TARTU ÜLIKOOL
MATEMAATIKATEADUSKOND
Arvutiteaduse instituut
Teoreetilise informaatika õppetool
informaatika eriala

Peeter Laud

Üldistatud kontekstivabad grammatikad

Magistritöö

Juhendaja: dots. Uno Kaljulaid

Autor: .... „ ... “ mai 1998

Juhendaja: .... „ ... “ mai 1998

Õppetooli juhataja: .... „ ... “ mai 1998

Tartu 1998

Sisukord

1 Sissejuhatus

Selleks, et arvuti mõistaks sisendit, mis on talle antud mõnes loomulikus keeles, tuleb talle kirjeldada selle keele struktuur. Vastav kirjeldusvahend peab olema küllalt võimas, et sellega oleks võimalik adekvaatselt kirjeldada niivõrd keerukat objekti nagu loomulik keel, ja teisest küljest küllalt lihtne selleks, et arvuti suudaks oma sisendit kiiresti analüüsida.

Kirjeldamise võimsuse ja analüüsi lihtsusega on erinevate keeli kirjeldavate formalismide juures tavaliselt nii, et mida suurem on üks, seda väiksem on teine. Näiteks võiks siinkohal tuua N. Chomsky pakutud genereerivate grammatikate klassifikatsiooni, kus defineeritud neli grammatikateklassi — kolmandat kuni nullindat tüüpi — on järjest suurema kirjeldamisvõimsusega ja raskemini analüüsitavad.

Chomsky hierarhia sees pakuvad erilist huvi teist tüüpi grammatikad — kontekstivabad grammatikad. Nende grammatikate heade omaduste hulka kuulub kiirete analüüsialgoritmide olemasolu, samuti on nad küllalt võimsad selleks, et nendega oleks mugav kirjeldada näiteks programmeerimiskeelte süntaksi. Kahjuks on kontekstivabade grammatikate võimsus siiski liiga väike selleks, et kirjeldada loomulikke keeli.

Chomsky hierarhia esimest tüüpi grammatikad — kontekstist sõltuvad grammatikad — on seevastu piisavalt võimsad kirjeldamaks loomulikke keeli. Nende puuduseks on analüüsialgoritmide keerulisus — kontekstist sõltuvate keelte analüüsiprobleem on PSPACE-täielik. Seetõttu ei tule ilmselt kõne alla võimalus, et loomulikus keeles suhtlevale arvutile oleks keel kirjeldatud kontekstist sõltuva grammatikaga.

Sel põhjusel tekib küsimus, kas ei leidu sobivat formalismi, mis oleks *pisut* võimsam kui kontekstivabad grammatikad, ja samal ajal omaks siiski nende häid omadusi. Need *nõrgalt kontekstist sõltuvad* formalismid peaksid tagama polünomiaalses ajas töötavad analüüsialgoritmid, piiratud „ristuvad sõltuvused“ lause erinevate osade vahel ja poollineaarsuse.

Kontekstivabasid grammatikaid võib üldistada sel viisil, et lisada grammatikale mingi formalism, mis ühel või teisel viisil piirab kontekstivaba tuletamist. Üks võimalus seda teha on näiteks moodustada kontekstivaba grammatika produktsioonidest järjendid ja lubada kasutada mingisse järjendisse kuuluvaid produktsioone ainult korraga.

Sulugrammatikad, mida vaadeldakse näiteks [GHR92]-s ja [Pit93]-s, on defineeritud sarnasel viisil. Enamgi veel — sulugrammatikatel on mitteterminalide hulk tükeldatud ja iga produktsioonide järjendi puhul leidub selline tükk, et nende produktsioonide vasakud pooled kuuluvad sellesse tükki ja iga sellesse tükki kuuluva mitteterminali puhul leidub täpselt üks produktsioon sellest järjendist, mille vasakuks pooleks on see mitteterminal. Iga mitteterminalide tükelduse tükki vaadeldakse kui *üldistatud sulgu*, mille komponentideks on tüki elemendid. Produktsioonide järjendite paremad pooled peavad moodustama *korrektsed sulgavaldised*. Samuti peab igal tuletuskäigu sammul olema tuletatud korrektne sulgavaldis. Selles sulgavaldises osad sümbolid „kuuluvad kokku“ (nagu avav ja sulgev sulg) ja produktsioone tohib rakendada vaid sümboleile, mis omavahel kokku kuuluvad. See tähendab seda, et paralleelselt võib produktsioone rakendada vaid sümboleile, mis tekkisid tuletuse samal sammul. Täpsed definitsioonid võib leida [GHR92]-st ja [Pit93]-st, samaväärsed definitsioonid on toodud käesoleva töö osas 2.2.

Sulugrammatikate klassil on defineeritav hierarhia, mille klassid on defineeritud produktsioonijärjendite maksimaalse komponentide arvu kaudu. Sulugrammatikate kirjeldusvõimsus kasvab koos grammatika *järgu* kasvamisega — esimest järku sulugrammatikad on täpselt kontekstivabad grammatikad; iga $l \in \mathbb{N}$ korral on keel $\{a_1^n a_2^n \cdots a_{2l}^n \mid n \geq 0\}$ defineeritav l-järku grammatikaga.

Samal ajal on sulugrammatikate jaoks olemas polünomiaalses ajas töötavad analüüsialgoritmid, nagu on näidatud [Pit93]-s.

1.1 Töö struktuur ja seos varasemate tulemustega

Käesolevas töös ei ole siiski vaadeldud sulugrammatikate kasutusvõimalusi kompuuterlingvistikas, vaid on üldistatud mõned kontekstivabu grammatikaid puudutavad tulemused. Sulugrammatikate põhiomaduseks on siin peetud seda, et produktsioone saab rakendada vaid sümboleile, mis üheaegselt tuletatud, ja grammatika definitsioonis sisaldab sellist omadust küllaltki ilmutatud kujul. Seejuures on loobutud nõudest, et tuletuskäigu igal sammul peaks olema tuletatud korrektne sulgavaldis, ja niimoodi saadud sulugrammatikatest pisut üldisem formalism, mida on nimetatud lihtsalt üldistatud kontekstivabadeks grammatikateks.

Osas 2.1 ongi defineeritud sulgude tähestik ja kirjeldatud, mida me mõistame „sõna“ all üle selle tähestiku ning millised on operatsioonid, mida nende sõnadega teha saab. Järgnevalt on defineeritud üldistatud kontekstivaba grammatika, keel mille ta defineerib ning peale seda üldistatud kontekstivaba grammatika derivatsioonipuu ([AU72]) mõistet.

Töö osas 3.1 on esitatud uus tulemus — üldistatud Ogdeni lemma (mille kontekstivabade grammatikate jaoks võib leida näiteks [AU72]-st). Töö osas 3.2 on Ogdeni lemmat kasutatud tõestamaks, et Lyndoni sõnade hulk üle kahe- (või enama-)tähelise tähestiku pole üldistatud kontekstivaba. Sellega on üldistatud [BB97]-s esitatud tulemus kontekstivabade keelte kohta. Abitulemusena kasutatav lemma 3.3 on sulugrammatikate kontekstis tõestatud [Pit93]-s ning kontekstivabal juhul raamatus [AU72].

Töö neljas osa sisaldab uue tulemuse — defineeritakse automaatide klass — puuautomaadid —, mis aktsepteerivad parajasti üldistatud kontekstivabu keeli. Seejuures on automaadi definitsioon esitatud sarnaselt pinuautomaadi definitsioonile raamatus [Har78]. Lisaks sellele näidatakse puuautomaatidele selline aktsepteerimisviis, et puuautomaadid aktsepteeriksid sel viisil parajasti sulugrammatikate genereeritud keeli. Saadud tulemuste rakendusena tõestatakse, et üldistatud kontekstivaba keele ja regulaarse keele ühisosa on üldistatud kontekstivaba ning sulugrammatikaga genereeritava keele ja regulaarse keele ühisosa on genereeritav sulugrammatikaga.

1.2 Töös kasutatavatest tähistustest

Me kasutame sümbolit $\mathbb{N}$ hulga $\{1, 2, 3, \dots\}$ tähistamiseks. Naturaalarve koos 0-ga tähistame $\mathbb{N}_0$-ga. Hulkade sisalduvust tähistame sümboliga $\subseteq$ ja ranget sisalduvust sümboliga $\subset$.

Kõik puud, mida me vaatleme, omavad juurt. Peale selle on puu tipu vahetud järglased järjestatud — tipul on vasakpoolseim, vasakult teine, jne. järglane. Samamoodi on järjestatud puud, mis moodustavad metsa — metsas on vasakpoolseim, vasakult teine, jne. puu. Puu kõrguseks nimetame maksimaalset kaugust selle puu juurest mõne selle puu leheni. Metsa kõrguseks nimetame selle metsa kõrgeima puu kõrgust.

„Puu läbimine eeljärjestuses“ tähendab puu tippude panekut teatavasse järjekorda (ja nende tippude vaatamist selles järjekorras):

1. Järjekorras esimene tipp on puu juurtipp.

2. Juurtipule järgnevad tema vasakpoolseimast vahetust järglasest algava alampuu tipud, seejärel vasakult teisest vahetust järglasest algava alampuu tipud, jne., viimasena tulevad juurtipu parempoolseimast vahetust järglasest algava alampuu tipud.

3. Juurtipu mingist vahetust järglasest algava alampuu tipud on selle alampuu eeljärjestuses läbimise järjekorras.

2 Definitsioonid

2.1 Sulutähestikud ja üldistatud kontekstivabad grammatikad

Definitsioon 2.1. Olgu $n \in \mathbb{N}_0$ ja tähistagu $[1..n]$ hulka $\{k \in \mathbb{N} \,|\, 1 \leq k \leq n\}$. Kui Σ on lõplik hulk — tähestik —, siis hulka

$$\Sigma^* := \{w : [1..n] \longrightarrow \Sigma \,|\, n \in \mathbb{N}_0\}$$

nimetatakse *sõnade* hulgaks üle tähestiku Σ. Sõna $w : [1..n] \longrightarrow \Sigma$ *pikkus* on defineeritud võrdusega $|w| := n$. Sõnade w_1 ja w_2 *konkatenatsiooniks* nimetatakse sõna $w_1w_2 : [1..|w_1| + |w_2|] \longrightarrow \Sigma$, mis on antud reegliga

$$(w_1w_2)(n) := \begin{cases} w_1(n), & \text{kui } n \leq |w_1| \\ w_2(n - |w_1|), & \text{kui } n > |w_1|. \end{cases}$$

$(\Sigma^*, \cdot)$ on vaba monoid vaba baasiga Σ.

Definitsioon 2.2. Olgu N lõplik hulk ja $g : N \longrightarrow \mathbb{N}$ kujutis. Hulka

$$[N, g] := \{(x, n) \in N \times \mathbb{N} \,|\, 1 \leq n \leq g(x)\}$$

nimetame *sulutähestikuks* märgenditega hulgast N ja aarsustega g.

Sulutähestiku elementi (x, n) tähistame edaspidi ka x_n-ga. Kui $x \in N$ aarsus on 1, siis me samastame elemendid x ja $(x, 1)$.

Kui $X \subseteq \mathbb{N}$ ja $k \in \mathbb{N}$, siis olgu

$$k + X := \{k + n \,|\, n \in X\}.$$

Olgu $X \subset \mathbb{N}$ lõplik ja $n \in \mathbb{N}$. Olgu $\mathcal{O}rd_X(n) := |X \cap [1..n]|$. Kui $n \in [1..|X|]$, siis olgu $\mathcal{E}l_X(n) := \min\{k \in \mathbb{N} \,|\, |X \cap [1..k]| = n\}$ (S.t. hulga X suuruselt n-s element, kui vähim element on esimene).

Definitsioon 2.3. Olgu $[N, g]$ sulutähestik. Olgu $n \in \mathbb{N}_0$, olgu $W = \{W_1, \ldots, W_k\}$ hulga $[1..n]$ tükeldus, olgu $l_w : W \longrightarrow N$ kujutus, kusjuures kehtigu iga $i \in [1..k]$ korral $|l_w(W_i)| = g(W_i)$. Olgu $w \in [N, g]^*$ defineeritud järgmiselt:

- $|w| := n$;
- iga $i \in [1..n]$ korral $w(i) := (l_w(W_j), \mathcal{E}l_{W_j}(i))$, kus j on selline, et $i \in W_j$.

Seega on w-d ja W-d teades leitav l_w. Paari $\langle w; W\rangle$ nimetame *sulustuseks* üle $[N, g]$. Kõigi sulustuste hulka üle sulutähestiku $[N, g]$ tähistame N_g^*.

Definitsioon 2.4. Sulustuste $\langle w_1; W_1\rangle$ ja $\langle w_2; W_2\rangle$ *konkatenatsiooniks* nimetame sulustust $\langle w; W\rangle$, kus $w = w_1w_2$ ja $W = W_1 \cup \{|w_1| + X \mid X \in W_2\}$; on ilmne, et W on hulga $[1..|w_1| + |w_2|]$ sobiv tükeldus.

Definitsioon 2.5. *Sulustatud järjendiks* üle N_g^* nimetame järjendit $(w_1, \ldots, w_n) \in ([N, g]^*)^n$ koos hulga $[1..\sum_{i=1}^n |w_i|]$ tükeldusega W nii, et $\langle w_1 \cdots w_n; W\rangle \in N_g^*$. Tähistame $\langle (w_1, \ldots, w_n); W\rangle$. Kõigi n-elemendiliste sulustatud järjendite hulka üle N_g^* tähistame $n{-}N_g^*$. Me samastame hulgad N_g^* ja $1{-}N_g^*$.

Definitsioon 2.6. Olgu Σ tähestik, olgu $w, u \in \Sigma^*$, $n \in [1..|w|]$. *Sümboli $w(n)$ sõnas w kohal n asenduseks sõnaga u* nimetame sõna $w[n \longleftarrow u] : [1..|w| + |u| - 1] \longrightarrow \Sigma$, mis on defineeritud kui

$$(w[n \longleftarrow u])(i) := \begin{cases} w(i), & \text{kui } i < n \\ u(i - n + 1), & \text{kui } n \leq i < n + |u| \\ w(i - |u| + 1), & \text{kui } n \geq n + |u|. \end{cases}$$

Definitsioon 2.7. Olgu T tähestik, olgu $w_1, \ldots, w_k, u \in \Sigma^*$, olgu $n \in [1..|w_1 \cdots w_k|]$. Järjendi $(w_1, \ldots, w_k)$ *n-nda sümboli asenduseks sõnaga u* nimetame järjendit $(w_1, \ldots, w_n)[n \longleftarrow u] := (w_1', \ldots, w_k')$, kus iga $j \in [1..k]$ korral

$$w_j' = \begin{cases} w_j[(n - \sum_{i=1}^{j-1} |w_i|) \longleftarrow u], & \text{kui } \sum_{i=1}^{j-1} |w_i| < n \leq \sum_{i=1}^{j} |w_i| \\ w_j, & \text{muidu.} \end{cases}$$

Definitsioon 2.8. Olgu $w_1, \ldots, w_k, u_1, \ldots, u_m \in \Sigma^*$ ja $X \subseteq [1..|w_1 \cdots w_k|]$, $|X| = m$, siis

$$(w_1, \ldots, w_k)[X \longleftarrow u_1, \ldots, u_m] := (w_1, \ldots, w_k),$$

juhul $m = 0$ ning edasi rekursiivselt

$$\begin{gathered}(w_1, \ldots, w_k)[X \longleftarrow u_1, \ldots, u_m] := \\ ((w_1, \ldots, w_k)[(\min X) \longleftarrow u_1])[\big((|u_1| - 1) + (X \setminus \{\min X\})\big) \longleftarrow u_2, \ldots, u_m].\end{gathered}$$

Definitsioon 2.9. Olgu $\langle (w_1, \ldots, w_k); W\rangle$ ja $\mathcal{U} = \langle (u_1, \ldots, u_n); U\rangle$ sulustatud järjendid üle N_g^*, olgu $X \in W$, $|X| = n$. *Sümboli $l_{w_1 \cdots w_k}(X)$ järjendis*

$\langle (w_1,\ldots,w_k);W\rangle$ *kohal* X *asenduseks järjendiga* $\mathcal{U}$ nimetame sulustatud järjendit $\langle (w_1,\ldots,w_k);W\rangle[X \longleftarrow \mathcal{U}] := \langle (v_1,\ldots,v_k);V\rangle$, kus

$$(v_1,\ldots,v_k) = (w_1,\ldots,w_k)[X \longleftarrow u_1,\ldots,u_n]$$

ja

$$V = \{\{y + \sum_{i=1}^{\mathcal{O}rd_X(y)} (|u_i|-1) \mid y \in Y\} \mid Y \in W \backslash \{X\}\} \cup \\ \{\{y + \mathcal{E}l_X(\mathcal{O}rd_Z(y)) - \mathcal{O}rd_Z(y) \mid y \in Y\} \mid Y \in U\},$$

kus $Z = \{1 + \sum_{i=1}^{l} |u_i| \mid 0 \le l \le n-1\}$.

Selgituseks — me asendame järjendis $(w_1,\ldots,w_k)$ positsioonidel hulgast X seisvad tähed vastavalt $u_1,\ldots,u_n$-ga ja loeme tulemuses kokkukuuluvaks need positsioonid, mis kuulusid kokku $(w_1,\ldots,w_k)$-s, ning need positsioonid, mis kuulusid kokku $(u_1,\ldots,u_n)$-s.

Definitsioon 2.10. *Üldistatud kontekstivabaks grammatikaks* nimetame nelikut $G = ([V,g],\Sigma,P,s)$, kus

- $[V,g]$ on sulutähestik;
- $\Sigma \subset V$, on terminaalne tähestik, seejuures $\forall t \in \Sigma: \ g(t) = 1$;
- P on produktsioonide — paaride $x \to y$ — hulk, kus $x \in V\backslash\Sigma$ ja $y \in g(x){-}V_g^*$;
- $s \in V\backslash\Sigma$ on aksioom, seejuures $g(s) = 1$.

Kui $G = ([V,g],\Sigma,P,s)$, siis tähistame $N := V\backslash\Sigma$ grammatika G mitteterminalide hulka. Edaspidi, kui me räägime grammatikast, mille terve tähestik on $[V,g]$ ja terminaalne tähestik Σ, siis kasutame N-i mitteterminalide hulga tähistamiseks, seda iga kord ilmutatult defineerimata.

Definitsioon 2.11. Olgu $G = ([V,g],\Sigma,P,s)$ üldistatud kontekstivaba grammatika. G *järk* on defineeritud kui arv

$$\mathbf{grad}\, G := \max\{g(x) \mid x \in V\}.$$

Definitsioon 2.12. Olgu $G = ([V,g],\Sigma,P,s)$ üldistatud kontekstivaba grammatika, $k \in \mathbb{N}$ ja $\langle (u_1,\ldots,u_k);U\rangle, \langle (v_1,\ldots,v_k);V\rangle \in k - V_g^*$. Järjendist $\langle (u_1,\ldots,u_k);U\rangle$ on *vahetult tuletatav* järjend $\langle (v_1,\ldots,v_k);V\rangle$ (tähistame $\langle (u_1,\ldots,u_k);U\rangle \Rightarrow \langle (v_1,\ldots,v_k);V\rangle$), kui leiduvad selline $X \in U$ ja $x \to y \in P$, et $l_{(u_1,\ldots,u_k)}(X) = x$ ja $\langle (v_1,\ldots,v_k);V\rangle = \langle (u_1,\ldots,u_k);U\rangle[X \longleftarrow y]$.

Definitsioon 2.13. Olgu $G = ([V,g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika, $n \in \mathbb{N}_0$, $k \in \mathbb{N}$ ja iga $i \in [0..n]$ korral $\langle u_i; U_i \rangle \in k - V_g^*$. Jada $\langle u_0; U_0 \rangle, \langle u_1; U_1 \rangle, \ldots, \langle u_n; U_n \rangle$ nimetame *derivatsiooniks grammatikas* G, kui iga $i \in [1..n]$ korral $\langle u_{i-1}; U_{i-1} \rangle \Rightarrow \langle u_i; U_i \rangle$. Arvu n nimetame *derivatsiooni pikkuseks.* Olgu iga $i \in [0..n]$ korral $u_i = (u_{i,1}, \ldots, u_{i,k})$. Derivatsiooni nimetame *vasakpoolseks derivatsiooniks*, kui iga $i \in [0..n-1]$ korral vastav $X_i \in U_i$, mille kohal olev sümbol asendatakse, et saada $\langle u_i; U_i \rangle$-st $\langle u_{i+1}; U_{i+1} \rangle$, sisaldab elementi x_i, mille korral

$$x_i = \min\{r \in [1..|u_{i,1} \cdots u_{i,k}|] \mid (u_{i,1} \cdots u_{i,k})(r) \in [N,g]\}.$$

Olgu $\stackrel{*}{\Rightarrow}$ relatsiooni $\Rightarrow$ refleksiivne transitiivne sulund.

Definitsioon 2.14. *Üldistatud kontekstivaba grammatika* $G = ([V,g], \Sigma, P, s)$ *defineerib keele*

$$\mathcal{L}(G) := \{w \in \Sigma^* \mid \langle s; \{\{1\}\} \rangle \stackrel{*}{\Rightarrow} \langle w; \{\{1\}, \{2\}, \ldots, \{|w|\}\} \rangle\}.$$

Keelt $L \subseteq \Sigma^*$, mille korral leidub selline üldistatud kontekstivaba grammatika G, et $\mathcal{L}(G) = L$, nimetame *üldistatud kontekstivabaks keeleks.*

Definitsioon 2.15. Olgu $G = ([V,g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika ja $A \in N$. Grammatika G *derivatsioonimetsaks juurega* A nimetatakse märgendatud tippudega metsa (märgenditega hulgast $[V,g] \cup [P,g]$, kus kui $x \to y \in P$, siis $g(x \to y) := g(x)$) koos tükeldusega tippudel, mis defineeritakse induktiivselt:

- $g(A)$ puud, millest igaühel on üksainus tipp märgendiga vasakult paremale vastavalt $A_1, \ldots, A_{g(a)}$, on derivatsioonimets juurega A. (klassijaotus koosneb ühest klassist, millesse need tipud kõik kuuluvad).

- Olgu T derivatsioonimets juurega A ja $N_1, N_2, \ldots, N_k$ tema tipud, mis moodustavad T tippude klassijaotuse ühe klassi ja mille märgendid on vastavalt $(x,1), (x,2), \ldots, (x,k)$, kus $x \in V \backslash T$ ja $g(x) = k$. Olgu $p = x \to \langle (y_1, \ldots, y_k); Y \rangle \in P$, kus $y_i = a_{i1} \ldots a_{il_i}$ ja $a_{ij} \in [V,g]$. Siis on derivatsioonimets juurega A ka T', mille saame puust T, $i = 1, \ldots, k$ korral märgendades tipu N_i ümber (p,i)-ga ja lisades tipule N_i l_i järglast (vasakult paremale lugedes) $N_{i1}, \ldots, N_{il_i}$.
  Lisatud tipu N_{ij}, kus $1 \leq i \leq k$ ja $1 \leq j \leq l_i$, märgendiks loeme sümboli a_{ij}. Tippe $N_{i_1 j_1}$ ja $N_{i_2 j_2}$ loeme kuuluvaks T' tippude tükelduse samasse klassi parajasti siis, kui arvud $j_1 + \sum_{k=1}^{i_1 - 1} l_k$ ja $j_2 + \sum_{k=1}^{i_2 - 1} l_k$ kuuluvad tükelduse Y samasse klassi. (Klassidesse, kuhu kuuluvad lisatud tipud, ei kuulu T tippe.)

Grammatika G *derivatsioonipuuks* nimetame G derivatsioonimetsa juurega s.

Kui N on metsa T tippude tükelduse klass ja $|N| = n$, siis tähistame N-i kuuluvaid tippe $N^{(1)}, \dots, N^{(n)}$, kus indeksid on metsa T eeljärjestuses läbimise järjekorras. Kirjutame $N = (N^{(1)}, \dots, N^{(n)})$.

Definitsioon 2.16. Grammatika $G = ([V, g], \Sigma, P, s)$ analüüsipuuks nimetatakse derivatsioonipuud, mille tippude märgendid kuuluvad hulka $\Sigma \cup [P, g]$.

Definitsioon 2.17. Märgendatud puu krooniks nimetame sõna, mille saame, kirjutades kõrvuti puu lehtede märgendid järjekorras vasakult paremale.

Kui T on üldistatud kontekstivaba grammatika derivatsioonimets, siis võime rääkida ka T tippude tükelduse klasside eellastest ja järglastest. On ilmne, et kui $M = (M^{(1)}, M^{(2)}, \dots, M^{(m)})$ja $N = (N^{(1)}, N^{(2)}, \dots, N^{(n)})$, siis mõni tipp N-st on mõne tipu M-st vahetu järglane parajasti siis kui iga tipp N-st on mõne tipu M-st vahetu järglane. Sel juhul ütleme, et N on M-i vahetu järglane ja M on N-i vahetu eellane. Kui X on T tipp, siis nimetame X-i krooniks T selle alampuu krooni, mille juureks on X. Kui $N = (N^{(1)}, N^{(2)}, \dots, N^{(n)})$ on T tippude tükelduse klass, siis nimetame N krooniks järjendit $(w^{(1)}, w^{(2)}, \dots, w^{(n)})$, kus $w^{(i)}$ on $N^{(i)}$ kroon.

2.2 Üldistatud sulud ja sulugrammatikad

Definitsioon 2.18. Olgu $[N, g]$ sulutähestik. Monoidi $[N, g]$ kongruentsi $\equiv_N$, mille genereerivad seosed

$$\{\prod_{i=1}^{g(k)} k_i \equiv_N \epsilon \,|\, k \in N\},$$

nimetame sulutähestiku $[N, g]$ *sulukongruentsiks.*

Definitsioon 2.19. *Sulgude keeleks* üle sulutähestiku $[N, g]$ nimetatakse keelt

$$\mathrm{D}(N, g) := \{w \in [N, g]^* \,|\, w \equiv_N \epsilon\}.$$

On ilmne, et kehtib järgmine

Lause 2.1. *Olgu $[N, g]$ sulutähestik. $\mathrm{D}(N, g)$ on $[N, g]^*$ vähim selline alamhulk X, mille korral on täidetud järgmised tingimused:*

1. $\epsilon \in X$;

2. *kui* $u, v \in X$*, siis* $uv \in X$*;*

3. *kui* $k \in N$ *ja* $u_1, \ldots, u_{g(k)-1} \in X$*, siis* $k_1 u_1 k_2 u_2 \cdots k_{g(k)-1} u_{g(k)-1} k_{g(k)} \in X$.

Definitsioon 2.20. Olgu $[N, g]$ on sulutähestik. *Sulujärjendiks* nimetame järjendit $(w_1, \ldots, w_n) \in ([N, g]^*)^n$, mille korral $w_1 \cdots w_n \in \mathrm{D}(N, g)$. Kõigi n-elemendiliste sulujärjendite hulka üle tähestiku $[N, g]$ tähistame $n-\mathrm{D}(N, g)$. Me samastame hulgad $\mathrm{D}(N, g)$ ja $1-\mathrm{D}(N, g)$.

Definitsioon 2.21. Olgu $[N, g]$ sulutähestik, olgu $w \in \mathrm{D}(N, g)$. Sõna w *genereeritud tükelduseks* $\mathcal{S}(w)$ nimetame hulga $[1..|w|]$ tükeldust, mille defineerime induktiivselt:

1. $\mathcal{S}(\epsilon) := \emptyset$.

2. Kui $u, v \in \mathrm{D}(N, g)$, siis

$$\mathcal{S}(uv) := \mathcal{S}(u) \cup (\{|u| + X \mid X \in \mathcal{S}(v)\}).$$

3. Kui $k \in N$ ja $u_1, \ldots, u_{g(k)-1} \in \mathrm{D}(N, g)$, siis

$$\begin{aligned}&\mathcal{S}(k_1 u_1 k_2 u_2 \cdots k_{g(k)-1} u_{g(k)-1} k_{g(k)}) := \\ &\{\{i + \sum_{j=0}^{i-1} |u_j| \mid i \in [1..g(k)]\}\} \cup \bigcup_{i=1}^{k-1} \{(i + \sum_{j=1}^{i-1} |u_j|) + X \mid X \in \mathcal{S}(u_i)\}.\end{aligned}$$

Sulujärjendi $(w_1, \ldots, w_n) \in n\text{–}\mathrm{D}(N, g)$ *genereeritud tükeldus* $\mathcal{S}((w_1, \ldots, w_n))$ on defineeritud kui

$$\mathcal{S}((w_1, \ldots, w_n)) := \mathcal{S}(w_1 \cdots w_n).$$

Definitsioon 2.22. Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika. Grammatikat G nimetame *sulugrammatikaks*, kui

$$\begin{aligned}&\forall p = x \to \langle (y_1, \ldots, y_{g(x)}); Y \rangle \in P : \\ &\qquad (y_1, \ldots, y_{g(x)}) \in g(x)-\mathrm{D}(V, g) \wedge \mathcal{S}((y_1, \ldots, y_{g(x)})) = Y.\end{aligned}$$

Definitsioon 2.23. Keelt $L \subseteq \Sigma^*$ nimetame *sulukeeleks*, kui leidub selline sulugrammatika G, et $\mathcal{L}(G) = L$.

3 Ogdeni lemma

3.1 Ogdeni lemma üldistatud kontekstivabade grammatikate jaoks

Lemma 3.1. *Olgu* $f_1, \ldots, f_n : [1..m] \longrightarrow [1..m]$ *suvalised kujutused. Olgu* $I(f_t)$ *kujutuse* f_t *kõigi püsipunktide hulk ning olgu* $1 \leq i \leq j \leq n$ *korral* $f_{i,j} = f_j \circ f_{j-1} \circ \cdots \circ f_i$. *Olgu antud suvalised* $X \subseteq \bigcap_{t=1}^{n} I(f_t) \subseteq [1..m]$ *ja* $k \in \mathbb{N}$. *Leidub arv* $\mathcal{N}(m, |X|, k)$ *nii, et kui* $n \geq \mathcal{N}(m, |X|, k)$, *siis leiduvad arvud* $0 \leq i_0 < i_1 < \ldots < i_k \leq n$ *nii, et* $f_{i_0+1,i_1} = f_{i_1+1,i_2} = \cdots = f_{i_{k-1}+1,i_k}$ *ja* $f_{i_0+1,i_1} = f^2_{i_0+1,i_1}$.

Tõestus. Kasutame induktsiooni $m - |X|$ järgi.
Baas. $m - |X| = 0$ ehk $|X| = m$. Sel juhul on $f_1 = \cdots = f_n = \mathbf{1}_{[1..m]}$ ja seega kui $n \geq k$, siis võime võtta $i_j = j$ $(j \in [0..k]$.
Samm. Olgu $m - |X| > 0$, eeldame, et leidub arv $\mathcal{N}(m, |X| + 1, k)$. Kui

$$n \geq ((k-1)|X|^{m-|X|} + 1)(m - |X|)\mathcal{N}(m, |X| + 1, k)) + 1,$$

siis leiab aset (vähemalt) üks järgmistest võimalustest:

1. Leidub arv l nii, et $l \geq (k-1)|X|^{m-|X|} + 1$ ja arvud $0 = j_0 \leq j_1 < j_2 < \cdots < j_l \leq n$ nii, et $\mathrm{Im} f_{1,j_1} = \mathrm{Im} f_{j_1+1,j_2} = \cdots = \mathrm{Im} f_{j_{l-1}+1,j_l} = X$.

2. Leiduvad arvud $1 \leq p \leq q \leq n$ nii, et $q - p \geq (m - |X|)\mathcal{N}(m, |X| + 1, k)$ ja $X \subset \mathrm{Im} f_{p,q}$.

Tõepoolest, oletame, et esimene võimalustest ei leia aset. Sel juhul olgu $l \leq (k-1)|X|^{m-|X|} + 1$ maksimaalne arv, mille korral leiduvad arvud $0 = j_0 \leq j_1 < \cdots < j_l$, et $\mathrm{Im} f_{j_0+1,j_1} = \mathrm{Im} f_{j_1+1,j_2} = \cdots = \mathrm{Im} f_{j_{l-1}+1,j_l} = X$. Arvud $j_0, \ldots, j_l$ on võimalik valida nii, et iga $s \in [1..l]$ korral kui $j_s - j_{s-1} \geq 2$, siis $\mathrm{Im} f_{j_{s-1}+1,j_s-1} \neq X$. Vaatame $(l+1)$ vahet $j_1 - j_0, j_2 - j_1, \ldots, j_l - j_{l-1}, n - j_l$. Kuna nende vahede summa on n ja neid vahesid pole rohkem kui $((k-1)|X|^{m-|X|} + 1)$ tükki, siis peab leiduma selline $s \in [1..l+1]$, et $j_s - j_{s-1} \geq (m - |X|)\mathcal{N}(m, |X| + 1, k) + 1$, kus $j_{l+1} := n$. Seega $(j_s - 1) - j_{s-1} \geq (m - |X|)\mathcal{N}(m, |X| + 1, k)$ ja $\mathrm{Im} f_{j_{s-1},j_s-1} \neq X$.

Esimesel juhul leiduvad arvud $1 \leq r_1 < r_2 < \cdots < r_k \leq l$ nii, et $f_{j_{r_1-1}+1,j_{r_1}} = f_{j_{r_2-1}+1,j_{r_2}} = \cdots = f_{j_{r_k-1}+1,j_{r_k}}$, sest iga funktsioon f, mis rahuldab lemmas funktsioonidele f_i esitatud tingimusi, ja mille korral $\mathrm{Im} f = X$, on määratud oma väärtustega hulgal $[1..m] \backslash X$ ja need väärtused kuuluvad kõik hulka X, seega on meil (erinevaid) selliseid funktsioone f kokku $|X|^{m-|X|}$ tükki. Kuna iga $0 \leq p < q \leq l$ korral $f_{j_p+1,j_q} = f_{j_p+1,j_{p+1}}$, siis võime võtta $i_p = j_{r_p-1} + 1$ kui $p \in [0..k-1]$ ja $i_k = j_{r_k}$.

Teisel juhul leidub arv r' nii, et $f_{p,q}(r') \notin X$. Kui $q' \in [p..q]$, siis $f_{p,q'}(r') \in [1..m]\backslash X$. Et oleme valinud $|[p..q]| \geq (m-|X|)\mathcal{N}(m,|X|+1,k)+1$, siis leidub selline $r \in [1..m]\backslash X$ ja arvud $p \leq j_1 < j_2 < \cdots < j_l \leq q$, kus $l \geq \mathcal{N}(m,|X|+1,k)+1$, et $\forall s \in [1..l] : f_{p,j_s}(r') = r$.

Induktsiooni eelduse kasutamiseks defineerime nüüd funktsioonid $f'_1, \ldots, f'_{l-1}$ järgmiselt: iga $s \in [1..l-1]$ korral olgu $f'_s = f_{j_s+1,j_{s+1}}$. Sel juhul iga $s \in [1..l-1]$ korral $f'_s(r) = r$ ja seega on f'_s püsipunktid kõik hulga X elemendid ja peale nende veel r. Kuna $l-1 \geq \mathcal{N}(m,|X\dot{\cup}\{r\}|,k)$, siis induktsiooni eelduse järgi leiduvad arvud $0 \leq i'_0 < i'_1 < \ldots < i'_k \leq l-1$ nii, et $f'_{i'_0+1,i'_1} = f'_{i'_1+1,i'_2} = \cdots = f'_{i'_{k-1}+1,i'_k}$ ja $f'_{i'_0+1,i'_1} = f'^2_{i'_0+1,i'_1}$. Kuna iga $s \in [1..k]$ korral $f'_{i'_{s-1}+1,i'_s} = f_{j_{i'_{s-1}+1}+1,j_{i'_s+1}}$, siis võttes iga $s \in [0..k]$ korral $i_s = j_{i'_s+1}$, rahuldavad arvud $i_0, \ldots, i_k$ lemma väidet.

Defineerime $\mathcal{N}(m,|X|,k)$-i kui vähima arvu, mis rahuldab selle lemma tingimusi. Me saame seda teha, kuna oleme näidanud, et neid tingimusi rahuldab arv $((k-1)|X|^{m-|X|}+1)(m-|X|)\mathcal{N}(m,|X|+1,k))+1$. $\square$

Selleks, et tõestada Ogdeni lemmat üldistatud kontekstivabade grammatikate jaoks, läheb meil veel tarvis järgmisi definitsioone:

Definitsioon 3.1. Olgu $G = ([V,g],\Sigma,P,s)$ üldistatud kontekstivaba grammatika, olgu $z \in \mathcal{L}(G)$, olgu sõnas z mõned positsioonid märgistatud, olgu T sõna z analüüsipuu. Olgu N puu T tippude tükelduse klass. N-i nimetame *hargnemisklassiks*, kui tal on vähemalt kaks vahetut järglast, mille kroonid sisaldavad märgistatud positsioone.

Definitsioon 3.2. Olgu G, z, T defineeritud samuti kui eelmises definitsioonis, olgu sõnas z mõned positsioonid märgistatud. Olgu $N_i = (N_i^{(1)}, \ldots, N_i^{(e_i)})$ hargnemisklass, olgu fikseeritud üks klassi N_i vahetutest järglastest — klass $N_{i+1} = (N_{i+1}^{(1)}, \ldots, N_{i+1}^{(e_{i+1})})$, sisaldagu N_{i+1} kroon märgistatud positsioone. Olgu $M = (M^{(1)}, \ldots, M^{(f)})$ klassi N_i mõni teine vahetu järglane, mille kroonis on märgistatud positsioone. Olgu tähestikus $\{a, \bar{a}, b, \bar{b}, \#\}$ antud sõna h, milles on täpselt e_{i+1} tähte hulgast $\{a, \bar{a}\}$, täpselt f tähte hulgast $\{b, \bar{b}\}$ ja täpselt $(e_i - 1)$ tähte #. Ütleme, et klassi N_i *hargnemisviis* on h, kui

- Suvaliste $j \in [1..e_{i+1}]$ ja $k \in [1..f]$ korral on sõnas h j-s a või $\bar{a}$ eespool k-ndat b-d või $\bar{b}$-d parajasti siis, kui analüüsipuus T-s on puu läbimisel eeljärjestuses $N_{i+1}^{(j)}$ enne $M^{(k)}$-d.

- Suvaliste $j \in [1..e_{i+1}]$ $[\![j \in [1..f]]\!]$ ja $k \in [0..e_i - 1]$ korral on sõnas h j-nda a või $\bar{a}$ $[\![b$ või $\bar{b}]\!]$ ees täpselt k tähte # parajasti siis, kui tipp $N_{i+1}^{(j)}$ $[\![M^{(j)}]\!]$ on tipu $N_i^{(k+1)}$ vahetu järglane.

- Suvalise $j \in [1..e_{i+1}]$ **[**$j \in [1..f]$**]** korral on sõnas h j-s a või $\bar{a}$ **[**b või $\bar{b}$**]** võrdne $\bar{a}$-ga **[**$\bar{b}$-ga**]** parajasti siis, kui tipu $n_{i+1}^{(j)}$ **[**$m^{(j)}$**]** kroon sisaldab märgistatud positsioone.

Tähistagu

$$\mathcal{H}(\bar{g}) := |\{h \in \{a, \bar{a}, b, \bar{b}, \#\} \mid \text{leidub üldistatud kontekstivaba grammatika } G\text{, selle analüüsipuu } T \text{ ja selle hargnemisklass } N_i \text{ nii, et } \mathbf{grad}\, G \leq \bar{g} \text{ ja } h \text{ on } N_i \text{ hargnemisviis}\}$$

Märkus 3.1. Kui hargnemisklassil n_i on kolm või enam vahetut järglast, mille kroonides on märgistatud positsioone, siis võib n_i-l olla mitu hargnemisviisi.

Teoreem 3.2 (Üldistatud Ogdeni lemma). *Iga üldistatud kontekstivaba grammatika $G = ([V, g], \Sigma, P, s)$ jaoks leidub $\mathbf{k} \in \mathbb{N}$ nii, et kui $z \in \mathcal{L}(G)$ ja $|z| \geq \mathbf{k}$ ja z-s on märgistatud $\mathbf{k}$ või rohkem positsiooni, siis leidub $\bar{g} \in \mathbb{N}$, $1 \leq \bar{g} \leq \mathbf{grad}\, G$ ja $k_1, \ldots, k_{\bar{g}} \in \mathbb{N}_0$, $k_1 + \cdots + k_{\bar{g}} = \bar{g}$ nii, et*

1. $$z = t_0 \prod_{j=1}^{\bar{g}} (v_{j,0} \prod_{i=1}^{k_j} (w_{l_j+i} v_{j,i}) t_j), \tag{3.1}$$

   kus $l_j := \sum_{i=1}^{j-1} k_j$ ja $t_j, v_{j,i}, w_j \in \Sigma^*$, seejuures, kui $k_j \neq 0$, siis $l_j < j \leq l_j + k_j$ $(= l_{j+1})$;

2. *leidub selline $j \in [1..\bar{g}]$, et w_j sisaldab märgistatud positsioone;*

3. *leiduvad sellised $j \in [0..\bar{g}]$, $i_1 \in [1..\bar{g}]$ ja $i_2 \in [1..k_{i_1}]$, et sõna z dekompositsioonis (3.1) asub v_{i_1,i_2} sõnade w_j ja w_{j+1} vahel (või enne w_1-e, kui $j = 0$; või peale $w_{\bar{g}}$-d, kui $j = \bar{g}$) ja v_{i_1,i_2} ja t_j mõlemad sisaldavad märgistatud positsioone;*

4. *iga $j \in [1..\bar{g}]$ korral sisaldab w_j märgistatud positsioone parajasti siis, kui $v_{j,0} \prod_{i=1}^{k_j} (w_{l_j+i} v_{j,i})$ sisaldab märgistatud positsioone; kui w_j sisaldab märgistatud positsioone, siis*

   $$t_{l_j} \prod_{r=l_j+1}^{\bar{l}_{j+1}} (v_{r,0} \prod_{i=1}^{k_r} (w_{l_r+i} v_{r,i}) t_r)$$

   sisaldab märgistatud positsioone;

5. *Leidub selline mitteterminal A, et $g(A) = \bar{g}$ ja*

$$\begin{aligned} s &\overset{*}{\Rightarrow} t_0 A_1 t_1 A_2 \cdots t_{\bar{g}-1} A_{\bar{g}} t_{\bar{g}} \\ (A_j)_{j=1}^{\bar{g}} &\overset{*}{\Rightarrow} (v_{j,0} A_{l_j+1} v_{j,1} A_{l_j+2} \cdots v_{j,k_j-1} A_{l_j+k+j} v_{j,k_j})_{j=1}^{\bar{g}} \\ (A_j)_{j=1}^{\bar{g}} &\overset{*}{\Rightarrow} (w_j)_{j=1}^{\bar{g}}. \end{aligned}$$

Tõestus. Olgu $l = \max\{|y| \mid x \to \langle y; Y\rangle \in P\}$ (produktsioonide parema poole maksimaalne pikkus), olgu

$$\mathbf{k} = l^{\mathcal{N}(\mathbf{grad}\, G,0,2)\mathcal{H}(\mathbf{grad}\, G)|N|+1}.$$

Olgu $z \in \mathcal{L}(G)$, olgu $|z| \geq \mathbf{k}$, märgistame z-s $\mathbf{k}$ või enam positsiooni. Olgu T sõna z mingi analüüsipuu.

Konstrueerime järjendi $N_1, N_2, \ldots$, mille komponentideks on T tippude tükelduse klassid, järgmiselt:

1. N_1 on klass, mille ainsaks elemendiks on T juur;

2. kui klassil N_i on täpselt üks vahetu järglane, mille kroon sisaldab märgistatud positsioone, siis võtame selle järglase N_{i+1}-ks;

3. kui klass n_i on hargnemisklass, siis võtame n_{i+1}-ks tema vahetu järglase, mille kroonis on enim märgistatud positsioone;

4. kui klass n_i sisaldab täpselt ühe tipu ja see tipp on T leht, siis on järjend konstrueeritud.

Olgu r äsjakonstrueeritud järjendi komponentide arv.

Väide 3.2.1. *Kui klasside $N_1, \ldots, N_i$ hulgas on s hargnemisklassi, siis klassi N_{i+1} kroonis on vähemalt $\frac{\mathbf{k}}{l^s}$ märgistatud positsiooni.*

Väite võime tõestada induktsiooniga i järgi. Tõepoolest, kui $i = 0$, siis $s = 0$ ja N_i on klass, mis koosneb T juurtipust ja seega on tema kroonis vähemalt $\mathbf{k}$ märgistatud positsiooni. Kehtigu väide $(i-1)$ puhul. Kui N_i ei ole hargnemisklass, siis on induktsiooni oletuse järgi tema kroonis vähemalt $\frac{\mathbf{k}}{l^s}$ märgistatud positsiooni ja seega on ka N_{i+1} kroonis vähemalt $\frac{\mathbf{k}}{l^s}$ märgistatud positsiooni. Kui N_i on hargnemisklass, siis on klasside $N_1, \ldots, N_{i-1}$ seas $(s-1)$ hargnemisklassi ja seega on induktsiooni oletusest tulenevalt N_i kroonis vähemalt $\frac{\mathbf{k}}{l^{s-1}}$ märgistatud positsiooni. Kuna klassil N_i pole üle l vahetu järglase, siis peab tal leiduma vahetu järglane, mille kroonis on vähemalt $\frac{\mathbf{k}}{l^s}$ märgistatud positsiooni. Kuna järjendi $N_1, \ldots, N_r$ konstruktsiooni tõttu on N_{i+1} klassi N_i vahetu järglane, mille kroonis on enim märgistatud positsioone, siis peab ka N_{i+1} kroonis olema vähemalt $\frac{\mathbf{k}}{l^s}$ märgistatud positsiooni.

Väitest 3.2.1 nähtub, et järjendis $N_1, \ldots, N_r$ on vähemalt $(\mathcal{N}(\mathbf{grad}\, G, 0, 2)\mathcal{H}(\mathbf{grad}\, G)|N| + 1)$ hargnemisklassi.

Järelikult on vähemalt $(\mathcal{N}(\mathbf{grad}\, G, 0, 2)\mathcal{H}(\mathbf{grad}\, G) + 1)$ hargnemisklassi märgendiks sama mitteterminal, mis olgu tähistatud A-ga. Kuna neil hargnemisklassidel ei saa olla enam kui $\mathcal{H}(\mathbf{grad}\, G)$ erinevat hargnemisviisi, siis on nende seas omakorda vähemalt $(\mathcal{N}(g(A), 0, 2) + 1)$ hargnemisklassi, millel on sama hargnemisviis h. Olgu järjendi viimased $(\mathcal{N}(g(A), 0, 2) + 1)$ hargnemisklassi, mis on märgistatud mitteterminaliga A ja mille hargnemisviis on h, tähistatud kui $X_1, X_2, \ldots, X_{\mathcal{N}(g(A),0,2)+1}$.

Defineerime funktsioonid $f_i : [1..g(A)] \longrightarrow [1..g(A)]$, kus $i \in [1..\mathcal{N}(g(A), 0, 2)]$ järgmiselt: $f_i(k) = j$ parajasti siis, kui tipp $X_{i+1}^{(k)}$ asub T sellises alampuus, mille juureks on tipp $X_i^{(j)}$. Olgu $1 \le i \le j \le \mathcal{N}(g(A), 0, 2)$, olgu $f_{i,j} := f_j \circ f_{j-1} \circ \cdots \circ f_i$, siis tipp $X_{j+1}^{(k)}$ asub T sellises alampuus, mille juureks on tipp $X_i^{(f_{i,j}(k))}$. Lemmast 3.1 saame, et leiduvad sellised i_0, i_1, i_2, et $1 \le i_0 < i_1 < i_2 \le \mathcal{N}(g(A), 0, 2)$ ja $f_{i_0+1,i_1} = f_{i_1+1,i_2} = f^2_{i_1+1,i_2}$. Tähistame $Y_0 = X_{i_0+1}, Y_1 = X_{i_1+1}, Y_2 = X_{i_2+1}$.

Olgu $\bar{g} = g(A)$, $k_j = |f^{-1}_{i_1+1,i_2}(j)|$ $(j \in [1..\bar{g}])$. Olgu Y_2 kroon tähistatud kui $(w_j)_{j=1}^{\bar{g}}$, Y_1 kroon tähistatud kui

$$(v_{j,0} \prod_{i=1}^{k_j} (w_{l_j+i} v_{j,i}))_{j=1}^{\bar{g}},$$

juurtipu kroon (terve sõna z) tähistatud kui

$$t_0 \prod_{j=1}^{\bar{g}} (v_{j,0} \prod_{i=1}^{k_j} (w_{l_j+i} v_{j,i}) t_j).$$

Selline sõna z dekompositsioon rahuldab lemma väidet. Tõepoolest:

1. Sõna z on avaldatud soovitud kujul. See, et $l_j < j \le l_{j+1}$ kui $k_j > 0$, järeldub sellest, et $f_{i_1+1,i_2} = f^2_{i_1+1,i_2}$, s.t. selle funktsiooni kujutiseks on parajasti tema püsipunktid. Järelikult kui mingi $p, q \in [1..\bar{g}]$ korral $Y_2^{(p)}$ on alampuus, mille juureks on $Y_1^{(q)}$, siis $k_q > 0$ ja $f_{i_1+1,i_2}(p) = q$ ning seega $f_{i_1+1,i_2}(q) = q$, mis tähendab seda, et $Y_2^{(q)}$ on alampuus, mille juureks on $Y_1^{(q)}$ Kuna $f_{i_1+1,\ldots,i_2}$ on monotoonne, siis $f^{-1}_{i_1+1,i_2} = [l_q + 1..l_{q+1}]$ ning seega $l_q + 1 \le q \le l_{q+1}$.

2. Kuna Y_2 on üks järjendi $N_1, \ldots, N_r$ komponentidest, siis Y_2 kroon sisaldab märgistatud positsioone.

3. Kuna Y_1 on hargnemisklass, siis leiduvad sellised j_1 ja j_2, et v_{j_1,j_2} sisaldab märgistatud positsioone. Kuna Y_0 on hargnemisklass, siis leidub selline j, et t_j sisaldab märgistatud positsioone. Kuna Y_0 ja Y_1 hargnemisviis on sama ja $f_{i_0+1,i_1} = f_{i_1+1,i_2}$, siis võime valida need j_1, j_2, j nii nagu kirjeldatud lemma väites.

4. Kuna Y_0, Y_1 ja Y_2 hargnemisviis on sama, siis on $Y_0^{(j)}$ kroonis märgistatud positsioone parajasti siis, kui neid on $Y_1^{(j)}$ kroonis, ning parajasti siis, kui neid on $Y_2^{(j)}$ kroonis.

5. On ilmne, et mitteterminali A – klasside Y_0, Y_1, Y_2 märgendi — jaoks leiduvad sellised tuletuskäigud, mis on kirjeldatud lemma väites. □

Märkus 3.2. Tõestusest on näha, et kui sõnal z on mitu analüüsipuud, siis me võime neist valida ükskõik millise. Seega võiks Ogdeni lemma sõnastuses „ ... kui $z \in \mathcal{L}(G)$ ja $|z| \geq \mathbf{k}$ ja z-s on märgistatud $\mathbf{k}$ või rohkem positsiooni ... " asemel olla „ ... kui $z \in \mathcal{L}(G)$ ja $|z| \geq \mathbf{k}$ ja z-s on märgistatud $\mathbf{k}$ või rohkem positsiooni, siis z iga analüüsipuu korral ... ".

3.2 Järeldus Lyndoni sõnade hulga kohta

Olgu Σ tähestik.

Definitsioon 3.3. Sõna $w \in \Sigma^*$ nimetatakse *primitiivseks*, kui ei leidu selliseid $v \in \Sigma^*$ ja $n \in \mathbb{N}$, $n \geq 2$, et $w = v^n$.

Olgu hulk Σ lineaarselt järjestatud. Järjestus Σ-l defineerib leksikograafilise järjestuse Σ^*-l.

Definitsioon 3.4. Primitiivset sõna üle tähestiku Σ nimetatakse *Lyndoni sõnaks*, kui ta pole suurem ühestki oma tsüklilisest permutatsioonist (sõnad $w_1, w_2 \in \Sigma^*$ on teineteise tsüklilised permutatsioonid, kui leiduvad sellised $u, v \in \Sigma^*$, et $w_1 = uv$ ja $w_2 = vu$).

Definitsioon 3.5. Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika, olgu $p = x \rightarrow \langle (y_1, \ldots, y_{g(x)}); Y \rangle \in P$. Produktsiooni P nimetame *ϵ-produktsiooniks*, kui leidub selline $i \in [1..g(x)]$, et $y_i = \epsilon$.

Definitsioon 3.6. Üldistatud kontekstivaba grammatikat $G = ([V, g], \Sigma, P, s)$ nimetame *ϵ-vabaks*, kui tema produktsioonide seas pole ϵ-produktsioone või kui tema produktsioonide seas on täpselt üks ϵ-produktsioon, mis on kujul $s \rightarrow \langle \epsilon; \{\emptyset\} \rangle$ ja sel juhul s ei esine produktsioonide paremas pooles.

Lemma 3.3. *Iga üldistatud kontekstivaba grammatika $G = ([V, g], \Sigma, P, s)$ korral leidub ϵ-vaba üldistatud kontekstivaba grammatika $G' = ([V', g], \Sigma, P', s')$ nii, et $\mathcal{L}(G) = \mathcal{L}(G')$.*

Tõestus. Olgu

$$V' := \Sigma \dot\cup \{s'\} \dot\cup \{\langle A\rangle^{(x)} \mid A \in N, x \in \{0,1\}^{g(x)} \backslash \{0\}^{g(x)}\}.$$

Olgu $g(s') := 1$. Kui $x \in \{0,1\}^*$, siis tähistagu $w_{\mathrm{h}}(x)$ sõna x Hammingi kaalu — ühega võrdsete bittide arvu sõnas x. Iga $\langle A\rangle^{(x)} \in V'$ korral olgu $g(\langle A\rangle^{(x)}) := w_{\mathrm{h}}(x)$. Olgu $n \in \mathbb{N}$ ja $x \in \{0,1\}^n \backslash \{0\}^n$. Defineerime seose $\rightsquigarrow_x \subseteq (n{-}V'^*_g)^2$ järgmiselt:

1. olgu $\langle (u_1, \ldots, u_n); U\rangle \in n{-}V'^*_g$;
2. olgu $i \in [1..n]$ selline, et $x(i) = 0$ ($x(i)$ tähistab i-ndat bitti sõnas x) ja $u_i \neq \epsilon$;
3. olgu $j \in [1..|u_i|]$, olgu $z = j + \sum_{k=1}^{j-1} |u_i|$, olgu $Z \in U$ selline, et $z \in Z$;
4. olgu $B \in N$ ja $y \in \{0,1\}^*$ sellised, et $l_{u_1 \cdots u_n}(Z) = \langle B\rangle^{(y)}$, olgu y' saadud y-st, võttes seal $\mathcal{E}l_Z(z)$-nda ühega võrduva biti võrdseks nulliga.

Kui meil on õnnestunud sel viisil valida z, Z, B, y', siis loeme

$$\begin{aligned}\langle (u_1, \ldots, u_n); U\rangle \rightsquigarrow_x \langle (u_1, \ldots, u_n); U\rangle[Z \longleftarrow \\ \langle(((\langle B\rangle^{(y')}, 1), \ldots, (\langle B\rangle^{(y')}, \mathcal{E}l_Z(z) - 1), \epsilon, \\ (\langle B\rangle^{(y')}, \mathcal{E}l_Z(z)), \ldots, (\langle B\rangle^{(y')}, w_{\mathrm{h}}(y'))); \{[1..w_{\mathrm{h}}(y)] \backslash \mathcal{E}l_Z(z)\}\rangle]\end{aligned}$$

See tähendab, et märgist $\rightsquigarrow_x$ paremal seisab sulustatud järjend, mis on saadud, kustutades järjendist $(u_1, \ldots, u_n)$ mitteterminalikomponendi, mis seisab positsioonil z, ja asendades vastava mitteterminali sellisega, mis seda komponenti ei sisalda.

Olgu $\overset{*}{\rightsquigarrow}_x$ seose $\rightsquigarrow_x$ refleksiivne transitiivne sulund. On ilmne, et iga $\langle (u_1, \ldots, u_n); U\rangle \in n{-}V'^*_g$ korral leidub täpselt üks selline $\langle (w_1, \ldots, w_n); W\rangle \in n{-}V'^*_g$, et $\langle (u_1, \ldots, u_n); U\rangle \overset{*}{\rightsquigarrow}_x \langle (w_1, \ldots, w_n); W\rangle$ ja iga $i \in [1..n]$ korral võrdusest $x(i) = 0$ järeldub $w_i \in \Sigma^*$. Juhul kui iga $i \in [1..n]$ korral võrdusest $x(i) = 0$ järeldub $w_i = \epsilon$, siis olgu $\{j_1, \ldots, j_k\} = \{i \in [1..n] \mid x(i) = 1\}$ $(j_1 < \ldots < j_k)$ ja

$$\mathcal{X}_x(\langle (u_1, \ldots, u_n); U\rangle) := \langle (w_{j_1}, w_{j_2}, \ldots, w_{j_k}); W\rangle.$$

Iga $n \in \mathbb{N}$ korral defineerime $I : n - V_g^* \longrightarrow n - {V'}_g^*$ järgmiselt: kui $\langle (u_1, \ldots, u_n); U\rangle \in n - V_g^*$, siis $I(\langle (u_1, \ldots, u_n); U\rangle)$ saamiseks asendame sõnades $u_1, \ldots, u_n$ iga tähe A_i, kus $A \in N$ ja $i \in [1..g(x)]$, tähega $(\langle A\rangle^{(1^{g(x)})}, i)$. Nüüd saame defineerida produktsioonide hulga

$$\begin{aligned}P' := \{\langle A\rangle^{(x)} \to \mathcal{X}_x(I(\mathcal{U})) \mid A \to \mathcal{U} \in P, x \in \{0,1\}^{g(A)} \setminus \{0\}^{g(A)},\\ \mathcal{X}_x(I(\mathcal{U})) \text{ eksisteerib ja ei sisalda } \epsilon\text{-ga võrdseid komponente}\}\\ \cup \{s' \to s\} \cup \{s' \to \epsilon \mid \text{kui } \epsilon \in \mathcal{L}(G)\}\end{aligned}$$

On ilmne, et G' on ϵ-vaba. Peale selle $\mathcal{L}(G) = \mathcal{L}(G')$. Tõepoolest, induktsiooniga üle derivatsiooni pikkuse on kerge näidata, et kehtib

Väide 3.3.1. *Olgu $A \in N$, olgu $w_1, \ldots, w_{g(A)} \in \Sigma^*$ sellised, et $w_1 \cdots w_{g(A)} \neq \epsilon$. Olgu $x \in \{0,1\}^{g(A)}$ selline, et iga $i \in [1..g(A)]$ korral $x(i) = 1$ parajasti siis, kui $w_i \neq \epsilon$. Olgu $\{j_1, \ldots, j_k\} = \{i \in [1..n] \mid x(i) = 1\}$. Siis on grammatikas G olemas derivatsioon $(A_i)_{i=1}^{g(A)} \overset{*}{\Rightarrow} (w_i)_{i=1}^{g(A)}$ parajasti siis, kui grammatikas G' on olemas derivatsioon $((\langle A\rangle^{(x)}, i))_{i=1}^{k} \overset{*}{\Rightarrow} (w_{j_i})_{i=1}^{k}$.*

Lisaks sellele on ilmne, et $\epsilon \in \mathcal{L}(G)$ parajasti siis, kui $\epsilon \in \mathcal{L}(G')$. $\square$

Lause 3.4. *Olgu $\Sigma = \{a, b\}$, kus loeme $a < b$. Lyndoni sõnade hulk L üle Σ pole üldistatud kontekstivaba.*

Tõestus. Oletame, et leidub selline üldistatud kontekstivaba grammatika G, et $\mathcal{L}(G) = L$. Olgu G ϵ-vaba, olgu $\mathbf{k}$ selle grammatika Ogdeni konstant. Olgu

$$\mathbf{c} = 6(\mathbf{k} + 2)(2\mathbf{k} + 3)(\mathbf{grad}\, G).$$

Vaatame sõna

$$z = (a^{\mathbf{k}+1}b)^2(a^{\mathbf{k}}b)^{\mathbf{c}},$$

mis on Lyndoni sõna. Iga $x \in [1..\mathbf{c}]$ korral tähistagu $\mathcal{A}[x]$ lõiku

$$[\mathbf{k} + 4 + x(\mathbf{k} + 1)..2\mathbf{k} + 3 + x(\mathbf{k} + 1)].$$

Sõnas z asub positsioonidel lõigust $\mathcal{A}[x]$ parajasti x-s alamsõna $a^{\mathbf{k}}$, millele vahetult eelneb ja järgneb b-täht. Fikseerime selle sõna mingi analüüsipuu T. Fikseerime mingi $x \in [1..\mathbf{c}]$ ja olgu sõnas z märgistatud positsioonid lõigust $\mathcal{A}[x]$. Olgu

$$z = t_0 \prod_{j=1}^{\bar{g}} (v_{j,0} \prod_{i=1}^{k_j} (w_{l_j+i} v_{j,i}) t_j)$$

sõna z dekompositsioon vastavalt Ogdeni lemmale. Vastavalt Ogdeni lemma teisele punktile leidub selline $\bar{j} \in [1..\bar{g}]$, et $w_{\bar{j}}$ sisaldab märgistatud positsioone. Arvestades, et sõnas z on märgistatud positsioonid järjest, tuleneb Ogdeni lemma kolmandast ja neljandast punktist, et leidub selline $j \in [1..\bar{g}]$, et

1. $k_j > 0$ ja w_j sisaldab märgistatud positsioone.
2. Leiab aset üks järgnevatest juhtudest:
   (a) t_{l_j} sisaldab märgistatud positsioone ning seega

   $$\prod_{s=l_j+1}^{j-1} (v_{s,0} t_s) v_{j,0} \prod_{s=1}^{j-l_j-1} (w_{l_j+s+1} v_{j,s})$$

   koosneb vaid märgistatud positsioonidest. Peale selle leidub selline $s \in [l_j+1..j-1]$, et $v_{s,0} \neq \epsilon$, või selline $s \in [0..j-l_j-1]$, et $v_{j,s} \neq \epsilon$.

   (b) $t_{l_{j+1}}$ sisaldab märgistatud positsioone ning seega

   $$\prod_{s=j-l_j+1}^{k_j} (v_{j,s-1} w_{l_j+s}) v_{j,k_j} \prod_{s=j+1}^{l_{j+1}} (t_{s-1} v_{s,0})$$

   koosneb vaid märgistatud positsioonidest. Peale selle leidub selline $s \in [j+1..l_{j+1}]$, et $v_{s,0} \neq \epsilon$, või selline $s \in [j-l_j..k_j]$, et $v_{j,s} \neq \epsilon$.

Seega saame sõna z vastavalt Ogdeni lemmale üles pumbates sõnad, milles leidub tõkestamata pikkusega lõik a-sid ning mis ei ole esimene lõik a-sid selles sõnas.

Olgu $\mathbf{j}$ vähim selline j, mille korral on täidetud eeltoodud tingimused 1 ja 2. Olgu $H_{\mathrm{a}}(x)$ võrdne $w_{\mathbf{j}}$ esimese positsiooni järjekorranumbriga sõnas z ja $H_{\mathrm{l}}(x)$ võrdne $w_{\mathbf{j}}$ viimase positsiooni järjekorranumbriga sõnas z ning olgu $d(x) := \mathbf{j}$.

Järelikult peab sõna z dekompositsioon olema selline, et üles pumbates kasvaks ka esimese a-de lõigu pikkus tõkestamatult. Seega peab leiduma selline $j' \in [1..\bar{g}]$, et

1. $k_{j'} > 0$ ja $w_{j'}$ esimene positsioon pole sõna z enam kui $(\mathbf{k}+2)$-s positsioon.
2. Mõni sõnadest $v_{j',0}, v_{j',1}, \ldots, v_{j',k_{j'}}$ ja $w_{l_{j'}+1}, \ldots, w_{j'-1}, w_{j'+1}, \ldots, w_{l_{j'+1}}$ ei ole tühi sõna.

3. Kui $w_{j'}$ esimene positsioon on sõna z täpselt $(\mathbf{k}+2)$-s positsioon, siis mõni sõnadest $v_{j',0}, \ldots, v_{j',j'-l_{j'}-1}$ ja $w_{l_{j'}+1}, \ldots, w_{j'-1}$ pole tühi sõna.

Kui $w_{j'}$ viimane positsioon oleks sõna z enam kui $(2\mathbf{k}+3)$-s positsioon, siis alla pumbates esimene lõik a-sid sõnas z lüheneks, kuid teine lõik oleks ikkagi $(\mathbf{k}+1)$-täheline. See tähendab, et alla pumbates saaksime sõna, mis pole Lyndoni sõna. Seega ei ole $w_{j'}$ viimane positsioon sõna z enam kui $(2\mathbf{k}+3)$-s positsioon.

Olgu $\mathbf{j}'$ vähim selline j', mille korral on täidetud eeltoodud tingimused 1 kuni 3. Olgu $H'_{\mathrm{a}}(x)$ võrdne $w_{\mathbf{j}'}$ esimese positsiooni järjekorranumbriga sõnas z ja $H'_{\mathrm{l}}(x)$ võrdne $w_{\mathbf{j}'}$ viimase positsiooni järjekorranumbriga sõnas z ning olgu $d'(x) := \mathbf{j}'$. $(H'_{\mathrm{l}}(x) \leq 2\mathbf{k}+3)$

Kuna iga $x \in [1..\mathbf{c}]$ korral on $H_{\mathrm{a}}(x) > 2(\mathbf{k}+2)+(x-1)(\mathbf{k}+1)+1$ või $H_{\mathrm{l}}(x) < 2(\mathbf{k}+2)+x(\mathbf{k}+1)-1$, siis

$$|\{(H_{\mathrm{a}}(x), H_{\mathrm{l}}(x)) \mid x \in [1..\mathbf{c}]\}| \geq 3(\mathbf{k}+2)(2\mathbf{k}+3)(\mathbf{grad}\, G).$$

Seega leidub meil hulk $X \subseteq [1..\mathbf{c}]$ nii, et $|X| \geq 3(\mathbf{grad}\, G)$ ja iga $x_1, x_2 \in X$ korral, kus $x_1 \neq x_2$ kehtib $(H_{\mathrm{a}}(x_1), H_{\mathrm{l}}(x_1)) \neq (H_{\mathrm{a}}(x_2), H_{\mathrm{l}}(x_2))$ ja $(H'_{\mathrm{a}}(x_1), H'_{\mathrm{l}}(x_1)) = (H'_{\mathrm{a}}(x_2), H'_{\mathrm{l}}(x_2))$, sest

$$|\{H'_{\mathrm{a}}(x) \mid x \in [1..\mathbf{c}]\}| \leq \mathbf{k}+2 \text{ ja } |\{H'_{\mathrm{l}}(x) \mid x \in [1..\mathbf{c}]\}| \leq 2\mathbf{k}+3.$$

Kui $x \in X$, siis olgu $Y_2(x)$ puu T tippude tükelduse see klass, mis on Ogdeni lemma tõestuses tähistatud Y_2-ga, juhul kui sõnas z on märgistatud positsioonid lõigust $\mathcal{A}[x]$. Kuna kõigi $x \in X$ korral on $Y_2(x)^{(d'(x))}$ kroon sama, siis iga $x_1, x_2 \in X$ korral kas $Y_2(x_1) = Y_2(x_2)$ või on klass $Y_2(x_1)$ klassi $Y_2(x_2)$ järglane või on klass $Y_2(x_2)$ klassi $Y_2(x_1)$ järglane.

Olgu $x_0 \in X$ selline, et $Y_2(x_0)$ on kõige juurepoolsem klass hulgast $\{Y_2(x) \mid x \in X\}$. Kuna T tippude tükelduse iga klassi korral ei leidu rohkem kui $(\mathbf{grad}\, G-1)$ hulga X elementi x nii, et $Y_2(x)$ oleks võrdne selle tükelduse klassiga, siis on klassi $Y_2(x_0)$ tippudel vähemalt $2\,\mathbf{grad}\, G$ järglast kujul $Y_2(x)^{(d(x))}$, kus $x \in [1..\mathbf{c}]$. Kuna tipp $Y_2(x_0)^{(d'(x_0))}$ järglaste hulgas pole sellisel kujul tippe, siis leidub $d_0 \in [1..|Y_2(x_0)|]$ nii, et leiduvad $x_1, x_2, x_3 \in X$ nii, et $x1 < x_2 < x_3$ ja puu T tipud $Y_2(x_1)^{(d(x_1))}$, $Y_2(x_2)^{(d(x_2))}$ ja $Y_2(x_3)^{(d(x_3))}$ on tipu $Y_2(x_0)^{(d_0)}$ järglased. Olgu $Y_1(x_2)$ puu T tippude tükelduse see klass, mis on Ogdeni lemma tõestuses tähistatud Y_1-ga, juhul kui sõnas z on märgistatud positsioonid lõigust $\mathcal{A}[x_2]$. Üks klassidest $Y_1(x_2)$ või $Y_2(x_0)$ peab olema teise järglane (või peavad nad kokku langema). Kui $Y_1(x_2)$ on $Y_2(x_0)$ järglane, siis on tipp $Y_1(x_2)^{(d'(x_2))}$ tipu $Y_2(x_0)^{(d'(x_0))}$ järglane ja $Y_2(x_2)^{(d'(x_2))}$ eellane. Kuna tipu $Y_1(x_2)^{(d'(x_2))}$ kroonis on rohkem tähti kui $Y_2(x_2)^{(d'(x_2))}$ kroonis, siis ei saa tippude $Y_2(x_2)^{(d'(x_2))}$ ja $Y_2(x_0)^{(d'(x_0))}$ kroonid võrdsed olla, järelikult peab $Y_1(x_2)$ olema $Y_2(x_0)$ eelane või peavad need kaks klassi kokku langema. Kuid sel juhul sisaldab tipu $Y_2(x_0)^{(d_0)}$ kroon (ja seega ka klassi $Y_1(x_2)$ kroon) sõna z kõik positsioonid lõigust $\mathcal{A}[x_2]$ ning järelikult sõna z dekompositsioon juhul, kui märgistatud olid positsioonid lõigust $\mathcal{A}[x_2]$, ei rahulda Ogdeni lemmat (nimelt ei leidu sellist $\bar{j} \in [0..\bar{g}]$, et $t_{\bar{j}}$ sisaldaks märgistatud positsioone). Seega ei saa leiduda üldistatud kontekstivaba grammatikat G nii, et $\mathcal{L}(G) = L$. $\square$

4 Üldistatud kontekstivabu grammatikaid aktsepteerivad automaadid

4.1 Puuautomaadi definitsioon

Definitsioon 4.1. Olgu $[K, g]$ sulutähestik. *Üldiseks sulustatud metsaks* üle $[K, g]$ nimetame metsa T, mille tipud on märgendatud $[K, g]$ elementidega ja mille tippudel on antud tükeldus $\{W_1, W_2, \dots, W_n\}$, kusjuures kehtib

- tükelduse iga klassi W_i korral leidub selline $x \in K$ (mida nimetame klassi W_i *märgendiks*), et W_i-s on täpselt $g(x)$ tippu ja nende märgendid on (puud T eeljärjestuses läbides) vastavalt $(x, 1), (x, 2), \dots, (x, g(x))$;
- tükelduse iga klassi W_i korral, mis ei sisalda T juurtippe, leidub tükelduse klass W_j (mida nimetame klassi W_i *vahetuks eellaseks*; klassi W_i nimetame klassi W_j *vahetuks järglaseks*), nii et klassi W_i kuuluva iga tipu vahetu eellane kuulub klassi W_j;
- T juurtipud moodustavad omaette tükeldusklassi.

Üldist sulustatud metsa, mis koosneb ühest puust, nimetame üldistatud sulustatud puuks.

Kui üldise sulustatud metsa tippude tükelduses pole selliseid klasse, millel on kaks vahetut järglast, millel on sama märgend ja mille komponentide arv on suurem kui üks, siis on sellel metsal ainult üks tipuhulga tükeldus, mis rahuldab eelmist definitsiooni. Selliseid üldiseid sulustatud metsi nimetame *sulustatud metsadeks* ja selliseid üldiseid sulustatud puid sulustatud puudeks.

Kui $\langle\!\langle \notin K$ on täht, siis saame puid T üle $[K, g]$ (mille tippudel pole defineeritud tükeldust; kuna sulustatud puu korral defineerib puu tükelduse, siis on ka nemad niimoodi kujutatavad) kujutada (mittesulustatud) sõnana $\mathrm{Wrd}(T) \in ([K, g] \cup \{\langle\!\langle\})^*$ järgmiselt:

- Kui T on ühe tipuga puu, mille märgendiks on $(x, i) \in [K, g]$, siis T kujutiseks $\mathrm{Wrd}(T)$ on (x, i).
- Kui T-s on enam kui üks tippu ja T juurtipul on n järglast, millelele vastavad T alampuud on $T_1, T_2, \dots, T_n$, siis

$$\mathrm{Wrd}(T) := (x, i)\langle\!\langle^n \mathrm{Wrd}(T_1)\mathrm{Wrd}(T_2) \cdots \mathrm{Wrd}(T_n).$$

Olgu $\mathrm{Tr}(K, g) := \{\mathrm{Wrd}(T) \,|\, T \text{ on sulustatud puu üle } [K, g]\}$.

Definitsioon 4.2. *Puuautomaadiks* nimetame järjendit $(Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$, kus

- Q on automaadi *olekute* hulk;
- Σ on *terminaalne tähestik*;
- Γ on *pinutähestik*;
- $[K, g]$ on *puutähestik*;
- $q_0 \in Q$ on automaadi *algolek*;
- $Z_0 \in \Sigma$ on *pinu alumine sümbol*;
- $F \subseteq Q$ on automaadi *lõppolekute* hulk;
- $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \longrightarrow \mathcal{P}(Q \times \Gamma^* \times [K, g])$ on automaadi *ülemineku-funktsioon*.

Definitsioon 4.3. Puuautomaadi (tähistused samad kui eelmises definitsioonis) *olekuks* nimetatakse nelikut (q, w, z, t), kus $q \in Q$, $w \in \Sigma^*$, $z \in \Gamma^*$ ja $t \in ([K, g] \cup \{\langle\!\langle\})^*$.

Definitsioon 4.4. Olgu $(Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ puuautomaat ja $S_1 = (q_1, w_1, z_1, t_1)$ ning $S_2 = (q_2, w_2, z_2, t_2)$ tema olekud. Ütleme, et olekust S_1 on *vahetult tuletatav* olek S_2 (tähistame $S_1 \vdash S_2$), kui

- $w_1 = aw_2$, kus $a \in \Sigma \cup \{\epsilon\}$;
- $t_2 = t_1 x \langle\!\langle^n$, kus $x \in [K, g]$ ja $n \in \mathbb{N}_0$;
- $z_1 = uz'$ ja $z_2 = uv$, kus $u, v \in \Gamma^*, z' \in \Gamma$;
- $(q_2, v, x) \in \delta(q_1, a, z')$ ja $|v| = n$

Puuautomaadi olekute järjendit $((q_i, w_i, z_i, t_i))_{i=1}^n$, kus $n \in \mathbb{N}$, nimetame puuautomaadi *arvutuskäiguks*, kui iga $i \in [1..n-1]$ korral $(q_i, w_i, z_i, t_i) \vdash (q_{i+1}, w_{i+1}, z_{i+1}, t_{i+1})$. Arvu $(n-1)$ nimetame selle arvutuskäigu *pikkuseks*.

Olgu $\overset{*}{\vdash}$ relatsiooni $\vdash$ refleksiivne transitiivne sulund.

Definitsioon 4.5. Puuautomaat $(Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ *aktsepteerib sõna* $w \in \Sigma^*$

- ... *tühja pinuga*, kui leiduvad sellised $q \in Q$ ja $t \in \mathrm{Tr}(K, g)$, et $(q_0, w, Z_0, \epsilon) \overset{*}{\vdash} (q, \epsilon, \epsilon, t)$;

- ... *lõppolekuga*, kui leiduvad sellised $q \in F$, $z \in \Gamma^*$ ja $t \in \mathrm{Tr}(K, g)$ ning $u, v \in ([K, g] \cup \{\langle\!\!\rangle\})^*$ nii, et $t = uv$ ja $q_0, w, Z_0, \epsilon) \models^* (q, \epsilon, z, u)$;
- ... *tühja pinu ja lõppolekuga*, kui leiduvad sellised $q \in F$ ja $t \in \mathrm{Tr}(K, g)$, et $(q_0, w, Z_0, \epsilon) \models^* (q, \epsilon, \epsilon, t)$.

Olgu $A = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ puuautomaat. Defineerime

$$\begin{aligned}
\mathcal{L}_{\mathcal{E}}(A) &:= \{w \in \Sigma^* \mid A \text{ aktsepteerib } w \text{ tühja pinuga}\};\\
\mathcal{L}_{\mathcal{F}}(A) &:= \{w \in \Sigma^* \mid A \text{ aktsepteerib } w \text{ lõppolekuga}\};\\
\mathcal{L}(A) &:= \{w \in \Sigma^* \mid A \text{ aktsepteerib } w \text{ tühja pinu ja lõppolekuga}\}.
\end{aligned}$$

Lause 4.1. *Olgu Σ tähestik ja $L \subseteq \Sigma^*$. Siis leidub puuautomaat A_1 nii, et $\mathcal{L}_{\mathcal{E}}(A_1) = L$, parajasti siis, kui leidub puuautomaat A_2 nii, et $\mathcal{L}_{\mathcal{F}}(A_2) = L$.*

Tõestus. Kui $A_1 = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ on puuautomaat, mis aktsepteerib keele L tühja pinuga, siis defineerime puuautomaadi A_2 järgmiselt:

$$A_2 := (Q \dot\cup \{q_0', f\}, \Sigma, \Gamma \dot\cup \{Y_0\}, [K \dot\cup \{k_0\}, g], \delta', q_0', Y_0, \{f\}),$$

kus $g(k_0) := 1$ ja δ' on defineeritud järgmiselt:

- $\delta'(q, a, z) := \delta(q, a, z)$, kui $q \in Q$, $z \in \Gamma$;
- $\delta'(q_0', \epsilon, Y_0) := \{(q_0, Y_0 Z_0, k_0)\}$;
- $\delta'(q, \epsilon, Y_0) := \{(f, \epsilon, k_0)\}$, kui $q \in Q$;
- $\delta' = \emptyset$ ülejäänud juhtudel.

Sel juhul $\mathcal{L}_{\mathcal{F}}(A_2) = L$.

Teistpidi, olgu $A_2 = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ on puuautomaat, mis aktsepteerib keele L lõppolekuga. Olgu $\bar{g} = \max\{g(x) \mid x \in K\}$ ja

$$\bar{z} = \max\{|v| \mid \text{leiduvad sellised } q_1, q_2, a, z, t, \text{ et } (q_2, v, t) \in \delta(q_1, a, z)\}.$$

Defineerime nüüd puuautomaadi

$$A_1 := (Q \dot\cup \{q_0', f\}, \Sigma, \Gamma \dot\cup \{Y_0, Z_f\}, [K \dot\cup \{k_0\}, g], \delta', q_0', Y_0, \emptyset),$$

kus $g(k_0) := 1$ ja δ' on defineeritud järgmiselt:

- $\delta'(q_0', \epsilon, Y_0) := \{(q_0, Y_0 Z_0, k)\}$;
- $\delta'(q, a, z) := \delta(q, a, z)$, kui $q \in Q \backslash F$ ja $z \in \Gamma$;

- $\delta'(q, a, z) := \delta(q, a, z)$, kui $q \in F$, $a \in \Sigma$ ja $z \in \Gamma$;
- $\delta'(q, \epsilon, z) := \delta(q, \epsilon, z) \cup (\{f\} \times \{Z_f^k \,|\, 0 \leq k \leq (\bar{g}-1)^2\bar{z}\} \times [K, g])$, kui $q \in F$ ja $z \in \Gamma$;
- $\delta'(f, \epsilon, z) := \{f\} \times \{Z_f^k \,|\, 0 \leq k \leq (\bar{g}-1)^2\bar{z}\} \times [K, g]$, kui $z \in \Gamma \cup \{Z_f\}$;
- $\delta'(f, \epsilon, Y_0) := \{(f, \epsilon, k)\}$;
- $\delta' = \emptyset$ ülejäänud juhtudel.

Sel juhul $\mathcal{L}_{\mathcal{E}}(A_1) = L$. Tõepoolest, automaadi A_1 võimalikud (olekust (q_0, w, Z_0, ϵ) algavad) arvutuskäigud on järgmise kujuga:

1. Automaat A_1 märgib oma pinu põhja ära Y_0-ga ja läheb automaadi A_2 algolekusse.
2. Automaat A_1 simuleerib automaati A_2 senikaua, kuni A_2 läheb lõppolekusse. Kui A_2 jätkab peale lõppolekusse minekut arvutamist, siis jätkab arvutamist ka A_1.
3. Kui A_2 peatub olekus (f', ϵ, z, u), kus $f' \in F$, siis A_1 läheb olekusse f. Järgnevalt tuleb konstrueerida selline v, et uv oleks mõne sulustatud puu kujutis ning selle käigus ka pinu tühjaks teha. Selleks lisatakse u-le tippe, millel on kuni $(\bar{g}-1)^2\bar{z}$ järglast. Kui puu on konstrueeritud, on automaat A_1 jõudnud seisu $(f, \epsilon, Y_0, k\llparenthesis^2 uv)$.
4. Lõpuks kustutatakse pinust Y_0. $\square$

Selle lause tõestusest on näha, et kui keele L jaoks leiduvad puuautomaadid, mis aktsepteerivad ta tühja pinuga või lõppolekuga, siis leidub ka puuautomaat, mis aktsepteerib selle keele tühja pinu ja lõppolekuga, sest konstrueeritud puuautomaadil A_2, mis aktsepteerib keele L lõppolekuga, on lõppolekus tühi pinu.

Järgnevates osades me näitame, et keel on puuautomaadiga aktsepteeritav parajasti siis, kui ta on genereeritav üldistatud kontekstivaba grammatikaga.

4.2 Üldistatud kontekstivaba keele aktsepteerimine puuautomaadiga

Lemma 4.2. *Iga üldistatud kontekstivaba grammatika $G = ([V, g], \Sigma, P, s)$ korral leidub sama järku üldistatud kontekstivaba grammatika $G' =$*

$([V', g], \Sigma, P', s^{(1)})$ nii, et $\mathcal{L}(G) = \mathcal{L}(G')$ ja grammatika G' ühegi produktsiooni paremas pooles ei esine kahte samasugust mitteterminali, teisisõnu kehtib

$$\forall x \to \langle y; Y\rangle \in P' \; \forall z \in N' : \; |\{X \in Y \mid l_y(X) = z\}| \leq 1.$$

Tõestus. Olgu $h : N \longrightarrow \mathbb{N}_0$ defineeritud järgmiselt:

$$h(z) = \max\{(|\{X \in Y \mid l_y(Y) = z\}|) \mid x \to \langle y; Y\rangle \in P\}$$

(maksimaalne mitteterminalide z arv produktsioonide paremates pooltes). Defineerime nüüd grammatika G' järgnevalt. Loeme, et

$$V' := \Sigma \cup \bigcup_{x \in N} \{x^{(1)}, x^{(2)}, \ldots, x^{(h(x))}\},$$

kus $g(x^{(i)}) := g(x)$. Produktsioonide hulgaks P' loeme

$$P' = \{x^{(i)} \to \langle y'; Y\rangle \mid x \to \langle y; Y\rangle \in P, 1 \leq i \leq h(x)\},$$

kus $y' \in [V', g]^*$ on selline, et kui $X \in Y$, siis

$$l_{y'}(X) = (l_y(X))^{(|\{Z \in Y \mid \min Z \leq \min X, l_y(Z) = l_y(X)\}|)}.$$

Tänu V' ja h definitsioonile on meil küllalt palju $l_y(X)$ koopiaid, et y' sel viisil defineerida.

On ilmne, et G ja G' genereerivad sama keele. (Enamgi veel, G ja G' analüüsipuude vahel on võimalik korraldada loomulik üksühene vastavus.)

$\square$

Definitsioon 4.6. Üldistatud kontekstivaba grammatika $G = ([V, g], \Sigma, P, s)$ *kontekstivabaks alusgrammatikaks* nimetame kontekstivaba grammatikat $G' := ([V, g], \Sigma, P', s)$, kus

$$P' := \bigcup_{x \to \langle (y_1, \ldots, y_{g(x)}); Y\rangle \in P} \{x_i \to y_i \mid i \in [1..g(x)]\}.$$

On ilmne, et $\mathcal{L}(G) \subseteq \mathcal{L}(G')$.

Definitsioon 4.7. Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika ja $G' = ([V, g], \Sigma, P', s)$ tema kontekstivaba alusgrammatika. Iga $p = x \to \langle y; Y\rangle \in P$ korral defineerime $g(p) := g(x)$. Grammatika G *alusgrammatika sulustuseks* nimetame kujutist $\mathrm{Kp} : P' \longrightarrow [P, g]$, mis on defineeritud järgmiselt:

Iga $p = x \to \langle (y_1, \ldots, y_{g(x)}); Y\rangle \in P$ ja $i \in [1..g(x)]$ korral $\mathrm{Kp}(x_i \to y_i) := (p, i)$.

Definitsioon 4.8. Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika ja olgu $G' = ([V, g], \Sigma, P', s)$ tema kontekstivaba alsugrammatika. Olgu T grammatika G' derivatsioonipuu. Puu T *sulustuseks* $\mathrm{Kp}(T)$ nimetame märgendatud puud, mille saame T-st, jättes tema lehtede märgendid samaks ja märgendades iga tema tipu v, mis ei ole leht, ümber $\mathrm{Kp}(p)$-ga, kus p on tipu v märgend puus T.

On ilmne, et kehtib

Lemma 4.3. *Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika ja G' tema alusgrammatika. Olgu $w \in \mathcal{L}(G')$, siis $w \in \mathcal{L}(G)$ parajasti siis, kui leidub grammatika G' selline analüüsipuu T, mille krooniks on sõna w ja mille tippudel leidub selline tükeldus, et* $\mathrm{Kp}(T)$ *koos selle tükeldusega on üldine sulustatud puu üle $[P \dot\cup \Sigma, g]$.*

Lause 4.4. *Iga üldistatud kontekstivaba keel on aktsepteeritav puuautomaadiga.*

Tõestus. Olgu $G = ([V, g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika. Olgu selle grammatika produktsioonid sellised, et ühegi produktsiooni paremas pooles ei esine kahte samasugust mitteterminali. Defineerime järgmise puuautomaadi:

$$A := (\{q\}, \Sigma, [V, g], [P \dot\cup \Sigma, g], \delta, q, s, \emptyset),$$

kus

1. iga $x \to \langle y; Y \rangle \in P$ korral $g(x \to \langle y; Y \rangle) := g(x)$;
2. $\delta(q, \epsilon, x_i) := \{(q, (y_i)^{\mathrm{R}}, p_i \,|\, p = x \to \langle y; Y \rangle \in P\}$, kui $x \in N$ ja $i \in [1..g(x)]$ ning sõna w korral tähistab $(w)^{\mathrm{R}}$ sedasama sõna paremalt vasakule loetuna;
3. $\delta(q, a, a) := \{(q, \epsilon, a)\}$, kui $a \in \Sigma$;
4. $\delta = \emptyset$ ülejäänud juhtudel.

Kuna G produktsioonide paremates pooltes ei esine ühesuguseid mitteterminale, siis on G derivatsioonipuud on sulustatud puud. Samuti on G kontekstivaba alusgrammatika derivatsioonipuude sulustused sulustatud puud.

Olgu $AG = ([V, g], \Sigma, P', s)$ grammatika G kontekstivaba alusgrammatika.

Väide 4.4.1. *Olgu $B \in [V, g]$, olgu $w \in \Sigma^*$, olgu T märgendatud puu märgenditega hulgast $[P \dot\cup \Sigma, g]$. Siis T on grammatika AG derivatsioonimetsaks juurega B ja krooniga w parajasti siis, kui $(q, w, B, \epsilon) \overset{*}{\models} (q, \epsilon, \epsilon, \mathrm{Wrd}(T))$.*

Väite „siis“-osa tõestame induktsiooniga üle puu T kõrguse.

Baas: T kõrgus on 0, see tähendab, et T koosneb ühest tipust. Sel juhul $w \in \Sigma$ ja $B = w$. Automaadi A konstruktsiooni kolmanda punkti järgi siis $(q, B, B, \epsilon) \vdash (q, \epsilon, \epsilon, B)$.

Samm: Olgu T kõrgus n, kehtigu väide puude jaoks, mille kõrgus on väiksem kui n. Leiduvad $B_1, \ldots, B_k \in [V, g]$ ja $w_1, \ldots, w_k \in \Sigma^*$ nii, et $w = w_1 \cdots w_k$, grammatikas AG on produktsioon $p = B \to B_1 \cdots B_k$ ja iga $i \in [1..k]$ korral leidub AG derivatsioonimets T_i juurega B_i ja krooniga w_i ning $\mathrm{Wrd}(T) = p\llparenthesis^k \mathrm{Wrd}(T_1) \cdots \mathrm{Wrd}(T_k)$.

Induktsiooni oletuse tõttu on iga $i \in [1..k]$ korral

$$(q, w_i, B_i, \epsilon) \vdash^* (q, \epsilon, \epsilon, \mathrm{Wrd}(T_i)). \tag{4.1}$$

Automaadi A konstruktsiooni teise punkti järgi

$$(q, w_1 \cdots w_k, B, \epsilon) \vdash (q, w_1 \cdots w_k, B_k \cdots B_1, p\llparenthesis^k)$$

ja edasi seose (4.1) järgi

$$\begin{aligned} &\vdash^* (q, w_2 \cdots w_k, B_k \cdots B_2, p\llparenthesis^k \mathrm{Wrd}(T_1)) \\ &\vdash^* (q, w_3 \cdots w_k, B_k \cdots B_3, p\llparenthesis^k \mathrm{Wrd}(T_1)\mathrm{Wrd}(T_2)) \\ &\vdash^* \cdots \vdash^* (q, \epsilon, \epsilon, \mathrm{Wrd}(T)). \end{aligned}$$

Väite „ainult siis“-osa tõestame induktsiooniga üle A arvutuskäigu pikkuse.

Baas: Arvutuskäigu pikkus on 1, see tähendab, et $(q, w, B, \epsilon) \vdash (q, \epsilon, \epsilon, \mathrm{Wrd}(T))$. Sel juhul $B = w \in \Sigma$ ja $\mathrm{Wrd}(T) = B$. Seega on T ühe tipuga puu, mille ainsa tipu märgendiks on B.

Samm: Olgu arvutuskäigu pikkus $n > 1$, kehtigu väide arvutuskäikude jaoks, mille pikkus on väiksem kui n. Olgu $w' \in \Sigma$, $B_1, \ldots, B_k \in [V, g]$ ja $p \in [P \dot\cup \Sigma, g]$ sellised, et

$$(q, w, B, \epsilon) \vdash (q, w', B_k \cdots B_1, p\llparenthesis^k) \vdash^* (q, \epsilon, \epsilon, \mathrm{Wrd}(T))$$

ja selle arvutuskäigu pikkus olgu n. Kuna arvutuskäigu pikkus on suurem kui üks, siis peab olema $p \in [P, g]$ ja $w = w'$ ning järelikult on grammatikas AG olemas produktsioon $p = B \to B_1 \cdots B_k$.

Leiduvad sellised $w_1, \ldots w_k \in \Sigma^*$, et $w = w_1 \cdots w_k$, ning sellised märgendatud puud $T_1, \ldots, T_k$ märgenditega hulgast $[P \dot\cup \Sigma, g]$, et
$\mathrm{Wrd}(T) = p\llparenthesis^k \mathrm{Wrd}(T_1) \cdots \mathrm{Wrd}(T_k)$, ja

$$\begin{aligned} (q, w_1 \cdots w_k, B_k \cdots B_1, p\llparenthesis^k) &\vdash^* (q, w_2 \cdots w_k, B_k \cdots B_2, p\llparenthesis^k \mathrm{Wrd}(T_1)) \\ &\vdash^* \cdots \vdash^* (q, w_k, B_k, p\llparenthesis^k \mathrm{Wrd}(T_1) \cdots \mathrm{Wrd}(T_{k-1})) \\ &\vdash^* (q, \epsilon, \epsilon, \mathrm{Wrd}(T)), \end{aligned}$$

seejuures on iga $i \in [1..k]$ korral olek

$$(q, w_{i+1} \cdots w_k, B_k \cdots B_{i+1}, p\langle\!\langle^k \mathrm{Wrd}(T_1) \cdots \mathrm{Wrd}(T_i))$$

selles arvutuskäigus esimene olek, kus pinu sisu (e. oleku kolmas komponent) koosneb mitte enam kui $(k - i)$ sümbolist. Järelikult on meil iga $i \in [1..k]$ korral olemas arvutuskäik

$$(q, w_i, B_i, \epsilon) \vdash^* (q, \epsilon, \epsilon, \mathrm{Wrd}(T_i)),$$

mille pikkus on väiksem kui n. Järelikult on T_i grammatika AG derivatsioonimets juurega B_i ja krooniga w_i. Kuna T juurtipu märgend oli $B \to B_1 \cdots B_k$ ja tal oli k vahetut järglast, millest lähtuvad alampuud olid vastavalt $T_1, \ldots, T_k$, siis on T grammatika AG derivatsioonimets juurega B ja krooniga $w_1 \cdots w_k = w$.

Sellest väitest ja lemmast 4.3 saame, et $\mathcal{L}_{\mathcal{E}}(A) = \mathcal{L}(G)$. □

4.3 Puuautomaadiga aktsepteeritava keele genereerimine üldistatud kontekstivaba grammatikaga

Lemma 4.5. *Iga puuautomaadiga aktsepteeritava keele korral leidub puuautomaat, millel on täpselt üks lõppolek ja mis aktsepteerib selle keele tühja pinu ja lõppolekuga.*

Tõestus. Olgu $A = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ puuautomaat. Defineerime automaadi $A' = (Q \dot\cup \{f\}, \Sigma, \Gamma, [K, g], \delta', q_0, Z_0, \{f\})$, kus

$$\begin{aligned}\delta'(q, a, z) := \delta(q, a, z) \cup \{(f, v, t) \in \{f\} \times \Gamma^* \times [K, g] \mid \\ \exists f' \in F : (f', v, t) \in \delta(q, a, z)\},\end{aligned}$$

kui $q \in Q$, $a \in \Sigma$ ja $z \in \Gamma$, muudel väärtustel on δ' võrdne tühja hulgaga. Siis $\mathcal{L}(A) = \mathcal{L}(A')$. Tõepoolest, automaat A' läheb lõppolekusse parajasti siis, kui automaat A läheb lõppolekusse ja lõpetab töö. □

Lause 4.6. *Iga keel, mille jaoks leidub teda aktsepteeriv puuautomaat, on üldistatud kontekstivaba.*

Tõestus. Olgu $A = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, \{f\})$ puuautomaat. Järgnevas me konstrueerime üldistatud kontekstivaba grammatika, mis genereerib keele, mille see puuautomaat tühja pinu ja lõppolekuga aktsepteerib.

Teostame selle konstruktsiooni järgnevate sammudena:

1. Olgu $\mathcal{Q} := Q \times \Gamma \times Q$. Olgu $\pi_1, \pi_3 : \mathcal{Q} \longrightarrow Q$ ja $\pi_2 : \mathcal{Q} \longrightarrow \Gamma$ vastavad projektsioonid.

2. Olgu $N := \{(k,\tilde{q}) \,|\, k \in K, \tilde{q} \in \mathcal{Q}^{g(k)}\}$. Iga $x = (k,\tilde{q}) \in N$ puhul defineerime $g(x) := g(k)$. Samuti defineerime $\bar{K} : N \longrightarrow K$ ja $\bar{Q} : [N,g] \longrightarrow \mathcal{Q}$ järgmiselt:

$$\bar{K}((k,((q_1^{(1)},z_1,q_1^{(2)}),\ldots,(q_{g(k)}^{(1)},z_{g(k)},q_{g(k)}^{(2)})))) := k$$
$$\bar{Q}(((k,((q_1^{(1)},z_1,q_1^{(2)}),\ldots,(q_{g(k)}^{(1)},z_{g(k)},q_{g(k)}^{(2)}))),i)) := (q_i^{(1)},z_i,q_i^{(2)}).$$

3. Olgu $V = N\dot{\cup}\Sigma\dot{\cup}\{s\}$, kus s on uus sümbol. Olgu $g(s) := 1$ ja iga $a \in \Sigma$ korral $g(a) := 1$.

4. Olgu $\mathcal{K} : Q \times \Gamma \longrightarrow \mathcal{P}([K,g])$ defineeritud järgmiselt:

$$\mathcal{K}(q,z) := \{k_i \in [K,g] \,|\, \exists a \in \Sigma \cup \{\epsilon\}, v \in \Gamma^*, q' \in Q : (q',v,k_i) \in \delta(q,a,z)\}.$$

5. Olgu $\mathcal{T} : \mathcal{Q} \times [K,g] \longrightarrow \mathcal{P}([V,g]^*)$ defineeritud järgmiselt:

$$\begin{aligned}\mathcal{T}((q_1,z,q_2),k_i) := \{&a(x_n,j_n)(x_{n-1},j_{n-1})\cdots(x_1,j_1) \,|\\ &n \in \mathbb{N}, x_1,\ldots,x_n \in N, \forall l \in [1..n] : j_l \in [1..g(x_l)], a \in \Sigma \cup \{\epsilon\},\\ &\exists z_1,\ldots,z_n \in \Gamma, q' \in Q : \big[(q',z_1\cdots z_n,k_i) \in \delta(q_1,a,z),\\ &\pi_3(\bar{Q}((x_1,j_1))) = q_2, \pi_1(\bar{Q}((x_n,j_n))) = q',\\ &\forall l \in [1..n-1] : \pi_3(\bar{Q}((x_{l+1},j_{l+1}))) = \pi_1(\bar{Q}((x_l,j_l))),\\ &\forall l \in [1..n] : [\pi_2(\bar{Q}((x_l,j_l))) = z_l \wedge (\bar{K}(x_l),j_l) \in \mathcal{K}(\pi_1(\bar{Q}((x_l,j_l))),z_2)]\big]\}\\ &\cup \{a \in \Sigma \cup \{\epsilon\} \,|\, (q_2,\epsilon,k_i) \in \delta(q_1,a,z)\}.\end{aligned}$$

6. Olgu iga $x \in N$ puhul $\mathcal{R}'(x) \subseteq ([V,g]^*)^{g(x)}$ defineeritud järgmiselt:

$$\begin{aligned}\mathcal{R}'(x) := \{&(w_1,\ldots,w_{g(x)}) \in\\ &\mathcal{T}(\bar{Q}((x,1)),(\bar{K}(x),1)) \times\cdots\times \mathcal{T}(\bar{Q}((x,g(x))),(\bar{K}(x),g(x))) \,|\\ &\text{iga } y \in N \text{ ja } i_1,i_2 \in [1..g(y)] \text{ korral sisaldab sõna } w_1\cdots w_{g(x)} \text{ tähte}\\ &y_{i_1} \text{ parajasti siis, kui ta sisaldab tähte } y_{i_2}\text{; iga } k \in K \text{ korral leidub}\\ &\text{sõnas } w_1\cdots w_{g(x)} \text{ parajasti 0 või } g(k) \text{ sellist}\\ &\text{tähte } y_i = (y,i)\text{, et } \bar{K}(y) = k\}.\end{aligned}$$

7. Olgu iga $x \in N$ puhul $\mathcal{R} \subseteq g(x){-}V_g^*$ defineeritud järgmiselt:

$$\begin{aligned}\mathcal{R}(x) := \{&\langle(w_1,\ldots,w_{g(x)});W\rangle \,|\, (w_1,\ldots,w_{g(x)}) \in \mathcal{R}'(x), W \text{ on}\\ &\text{hulga } [1..|w_1\cdots w_{g(x)}|] \text{ selline tükeldus, et}\\ &\langle(w_1,\ldots,w_{g(x)});W\rangle \in g(x){-}V_g^*\}.\end{aligned}$$

8. Olgu $G = ([V,g], \Sigma, P, s)$ üldistatud kontekstivaba grammatika, kus produktsioonide hulk P on defineeritud järgmiselt:

$$P := \{s \to (k_0, ((q_0, Z_0, f))) \mid k_0 \in K, g(k_0) = 1\} \\ \cup \{x \to \langle y; Y\rangle \mid x \in N, \langle y; Y\rangle \in \mathcal{R}(x)\}.$$

Olgu T grammatika G derivatsioonimets (mingi juurega), mis ei koosne vaid juurtippudest. Olgu T_i selle metsa vasakult lugedes i-s puu. Tähistame $\bar{K}(T_i)$-ga puud, mille saame, kui

1. kustutame puust T_i lehed ja

2. märgendame T_i tipud, mis ei ole lehed, ümber — kui tipu v märgend puus T on $(x \to \langle y; Y\rangle, j) \in [P, g]$, kus $j \in [1..g(x)]$, siis tipu v märgendiks puus $\bar{K}(T_i)$ võtame $(\bar{K}(x), j) \in [K, g]$.

Kui $x \in N$ ja $i \in [1..g(x)]$, siis tähistame $\bar{K}(x_i)$-ga tähte $(\bar{K}(x), i) \in [K, g]$.

Väide 4.6.1. *Olgu $B \in N$, olgu $w_1, \dots, w_{g(B)} \in \Sigma^*$. Olgu T sulustatud mets märgenditega tähestikust $[N \cup T, g]$. Siis on T grammatika derivatsioonimets juurega B, mille puude kroonideks on vasakult paremale vastavalt $w_1, \dots, w_{g(B)}$ siis ja ainult siis, kui iga $i \in [1..g(B)]$ korral*

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \models^* (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_i))), \tag{4.2}$$

kus T_i on metsa T vasakult lugedes i-s puu.

Väite „siis"-osa tõestame induktsiooniga üle metsa T kõrguse.

Baas: metsa kõrgus on 1. Siis leidub grammatikas G produktsioon $B \to (w_1, \dots, w_{g(x)})$. Siis on grammatika konstruktsiooni punktide 6 kuni 8 järgi iga $i \in [1..g(B)]$ korral $w_i \in \mathcal{T}(\bar{Q}(B_i), \bar{K}(B_i))$. Järelikult konstruktsiooni punkti 5 järgi iga $i \in [1..g(B)]$ korral

$$(\pi_3(\bar{Q}(B_i)), \epsilon, \bar{K}(B_i)) \in \delta(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)))$$

ja seega iga $i \in [1..g(B)]$ korral

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \models (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_i))).$$

Samm: Olgu metsa kõrgus $n > 1$, kehtigu väide metsade puhul, mille kõrgus on väiksem kui n. Olgu metsa T juurklassi märgend

$$p = B \to \langle (y_1, \dots, y_{g(x)}); Y\rangle \in P, \tag{4.3}$$

kus iga $i \in [1..g(B)]$ korral $y_i = y_{i,1}y_{i,2}\cdots y_{i,l_i}$ ja $y_{i,j} \in [N \cup \Sigma, g]$. Y on tükeldus, mis kirjeldab, millised positsioonid produktsiooni paremas pooles kokku kuuluvad. Olgu $Y = \{Y_1, \ldots, Y_k\}$ ja olgu iga $j \in [1..k]$ korral $l_{y_1\cdots y_{g(x)}}(Y_j) = C^{(j)}$. Leiduvad sõnad $w_{i,j} \in \Sigma^*$, kus $i \in [1..g(B)]$ ja $j \in [1..l_i]$, et iga $i \in [1..g(B)]$ korral $w_i = w_{i,1}\cdots w_{i,l_i}$ ning iga $i \in [1..g(B)]$ ja $j \in [1..l_i]$ korral on $w_{i,j}$ puu T_i juure j-nda alamtipu kroon. Kuna metsa T juurklassi vahetutest järglastest algavad alammetsad on kõik väiksema kõrgusega kui n, siis iga $j \in [1..k]$ korral kui $C^{(j)} \in N$, siis induktsiooni oletuse kohaselt kõigi $i' \in [1..g(C^{(j)})]$ korral

$$(\pi_1(\bar{Q}(C^{(j)}_{i'})), w'_{j,i'}, \pi_2(\bar{Q}(C^{(j)}_{i'})), \epsilon) \models^* (\pi_3(\bar{Q}(C^{(j)}_{i'})), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_{j,i'}))),$$

kus $w'_{j,i'}$ on w_{i_1,i_2} mis vastab Y_j i'-ndale elemendile, ja $T_{j,i'}$ on vastav puu, mis algab mõne T juurtipu mõnest alamtipust ja mille krooniks on w_{i_1,i_2}.

Kuna G sisaldab produktsiooni (4.3), siis grammatika konstruktsiooni punktide 6 kuni 8 järgi iga $i \in [1..g(B)]$ korral

$$y_{i,1}y_{i,2}\cdots y_{i,l_i} \in \mathcal{T}(\bar{Q}(B_i), \bar{K}(B_i)),$$

kui $y_{i,1}\cdots y_{i,l_i} \notin \Sigma \cup \{\epsilon\}$, ja $w_i \in \mathcal{T}(\bar{Q}(B_i), \bar{K}(B_i))$ muidu. Teisel juhul

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \vdash (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_i))).$$

Esimesel juhul olgu $x = 1$ ja $a = \epsilon$, kui $y_{i,1} \notin \Sigma$, ja $x = 2$ ja $a = y_{i,1}$ muidu. Grammatika konstruktsiooni punkti 5 järgi

$$(\pi_1(\bar{Q}(y_{i,x})), \pi_2(\bar{Q}(y_{i,l_i}))\cdots\pi_2(\bar{Q}(y_{i,x})), \bar{K}(B_i)) \in \delta(\pi_1(\bar{Q}(B_i)), a, \pi_2(\bar{Q}(B_i))).$$

Olgu iga $j \in [x..l_i - 1]$ korral $q_j = \pi_3(\bar{Q}(y_{i,j}))$ (ja ühtlasi $= \pi_1(\bar{Q}(y_{i,j+1}))$). Seega on automaadil A olemas arvutuskäik

$$\begin{aligned}
&(\pi_1(\bar{Q}(B_i)), w_{i,1}\cdots w_{i,l_i}, \pi_2(\bar{Q}(B_i)), \epsilon)\\
&\vdash (\pi_1(\bar{Q}(y_{i,x})), w_{i,x}\cdots w_{i,l_i}, \pi_2(\bar{Q}(y_{i,l_i}))\cdots\pi_2(\bar{Q}(y_{i,x})), p\llparenthesis^{l_i-(x-1)})\\
&\models^* (q_x, w_{i,x+1}\cdots w_{i,l_i}, \pi_2(\bar{Q}(y_{i,l_i}))\cdots\pi_2(\bar{Q}(y_{i,x+1})),\\
&\qquad\qquad p\llparenthesis^{l_i-(x-1)}\mathrm{Wrd}(\bar{K}(T_{i,x})))\\
\models^* \cdots &\models^* (q_{l_i-1}, w_{i,l_i}, \pi_2(\bar{Q}(y_{i,l_i})),\\
&\qquad\qquad p\llparenthesis^{l_i-(x-1)}\mathrm{Wrd}(\bar{K}(T_{i,x}))\cdots\mathrm{Wrd}(\bar{K}(T_{i,l_i-1})))\\
&\models^* (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_i))).
\end{aligned}$$

Väite „ainult siis“-osa tõestame induktsiooniga üle arvutuskäikude pikkuste summa. Olgu T sulustatud mets märgenditega tähestikust $[K, g]$, näitame, et leidub selline grammatika G derivatsioonimets $\dot{T}$, mille juur on B ja puude kroonid $w_1, \ldots, w_{g(x)}$, et $\bar{K}(\dot{T}) = T$.

Baas: Arvutuskäikude pikkuste summa on $g(B)$. Sel juhul iga $i \in [1..g(B)]$ korral

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \vdash (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \bar{K}(B_i)),$$

ja seega iga $i \in [1..g(B)]$ korral

$$(\pi_3(\bar{Q}(B_i)), \epsilon, \bar{K}(B_i)) \in \delta(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i))).$$

Grammatika G konstruktsiooni punkti 5 järgi siis iga $i \in [1..g(B)]$ korral $w_i \in \mathcal{T}(\bar{Q}(B_i), \bar{K}(B_i))$ ning konstruktsiooni punktide 6 kuni 8 järgi

$$p = (B_1, \ldots, B_{g(B)}) \to \langle (w_1, \ldots, w_{g(B)}; \{\{i\} \mid w_i \neq \epsilon\}\rangle \in P.$$

Koosnegu $\dot{T}$ puudest $\dot{T}_1, \ldots, \dot{T}_{g(B)}$, olgu iga $i \in [1..g(B)]$ korral $\dot{T}_i$ juure märgend (p, i) ja olgu juurel üks vahetu järglane märgendiga w_i, kui $w_i \neq \epsilon$ ja olgu juur vahetute järglasteta, kui $w_i = \epsilon$. Sel juhul $\bar{K}(\dot{T}) = T$.

Samm: Olgu arvutuskäikude pikkuste summa $n > g(B)$, kehtigu väide arvutuskäikude komplektide puhul, mille pikkuste summa on väiksem kui n. Olgu iga $i \in [1..g(B)]$ korral arvutuskäigu (4.2) esimene samm

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \vdash (q_{i,0}, w'_i, z_{i,l_i} \cdots z_{i,1}, \bar{K}(B_i)\langle\!\!\langle^{l_i})$$

(parema poole neljas komponent algab $\bar{K}(B_i)$-ga grammatika konstruktsiooni punktide 5 ja 6 tõttu), kus $q_{i,0} \in Q$, $w_i = a_i w'_i$ ning $a_i \in \Sigma \cup \{\epsilon\}$, $l_i \in \mathbb{N}_0$ ning $z_{i,1}, \ldots, z_{i,l_i} \in \Gamma$. Iga $i \in [1..g(B)]$ korral olgu $q_{i,1}, \ldots, q_{i,l_i} = \pi_3(\bar{Q}(B_i)) \in Q$, $w'_{i,1}, \ldots, w'_{i,l_i} \in \Sigma^*$ ja $T_{i,1}, \ldots, T_{i,l_i}$ märgendatud puud märgenditega hulgast $[K, g]$ nii, et $w'_i = w'_{i,1} \cdots w'_{i,l_i}$ ja arvutuskäik (4.2) on kujul

$$\begin{aligned}
&(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon)\\
&\vdash (q_{i,0}, w'_{i,1} \cdots w'_{i,l_i}, z_{i,l_i} \cdots z_{i,1}, \bar{K}(B_i)\langle\!\!\langle^{l_i})\\
&\vdash^* (q_{i,1}, w'_{i,2} \cdots w'_{i,l_i}, z_{i,l_i} \cdots z_{i,2}, \bar{K}(B_i)\langle\!\!\langle^{l_i}\mathrm{Wrd}(T_{i,1}))\\
\vdash^* \cdots &\vdash^* (q_{i,l_i-1}, w'_{i,l_i}, z_{i,l_i}, \bar{K}(B_i)\langle\!\!\langle^{l_i}\mathrm{Wrd}(T_{i,1}) \cdots \mathrm{Wrd}(T_{i,l_i-1}))\\
&\vdash^* (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(T_i)),
\end{aligned}$$

kus $\mathrm{Wrd}(T_i) = \bar{K}(B_i)\langle\!\!\langle^{l_i}\mathrm{Wrd}(T_{i,1}) \cdots \mathrm{Wrd}(T_{i,l_i})$ on metsa T vasakult lugedes i-s puu ja iga $j \in [1..l_i]$ korral on olek

$$(q_{i,j}, w'_{i,j+1} \cdots w'_{i,l_i}, z'_{i,l_i} \cdots z'_{i,j+1}, \bar{K}(B_i)\langle\!\!\langle^{l_i}\mathrm{Wrd}(T_{i,1}) \cdots \mathrm{Wrd}(T_{i,j}))$$

esimene olek selles arvutuskäigus, kus pinu sisu koosneb mitte enam kui $(l_i - j)$ sümbolist. Järelikult on meil iga $i \in [1..g(B)]$ ja $j \in [1..l_i]$ korral olemas arvutuskäik

$$(q_{i,j-1}, w'_{i,j}, z_{i,j}, \epsilon) \vdash^* (q_{i,j}, \epsilon, \epsilon, \mathrm{Wrd}(T_{i,j})).$$

Olgu $U_1, \dots, U_k$ metsa T tippude tükelduse need klassid, mis koosnevad T puude juurtippude vahetutest järglastest ($U_1, \dots, U_k$ tipud on täpselt puude $T_{i,j}$ juurtipud). Olgu $r \in [1..k]$, olgu $s_r = |U_r|$, olgu $i'_{r,1}, \dots, i'_{r,s_r}$ ja $j'_{r,1}, \dots j'_{r,s_r}$ sellised, et $T_{i'_{r,1},j'_{r,1}}, \dots T_{i'_{r,s_r},j'_{r,s_r}}$ on puud, mille juurtippudeks on tipud U_r-st. Olgu $\bar{k}^{(r)} \in K$ klassi U_r märgend.

Iga $r \in [1..k]$ korral kehtib, et iga $p \in [1..s_r]$ korral

$$(q_{i'_{r,p},j'_{r,p}-1}, w'_{i'_{r,p},j'_{r,p}}, z_{i'_{r,p},j'_{r,p}}, \epsilon) \models^* (q_{i'_{r,p},j'_{r,p}}, \epsilon, \epsilon, \mathrm{Wrd}(T_{i'_{r,p},j'_{r,p}})).$$

Seega, iga $r \in [1..k]$ korral defineerides

$$C^{(r)} := (\bar{k}^{(r)}, ((q_{i'_{r,p},j'_{r,p}-1}, z_{i'_{r,p},j'_{r,p}}, q_{i'_{r,p},j'_{r,p}}))_{p=1}^{s_r}),$$

saame induktsiooni oletusest, et leidub grammatika G derivatsioonimets $\dot{T}^{(r)}$, mille juureks on $C^{(r)}$ ja mille puude kroonideks on $w'_{i'_{r,p},j'_{r,p}}$, kus $p \in [1..s_r]$. Peale selle, kui $\dot{T}_p^{(r)}$ on metsa $\dot{T}^{(r)}$ vasakult lugedes p-s puu, siis $\bar{K}(\dot{T}_p^{(r)}) = T_{i'_{r,p},j'_{r,p}}$.

Olgu $\langle (a_1u_1, \dots, a_{g(B)}u_{g(B)}; U\rangle \in g(B)\text{–}V_g^*$ selline, et iga $i \in [1..g(B)]$ korral $|u_i| = l_i$, tükelduse U, klasside hulka kuuluvad $U'_1, \dots, U'_k$, mis vastavad metsa T tippude tükelduse klassidele $U_1, \dots, U_k$ (kui mets T läbida eeljärjestuses), tükelduse U ülejäänud klassid vastavad tähtedele a_i ja iga $i \in [1..g(B)]$ korral $l_{a_1u_1 \cdots a_{g(B)}u_{g(B)}}(U'_i) = C^{(r)}$. Vastavalt grammatika G konstruktsioonile sisaldab G produktsiooni $p = B \to \langle (a_1u_1, \dots, a_{g(B)}u_{g(B)}; U\rangle$.

Konstrueerime nüüd sulustatud metsa $\dot{T}$ järgmiselt:

1. Mets koosneb $g(B)$-st puust, iga $i \in [1..g(B)]$ korral on vasakult lugedes i-nda puu $\dot{T}_i$ juurtipu märgendiks (p, i).

2. Iga $i \in [1..g(B)]$ korral on $\dot{T}_i$ juurtipul $|a_iu_i|$ vahetut järglast. Kui $a_i \neq \epsilon$, siis on vasakpoolseima vahetu järglase märgend a_i. Olgu $j \in [1..l_i]$, olgu r, p sellised, et $(i'_{r,p}, j'_{r,p}) = (i, j)$. Puu $\dot{T}_i$ juurtipu $(j + |a_i|)$-s vahetu järglane olgu puu $\dot{T}_p^{(r)}$ juurtipuks.

Sel juhul $\dot{T}$ juureks on B ja tema puude kroonideks on $w_1, \dots, w_{g(B)}$ ning $\bar{K}(\dot{T}) = T$. Sellega on väide 4.6.1 tõestatud.

Väitest 4.6.1 saame nüüd, et, kui $w \in \Sigma^*$ ja $k_0 \in K$, kus $g(k_0) = 1$, siis leidub grammatika G derivatsioonimets T juurega $(k_0, (q_0, Z_0, f))$ ja oma ainsa puu krooniga w parajasti siis, kui

$$(q_0, w, Z_0, \epsilon) \models^* (f, \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}T)).$$

Seega $\mathcal{L}(A) = \mathcal{L}(G)$. $\square$

Lausetest 4.4 ja 4.6 saame, et kehtib

Teoreem 4.7. *Keele $L \subseteq \Sigma^*$ jaoks leidub teda genereeriv üldistatud kontekstivaba grammatika parajasti siis, kui leidub teda aktsepteeriv puuautomaat.*

4.4 Sulukeeli aktsepteerivad automaadid

Definitsioon 4.9. Olgu T üldistatud sulustatud mets üle $[K, g]$. Metsa T nimetame *sulumetsaks*, kui T tippude tükelduse iga klassi W, millel on kaks või enam vahetut järglast, ja klassi W iga kahe vahetu järglase W_1 ja W_2 korral leiab puud T eeljärjestuses läbides aset üks järgmistest juhtudest:

- ei leidu selliseid tippe $W_{1,i}, W_{1,j} \in W_1$ ja $W_{2,k} \in W_2$, et tipp $W_{1,i}$ on enne tippu $W_{2,k}$, mis on omakorda enne tippu $W_{1,j}$.
- ei leidu selliseid tippe $W_{2,i}, W_{2,j} \in W_2$ ja $W_{1,k} \in W_1$, et tipp $W_{2,i}$ on enne tippu $W_{1,k}$, mis on omakorda enne tippu $W_{2,j}$.

Puud, mis on sulumets, nimetame sulupuuks.

Olgu $\mathrm{BTr}(K, g) := \{\mathrm{Wrd}(T) \mid T$ on sulupuu ja sulustatud puu üle $[K, g]\}$.

Definitsioon 4.10. Olgu $A = (Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$ puuautomaat ja olgu $w \in \Sigma^*$. Ütleme, et A *sulg-aktsepteerib* sõna w

- ... *tühja pinuga*, kui leiduvad sellised $q \in Q$ ja $t \in \mathrm{BTr}(K, g)$, et $(q_0, w, Z_0, \epsilon) \models^* (q, \epsilon, \epsilon, t)$;
- ... *lõppolekuga*, kui leiduvad sellised $q \in F$, $z \in \Gamma^*$ ja $t \in \mathrm{BTr}(K, g)$ ning $u, v \in ([K, g] \cup \{\langle\!\langle\})^*$ nii, et $t = uv$ ja $q_0, w, Z_0, \epsilon) \models^* (q, \epsilon, z, u)$;
- ... *tühja pinu ja lõppolekuga*, kui leiduvad sellised $q \in F$ ja $t \in \mathrm{BTr}(K, g)$, et $(q_0, w, Z_0, \epsilon) \models^* (q, \epsilon, \epsilon, t)$.

Analoogiliselt keelte aktsepteerimisele puuautomaadiga saame defineerida keelte sulg-aktsepteerimise puuautomaadiga. Samuti on võimalik tõestada (täpselt samamoodi nagu lauses 4.1), et keelte aktsepteerimisvõimsuse koha pealt on kõik kolm sulg-aktsepteerimise viisi võrdsed.

Lause 4.8. *Keele $L \subseteq \Sigma^*$ jaoks leidub teda genereeriv sulugrammatika parajasti siis, kui leidub teda sulg-aktsepteeriv puuautomaat.*

Tõestus. Olgu $G = ([V,g],\Sigma,P,s)$ sulugrammatika. Kuna konstruktsioon, mis on esitatud lemmas 4.2, viib sulugrammatika sulugrammatikaks, siis võime eeldada, et G produktsioonide paremad pooled ei sisalda sama mitteterminali mitu korda. Konstrueerides grammatika G järgi puuautomaadi A nagu lauses 4.4, saame automaadi, mis aktsepteerib mingi sõna tühja pinuga parajasti siis, kui ta sulg-aktsepteerib selle sõna tühja pinuga.

Teist pidi, olgu $A = (Q,\Sigma,\Gamma,[K,g],\delta,q_0,Z_0,F)$ puuautomaat. Kuna lemmas 4.5 esitatud konstruktsioon viib selle automaadi selliseks, mis sulg-aktsepteerib tühja pinu ja lõppolekuga sama keele, mis A, siis võime eeldada, et $|F| = 1$. Modifitseerime lauses 4.6 toodud konstruktsiooni, muutes seal kuuendat ja seitsmendat punkti järgnevalt:

6. Olgu iga $x \in N$ puhul $\mathcal{R}'(x) \subseteq ([V,g]^*)^{g(x)}$ defineeritud järgmiselt:

$$\begin{aligned}
&\mathcal{R}'(x) := \{(w_1,\ldots,w_{g(x)}) \in \\
&\quad \mathcal{T}(\bar{Q}((x,1)),(\bar{K}(x),1)) \times \cdots \times \mathcal{T}(\bar{Q}((x,g(x))),(\bar{K}(x),g(x))) \mid \\
&\text{iga } y \in N \text{ ja } i_1,i_2 \in [1..g(y)] \text{ korral sisaldab sõna } w_1\cdots w_{g(x)} \text{ tähte} \\
&y_{i_1} \text{ parajasti siis, kui ta sisaldab tähte } y_{i_2}\text{; iga } k \in K \text{ korral leidub} \\
&\qquad\qquad \text{sõnas } w_1\cdots w_{g(x)} \text{ parajasti 0 või } g(k) \text{ sellist} \\
&\qquad\qquad\qquad \text{tähte } y_i = (y,i)\text{, et } \bar{K}(y) = k, \\
&\qquad\qquad\qquad\quad \underline{\text{peale selle, } (w_1,\ldots,w_{g(x)}) \in g(x){-}\mathrm{D}(V,g)}\}.
\end{aligned}$$

7. Olgu iga $x \in N$ puhul $\mathcal{R} \subseteq g(x){-}V_g^*$ defineeritud järgmiselt:

$$\begin{aligned}
\mathcal{R}(x) := \{\langle (w_1,\ldots,w_{g(x)}); \mathcal{S}((w_1,\ldots,w_{g(x)}))\rangle \mid& \\
(w_1,\ldots,w_{g(x)}) \in \mathcal{R}'(x)\}&.
\end{aligned}$$

Analoogiliselt väitega 4.6.1 võime tõestada, et kehtib

Väide 4.8.1. *Olgu $B \in N$, olgu $w_1,\ldots,w_{g(B)} \in \Sigma^*$. Olgu T sulumets märgenditega tähestikust $[N \cup T, g]$. Siis on T grammatika derivatsioonimets juurega B, mille puude kroonideks on vasakult paremale vastavalt $w_1,\ldots,w_{g(B)}$, siis ja ainult siis, kui iga $i \in [1..g(B)]$ korral*

$$(\pi_1(\bar{Q}(B_i)), w_i, \pi_2(\bar{Q}(B_i)), \epsilon) \models^* (\pi_3(\bar{Q}(B_i)), \epsilon, \epsilon, \mathrm{Wrd}(\bar{K}(T_i))),$$

kus T_i on metsa T vasakult lugedes i-s puu.

Sellest väitest järeldub, et $w \in \mathcal{L}(G)$ parajasti siis, kui A sulg-aktsepteerib w tühja pinu ja lõppolekuga. □

4.5 Järeldus üldistatud kontekstivaba keele ja regulaarse keele ühisosa kohta

Definitsioon 4.11. *Lõplikuks automaadiks* nimetatakse viisikut $(Q, \Sigma, \delta, q_0, F)$, kus

1. Q on automaadi *olekute hulk*;
2. Σ on *terminaalne tähestik*;
3. $q_0 \in Q$ on automaadi *algolek*.
4. $F \subseteq Q$ on automaadi *lõppolekute* hulk;
5. $\delta : Q \times \Sigma \longrightarrow \mathcal{P}(Q)$ on automaadi *üleminekufunktsioon*.

Definitsioon 4.12. Lõpliku automaadi (tähistused samad kui eelmises definitsioonis) *olekuks* nimetatakse paari (q, w), kus $q \in Q$ ja $w \in \Sigma^*$.

Definitsioon 4.13. Olgu $(Q, \Sigma, \delta, q_0, F)$ lõplik automaat ja $S_1 = (q_1, w_1)$ ning $S_2 = (q_2, w_2)$ tema olekud. Ütleme, et olekust S_1 on *vahetult tuletatav* olek S_2 (tähistame $S_1 \models S_2$), kui leidub selline $a \in \Sigma$, et $w_1 = aw_2$ ja $q_2 \in \delta(q_1, a)$.

Olgu $\models^*$ relatsiooni $\models$ refleksiivne transitiivne sulund.

Definitsioon 4.14. Ütleme, et lõplik automaat $(Q, \Sigma, \delta, q_0, F)$ *aktsepteerib* sõna $w \in \Sigma^*$, kui leidub selline $f \in F$, et $(q_0, w) \models^* (f, \epsilon)$.

Olgu $A = (Q, \Sigma, \delta, q_0, F)$ lõplik automaat. Defineerime

$$\mathcal{L}(A) := \{w \in \Sigma^* \mid A \text{ aktsepteerib } w\}.$$

Definitsioon 4.15. Olgu T tähestik. Keelt $L \subseteq \Sigma^*$ nimetame *regulaarseks*, kui leidub selline lõplik automaat A, et $L = \mathcal{L}(A)$.

Lause 4.9. *Olgu Σ tähestik, olgu $L_1 \subseteq \Sigma^*$ üldistatud kontekstivaba keel, olgu $L_2 \subseteq \Sigma^*$ regulaarne keel. Siis $L_1 \cap L_2$ on üldistatud kontekstivaba keel.*

Tõestus. Olgu $A_1 = (Q_1, \Sigma, \Gamma, [K, g], \delta_1, q_{0,1}, Z_0, F_1)$ puuautomaat, mille korral $\mathcal{L}_{\mathcal{F}}(A_1) = L_1$. Olgu $A_2 = (Q_2, \Sigma, \delta_2, q_{0,2}, F_2)$ lõplik automaat, mille korral $\mathcal{L}(A_2) = L_2$. Konstrueerime puuautomaadi

$$A = (Q_1 \times Q_2, \Sigma, \Gamma, [K, g], \delta, (q_{0,1}, q_{0,2}), Z_0, F_1 \times F_2),$$

kus, kui $q_1 \in Q_1$, $q_2 \in Q_2$, $a \in \Sigma$, $z \in \Gamma$, siis

$$\begin{aligned}
\delta((q_1,q_2),a,z) &:= \{((q_1',q_2'),z',k_i) \mid (q_1',z',k_i) \in \delta_1(q_1,a,z), q_2' \in \delta_2(q_2,a)\} \\
\delta((q_1,q_2),\epsilon,z) &:= \{((q_1',q_2),z',k_i \mid (q_1',z',k_i \in \delta_1(q_1,a,z)\}.
\end{aligned}$$

Siis $\mathcal{L}_{\mathcal{F}}(A) = L_1 \cap L_2$.

Tõepoolest, on ilmne, et kehtib väide

Väide 4.9.1. *Olgu $w \in \Sigma^*$, $q_1 \in Q_1$, $q_2 \in Q_2$, $z \in \Gamma^*$, $u \in ([K,g] \cup \{\langle\!\langle\})^*$. Siis*

$$(q_{0,1}, w, Z_0, \epsilon) \models^* (q_1, \epsilon, z, u) \text{ ja } (q_{0,2}, w) \models^* (q_2, \epsilon)$$

parajasti siis, kui

$$((q_{0,1}, q_{0,2}), w, Z_0, \epsilon) \models^* ((q_1, q_2), \epsilon, z, u). \qquad \square$$

Järeldus 4.10. *Sulukeele L_1 ja regulaarse keele L_2 ühisosa on sulukeel.*

Tõestus. Kui A_1 on puuautomaat, mis sulg-aktsepteerib keele L_1, ja A_2 on lõplik automaat, mis aktsepteerib keele L_2, siis puuautomaat A, mis on konstrueeritud nii nagu eelmise lause tõestuses, sulg-aktsepteerib keele $L_1 \cap L_2$. $\square$

Viited

[AU72] A. V. Aho, J. D. Ullman. The Theory of Parsing, Translation and Compiling. Vol. 1: Parsing. Prentice-Hall, Inc. Englewood Cliffs, N. J. 1972; tõlge vene keelde - Moskva, 1978.

[BB97] J. Berstel, L. Boasson. The Set of Lyndon Words is not Context-Free, Bulletin of the European Assoc. for Theor. Comp. Sci. **63** (1997), pp139-140.

[GHR92] Y. Guan, G. Hotz, A. Reichert. Tree Grammars with Multilinear Interpretation. Technischer Bericht, Universität des Saarlandes, 1992[1].

[Har78] M. Harrison. Introduction to Formal Language Theory. Addison-Wesley, 1978.

[Pit93] G. Pitsch. Analyse von Klammergrammatiken. Dissertation, Universität des Saarlandes, 1997.

[1] `http://www-hotz.cs.uni-sb.de/bib/Diplomarbeiten/Diplom-Reichert.ps.gz` (seisuga 20.05.1998)

Generalized Context-Free Grammars

Peeter Laud

Resume

In this paper we obtain some results for coupled context-free grammars ([GHR92],[Pit93]), that constitute generalizations of similar results for context-free grammars. Actually we obtain these results for a slightly more general generative formalism than coupled context-free grammars. Namely we drop the condition that right sides of productions must be correctly bracketed tuples and introduce explicit means for defining 'coupling of letters'. Let $[V, g]$ be the bracket alphabet as defined in [GHR92] and [Pit93]. We define *bracketing* over $[V, g]$ as a pair $\langle w; W\rangle$, where $w \in [V, g]^*$ and W is a partitioning of the set $[1..|w|]$ such that for every $X = \{x_1, \ldots, x_k\} \in W$ there exists some $A \in V$ such that $g(A) = k$ and the letters in w standing on positions $x_1, \ldots, x_k$ are $A_1, \ldots, A_k$, respectively. We define the *generalized context-free grammar* as a tuple $G = ([V, g], \Sigma, P, s)$, where $\Sigma \subset V$ and for every t in Σ holds $g(t) = 1$. The set of productions P consists of pairs of the form

$$(A_1, \ldots, A_{g(A)}) \rightarrow \langle (w_1, \ldots, w_{g(A)}); W\rangle,$$

where $A \in V\backslash\Sigma$ and $\langle w_1 \cdots w_{g(A)}; W\rangle$ is a bracketing over $[V, g]$. The language generated by such kind of grammar is defined in the usual way.

Our first new result is Ogden's iteration lemma (see theorem 3.2 in section 3). As its application and as a generalization of the main result in [BB97] we prove the following

Theorem. *The set of Lyndon words over two-letter alphabet is not generalized context-free.*

Recall that a word is called *Lyndon word* if it is a primitive word which is minimal under cyclic permutation; a word is *primitive* if it is not a proper power of a shorter word.

The second new result here is an automata model for generalized context-free languages and also for coupled context-free languages. The *tree-automaton* is defined as a tuple $(Q, \Sigma, \Gamma, [K, g], \delta, q_0, Z_0, F)$, where $(Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$ is a pushdown automaton (see [Har78]), $[K, g]$ is a bracket alphabet and δ is modified in such way that its image lies in $\mathcal{P}(Q \times \Gamma^* \times [K, g])$, but not in $\mathcal{P}(Q \times \Gamma^*)$. Every computation of this automaton *gives a tree* in preorder. Whenever a rule of the type $(q_2, z_1 \cdots z_n, k_i) \in \delta(q_1, a, z)$ is used during a

computation, a new node labeled by k_i is added to the tree and this new node is defined to have n descendants. A tree is called *bracketed* if a partitioning of its nodes can be defined in such a way that similarly to the bracketings, each class of the partitioning has a “label” in K and moreover, this partitioning is consistent with the ancestor-descendant relation of the nodes. A tree-automaton *accepts* a word w, if it has a computation that accepts w as pushdown automaton and gives a bracketed tree. We prove the following:

Theorem. *The accepting power of tree-automata and the generating power of generalized context-free grammars are the same.*

As an application of this result and as a generalization of the similar result for context-free grammars, we prove that the intersection of a generalized context-free language and a regular language is generalized context-free.