomaamme, että kieliopin G_2 säännöissä ei voi olla samaa oikeaa puolta kahdella tai useammalla apumerkillä. Numeroimme tietyt säännöt ja sääntöparit yksikäsitteiseen järjestykseen. Muotoa $S \rightarrow \alpha_1 B \alpha_2$ olevat säännöt, missä B on joukossa C_{bc}, tulee numeroiduksi, ja jos tämä numero on i, olkoon N_i kaikkien niiden arvojen n joukko, että jollekin arvolle m, on olemassa johto $S \Rightarrow \alpha_1 B \alpha_2 \Rightarrow^* a^n b^m c^m d^n$. Olkoon myös M_i niiden arvojen m joukko, että jollekin arvolle n, on olemassa johto $S \Rightarrow \alpha_1 B \alpha_2 \Rightarrow^* a^n b^m c^m d^n$. Sääntöjen pari $S \rightarrow \alpha$ ja $A \rightarrow \alpha_1 B \alpha_2$ tulee numeroiduksi, jos α sisältää joukon C_{ad} apumerkin, A on joukossa C_{ad} ja B on joukossa C_{bc}. Jos tälle parille osoitetaan numero i, silloin määritämme kokonaislukujoukkojen parin N_i olemaan kaikkien niiden arvojen n joukko, että jollekin arvolle m on olemassa johto $S \Rightarrow \alpha \Rightarrow^* x_1 A x_2 \Rightarrow x_1 \alpha_1 B \alpha_2 x_2 \Rightarrow^* a^n b^m c^m d^n$.

Määritämme myös kokonaislukujoukkojen parin M_i olemaan kaikkien niiden arvojen m joukko, että jollekin arvolle n on olemassa johto $S \Rightarrow \alpha \Rightarrow^* x_1 A x_2 \Rightarrow x_1 \alpha_1 B \alpha_2 x_2 \Rightarrow^* a^n b^m c^m d^n$. Mille tahansa n kokonaislukujoukkojen parissa N_i, ja luvulle m kokonaislukujoukkojen parissa M_i, on olemassa johto $S \Rightarrow^*_{G2} a^n b^m c^m d^n$, ja täten apulauseesta 6.1.4 seuraa, että G_2 tuottaa kaikki paitsi äärellisen määrän lauseita kielessä $\{a^n b^n c^n d^n | n \geq 1\}$. Johtopäätöksenä on, että jollekin n, $a^n b^n c^n d^n \in L(G_1)$ ja $a^n b^n c^n d^n \in L(G_2)$. Tällä lauseella on siis kaksi vasemmanpuoleista johtoa kieliopissa G. [Hopcroft and Ullman, 1979] □

Kielioppi kielelle $L = \{a^n b^n c^m d^m | n \wedge m \geq 1\} \cup \{a^n b^m c^m d^n | n \wedge m \geq 1\}$ on seuraava: $S \rightarrow S_0 | S_1$, $S_1 \rightarrow AB$, $A \rightarrow$ aAb|ab, $B \rightarrow$ cBd|cd, $\rightarrow S_1 \rightarrow$ a$S_1 d | aCd$, $C \rightarrow bCc | bc$.

Ogdenin lemmalla on kaksi tarkoitusta. Sitä voidaan siis käyttää osoittamaan, että jokin kieli L ei ole kontekstiton ja lisäksi lemmalla on todistettavissa joidenkin kontekstittomien kielten luontainen moniselitteisyys. Merkitseminen kannattaa tehdä ainoastaan yhdellä tavalla. Pumpatuilla merkkijonoilla voi olla enemmän kuin k merkittyä merkkiä.

6.2 Ogdenin lemman toteuttavista kielistä

Määritelmä 6.2.1. Kieli L on ogdenmainen, jos on olemassa sellainen kokonaisluku k, että jos missä tahansa kielen L sanassa z on merkitty k tai enemmän merkkejä, niin sanalla z on osamerkkijonoihin jako $z = uvwxy$, joka toteuttaa ehdot

1) $uv^i wx^i y \in$ L, kun $i \geq 0$.
2) joko kukin u, v ja w, tai w, x ja y sisältää merkityn merkin.
3) vwx sisältää korkeintaan k merkittyä merkkiä.

Boasson ja Horvath [1978] näyttävät, että erilaisia ei-kontekstittomien kielten tyyppisiä Ogdenin lemman toteuttavia kieliä on helppo muodostaa. Boasson ja Horvath

vastaavat Horvathin [1978] aiemmin esittämään kysymykseen seuraavalla väitteellä:

Väite 6.2.2. On olemassa kontekstisia, rekursiivisia, rekursiivisesti numeroituvia ja ei-rekursiivisesti numeroituvia kieliä, jotka ovat ogdenmaisia.

Todistus. Tarkastellaan joukon $\mathbb{N}$ alijoukkoa F ja määritellään $H_p = \{(ab)^n | n \in F\}$ ja $I_p = H_p \cup$ X* $\{aa, bb\}$ X*, missä molemmat ovat kieliä X = $\{a, b\}$.
Joukot ovat erillisiä. Kieli I_p on kontekstiton, mikäli H_p on. I_p on kontekstinen, jos H_p on ja I_p ei ole rekursiivisesti numeroituva, jos H_p ei ole.

Täten H_p määrittää kielen I_p, mutta I_p ei määritä kieltä H_p. Kieli I_p on ogdenmainen arvolla k = 4. Täten väite on todistettu valitsemalla H_p kontekstiseksi tai ei-rekursiivisesti numeroituvaksi kieleksi. [Boasson and Horvath, 1978] □

6.3 Ogdenin lemman yleistäminen

Tässä kohdassa ja luvussa 9 poissuljettujen merkkien lukumäärää sanassa z merkitään $\varrho(z)$.

Lause 6.3.1. (Baderin ja Mouran ehto) Olkoon L kontekstiton kieli. On olemassa sellainen kokonaisluku k, että kun d merkkiä on merkitty ja e merkkiä on poissuljettu, kun $d > k^{(e+1)}$, niin sana z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että

1. vx sisältää vähintään yhden merkityn merkin, eikä poissuljettuja merkkejä.
2. Jos d on merkittyjen merkkien lukumäärä ja ϱ on poissuljettujen merkkien lukumäärä osasanassa vwx, niin d $\leq k^{(\varrho+1)}$.
3. $uv^iwx^iy \in L$, kun $i \geq 0$.

Todistus. Oletetaan, että z $\neq \lambda$. Kielellä L - $\{\lambda\}$ on Chomskyn normaalimuodossa oleva kielioppi G. Olkoon kieliopilla G n apumerkkiä, ja olkoon $k = 2^{n+1}$. Tarkastelkaamme yhtä derivaatiopuuta sanalle z kieliopissa G. Jos puun solmun molemmilla jälkeläisillä on merkittyjä jälkeläisiä, kutsutaan tuota solmua Θ-solmuksi. Olkoon H polku, jolla on suurin lukumäärä Θ-solmuja. Koska sanalla z on vähintään $2^{(n+1)(e+1)}$ merkittyä merkkiä, polulla H on vähintään (n + 1)(e + 1) Θ-solmua. Jaamme polun H alimmaisen osan e + 1 osapolkuihin, joista kukin sisältää n + 1 Θ-solmua. Jokaisessa osapolussa täytyy olla kaksi Θ-solmua, joissa on sama tunniste, esimerkiksi A. Tällöin on olemassa kaksi perusmerkkien merkkijonoa v' ja x' ja kaksi sellaista apumerkkiä, B ja C, että $A \Rightarrow BC \Rightarrow^*$ v'Ax'. Koska ylempi A on Θ-solmu, apumerkeillä B ja C on merkittyjä jälkeläisiä. Apumerkit B ja C eivät voi molemmat hallita alempaa Θ-solmua A, joten perusmerkkien $v'x$' täytyy sisältää vähintään yksi merkitty merkki.

Alkaen lehdestä, jossa H päättyy, jatketaan osapolkujen kautta kunnes löydetään yk-

si osapolku, jonka pumpattujen osamerkkijonojen pari, eli v ja x, ei sisällä poissuljettuja merkkejä. Sellaisen parin tiedetään olevan olemassa, koska on e + 1 erillistä paria, mutta ainoastaan e poissuljettua merkkiä. Olkoon tämä pari v, x. Täten olemme todistaneet lauseen 6.3.1. kohdat 1. ja 3. Oletetaan, että meidän täytyisi kiivetä osapolulle, jonka pituus on (g + 1) löytääksemme parin v, x. Tämän osapolun alla jokaisessa osapolussa pumpattujen osamerkkijonojen pari sisältää vähintään yhden poissuljetun merkin. Täten jos poissuljettujen merkkien lukumäärä osamerkkijonossa vwx on ϱ, niin $\varrho \geq$ g. Osamerkkijonoa vwx hallitsee yksittäinen solmu osapolulla, jonka pituus on (g + 1). Polun H määrityksen nojalla mikään tämän solmun alapuolella oleva polku ei sisällä enempää kuin (n + 1)(g + 1) Θ-solmua. Näin ollen, jos merkittyjen merkkien lukumäärä osamerkkijonossa vwx on d, niin d $\leq 2^{(n+1)(g+1)} = k^{(g+1)} \leq k^{(\varrho+1)}$, mikä todistaa ehdon 2. [Bader and Moura, 1982] □

Seuraavaksi esitämme huomioita todistuksesta. Sanalla z on vähintään $k^{(e+1)}$ merkittyä merkkiä. Apumerkit B ja C sijaitsevat ylemmän Θ-solmun alapuolella. Merkityt merkit sijaitsevat kaikkien Θ-solmujen alapuolella. Jos polulla H on vähintään 2(n + 1)(e + 1) Θ-solmua, niin se koostuu vähintään (n + 1)(e + 1) vasemmasta Θ-solmusta ja vähintään (n + 1)(e + 1) oikeasta Θ-solmusta. Erillisten parien kohdalla e + 1 tarkoittaa, että on yksi kappale pareja v, x, joka ei siis sisällä poissuljettuja merkkejä. Ja koska pariin v, x liittyy apumerkit B ja C, niin jokaisen parin solmut on nimetty samalla apumerkillä ja täten parissa v, x on oltava ainakin yksi merkitty merkki. Pari v, x kuuluu joukkoon T^*. Pumpattujen osamerkkijonojen pari v, x voi siis sijaita polulla, jonka pituus on (g + 1) ja tällä polulla on solmu, joka hallitsee osamerkkijonoa vwx. Pumppaus kasvattaa merkittyjen merkkien lukumäärää. Pituus g ei voi ylittää osamerkkijonon vwx sisältämien poissuljettujen merkkien lukumäärää.

Esimerkki 6.3.2. Olkoon kieli $L = \{z \in \{a,b\}^* \mid$ jos $z = ab^q$, niin q on alkuluku$\}$. Kieli toteuttaa Ogdenin lemman, mutta ei Baderin ja Mouran lemmaa. Mikä tahansa kielen $\{a,b\}^*$ merkkijono kuuluu kieleen L, jos se ei ole muotoa ab^q. Riippumatta merkittyjen merkkien sijainneista mikä tahansa merkkijono, joka ei ole muotoa ab^q, voidaan pumpata siten, että se ei edelleenkään ole muotoa ab^q. On huomioitavaa, että ei ole mahdollista rajata a-merkki pois pumpattavista merkkijonoista. Mikäli a-merkin sisältävä osamerkkijono pumpataan, on tuloksena merkkijono, joka ei ole muotoa ab^q, joka siten kuuluu kieleen L. Täten kieli L toteuttaa Ogdenin lemman. Olkoon k yleistetyn lemman vakio, ja q pienin sellainen alkuluku, että $q > k^2$.

Havaitsemme, että ab^q kuuluu kieleen L. Olkoon a poissuljettu ja olkoot kaikki b-merkit merkittyjä. On siis d=q ja e=1. Koska d = $q > k^2 = n^{(e+1)}$, lemman mukaan ab^q sisältää parin merkittyjä osamerkkijonoja, jotka sisältävät ainoastaan b-merkkejä. Koska vx sisältää ainoastaan b-merkkejä, saadaan pumppaamalla lauseita, joissa b-merkkien lukumäärä ei ole alkuluku. Oletetaan, että b-merkin lukumäärä kyseisessä parissa on p. Pumpattaessa ab^q q valinnalla q=11 , saamme $ab^{q(p+1)}$, joka ei kuulu kieleen L. Täten L ei toteuta yleistetyn lemman ehtoja, joten kieli L ei ole kontekstiton. [Bader and Moura, 1982]

Havaitsemme esimerkissä yhteneväisyyksiä esimerkkiin 4.1.1, jossa kieleen kuuluu ainoastaan merkkijonoja, joiden pituudet ovat alkulukuja. Yhtenä eroavaisuutena huomaamme, että tässä esimerkissä q on alkuluku, kun taas esimerkissä 4.1.1 oli i > 0, mikä luonnollisesti sisältää merkkijonoja, joiden pituudet ovat alkulukuja, mutta ei ainoastaan niitä.

Yleistetyn lemman toteuttaminen ei ole riittävää kielen kontekstittomaksi toteamiselle. Tämän osoittamiseen tarvitsemme kaksi apulausetta. [Bader and Moura, 1982]

Apulause 6.3.3 Olkoon kieli $L_{11} = \{z \in \{a, b\}^* \mid \text{jos } z = (ab^q), \text{ niin q on alkuluku}\}$. Olkoon g mikä tahansa sellainen funktio $\mathbb{N} \to \mathbb{N}$, että kaikille x, g(x) > x ja g(x) > 1. Kaikki sanat z kuuluvat kieleen L_{11}. Jos d merkkiä sanassa z ovat merkittyjä ja e merkkiä ovat poissuljettuja, kun d > g(e), niin yleistetyn lemman tulokset seuraavat valitsemalla d ≤ g(ϱ) yleistetyn lemman ehdossa 2. [Bader and Moura, 1982]

Todistus. Olkoon $z \in L_{11}$. Tarkastellaan kahta tapausta. Ensiksi oletetaan, että jos mikä tahansa yksi merkki poistetaan sanasta z, niin tuotettu merkkijono edelleen kuuluu kieleen L_{11}. Koska d > g(e) > e, on vähintään kaksi merkkiä sanassa z, jotka ovat merkittyjä, mutta eivät poissuljettuja. Olkoon näistä yksi a ja olkoon a = v ja olkoon w = x = λ. Näin ollen ehto 1 toteutuu. Ehto 2 toteutuu, koska g(0) ≥ 1. Kun kaikille i ≥ 1, uv^iwx^iy sisältää osamerkkijonon muotoa aa. Täten ehto 3 toteutuu, kun i ≥ 1. Kun i=0, alussa tehdyn oletuksen nojalla $uv^iwx^iy \in L_{11}$. Täten ehto 3 toteutuu. Muutoin, oletetaan, että sanassa z on sellainen merkki, että jos se poistetaan, niin saadaan merkkijono, joka ei ole kielessä L_{11}. Olkoon tämä merkki a. Kuten kielen L_{11} määrityksestä käy ilmi, ainoat kieleen L_{11} kuulumattomat merkkijonot ovat muotoa ab^q, missä q ei ole alkuluku. Täten sanan z täytyy olla muotoa $(ab)^p$ a$(ab)^d$, missä p ≥ 0, d ≥ 0, ja p + d = q. Edelleen d > g(e) > e, joten saadaan vähintään kaksi merkkiä sanassa z, jotka ovat merkittyjä, mutta eivät poissuljettuja. Oletetaan, että d = 0. Silloin sanassa z on viimeisenä merkkinä a. Koska z sisältää vähintään kaksi merkkiä jotka ovat merkittyjä, mutta eivät poissuljettuja, sanassa z on jokin muu merkki kuin a, joka on merkitty, mutta ei poissuljettu. Olkoon tämä merkki v, ja ja olkoon w = x = λ. Näin ollen ehdot 1 ja 2 toteutuvat. Kun kaikille i ≥ 0, merkkijonossa uv^iwx^iy on viimeisenä merkkinä a ja täten merkkijono kuuluu kieleen L_{11}. Täten ehto 3 toteutuu.

Muutoin, d > 0. Silloin sanassa z on muotoa aa oleva osamerkkijono. Jos tämän osamerkkijonon ulkopuolella on jokin merkki, joka on merkitty, mutta ei poissuljettu, niin tilanne on vastaava kuin tapauksessa, jossa d = 0. Muutoin, molempien a-merkkien täytyy olla merkittyjä osamerkkijonossa, mutta ei poissuljettuja. Olkoon v = aa ja olkoon $w = x = \lambda$. Selvästi ehto 1 toteutuu. Ehto 2 toteutuu, koska g(0) ≥ 2. Kun i=0, merkkijonolla uv^iwx^iy on muotoa bb oleva osamerkkijono tai se alkaa merkillä b. Joka tapauksessa $uv^iwx^iy \in L_{11}$, joka toteuttaa ehdon 3. [Bader and Moura, 1982] □

Todistuksessa siis käsitellään tapaukset, joista ensimmäisessä sanasta z poistetaan mikä tahansa merkki, jonka jälkeen kieli edelleen kuuluu kieleen L_{11} ja toisessa oletetaan, että sanassa z on jokin merkki niin, että jos se poistetaan, niin saamme merkkijonon, joka ei kuulu kieleen L_{11}. Todistuksen aluksi ehto 1 täyttyy, koska v sisältää merkityn merkin, mutta ei poissuljettuja. Todistuksen lopuksi ehto 2 täyttyy, koska v=aa, jolloin g(0) ≥ 2.

Apulause 6.3.4 Kieli L_{11} ei ole kontekstiton. [Bader and Moura, 1982]

Todistus. Tehdään oletus, että kieli on kontekstiton. Kontekstittomien kielten joukko on suljettu gsm-kuvausten ja peilikuvan suhteen [Salomaa, 1973]. Olkoon M kielen L_{11} kuva sellaisen gsm-kuvauksen suhteen, joka liittää muotoa ab^q olevien lauseiden perään merkkijonon a#at. Olkoon Q kielen M peilikuva ja olkoon I kielen Q kuva sellaisen gsm-kuvauksen suhteen, joka poistaa kaikki merkit, kunnes se löytää merkkijonon a# ja siitä eteenpäin poistaa merkit # ja b. L_{11} on kontekstiton, joten myös I on. Mutta $I = \{a^n \mid n$ on alkuluku tai $0\}$, joka ei ole kontekstiton (ks. esimerkki 4.1.1). Joten myöskään L_{11} ei ole kontekstiton.

Täten L_{11} toteuttaa lemman ehdot kaikilla mahdollisilla funktioilla g, huomaamiemme rajoitusten mukaisesti. Mutta L_{11} ei ole kontekstiton. Tällöin funktio, joka mahdollistaisi lemmamme luonnehtia kontekstittomia kieliä, ei voi olla olemassa. [Bader and Moura, 1982] □

Lemmassa pumpataan ainoastaan merkittyjä merkkijonoja, mutta ei yhtäkään poissuljettua. Eroavaisuutena Ogdenin lemmaan mukaan otetaan poissuljettavuus, joka on Ogdenin lemmaa vahvistava tekijä. Poissuljettavuudella pystytään hallitsemaan kieltä, sillä siten voidaan merkitä esimerkiksi jokin merkki poissuljetuksi, joka voi olla c, jolloin se ei tule pumpatuksi. Muutoin tuloksena voi olla merkkijonoja, jotka eivät ole toivottuja. Totesimme esimerkissä 6.3.2, että Ogdenin lemma on riittämätön. Mutta Baderin ja Mouran lemmalla saimme todistettua, että tarkasteltava kieli ei ole kontekstiton, koska merkkien poissulkemisen ansiosta pumpattu merkkijono ei enää kuulu kieleen. Täten Baderin ja Mouran lemma on vahvempi kuin Ogdenin lemma.

7 Vaihtolemma

Tässä luvussa esitellään vaihtolemma (interchange lemma), johon liittyy täydellinen sekoitus.

Tässä luvussa käytetään merkintää ⧢, joka tarkoittaa täydellistä sekoitusta. Jos w = $\delta_1\delta_2\ldots\delta_n$ ja x = $\varpi_1\varpi_2\ldots\varpi_n$ ovat saman pituisia äärellisiä sanoja, niin δ⧢ϖ tarkoittaa sanaa $\delta_1\varpi_1\delta_2\varpi_2\ldots\ \delta_n\varpi_n$, sanojen w ja x täydellistä sekoitusta. Esimerkiksi ybwp ⧢ kgmf = ykbgwmpf.

Oletamme tunnetuksi käänteisen morfismin h^{-1} (ks. [Salomaa, 1973]). Oletamme tunnetuksi myös, että kontekstittomien kielten luokka on suljettu käänteisen morfismin suhteen ja säännöllisten kielten kanssa tehtävän leikkauksen suhteen. [Salomaa, 1973].

Neliö tarkoittaa muotoa xx olevaa merkkijonoa, esimerkiksi sana korolkorol on neliö. Jos w on äärellinen tai ääretön merkkijono, jossa ei ole ei-tyhjää osasanaa, joka on edellä kuvattua muotoa xx, niin sana on neliötön. Esimerkiksi merkkijono square on neliötön, mutta merkkijono squarefree ei ole, sillä osasana ee on neliö.

Määritelmä 7.1 Olkoon t = $t_0t_1t_2\ldots$ = 01101001 ... . Ääretön merkkijono t on Thuen-Morsen binäärisekvenssi.

Lause 7.2 Tarkoittakoon c_n, n ≥ 1, ykkösmerkkien lukumäärää n:nnen ja (n+1):nnen nollan välissä sanassa t. Asetetaan c = $c_1c_2c_3\ldots$. Tällöin c = 210201 on ääretön neliötön sana aakkostossa Σ_3.

Todistus. Sivuutetaan (ks. [Shallit, 2008]). □

Lause 7.3 (Vaihtolemma)
Olkoon kieli L kontekstiton. Silloin on sellainen kielestä riippuva kokonaisluku $k >$ 0, että kaikille luvuille n ≥ 2, kaikille alijoukoille R ⊆ $L \cap \Sigma^n$, ja kaikille luvuille n, 2 ≤ m ≤ n, on olemassa sellainen alijoukko Z ⊆ R, Z = $\{z_1, z_2, \ldots, z_p\}$, että p ≥ $\frac{|R|}{k(n+1)^2}$ ja on sellainen osamerkkijonoihin jako $z_i = w_ix_iy_i$, 1 ≤ i ≤ p, että

i) $|w_1| = |w_2| = \ldots = |w_p|$,
ii) $|y_1| = |y_2| = \ldots = |y_p|$,
iii) m/2 < $|x_1| = |x_2| = \ldots = |x_p| \leq$ m,
iv) $w_ix_jy_i \in L$, ∀ i, j, 1 ≤ i, j ≤ p.

Todistamme ensin kaksi apulausetta.

Apulause 7.4 Olkoon $G = (N,T,P,S)$ kontekstiton kielioppi Chomskyn normaalimuodossa ja olkoon $L(G) = L$. Olkoon m ≥ 2. Silloin kaikille merkkijonoille $z \in$ L, missä $|z| \geq$ m, on olemassa apumerkki $A \in N$ ja muotoa $S \Rightarrow^*$ wAy $\Rightarrow^*$ wxy = z oleva johto, missä $\frac{m}{2} <$ |x| ≤ m.

Todistus. Olkoon q sanan z derivaatiopuun T juuri. Juurella q on ≥ m jälkeläistä, jotka ovat perusmerkkejä. Jos juurella q on täsmälleen m jälkeläistä, jotka ovat perusmerkkejä, valitaan w = y = λ, $A = S$, ja x = z. Oletetaan, että juurella q on > m jälkeläistä. Nyt toistuvasti korvataan q jälkeläisellä, jolla on suurin määrä perusmerkkijälkeläisiä, kunnes juurella q on ≤ m jälkeläistä. Koska juuren q vanhemmalla on > m jälkeläistä, juurella q täytyy olla > m/2 jälkeläistä. Olkoon A juuren q tunniste. Silloin on olemassa sellaiset merkkijonot w ja y, että $S \Rightarrow^*$ wAy ja $A \Rightarrow^*$ x, missä $\frac{m}{2} <$ |x| ≤ m. [Shallit, 2008] □

Nyt määritetään tiettyjä johtoihin liittyviä joukkoja. Valitaan alijoukko R ⊆ L ∩ Σ^n. Luvuille n_1 ja n_2, missä $0 \leq n_1, n_2 \leq$ n, määritetään $Q_{n,R}$ $(n_1, A, n_2) = \{z \in$ R | on olemassa johto $S \Rightarrow^*$ wAy $\Rightarrow^*$ wxy = z, |w| = n_1, |y| = $n_2\}$.

Apulause 7.5 Olkoon $G = (N,T,P,S)$ kontekstiton kielioppi Chomskyn normaalimuodossa ja olkoon $L(G) = L$. Olkoon 2 ≤ m ≤ n. Silloin kaikille alijoukoille R ⊆ $L \cap \Sigma^n$, on olemassa sellaiset luvut $0 \leq n_1, n_2 \leq$ n, että $\frac{m}{2} <$ n - n_1 - $n_2 \leq$ m ja sellainen apumerkki $A \in N$, että $|Q_{n,R}\ (n_1, A, n_2)| \geq \frac{|R|}{|N|(n+1)^2}$.

Todistus. Saadaan

$$R = \bigcup_{\substack{A \in N \\ 0 \leq n_1, n_2 \leq n}} Q_{n,R}(n_1, A, n_2) = \bigcup_{\substack{A \in N \\ 0 \leq n_1, n_2 \leq n \\ \frac{m}{2} < n - n_1 - n_2 \leq m}} Q_{n,R}(n_1, A, n_2),$$

missä viimeistä yhdistettä käytettiin apulauseessa 7.4. Täten olemme kirjoittaneet alijoukon R yhdisteeksi, jossa on korkeintaan (n + 1)^2|N| joukkoa, joten näistä joukoista ainakin yhdellä on vähintään $\frac{|R|}{|N|(n+1)^2}$ alkiota. [Shallit, 2008] □

Nyt voimme todistaa vaihtolemman.

Todistus. Olkoon $G = (N,T,P,S)$ Chomskyn normaalimuodossa oleva kontekstiton kielioppi, joka tuottaa kielen L, johon ei kuulu tyhjää merkkijonoa λ. Olkoon kieliopin G apumerkkien lukumäärä k, ja valitaan alijoukko R ⊆ $L \cap \Sigma^n$. Silloin apulauseen 7.5 nojalla on olemassa sellaiset luvut $0 \leq n_1, n_2 \leq$ n, että $\frac{m}{2} <$ n - n_1

- $n_2 \leq$ m, ja sellainen apumerkki A, että $|Q_{n,R}(n_1, A, n_2)|, \geq \frac{|R|}{|N|(n+1)^2}$. Valitaan Z = $Q_{n,R}(n_1, A, n_2)$. Jokaisella merkkijonolla z_i, joka on joukossa $Q_{n,R}(n_1, A, n_2)$, on muotoa $S \Rightarrow^* w_i A y_i \Rightarrow^* w_i x_i y_i = z_i$ oleva johto, missä $|w_i| = n_1$ ja $|y_i| = n_2$. Täten $S \Rightarrow^* w_i A y_i \Rightarrow^* w_i x_j y_i \in L$. Näin vaihtolemma on todistettu. [Shallit, 2008] □

Seuraavaksi esitämme esimerkin lemman soveltamisesta.

Esimerkki 7.6 Olkoon kieli $L_i = \{xyyz | x, z \in \Sigma^*, y \in \Sigma^+\}$, missä $\Sigma = \{0, 1, \ldots, i - 1\}$. Näin ollen L_i on kaikkien sanojen kieli, jotka sisältävät neliöitä i-kirjaimisessa aakkostossa. [Shallit, 2008]

Käytämme vaihtolemmaa seuraavan lauseen todistamisessa.

Lause 7.7 Kieli L_i ei ole kontekstiton, kun i ≥ 3. [Shallit, 2008]

Todistus. Todistamme aluksi tuloksen arvolla i=6. Lopuksi selitämme, miten saamme tuloksen kaikille i ≥ 3. Oletetaan, että L_6 on kontekstiton. Olkoon k vaihtolemmassa annettu vakio ja valitaan n niin suureksi, että se on jaettavissa luvulla 8 ja $\frac{2^{n/4}}{k(n+1)^2} > 2^{n/8}$.

Lauseen 7.2 nojalla on olemassa neliötön mielivaltaisen pitkä merkkijono kolmen kirjaimen aakkostossa. Valitaan sellainen merkkijono r', jonka pituus on n / 4 - 1 aakkostossa {0,1,2} ja asetetaan r = 3r'. Olkoon $A_n = \{rr \sqcup\!\sqcup s | s \in \{4, 5\}^{n/2}\}$, missä $\sqcup\!\sqcup$ on luvun alussa esitelty täydellinen sekoitus. Täten jokainen merkkijono kielessä A_n on pituudeltaan n ja niillä on seuraavat ominaisuudet:

1. Jos $z_1 = w_1 x_1 y_1$ ja $z_2 = w_2 x_2 y_2$ ovat merkkijonoja kielessä A_n, missä $|w_1| = |w_2|$, $|x_1| = |x_2|$, $|y_1| = |y_2|$, niin $w_1 x_2 y_1$ ja $w_2 x_1 y_2$ ovat molemmat kielessä A_n. Kun z_1 = rr $\sqcup\!\sqcup$ s, z_2 = rr $\sqcup\!\sqcup$ s', osamerkkijonon x_1 korvaaminen osamerkkijonolla x_2 jättää merkkejä rr vastaavat merkit samaksi, kun taas merkkiä s vastaavat merkit voivat vaihtua. Mutta koska mikä tahansa s on sallittu, tämä muutos ei vaikuta kieleen A_n kuulumiseen.

2. Jos $z \in A_n$, niin sana z sisältää neliön, jos ja vain jos z on neliö. Jos z sisältäisi neliön, niin tarkasteltaisiin ainoastaan sanan z merkkejä aakkostossa {0,1,2,3}, ja meillä edelleen olisi neliö. Mutta tämä on mahdotonta, koska r on neliötön. Nyt määritetään kielen A_n alijoukko $B_n = L_6 \cap A_n = \{rr \sqcup\!\sqcup ss | s \in \{4, 5\}^{n/4}\}$.

Selvästi $|B_n| = 2^{n/4}$. Koska $B_n \subseteq L_6$, vaihtolemma soveltuu, kun m = n/2 ja R = B_n. Silloin on alijoukko Z ⊆ B_n, Z = $\{z_1, z_2, \ldots, z_k\}$, missä $z_i = w_i x_i y_i$, toteuttaa lemman johtopäätökset. Erityisesti, k = |Z| ≥ $\frac{2^{n/4}}{k(n+1)^2} > 2^{n/8}$.

Nyt on kaksi tapausta tarkasteltavana, ja kumpikin johtaa ristiriitaan.

Tapaus 1: On olemassa sellaiset indeksit g ja h, että $x_g \neq x_h$.
Tapaus 2: Ei ole olemassa sellaisia indeksejä.
Tapauksessa 1 tiedämme vaihtolemman mukaan, että $w_g x_h y_g \in L_6$. Koska $x_g \neq x_h$, sanan z_g puolikkaassa täytyy olla 4 tai 5, joka on muuttunut. Mutta koska $|x_h| \leq$ n/2, vastaava merkki ei ole muuttunut merkkijonon toisessa puolikkaassa. Näin ollen $w_g x_h y_g$ ei voi olla neliö. Koska $w_g x_h y_g \in A_n$, niin aiemmin havaitun perusteella $w_g x_h y_g$ ei voi sisältää neliötä. Täten $w_g x_h y_g \notin L_6$, joka on ristiriita.

Tapauksessa 2 kaikkien osamerkkijonojen x_i täytyy olla samoja. Joten on vähintään n/4 esiintymää, joihin kaikkiin z_i yhtyy. Tämä n/4 esiintymän joukko sisältää vähintään n/8 esiintymää merkkejä 4 ja 5. On ainoastaan korkeintaan n/8 esiintymää, joissa voimme vapaasti valita merkkien 4 ja 5 välillä. Täten, $|Z| \leq 2^{n/8}$, joka on ristiriita. Seurauksena on, että L_6 ei ole kontekstiton kieli.

Nyt osoitamme, miten saamme tuloksen, kun L_i, i $\geq$ 3. Osoitetaan, että seuraava morfismi h on neliöttömyyden säilyttävä, eli x on neliötön jos ja vain jos h(x) on neliötön:

h(0) $\rightarrow$ 0102012022012102010212
h(1) $\rightarrow$ 0102012022201210120212
h(2) $\rightarrow$ 0102012101202101210212
h(3) $\rightarrow$ 0102012101202120121012
h(4) $\rightarrow$ 0102012102010210120212
h(5) $\rightarrow$ 0102012102120210120212.

Oletetaan, että L_{13} on kontekstiton. Koska kontekstittomien kielten luokka on suljettu käänteisen morfismin suhteen, kieli $h^{-1}(L_{13})$ on kontekstiton. Mutta koska h on neliöttömyyden säilyttävä, $h^{-1}(L_{13})$ on kaikkien neliönsisältävien sanojen kieli kuuden kirjaimen aakkostossa, siis L_6, jonka juuri todistimme olevan ei-kontekstiton, mikä on ristiriita. Lopuksi oletetaan, että L_i on kontekstiton jollekin i $\geq$ 3. Silloin $L_i \cap \{0,1,2\}^*$ on kontekstiton, sillä kontekstittomien kielten luokka on suljettu säännöllisten kielten kanssa tehtävän leikkauksen suhteen. Mutta tämä on L_{13}, mikä on ristiriita. □

Lauseen 7.7 todistuksessa $2^{n/4}$ vastaa alijoukkoa R lauseessa 7.3. Meidän on siis valittava riittävän suuri n ja mikäli c=3, niin esimerkiksi n=132 toteuttaa yhtälön. Havaitsemme lauseen 7.7 todistuksen lopussa olevista riveistä 0-5, jotka siis edustavat kolmekirjaimista aakkostoa, että niissä ei ole neliöitä. Ainoastaan kyseisten rivien avulla ei vielä siis todistettu, että L_{13} ei ole kontekstiton, vaan se edellytti luvun alussa kerrottua tunnettua tulosta, jonka mukaan konteksittomien kielten luokka on suljettu käänteisen morfismin suhteen ja lisäksi kielen, jossa on kuusikirjaiminen aakkosto, osoittamista kontekstittomaksi.

Vaihtolemmassa on ideana, että mikäli kieli ei ole kontekstiton, niin voidaan korvata

kieleen kuuluvan sanan osasana jollakin toisella merkkijonolla ja korvattu sana edelleen kuuluu kieleen. Tämä edellyttää sanojen osajoukon riittävää suuruutta. Voidaan siis löytää vähintään kaksi pituudeltaan n olevaa merkkijonoa, jotka voidaan vaihtaa toisillaan kieleen edelleen kuuluvien uusien sanojen tuottamiseksi. Merkittävää vaihtolemmassa on, että siinä ei ole pumppaamista, kuten muissa tarkastelemissamme lemmoissa, vaan tarkasteltavana on täsmälleen samanpituisia merkkijonoja, eikä merkkijonojen pituuksia siis erikseen kasvateta pumppaamalla.

Käytettäessä lemmaa on tarkoituksena löytää sopivat R, n ja m. Shallit [2008] osoittaa, että toistuvat merkkijonot eivät ole kontekstittomia. Vaihtolemman avulla voidaan siis todistaa, että kieli ei ole kontekstiton, edellyttäen aakkoston olevan vähintään kolmikirjaiminen.

8 Parikhin lause

Tässä luvussa esittelemme Parikhin lauseen ja tarkastelemme siihen liittyviä esimerkkejä.

Määritelmä 8.1 Joukon $\mathbb{N}^q$ alijoukko A on lineaarinen, jos on olemassa sellaiset $u_0, u_1, \ldots, u_r \in \mathbb{N}^q$, että A = $\{u_0 + a_1u_1 + \ldots + a_ru_r \mid a_1, a_2, \ldots, a_r \in \mathbb{N}\}$.

Yllä siis u-merkit edustavat vektoreita, jotka voidaan merkitä myös niin, että kunkin u-merkin päällä on yläviiva.

Määritelmä 8.2 Joukon $\mathbb{N}^q$ alijoukko A on semilineaarinen, jos se on äärellisen monen lineaarisen joukon yhdiste.

Lause 8.3 Kun q $\geq$ 1, joukon $\mathbb{N}^q$ semilineaaristen joukkojen luokka on suljettu yhdisteen, leikkauksen ja komplementin suhteen.

Todistus. Sivuutetaan (ks. [Ginsburg and Spanier, 1966]). □

Seuraavaksi määrittelemme Parikhin kuvauksen, jota merkitään Ψ.

Määritelmä 8.4 Olkoon aakkosto $\Sigma = \{a_1, a_2, \ldots, a_q\}$. Silloin $\Psi : \Sigma^* \rightarrow^* \mathbb{N}^q$ kuvaa sanan $z \in \Sigma^*$ vektoriin $(|z|_{a_1}, |z|_{a_2}, \ldots, |z|_{a_q})$, jonka pituus on q.

Kyseessä on siis q-kirjaiminen aakkosto. Kuvaus Ψ voidaan laajentaa myös kieliin L, jolloin $\Psi(L)$ tarkoittaa kielen L Parikhin kuvaa, $\Psi(L) = \bigcup_{z \in L} \{\Psi(z)\}$.

Kaksi sanaa w ja v aakkostossa Σ ovat kirjainvastaavia, jos Ψ(w) = Ψ(v). Tällöin esimerkiksi qqroopo muodostaa kirjainvastaavuuden sanan oqporoq kanssa. Parikhin kuvaus on olennaisesti sanan kommutatiivinen kuva ja Ψ(xy) = Ψ(x) + Ψ(y) kaikille merkkijonoille x,y $\in \Sigma^*$. Jokainen kontekstiton kieli on kirjainvastaava säännölliseen kieleen. Esimerkiksi kontekstiton kieli $\{(a^nb^n | n \geq 1)\}$ ja säännöllinen kieli $\{(ab)^n | n \geq 1\}$ ovat kirjainvastaavia, sillä molemmilla kielillä on kommutatiivinen kuva, joka on $\{(n,n) | n \geq 1\}$. [Salomaa, 1973].

Esimerkki 8.5 Olkoon aakkosto $\Sigma = \{e_1, e_2, e_3, e_4\}$. Sanalle $z = e_4e_4e_1e_3e_2e_4e_1e_2e_1$ pätee $\Psi(z)$ = (3,2,1,3).

Esimerkki 8.6 Olkoon aakkosto $\Sigma = \{a, b, c, d, e, f, g, h, i\}$. Valitkaamme sanat $z1$ = giibcbcebgii ja $z2$ = iiiicbgecgbb. Toteamme, että Parikhin kuva on sanoissa sa-

ma, sillä Ψ(z1) = Ψ(z2) = (0,3,2,0,1,0,2,0,4). Valitkaamme vielä kolmanneksi sanaksi $z3$ = igbiebigbcie. Tällöin Parikhin kuva ei ole kaikissa sanoissa sama.

Esimerkki 8.7 Olkoon aakkosto $\Sigma = \{a,b,c,d,e,f,g,h,i\}$. Tällöin Ψ(giibcbcebgii) = (0,3,2,0,1,0,2,0,4). Kuvan alkioiden summa 3+2+1+2+4 on luonnollisesti sama kuin argumenttina olleen sanan pituus.

Seuraavassa esimerkissä merkinnöillä ⟨ ja ⟩ merkitään joukon virittämiä vektoreita, semilineaarisia joukkoja.

Esimerkki 8.8

a) $L_{10} = \{0,01\}^*$. Silloin $\Psi(L_{10}) = \langle(1,0), (1,1)\rangle$.
b) $L_{11} = \{0,1\}^2$. Silloin $\Psi(L_{11}) = \langle(0,2), (1,1), (2,0)\rangle$.
c) $L_{12} = \{0^n1^n | n \geq 1\}$. Silloin $\Psi(L_{12}) = (1,1) + \langle(1,1)\rangle$.
d) $L_{13} = \{0^n1^n2^n | n \geq 1\}$. Silloin $\Psi(L_{13}) = (1,1,1) + \langle(1,1,1)\rangle$.
e) $L_{14} = \{0^m1^m2^n3^n | m,n \geq 1\}$. Silloin $\Psi(L_{14}) = (1, 1, 1, 1) + \langle(1, 1, 0, 0), (0, 0, 1, 1)\rangle$. [Shallit, 2008].

Esimerkiksi tapauksen c) kohdalla pätee, että $\Psi(L_{12}) = \{(1,1),(2,2),(3,3),(4,4),\dots\}$, mikäli $0 \notin \mathbb{N}$. Havaitsemme, että $\Psi(L_{11}), \dots, \Psi(L_{14})$ ovat semilineaarisia. Seuraavaksi esiteltävä lause tarjoaa suhteen semilineaaristen joukkojen ja säännöllisten kielten välille.

Lause 8.9 Olkoon $X \subseteq \mathbb{N}^q$ semilineaarinen joukko. Silloin on olemassa säännöllinen kieli $L \subseteq \Sigma^*$, missä on sellainen aakkosto $\Sigma = \{a_1, a_2, \dots, a_q\}$, että $\Psi(L) = X$.

Todistus. Semilineaarinen joukko on yhdiste äärellisestä määrästä lineaarisia joukkoja. Näin ollen riittää osoittaa tulos lineaariselle joukolle T. Olkoon $T = u_0 + \langle u_1, u_2, \dots, u_t\rangle$, missä $u_i = (v_{i,1}, v_{i,2}, \dots, v_{i,q})$. Olkoon

$$L = a_1^{v_{0,1}} a_2^{v_{0,2}} \dots a_q^{v_{0,q}} \Big(\sum_{1 \leq i \leq t} a_1^{v_{i,1}} a_2^{v_{i,2}} \dots a_q^{v_{i,q}} \Big)^*.$$

Tällöin kieli L on säännöllinen ja $\Psi(L) = T$. [Shallit, 2008]. □

Ennen Parikhin lauseen todistamista tarvitsemme seuraavan apulauseen:

Apulause 8.10 Olkoon $G = (N,T,P,S)$ kontekstiton kielioppi Chomskyn normaalimuodossa, jossa on q apumerkkiä. Olkoon $k = 2^{q+1}$. Kaikille luvuille j ≥ 1, pätee että jos $z \in L(G)$ ja $|z| \geq k^j$, jokaisella johdolla $S \Rightarrow^* z$ on sama derivaatiopuu kuin johdolla, joka on muotoa

$$S \Rightarrow^* uAy$$
$$\Rightarrow^* uv_1Ax_1y$$

$$\Rightarrow^* uv_1v_2Ax_2x_1y$$
$$\vdots$$
$$\Rightarrow^* uv_1v_2 \ldots v_jAx_j \ldots x_2x_1y$$
$$\Rightarrow^* uv_1v_2 \ldots v_jwx_j \ldots x_2x_1y = z,$$

missä $A \in N$, $v_ix_i \neq \lambda$, kun $1 \leq \mathrm{i} \leq \mathrm{j}$, ja $|v_1v_2 \ldots \ v_jx_j \ldots x_2x_1| \leq k^j$.

Todistus. Sivuutetaan (ks. [Shallit, 2008]). □

Lause 8.11 (Parikhin lause) Jos kieli L on kontekstiton, niin $\Psi(L)$ on semilineaarinen.

Todistus. Oletetaan, että kieleen L ei kuulu λ. Olkoon $G = (N,T,P,S)$ kielen $L - \{\lambda\}$ tuottama kontekstiton kielioppi Chomskyn normaalimuodossa, ja olkoon k apulauseen 8.10 vakio. Olkoon $\mathrm{U} \subseteq N$ mikä tahansa apumerkkien joukko, jossa on S. Määritetään L_U kaikkien sanojen joukoksi, jotka G on tuottanut käyttäen joukon U apumerkkejä. Nyt on ainoastaan äärellisen monta joukkoa L_U, ja $L = \bigcup_{\{S\} \subseteq U \subseteq N} L_U$. Täten riittää osoittaa, että kukin L_U on semilineaarinen. Seuraavaksi kiinnitämme mielivaltaisen joukon U ja oletamme, että kaikki johdot käyttävät ainoastaan apumerkkejä joukosta U. Olkoon $\mu = |\mathrm{U}|$, ja määritetään kielet
$\mathrm{E} = \{\mathrm{w} \in L_U : |\mathrm{w}| < k^\mu\}$ ja
$\mathrm{F} = \{\mathrm{vx} \mid 1 \leq |\mathrm{vx}| \leq k^\mu$ ja $A \Rightarrow^* \mathrm{v}A\mathrm{x}$ jollekin apumerkille $A \in \mathrm{U}\}$.

E ja F ovat äärellisiä, joten $\Psi(\mathrm{EF}^*)$ on semilineaarinen. Seuraavaksi osoitamme, että $\Psi(L_U) = \Psi(\mathrm{EF}^*)$. Ensiksi osoitamme, että $\Psi(L_U) \subseteq \Psi(\mathrm{EF}^*)$. Olkoon $z \in L_U$; todistamme induktiolla merkkijonon z pituuden suhteen, että $\Psi(z) \in \Psi(\mathrm{EF}^*)$. Perusaskel on $|z| < \mathrm{k}^\mu$. Tässä tapauksessa, $z \in \mathrm{E} \subseteq \mathrm{EF}^*$, joten $\Psi(z) \in \Psi(\mathrm{EF}^*)$. Induktioaskeleen oletuksena on, että $|z| \geq k^\mu$ ja $z \in L_U$. Apulauseen 8.10 mukaan sanalle z on olemassa johto, joka voidaan kirjoittaa muodossa

$$S \Rightarrow^* uAy$$
$$(d_1) \Rightarrow^* uv_1Ax_1y$$
$$(d_2) \Rightarrow^* uv_1v_2Ax_2x_1y$$
$$\vdots$$
$$(d_\mu) \Rightarrow^* uv_1v_2 \ldots v_\mu Ax_\mu \ldots x_2x_1y$$
$$\Rightarrow^* uv_1v_2 \ldots v_\mu wx_\mu \ldots x_2x_1y = z,$$

missä johdot on merkitty tunnisteilla $d_1, d_2, \ldots, d_\mu$.

Kukin μ - 1 apumerkeistä $B \in$ U - {A} yhdistää johdon d_i, jos B esiintyy johdossa d_i. Koska on μ tunnistein merkittyä johtoa ja ainoastaan μ - 1 apumerkkiä joukossa $B \in$

U - $\{A\}$, täytyy olla vähintään yksi johto d_i, joka ei yhdisty mihinkään apumerkkiin. Täten voimme jättää johdon d_i pois saadaksemme johdon

$$S \Rightarrow^* uv_1 \dots v_{i-1}v_{i+1} \dots v_\mu w x_\mu \dots x_{i+1}x_{i-1} \dots x_1 y = z',$$

missä z' $\in L_U$. Koska |z'| < |z|, saamme induktion nojalla Ψ(z') $\in \Psi(\mathrm{EF}^*)$. Nyt $v_i x_i \in$ F, joten Ψ(z) = Ψ(z'$v_i x_i$) $\in \Psi(\mathrm{EF}^*)$. Nyt jatkamme todistusta toiseen suuntaan.
Olkoon z $\in \mathrm{EF}^*$; silloin $z = e_0 f_1 f_2 \dots f_i$, kun t $\geq$ 0, missä $e_0 \in$ E ja $f_i \in$ F, kun 1 $\leq$ i $\leq$ t. Todistamme induktiolla luvun t suhteen, että $\Psi(z) \in \Psi(L_U)$.

Perusaskel on t = 0. Siinä tapauksessa, $z = e_0 \in$ E. Silloin määritelmän nojalla, $z \in L_U$, joten $\Psi(z) \in \Psi(L_U)$.
Induktioaskeleen oletuksena on, että $z = e_0 f_i \dots f_t$, missä $e_0 \in$ E ja jokainen $f_i \in$ F. Koska $f_t \in$ F, voidaan kirjoittaa $f_t = vx$, missä $1 \leq |vx| \leq k^\mu$ ja $A \Rightarrow^*$ vAx. Induktio-oletuksen nojalla tiedetään, että $\Psi(e_0 f_1 f_2 \dots f_{t-1}) = \Psi(z')$ jollekin $z' \in L_U$. Mutta koska $z' \in L_U$, on olemassa sanan z' johto, jossa käytetään apumerkkiä A, esimerkiksi S $\Rightarrow^*$ v'Ax' $\Rightarrow^*$ $v'w'x' = z'$. Mutta silloin $S \Rightarrow^*$ v'A'x' $\Rightarrow^*$ v'vAxx' $\Rightarrow^*$ v'vw'xx' = z'', joten $z'' \in L_U$. Nyt $\Psi(z'') = \Psi(z') + \Psi(vx) = \Psi(e_0 f_1 f_2 \dots f_{t-1}) + \Psi(f_t) = \Psi(z)$, joten $\Psi(z) \in L_U$. [Shallit, 2008]. □

Todistuksessa siis osoitetaan, että kukin L_U on semilineaarinen, eli $\Psi(L_U) = \mathrm{EF}^*$, koska joukkoja L_U on ainoastaan äärellinen määrä. Todistuksessa μ - 1 tarkoittaa, että kaikkia apumerkkejä, paitsi S, käytetään johdoissa. Pumppauslemman ollessa riittämätön Parikhin lauseen avulla pystytään todistamaan joissakin tapauksissa, että kieli ei ole kontekstiton.

Esimerkki 8.12 Kieli $L = \{a^i b^j | j \neq i^2\}$ ei ole kontekstiton. Oletetaan, että kieli L on kontekstiton. Silloin Parikhin lauseen nojalla joukko $\Psi(L) = \{(i,j) | j \neq i^2\}$ on semilineaarinen. Lauseen 8.4 nojalla T = $\overline{\Psi(L)} = \{(i, i^2) | i \geq 0\}$ on semilineaarinen. Mutta lauseen 8.9 nojalla on olemassa sellainen säännöllinen kieli R $\subseteq \{a, b\}^*$, että Ψ(R) = T. Seuraavaksi tarkastellaan morfismia h : $\{a, b\}^* \rightarrow$ c*, jonka on määrittänyt h(a) = h(b) = c. Silloin h(R) = $\{c^{n^2+n} | n \geq 0\}$, jonka näemme olevan ei-säännöllinen kieli käyttämällä pumppauslemmaa, mikä on ristiriita, sillä säännölliset kielet ovat suljettuja morfismin sovelluksen suhteen. Täten, kieli L ei ole kontekstiton. [Shallit, 2008].

9 Baderin ja Mouran lemman yleistäminen vahvoilla iterointilemmoilla

Tässä luvussa esittelemme Baderin ja Mouran lemman yleistyksen useilla iterointilemmoilla säännöllisille, lineaarisille, kontekstittomille ja lineaarisille indeksoiduille kielille.

Dömösi ja Kudlek [1999] ovat kehittäneet Baderin ja Mouran lemmaa kahteen suuntaan parantamalla sanan z merkittyjen merkkien lukumäärän alarajaa sekä parantamalla vastaavasti osasanan vwx merkittyjen merkkien lukumäärän ylärajaa. Todistaaksemme uudet iteraatiolemmat tarvitsemme kolme lausetta, jotka esittelemme seuraavaksi.

Tässä luvussa valittujen merkkien lukumäärää sanassa z merkitään $\sigma(z)$ ja poissuljettujen merkkien lukumäärää sanassa z vastaavasti $\varrho(z)$.

Lause 9.1 Olkoon L kontekstiton kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k = k(L) \geq 2$, että kaikki sanat $z \in L$, missä $\mathrm{d}(z) > k^{\varrho(z+1)}$, voidaan kirjoittaa muodossa $z = uvwxy$ siten, että

1. $\varrho(vx) = 0$ ja joko $\mathrm{d}(u) > 0$, $\mathrm{d}(v) > 0$, $\mathrm{d}(w) > 0$
tai $\mathrm{d}(w) > 0$, $\mathrm{d}(x) > 0$, $\mathrm{d}(y) > 0$
2. $\mathrm{d}(vwx) \leq k^{\varrho(w)+1}$
3. $uv^iwx^iy \in L$, $i \geq 0$.

Todistus. Sivuutetaan (ks. [Dömösi et al., 1996]). □

On mahdollista, että poissuljettu merkki on myös merkitty merkki. Merkittävää on havaita, että $\mathrm{d}(vwx)$ - $\varrho(vx) = \varrho(w)$. Lauseen 9.1 ehto 1 on kirjoitettavissa merkittyjen merkkien osalta myös muodossa $\mathrm{d}(vx) \geq 1$. Valitsemalla $\varrho(z)$ =0, saamme seuraavan (valinta ilmenee ehdoissa 1 ja 2) lauseen:

Lause 9.2 Olkoon L kontekstiton kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k = k(L) \geq 2$, että kun $\mathrm{d}(z) > k$, niin z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että

1. Joko $\mathrm{d}(u) > 0$, $\mathrm{d}(v) > 0$, $\mathrm{d}(w) > 0$ tai $\mathrm{d}(w) > 0$, $\mathrm{d}(x) > 0$, $\mathrm{d}(y) > 0$
2. $\mathrm{d}(vwx) \leq k$
3. kaikille $i \geq 0$, $uv^iwx^iy \in L$.

Todistus. Sivuutetaan (ks. [Dömösi et al., 1996]). □

Eroavuuksina lauseeseen 9.1 nähden havaitsemme, että nyt lauseesta on poistettu eksponentti $^{\varrho(z+1)}$, ehdosta 1 on poistettu $\varrho(vx) = 0$ ja ehdosta 2 on poistettu $^{\varrho(w+1)}$.

Seuraavaksi esittelemme neljä uutta vahvaa iteraatiolemmaa. Lemmat kattavat neljä erityyppistä kieltä.

Lause 9.3 (Säännöllisille kielille tarkoitettu lemma.)
Olkoon L säännöllinen kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k \geq 2$, että kun $\mathrm{d}(z) > k \cdot \max(\varrho(z),1)$, niin sana z voidaan kirjoittaa muodossa $z = uvw$ siten, että

1. $\varrho(v)$=0, $\mathrm{d}(u) > 0$, $\mathrm{d}(v) > 0$
2. $\mathrm{d}(uv) \leq k \cdot (\varrho(u) + 1)$
3. kaikille $i \geq 0$, $uv^i w \in L$.

Todistus. Sivuutetaan (ks. [Dömösi ja Kudlek, 1999]). □

Merkittyjen merkkien määrä sanassa z on suurempi kuin lemmassa annettu vakio k kertaa korkeintaan poissuljettujen merkkien määrä sanassa z. Lopuksi oleva 1 on vakio, joka on positiivinen tässä ja seuraavissa lauseissa. Luku 1 tarkoittaa derivatiopuun merkittyä polkua, missä toteutuu se, että poissuljettu merkki voi olla myös merkitty. Havaitsemme, että $\mathrm{d}(uv)$ - $\varrho(v) = \varrho(u)$. Eroavuutena alkuperäiseen säännöllisten kielten lemmaan on merkkien merkitseminen ja poissulkeminen.

Toinen uusi lause antaa vastaavan tuloksen lineaarisille kielille.

Lause 9.4 (Lineaarisille kielille tarkoitettu lemma.)
Olkoon L lineaarinen kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k \geq 2$, että kun $\mathrm{d}(z) > k \cdot \max(\varrho(z), 1)$, niin sana z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että
1. $\varrho(vx) = 0$ ja joko $\mathrm{d}(u) > 0$, $\mathrm{d}(v) > 0$, $\mathrm{d}(w) > 0$
tai $\mathrm{d}(w) > 0$, $\mathrm{d}(x) > 0$, $\mathrm{d}(y) > 0$
2. $\mathrm{d}(uvxy) \leq k \cdot (\varrho(uy) + 1)$
3. kaikille $i \geq 0$, $uv^i wx^i y \in L$.

Todistus. Sivuutetaan (ks. [Dömösi ja Kudlek, 1999]). □

Havaitsemme, että d($uvxy$) - $\varrho(vx)$ = ϱ(uy).

Lause 9.5 (Kontekstittomille kielille tarkoitettu lemma.)
Olkoon L kontekstiton kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k = k(L) \geq 2$, että kun d(z) > $k \cdot \max(\varrho(z), 1)$, niin sana z voidaan kirjoittaa muodossa $z = uvwxy$ siten, että
1. $\varrho(vx) = 0$ ja joko d(u) > 0, d(v) > 0, d(w) > 0 tai d(w) > 0, d(x) > 0, d(y) > 0
2. d(vwx) $\leq k \cdot (\varrho(w) + 1)$
3. kaikille $i \geq 0$, $uv^i wx^i y \in L$.

Havaitsemme, että d(vwx) - $\varrho(vx)$ = $\varrho(w)$.

Todistus. Olkoon $G = (N, T, P, S)$ Chomskyn normaalimuodossa oleva kontekstiton kielioppi, joka tuottaa kielen L. Tarkastellaan kieliopin G mukaista sanan z derivaatiopuuta T_z. Määritetään solmu oksapisteeksi, jos sillä on kaksi jälkeläistä ja niistä molemmilla on poissuljettuja jälkeläisiä. Määritetään solmu vapaaksi, jos kummallakaan jälkeläisellä ei ole poissuljettuja jälkeläisiä. Polku (osittainen) on merkitty, jos yksikään sen risteävistä solmuista ei ole oksapiste, tai sillä ei ole risteäviä solmuja. Aloitussolmu on joko oksapiste tai puun juuri, ja päätesolmu on joko oksapiste tai poissuljettu merkki. Määritelmistä seuraa, että derivaatiopuulla T_z on täsmälleen $\varrho(z)$ - 1 oksapistettä, ja että merkittyjen polkujen lukumäärä on joko 2 · $\varrho(z)$ - 2, jos derivaatiopuun T_z juuri on oksapiste, tai 2 · $\varrho(z)$ - 1, jos juuri ei ole oksapiste. Nyt määritetään $k = 2k' \cdot (2|N| + 3) + 1$, missä k' on vakio lauseesta 9.2.

Jos $\varrho(z) = 0$, niin väite on kuten lause 9.2. Olkoon $\varrho(z) > 0$. On korkeintaan 2$\varrho(z)$ - 1 merkittyä polkua, ja enemmän kuin (k - 1) · $\varrho(z)$ valittua merkkiä. Kaikki nämä ovat merkityillä poluilla risteävien solmujen vapaiden jälkeläisten tuottamia. Tästä on seurauksena, että vähintään yhden merkityn polun p täytyy tuottaa enemmän kuin $\frac{1}{2}(k$ - 1) valittua merkkiä, koska muutoin $\frac{1}{2}(k$ - 1) · (2$\varrho(z)$ -1) $\leq$ (k - 1) · $\varrho(z)$. Erotamme kaksi tapausta.

Tapaus 1. On olemassa pääpolun p vapaa jälkeläinen, joka on sellaisen binääripuun juuri, joka tuottaa enemmän kuin k' valittua merkkiä. Silloin lausetta 9.2 voidaan soveltaa tuottamaan lause u"vwxy,' missä $\varrho(vx) = 0$, joko d(u") > 0, d(v) > 0, d(w) > 0, tai d(w) > 0, d(x) > 0, d(y') > 0, d(vwx) $\leq k' < k \cdot (\varrho$(w) + 1), ja lauseet $uv^i wx^i y$ kuuluvat kieleen L, kun u = u'u", y = y'y".

Tapaus 2. Jokainen pääpolun p vapaa jälkeläinen tuottaa korkeintaan k' valittua merkkiä. Tuottaakseen enemmän kuin $\frac{1}{2}(k$ - 1) valittua merkkiä täytyy olla olemassa enemmän kuin 2 · $|N|$ + 3 vapaata pääpolun p jälkeläistä, joista kukin tuottaa vähintään yhden valitun merkin, koska muutoin korkeintaan $k' \cdot (2\ |N| + 3) = \frac{1}{2}(k$ - 1) valittua merkkiä ovat pääpolun p tuottamia.

Kullakin näistä vapaista jälkeläisistä on edeltäjä polussa p, ja siellä on vähintään

$2 \cdot |N| + 4$ sellaista. Pääpolku p muodostuu osapoluista p_1, p_2, p_3 ja p_4, jotka tuottavat vastaavaasti vähintään 1, $|N| + 1$, $|N + 1|$ ja 1 valittua merkkiä. Täten p_2 ja p_3 sisältävät $|N| + 1$ vapaiden jälkeläisten edeltäjää, kukin tuottaen vähintään yhden valitun merkin. Tällöin kullakin polulla p_2 ja p_3 on olemassa kaksi sellaista edeltäjää, joilla on identtiset tunnisteet A (vastaavasti B).

Sallitaan polkujen p_i tuottaa a_i (b_i) vasemmalle (vastaavasti oikealle).
Tapaus 2.1. p_2p_3 sisältää kaksi sellaista edeltäjää, joilla on identtiset tunnisteet A, ja missä $A \Rightarrow^*$vAx, σ(v) > 0, σ(x) > 0. Koska $\sigma(a_1b_1) > 0$, on seurauksena, että joko d(u) > 0, d(v) > 0, d(w) > 0, tai d(w) > 0, d(x) > 0, d(y) > 0.
Tapaus 2.2. Tarkastellaan kuutta alitapausta, joissa L(R) on merkintä sille, että p_2 ja p_3 tuottavat valitut merkit ainoastaan vasemmalla (vastaavasti oikealla).

RR : $\sigma(a_1b_1) > 0$, $\sigma(a_2) = 0$, $\sigma(b_2) > 0$, $\sigma(a_3) = 0$, $\sigma(b_3) > 0$, $b_3 = c_3xd_3$, σ(x) > 0, ja p_2 vaikuttaa niin, että σ(y) > 0.
RL1 : $\sigma(a_1) > 0$, $\sigma(b_1) \geq 0$, $\sigma(a_2) = 0$, $\sigma(b_2) > 0$, $\sigma(a_3) > 0$, $\sigma(b_3) = 0$, $\sigma(a_3) = c_3vd_3$, σ(v) > 0, ja p_1 vaikuttaa niin, että σ(u) > 0.
RL2 : $\sigma(a_1) \geq 0$, $\sigma(b_1) > 0$, $\sigma(a_2) = 0$, $\sigma(b_2) > 0$, $\sigma(a_3) > 0$, $\sigma(b_3) = 0$, $\sigma(b_2) = c_2xd_2$, $\sigma(x) > 0$, ja p_1 vaikuttaa niin, että σ(y) > 0.
Muut alitapaukset ovat symmetrisiä.

Valitsemalla p lähimpänä lehteä olevaksi, joka sisältää 2N + 4 sellaista edeltäjää, joiden jälkeläiset tuottavat vähintään yhden valitun merkin, saadaan voimaan d(vwx) < $(2\varrho$(w) - 1) $\cdot \frac{1}{2}(k$ - 1) + ϱ(w) $\leq k \cdot \varrho$(w) $\leq k \cdot (\varrho$(w) + 1). [Dömösi and Kudlek, 1999] □

Määritelmä 9.6 Kieli on lineaarisesti indeksoitu, kun se on tuotettu lineaarisesti indeksoidulla kieliopilla $G = (N,T,I,P,S)$, missä N on apumerkkien joukko, T on perusmerkkien joukko, I on indeksien joukko, P on äärellinen parien (Af, a) joukko, missä $A \in N$, f $\in I \cup \{\lambda\}$, a $\in T^* \cup T^*\ N\ I\ ^*\ T^*$, äärellinen joukko sääntöjä, ja S on aloitusmerkki. [Duske and Parchmann, 1984]

Lineaarisesti indeksoidun kieliopin tuottama mekanismi kielen tuottamiseksi oletetaan tunnetuksi (ks. [Duske and Parchmann, 1984]).

Seuraavassa lauseessa z^R merkitsee sanan z peilikuvaa.

Lause 9.7
Jos $L' \subseteq$ Y* on lineaarinen indeksoitu kieli, niin on olemassa sellainen aakkosto T, kontekstiton kieli $L \subseteq T^*$ ja kaksi sellaista morfismia $h_1, h_2 : T^* \rightarrow$ Y*, että $L' = \{h_1(w)h_2(w)^R | w \in L\}$.

Todistus. Sivuutetaan (ks. [Duske and Parchmann, 1984]). □

Lause 9.8 (Lineaarisille indeksoiduille kielille tarkoitettu lemma.)
Olkoon L lineaarinen kieli. Silloin on olemassa sellainen kielestä L riippuva kokonaisluku $k \geq 2$, että kun d(z) > $k \cdot (\varrho(z) + 1)$, niin sana z voidaan kirjoittaa muodossa $z{=}u_1v_1w_1x_1y_1u_2v_2w_2x_2y_2$ siten, että

1. $\varrho(v_1x_1v_2x_2) = 0$ ja joko d(u_1u_2) > 0, d(v_1v_2) > 0, d(w_1w_2) > 0 tai d(w_1w_2) > 0, d(x_1x_2) > 0, d(y_1y_2) > 0
2. d($v_jw_jx_j$) $\leq k \cdot (\varrho(w_1w_2) + 1)$, j = 1,2
3. kaikille $i \geq 0$, $u_1v_1^iw_1x_1^iy_1y_2x_2^iw_2v_2^iu_2 \in L$.

Todistus. Tarkastellaan lineaarista indeksoitua kieltä $L \subseteq \mathrm{Y}^*$. Lauseen 9.5 nojalla on olemassa sellainen aakkosto T, että kontekstiton kieli $L' \subseteq T^*$ ja kaksi sellaista morfismia $h_1, h_2 : T^* \rightarrow \mathrm{Y}^*$, että $L = \{h_1(w)h_2(w)^R | w \in L'\}$. Olkoon a = $2 \cdot \max\{|h_i(x)| \mid \mathrm{i} \in \{1,2\}, x \in T\}$ ja k = an', missä k' on lauseen 9.5 vakio kielelle L'. Nyt a > 0, koska muutoin $L = \{\lambda\}$. Tarkastellaan sanaa $z \in L$, missä d(z) > $k \cdot (\varrho(z)+ 1)$. On olemassa sellainen sana p $\in L$', että $h_1(p)h_2(p)^R$. Olkoon sana p lyhin mahdollinen. Tällöin $z{=}h_1(p')h_2(p')^R$, $p' \in L$', johtaa siihen, että $|p| \leq |p'|$. Olkoon $p = s_1 \ldots s_t \in T^+$, $s_1, \ldots, s_t \in T$.

Tarkastellaan sanoja $z_1 = h_1(p)$ ja $z_2 = h_2(p)$. Silloin $\varrho(z) = \varrho(z_1) + \varrho(z_2)$. Kun i kuuluu joukkoon $\{1,\ldots,t\}$, suljetaan pois sanan p i. merkki = $s_1 \ldots s_t$, jos ja vain jos yksi merkeistä $h_1(s_i)h_2(s_i)$ on suljettu pois. Silloin ϱ(p) $\leq \varrho(z_1) + \varrho(z_2) = \varrho(z)$. Lisäksi, kun i kuuluu joukkoon $\{1,\ldots,t\}$, merkitään sanan p i. merkki = $s_1 \ldots s_t$, jos ja vain jos yksi merkeistä $h_1(x_i)h_2(x_i)$ on merkitty. Tämä johtaa siihen, että a $\cdot$ d(p) $\geq$ d(z) > $k \cdot \max(\varrho(z),1) = (ak') \cdot \max(\varrho(z),1) \geq a \cdot k' \cdot \max(\varrho(p),1)$, ja tästä päädytään d(p) > $k' \cdot \max(\varrho(p),1)$.

Nyt voidaan soveltaa lausetta 9.5 kieleen L' ja sanaan p $\in L$'. Täten sanalle p = $uvwxy$, missä joko d(u) > 0, d(v) > 0, d(w) > 0, tai d(w) > 0, d(x) > 0, d(y) > 0, $\varrho(vx) = 0$, d(vwx)$\leq k \cdot (\varrho(w) + 1)$, pätee, että kaikilla arvoilla i ≥ 0 sanat uv^iwx^iy kuuluvat kieleen L' ja
$z = h_1(u)h_1(v)h_1(w)h_1(x)h_1(y)h_2(y)^Rh_2(x)^Rh_2(w)^Rh_2(v)^Rh_2(u)^R$.

Oletetaan, että $h_1(v)h_1(x)h_2(x)h_2(v) = \lambda$. Nyt kaikille arvoille i ≥ 0 pätee uv^iwx^iy $\in$ L' ja erityisesti uwy kuuluu kieleen L', joten $z = h_1(\mathrm{uwy})h_2(\mathrm{uwy})^R$. Koska $|vx| >$ 0, tämä muodostaa ristiriidan sanan p pituuden minimaalisuudelle. Täten $|h_1(v)h_1(x)h_2(x)h_2(v)|$ > 0. Asetetaan $u_1 = h_1(u), v_1 = h_1(v), w_1 = h_1(w), x_1 = h_1(x), y_1 = h_1(y)$, ja $y_2 = h_2(y)^R, x_2 = h_2(x)^R, w_2 = h_2(w)^R, v_2 = h_2(v)^R, u_2 = h_2(u)^R$. Silloin $|v_1x_1x_2v_2| >$ 0.

Poissuljetuista merkeistä $\varrho(vx) = 0$ seuraa $\varrho(v_1x_1x_2v_2) = 0$. Merkityistä merkeistä d(u) > 0 , d(v) > 0, d(w) > 0 seuraa, että d(u_1u_2) > 0, d(v_1v_2) > 0, d(w_1w_2) > 0, ja ominaisuuksista d(w) > 0, d(x) > 0, d(y) > 0 seuraa, että d(w_1w_2) > 0, ja tästä

edelleen seuraa, että d(x_1x_2) > 0, d(y_1y_2) > 0. Lauseen 9.7 perusteella d(vwx) ≤ (k') · (ϱ(w) + 1). Koska ϱ(w) ≤ $\varrho(w_1w_2)$ seuraa, että d($v_jw_jx_j$) ≤ a · d(vwx) ≤ a · (k') · (ϱ(w) + 1) ≤ (ak') · (ϱ(w) + 1) ≤ k · (ϱ(w) + 1) ≤ k · ($\varrho(w_1w_2)$ + 1) kun j = 1,2. [Dömösi and Kudlek, 1999] □

Havaitsemme, että lauseen 9.8 ehdoissa 1-3 esiintyy indeksointia. Havaitsemme, että d($v_jw_jx_j$) - $\varrho(v_1x_1v_2x_2)$ = $\varrho(w_1w_2)$. Merkintä j siis sisältää kaksi merkkiä. Merkittävänä eroavaisuutena huomaamme, että $uvwxy$ on kahdesti peräkkäin jälkimmäisen osan ollessa käänteisessä järjestyksessä peilikuvan lailla kirjoitettu, mikä johtuu lauseessa 9.7 olevista morfismeista sanan $uvwxy$ rakenteiden siis säilyessä samanlaisina. Lauseessa 9.8 pumpataan poikkeuksellisesti jopa neljässä lohkossa. Aluksi esitellyistä lauseista todettakoon, että lauseen 9.5 todistuksessa siis käytettiin lauseita 9.1 ja 9.2 sekä lauseen 9.8 todistuksessa lausetta 9.7. Dömösi ja Kudlek [1999] siis esittelevät neljä uutta iteraatiolemmaa.

Alussa mainittua sanan z merkittyjen merkkien lukumäärän alarajan parantaminen toteutuu lauseissa 9.3 ja 9.4 sekä vastaavasti osasanan vwx merkittyjen merkkien lukumäärän ylärajan parantaminen toteutuu lauseissa 9.5 ja 9.8. Lauseen 9.5, joka siis on kontekstittomille kielille, todistuksessa kokonaisluvun k määrityksen yhteydessä esiintyvä k' on kokonaisluku lauseesta 9.2, joka myös koskee kontekstittomia kieliä, kun taas lauseen 9.8 todistuksessa, joka on lineaarisesti indeksoiduille kielille, vastaavasti k' on lauseesta 9.4, mutta joka koskee kontekstittomia kieliä. Lauseen 9.5 todistuksen lopussa on implikaatio, jonka alku- ja loppuosa muodostuvat lauseen 9.5 ehdosta 2 ja vastaavasti lauseen 9.8 kohdalla.

10 Tietojenkäsittelytieteellisiä sovelluksia

Tässä luvussa tarkastelemme ohjelmointikielten todistamista ei-konteksittomiksi ja esitämme esimerkkejä.

Apulause 10.1 Jos W on äärellinen puu, jossa millään solmulla ei ole kuin korkeintaan r jälkeläistä ja pisimmän polun pituus on h, niin puun W lehtien lukumäärä on korkeintaan r^h.

Todistus. Sivuutetaan (ks. [Sokolowski, 1978]). □

Olkoon $G = (N,T,P,S)$ kontekstiton kielioppi. Olkoon n perusmerkkien lukumäärä ja r kieliopin G pisimmän säännön pituus.

Apulause 10.2 Jos s on positiivinen kokonaisluku, z sana aakkostossa T, y on sanan z osasana, |y| ≤ s, $|z| \geq r^{(s+1)(n+1)}$ ja on olemassa apumerkki B ja johto $B \Rightarrow^* z$, silloin poistamalla joitakin sanan z kirjaimia osasanan y ulkopuolella voimme saada aidosti lyhyemmän sanan z', joka myös on johdettavissa apumerkistä B.

Todistus. Olkoon R pisin polku derivaatiopuun W apumerkistä B. Apulauseen 10.1 mukaan pisimmän polun pituus (R) ≥ (s+1)(n+1). Koska |y| ≤ s, puun W solmuja ei ole enempää kuin s, missä polku R kohtaa polun osasanasta y. Täten on olemassa polun R osapolku R', jonka pituus on n + 1 ja jossa ei ole tämän kaltaisia solmuja. Vähintään yhden apumerkin A täytyy tulla toistetuksi tämän polun varrella, joten on olemassa sellaiset sanat d_1, d_2, d_3, d_4 ja d_5, että
1) $B \Rightarrow^* d_1 A\, d_5$,
2) $A \Rightarrow^* d_2 A\, d_4$,
3) $A \Rightarrow^* d_3$,
missä vähintään toinen sanoista d_2 ja d_4 on ei-tyhjä ja pätee, että $d_2 \cap$ y = ∅ ja $d_4 \cap$ y = ∅. Yhdistämällä 1) ja 3) saamme johdon $B \Rightarrow^* d_1 d_3 d_5 = z$', missä tuloksena on lyhyempi sana ja täten apulause 10.2 on todistettu. [Sokolowski, 1978] □

Todistuksessa on siis lähtökohtana sana z, jonka pituus > z'. Toistimme saamassamme johdossa $B \Rightarrow^* d_1 d_3 d_5 = z$' vähintään yhtä apumerkkiä A osapolulla R'. Lisäksi poistamalla sanasta z joitakin kirjaimia saimme tavoittelemamme lyhyemmän sanan z'.

Lause 10.3 (Sokolowski) Jos L on kontekstiton kieli, niin jokaiselle perusmerkkien aakkoston T osajoukolle T', jossa on vähintään merkit a ja b, ja kaikille perusmerkkijonoille u_1, u_2 ja u_3 pätee, että jos $\{u_1 x u_2 x u_3 | x \in T'^+\}$ kuuluu kieleen L, niin on olemassa sellaiset joukon T' merkeistä muodostuvat merkkijonot x' ja x", että

$u_1x'u_2x''u_3$ kuuluu kieleen L.

Todistus. Olkoot nyt s = max($|u_1|$, $|u_2|$, $|u_3|$) ja m = $r^{r^{(s+1)(n+1)}}$. Oletetaan, että kaikki muotoa w = $u_1xu_2xu_3$, missä x $\in \sum'^{+}$, olevat sanat kuuluvat kieleen L. Jos asetamme x = $a^{(m)}b^{(m)}$, saamme kielen L sanan. Olkoon W sanan w derivaatiopuu kieliopissa G ja R sen pisin polku. Olkoon t_0 polun R sellainen "*nuorin*" solmu, että puun W täydellä osapuulla W_0, jonka juurena olevalla solmulla t_0 on vähintään $r^{(s+1)(n+1)}$ lehteä. Olkoon t_1 solmun t_0 välitön jälkeläinen polun R varrella - W_1 puun W täysi osapuu, missä W_1 on puun W juuri. Osapuulla W_1 on vähemmän kuin $r^{(s+1)(n+1)}$ lehteä, joten apulauseen 10.1 mukaan osapuussa W_1 olevan polun R osa on lyhyempi kuin $r^{(s+1)(n+1)}$. Tällöin osapuussa W_0 olevan polun R osan pituus on korkeintaan $r^{(s+1)(n+1)}$ ja siten osapuun W_0 lehtien lukumäärä ei ole suurempi kuin $r^{(s+1)(n+1)}$ = m.

Saimme arvion $r^{(s+1)(n+1)} \leq$ osapuun W_0 lehtien lukumäärä $\leq$ m. Olkoon A solmun t_0 apumerkki derivaatiopuussa W ja z osapuun W_0 perusmerkkisana. Johtuen sen pituudesta, sanalla z voi olla ei-tyhjä leikkaus, jossa ei ole enempää kuin yksi perusmerkkijonoista u_1, u_2 ja u_3. Olkoon y tämä leikkaus. Koska apulauseen 10.2 oletukset täyttyvät, voimme korvata sanan z lyhyemmällä sanalla z' vaikuttamatta perusmerkkijonoihin u_1, u_2, u_3. Verrataan vielä sanojen z ja w pituutta ja havaitsemme, että tämä korvaus muuttaa joko ensimmäisen tai toisen merkkijonon x sanassa w, tai b-merkin lukumäärän ensimmäisessä ja a-merkin lukumäärän toisessa sanassa. Tämän muutoksen jälkeen molemmat x-merkkijonot eivät voi olla samat. [Sokolowski, 1978] □

Havaitsemme, että Bar-Hillelin lemmaa mukaillen u_1 = u, u_2 = v ja u_3 = w, ja jolloin $u_1x'u_2x''u_3 = ux'vx''w$. Lähdetään liikkeelle kahdesta samasta merkkijonosta x = $a^{(m)}b^{(m)}$. Mutta toinen x muuttuu korvauksen jälkeen. Korvaus ei siis koske merkkijonoja u_1, u_2 ja u_3. Huolimatta siitä, että myös $^{(m)}$ säilyy samana, niin merkkijonot x eivät säily samoina, koska esimerkiksi a-merkkejä voi aluksi olla 3 ja b-merkkejä 5 tai a-merkkejä on vähemmän kuin b-merkkejä tai merkkejä on yhtä paljon. Ratkaisevaa oli poistaa kyseinen samuus.

Ohjelmointikieli Algol 60 todistettiin (vuonna 1962) kontekstittomaksi. Seuraavaksi Sokolowski todistaa, että Algol ei ole kontekstiton.

Esimerkki 10.5 Oletetaan, että Σ' on kirjainten joukko ja Σ on joukko Algolin perusmerkkejä ja tarkastellaan ohjelmaa
begin real p; p = 0 **end**.

Ohjelma säilyy oikeana jokaiselle korvaukselle aakkoston Σ'^{+} sanalle koskien merkin x kumpaakin esiintymää. Jos Algol olisi kontekstiton, niin se sisältäisi sanan **begin** real p'; p" := 0 **end**,
missä p' ≠ p" , jota se selvästi ei sisällä.

Algol ei siis ole kontekstiton. [Sokolowski, 1978]

Esimerkki 10.6 Olkoon kieli $L = \{xx|x \in \Sigma^*\}$, ja sisältäköön aakkosto Σ vähintään kolme merkkiä. Todistamme, että kieli L ei ole kontekstiton. Olkoot $\Sigma = \{a, b, c\}$, $\Sigma' = \{a, b\}$, $u_1 = u_2 = c, u_3 = \lambda$. Nyt $\{cxcx|x \in \{a, b\}^+\} \subseteq L$. Jos L olisi kontekstiton, niin se sisältäisi sanan cx'cx", missä x' $\neq$ x" ja x', x" $\in \{a, b^+\}$, jota se ei sisällä. [Sokolowski, 1978]

Esimerkki 10.7 Olkoon Σ aakkosto ja c $\notin \Sigma$. Olkoon $L = \Sigma^+ \cup \{\text{xc}^{(h)}\ \text{x}|\text{x} \in \Sigma^+$ ja h $> 0\}$. Mikä tahansa kielen L sana z muodostuu joko tunnisteen esittelystä tai molemmista - esittelystä ja sovelluksesta erotinmerkkijonon c erottamina. Sanan $z$ositukselle kielestä L valitsemme lauseen 10.3 mukaisesti seuraavan merkintätavan: jos z sisältää merkin "c" ($z = \text{xc}^{(h)}$ x), niin

x - ensimmäinen "c" - tyhjä - tyhjä - loput sanasta z,

mutta jos z ei sisällä merkkiä "c":

tyhjä - ensimmäinen sanan z merkki - tyhjä - tyhjä - loput sanasta z.

Kielessä L on välttämättömiä muuttujien esittelyjä. Yllä on kyse siitä poissulkevatko välttämättömät määritykset kontekstiriippumattomuuden. Bar-Hillelin lemmaa, eli lausetta 4.1, käyttäen ei voida todistaa kielen kontekstiriippumattomuutta. Mutta sen sijaan kieli ei toteuta Sokolowskin lausetta ja ei täten ole kontekstiton. [Sokolowski, 1978]

Sokolowski soveltaa siis esimerkkiinsä 10.7 Bar-Hillelin lemmaa, koska haluaa osoittaa lemman riittämättömyyden sekä oman lauseensa riittävyyden. Havaitsemme selvästi, että Bar-Hillelin lemman avulla esimerkin kielestä ei voida sanoa juuri mitään. Esimerkin kieli onkin täysin erilainen kuin kielet, joita Bar-Hillelin lemmalla yleensä tutkitaan. Sokolowskin lause on siis tällaisissa tapauksissa vahvempi kuin Bar-Hillelin lemma. Sokolowskin lauseella voi todistaa, että kieli L ei ole kontekstiton. Mutta on huomioitava, että lauseessa on merkittävä rajoitus, sillä lause toimii ainoastaan joissakin tapauksissa. Sokolowskin lause on suppeampi kuin Bar-Hillelin lemma.

Sokolowskin lauseeseen ei liity pumppaamista. Lause soveltuu ohjelmointikielille ja perustuu tosiasiaan, että ei ole olemassa menetelmää tarkistaa, ovatko sanan kaksi osasanaa samat, mikä liittyy kontekstittomien kielten ominaisuuksiin.

Algol ei ole ainoa ohjelmointikieli, josta Sokolowskin lauseen avulla voidaan todistaa, että ohjelmointikieli ei ole kontekstiton, vaan lause soveltuu muillekin Algolin kaltaisille ohjelmointikielille.

11 Yhteenveto

Säännöllisten kielten pumppauslemmassa merkkijonoja pumpataan aina yhdestä kohdasta. Ehrenfeucht ja muut ovat vahvistaneet lemmaansa lohkopumppausominaisuudella. Kielellä on myös peruutuksellisia ominaisuuksia. Bar-Hillelin lemmalla voidaan osoittaa, että jokin kieli ei ole kontekstiton, mutta on olemassa kieliä, joita ei voida osoittaa kontekstittomiksi käyttämällä lemmaa. Lisäksi esiteltiin kieliä, jotka toteuttavat Bar-Hillelin lemman. Luvussa 5 esiteltiin lemmat lineaarisille ja eilineaarisille kontekstittomille kielille. Totesimme, että Horvathin lemman ehdot eivät toteudu jossain kielessä ja toisinaan Bar-Hillelin lemman ehdot eivät toteudu jossain muussa kielessä. Joissakin tapauksissa on kokeiltava molempia lemmoja.

Bar-Hillelin lemmaa on siis vahvistettu ja käytetty perustana useissa lemmoissa, samoin Ogdenin lemmaa. Ogdenin lemmalla on kaksi tarkoitusta. Ogden ensin kehitti merkkien merkitsemisen, jonka avulla generoitavia sanoja pystytään hallitsemaan. Seuraavaksi Ogden alkoi käyttää lemmaa siten, että sillä voidaan todistaa jotkut kielet luontaisesti moniselitteiseksi. Luontaisessa moniselitteisyydessä samalla sanalla voi olla vähintään kaksi eri vasenta johtoa kaikissa mahdollisissa kieliopeissa.

Ogdenin lemmaa on yleistetty kehittämällä siihen lisää ominaisuuksia. Bader ja Moura ottivat käyttöön poissuljettavuuden, jonka avulla voidaan hallita saatavia sanoja vielä enemmän kuin Ogdenin lemmaa käyttämällä. Bar-Hillelin lemman ongelmana on, että koska kaikki merkit pidetään samanarvoisina, niin lemma jää usein riittämättömäksi ja näimme esimerkissä 6.0.2, että sama kieli voitiin osoittaa eikontekstittomaksi käyttämällä Ogdenin lemmaa, vaikka Bar-Hillelin lemma ei siis riittänyt. Ogdenin lemmassa merkit ovat jaettavissa kahteen luokkaan, jotka ovat merkitsemättömät ja merkityt, kun taas Baderin ja Mouran lemmassa merkit voidaan jakaa jopa kolmeen luokkaan, jotka ovat edellisten lisäksi poissuljetut. Täten Bar-Hillelin lemmalla voidaan hyvin vähän vaikuttaa siihen minkälaisia sanoja kieleen saadaan, Ogdenin lemmalla voidaan vaikuttaa selvästi enemmän ja Baderin ja Mouran lemmalla voidaan siis vaikuttaa vielä enemmän kuin Ogdenin lemmalla. Lisäksi esiteltiin kieliä, jotka ovat ogdenmaisia.

Vaihtolemmassa ei ole pumppausominaisuutta, vaihtolemmassa korvataan merkkijonoja. Poikkeuksellista on, että korvatut sanat edelleen kuuluvat kieleen. Todistettaessa, että kieli ei ole kontekstiton, aakkoston on oltava vähintään kolmikirjaiminen. Bar-Hillelin lemman ollessa riittämätön Parikhin lauseen avulla voidaan todistaa joissakin tapauksissa, että kieli ei ole kontekstiton. Parikhin lauseessa siis minkä tahansa kontekstittoman kielen Parikhin kuva on semilineaarinen joukko. Parikhin kuvauksessa aakkostoa koskeva järjestys voidaan sivuuttaa.

Dömösi ja Kudlek ovat laajentaneet Baderin ja Mouran lemmaa, joka perustuu Ogdenin lemmaan, parantamalla sanan z merkittyjen merkkien lukumäärän alarajaa se-

kä parantamalla vastaavasti osasanan *vwx* merkittyjen merkkien lukumäärän ylärajaa. Heidän lemmojaan voidaan käyttää säännöllisille, lineaarisille, kontekstittomille ja lineaarisille indeksoiduille kielille. Sokolowskin lauseessa ei ole pumppaamista. Vaikka Sokolowskin lause on siis joissakin tapauksissa vahvempi kuin Bar-Hillelin lemma, niin Sokolowskin lause on kuitenkin suppeampi kuin Bar-Hillelin lemma. Sokolowskin lause soveltuu Algolin kaltaisille ohjelmointikielille. Pumppauslemmoista Baderin ja Mouran lemma ja sitä kautta myös Dömösin ja Kudlekin lemmat ovat merkittävän vahvoja.

Viiteluettelo

[Automaatit-kurssi] Tampereen yliopisto, Talvi 2012.

[Bader and Moura, 1982] Christopher Bader and Arnaldo Moura. A generalization of Ogden's lemma. *The Journal of the ACM 29*, 1982, 404-407.

[Boasson and Horvath, 1978] L. Boasson and Sandor Horvath. On languages satisfying Ogden's lemma. *RAIRO - Theoretical Informatics and Applications - Informatique Théorique et Applications* **12**, 1978, 201-202.

[Duske and Parchmann, 1984] Jürgen Duske and Rainer Parchmann. Linear Indexed Languages. *Theoretical Computer Science* **32**, 1984, 47-60.

[Dömösi et al., 1996] Pál Dömösi, Masami Ito, Masashi Katsura and Chrystopher L. Nehaniv. A New Pumping Property of Context-Free Language. *Combinatorics, Complexity and Logic*, 1996, 187-193.

[Dömösi and Kudlek, 1999] Pál Dömösi and Manfred Kudlek. Strong iteration lemmata for regular, linear, context-free, and linear indexed languages. In: *LNCS* **1684**, 1999, 226-233.

[Ehrenfeucht et al., 1981] Andrzej Ehrenfeucht, Rohit J. Parikh and Grzegorz Rozenberg. Pumping Lemmas for Regular Sets. *SIAM Journal on Computing* **10**, 1981, 536-541.

[Ginsburg and Spanier, 1966] Seymour Ginsburg and Edwin H. Spanier. Finite-turn Pushdown Automata. *SIAM Journal on Computing* **4**, 1966, 429-453.

[Hopcroft and Ullman, 1979] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Language, and Computation*, Addison-Wesley, Reading MA, 1979.

[Hopcroft et al., 2001] John E. Hopcroft, Jeffrey D. Ullman and Rajeev Motwani. *Introduction to Automata Theory, Language, and Computation*, Addison-Wesley, Reading, Massachusetts, second edition, 2001.

[Horvath, 2006] Geza Horvath. New Pumping Lemma for Non-Linear Context-Free Languages. In: *Proc. of the 9th Symposium on Algebras, Languages and Computation*, 2006, 160-163.

[Horvath, 1978] Sandor Horvath. The family of languages satisfying Bar-Hillel's Lemma. *RAIRO - Theoretical Informatics and Applications - Informatique Théorique et Applications* **12** , 1978, 193-199.

[Howie, 1991] John Howie. *Automata and Languages*, Oxford Science Publications,

1991.

[Jaffe, 1978] Jeffrey Jaffe. A Necessary and Sufficient Pumping Lemma for Regular Languages. *Sigact News*, **2** , 1978, 48-49

[Martin, 2003] John Martin. *Introduction to Languages and the Theory of Computation*, McGraw-Hill Higher Education, third edition, 2003.

[Mateescu and Salomaa, 1997] Alexandru Mateescu and Arto Salomaa. Formal Languages: an Introduction and a Synopsis. In: Grzegorz Rozenberg and Arto Salomaa (eds.), *Handbook of Formal Languages*, Springer, 1997.

[Nerode, 1958] Anil Nerode. Linear automata transformations. *Proc. American Mathematical Society* **9**, 1958, 541-544.

[Ogden, 1968] William Ogden. A helpful result for proving inherent ambiguity. *Mathematical Systems Theory* **2**, 1968, 191-194.

[Ramsey, 1954] F. P. Ramsey. On a problem of formal logic. *The Foundations of Mathematics*, 1954, 82-111.

[Rosen, 2010] Kenneth Rosen. *Elementary Number Theory and its Applications*, Pearson Addison Wesley, 2010.

[Salomaa, 1973] Arto Salomaa. *Formal Languages*, Academic Press, 1973.

[Shallit, 2008] Jeffrey Shallit. *A Second Course in Formal Languages and Automata Theory*, Cambridge University Press, 2008.

[Simon, 1999] Matthew Simon. *Automata Theory*, World Scientific, 1999.

[Sokolowski, 1978] Stefan Sokolowski. A method for proving programming languages non-context-free. *Information Processing Letters* **7**, 1978, 151-153.