ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
"ΘΕΩΡΗΤΙΚΗ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΕΛΕΓΧΟΥ"

Το Θεώρημα των Chomsky-Schützenberger για αλγεβρικές γραμματικές με βάρη

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παύλος Φ. Μαραντίδης

Επιβλέπων: Γεώργιος Ραχώνης
Αν. Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2014

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
"ΘΕΩΡΗΤΙΚΗ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΕΛΕΓΧΟΥ"

Το Θεώρημα των Chomsky-Schützenberger για αλγεβρικές γραμματικές με βάρη

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παύλος Φ. Μαραντίδης

Επιβλέπων: Γεώργιος Ραχώνης
Αν. Καθηγητής Α.Π.Θ.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18 Δεκεμβρίου 2014.

.....	.....	.....
Α. Πάπιστας	Δ. Πουλάκης	Γ. Ραχώνης
Καθηγητής Α.Π.Θ.	Καθηγητής Α.Π.Θ.	Αν. Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2014

……..
Παύλος Φ. Μαραντίδης
Πτυχιούχος Μαθηματικός Α.Π.Θ.

ΠΕΡΙΛΗΨΗ

Το 1963 οι Chomsky και Schützenberger διατύπωσαν το ομώνυμο θεώρημα, το οποίο συνδέει τις γλώσσες χωρίς συμφραζόμενα με τις αναγνωρίσιμες γλώσσες και τις γλώσσες του Dyck. Τα τελευταία χρόνια η έρευνα έχει στραφεί σε αλγεβρικές γραμματικές με βάρη που επεκτείνουν την κλασσική θεωρία. Τα βάρη αυτά συνήθως θεωρούνταν στοιχεία ενός ημιδακτυλίου. Λόγω όμως της αδυναμίας των ημιδακτυλίων να περιγράψουν πράξεις όπως η μέση κατανάλωση πόρων, πρόσφατα οι Droste και Vogler θεώρησαν τα βάρη ως στοιχεία μιας γενικότερης αλγεβρικής δομής, του unital valuation monoid. Στην παρούσα εργασία, αφού ορίσουμε τις προαπαιτούμενες έννοιες, δίνουμε μια απόδειξη του κλασσικού θεωρήματος, κάνουμε μια περιγραφή των αλγεβρικών γραμματικών με βάρη από ένα unital valuation monoid και τέλος παρουσιάζουμε μια εκδοχή του παραπάνω θεωρήματος για τις γραμματικές αυτές.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Γραμματική χωρίς συμφραζόμενα, unital valuation monoid, Θεώρημα Chomsky-Schützenberger

ABSTRACT

In 1963, Chomsky and Schutzenberger proved the homonymous theorem, which connects context-free languages with recognizable languages and the languages of Dyck. Later, researchers focused on algebraic grammars with weights, in an attempt to extend the classical theory and provide a theoretical background for the applications. The weights were usually considered to be elements of a semiring. However, due to the inability of semirings to include operations, such as average consumption of resources, recently, Droste and Vogler considered taking weights from a more general algebraic structure, that of a unital valuation monoid. In this paper, after the definition of the prerequisite concepts, we give a proof of the classical theorem, we proceed to describe algebraic grammars with weights taken from a unital valuation monoid and, finally, we present a version of the above theorem for such grammars.

KEY WORDS

Context-free Grammar, Unital Valuation Monoid, Chomsky-Schützenberger Theorem

Στους γονείς και τους δασκάλους μου

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγικές Έννοιες

Στο κεφάλαιο αυτό δίνουμε βασικούς ορισμούς και έννοιες από την Άλγεβρα που θα χρησιμοποιούνται στην εργασία.

1.1 Αλφάβητα, Λέξεις και Αλγεβρικές Δομές

Έστω ένα σύνολο Σ, το οποίο θα καλούμε *αλφάβητο* και τα στοιχεία του *γράμματα*.

Μια πεπερασμένη ακολουθία στοιχείων του Σ καλείται *λέξη*. Θεωρούμε επίσης την *κενή λέξη* ε, μια ακολουθία χωρίς στοιχεία. Το σύνολο όλων των λέξεων με γράμματα από το αλφάβητο Σ μαζί με την κενή λέξη συμβολίζεται με Σ^*.

Στο Σ^* μπορούμε να ορίσουμε την *παράθεση* λέξεων. Πράγματι, αν θεωρήσουμε δύο λέξεις $a = (a_1, a_2, \ldots, a_n),\ b = (b_1, b_2, \ldots, b_m) \in \Sigma^*$ έχουμε

$$ab = (a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_m) = (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m) \in \Sigma^*.$$

Χάριν απλότητας, θα γράφουμε $a_1a_2 \ldots a_n$ αντί για $(a_1, a_2, \ldots, a_n)$.

Ένα σύνολο λέξεων L, ένα υποσύνολο δηλαδή του Σ^*, καλείται *γλώσσα*. Έστω $L, L_1, L_2 \subseteq \Sigma^*$ γλώσσες με γράμματα από το Σ. Μεταξύ γλωσσών ορίζονται οι κλασσικές συνολοθεωρητικές πράξεις:

• Η τομή τους αποτελείται από όλες τις λέξεις που είναι κοινές και στις δύο.

$$L_1 \cap L_2 = \{w \in \Sigma^* \mid w \in L_1 \text{ και } w \in L_2\}.$$

• Η ένωσή τους αποτελείται από όλες τις λέξεις που ανήκουν σε τουλάχιστον μία από τις γλώσσες.

$$L_1 \cup L_2 = \{w \in \Sigma^* \mid w \in L_1 \text{ ή } w \in L_2\}.$$

• Το συμπλήρωμα μιας γλώσσας αποτελείται από όλες τις λέξεις που δεν ανήκουν στη γλώσσα.

$$L^C = \{w \in \Sigma^* \mid w \notin L\}.$$

Ορίζονται όμως και νέες πράξεις:
• Η παράθεση δύο γλωσσών αποτελείται από όλες τις λέξεις που προκύπτουν από παράθεση μιας λέξης της μίας γλώσσας με μία της άλλης.

$$L_1L_2 = \{w \in \Sigma^* \mid \exists w_1 \in L_1 \text{ και } w_2 \in L_2 \text{ έτσι ώστε } w = w_1w_2\}.$$

• Η θήκη μιας γλώσσας αποτελείται από όλες λέξεις που προκύπτουν από παράθεση οσωνδήποτε λέξεων της γλώσσας και ορίζεται επαγωγικά ως εξής:

$$\begin{aligned} L^0 &= \{\varepsilon\} \\ L^{n+1} &= \left\{wv \in \Sigma^* \mid w \in L_n \text{ και } v \in L\right\} \\ L^* &= \bigcup_{i \in \mathbb{N}} L^i \end{aligned}$$

Ορισμός 1.1. *Μονοειδές καλείται ένα σύνολο M εφοδιασμένο με μια πράξη που συνήθως καλούμε πρόσθεση και συμβολίζεται με $+$, αν ικανοποιεί τις παρακάτω συνθήκες:*

1. *Η πράξη είναι προσεταιριστική*
2. *Υπάρχει ουδέτερο στοιχείο $0 \in M$ ως προς την πράξη*

Αν επιπλέον η πράξη είναι αντιμεταθετική, το μονοειδές θα λέγεται αντιμεταθετικό.

Ορισμός 1.2. *Έστω δύο μονοειδή $(M, +, 0)$ και $(N, \oplus, e)$. Μια απεικόνιση $\phi : M \to N$ λέγεται μορφισμός μονοειδών αν διατηρεί τις πράξεις και το ουδέτερο, δηλαδή αν $\phi(a + b) = \phi(a) \oplus \phi(b)$ και $\phi(0) = e$.*

Ορισμός 1.3. *Έστω ένα μονοειδές M. Το M λέγεται ελεύθερο μονοειδές αν υπάρχει σύνολο $K \subset M$ έτσι ώστε για κάθε μονοειδές N και κάθε απεικόνιση $\phi : K \to N$ να υπάρχει μοναδικός μορφισμός $\hat{\phi} : M \to N$ που να επεκτείνει την φ.*

Θα λέμε τότε ότι το M *παράγεται ελεύθερα* από το K.

Στο Σ^* η πράξη της παράθεσης είναι προσεταιριστική και η κενή λέξη λειτουργεί ως ουδέτερο στοιχείο, άρα το Σ^* εφοδιασμένο με την παράθεση και ουδέτερο στοιχείο την κενή λέξη ε είναι μονοειδές. Ισχύει μάλιστα ότι:

Θεώρημα 1.1. *Το Σ^* είναι το ελεύθερο μονοειδές που παράγεται από το Σ.*

Απόδειξη. Έστω μονοειδές $(N, \cdot, 1)$ και απεικόνιση $\phi : \Sigma \to N$. Θα δείξουμε αρχικά ότι υπάρχει μορφισμός $\hat{\phi} : \Sigma^* \to N$ που να επεκτείνει την φ και στη συνέχεια ότι είναι μοναδικός.
Με επαγωγή στο μήκος της λέξης $w \in \Sigma^*$ ορίζουμε:
• Για $w = \varepsilon$, $\hat{\phi}(\varepsilon) = 1$
• Για $w = a \in \Sigma$, $\hat{\phi}(a) = \phi(a)$
• Για $w = a_1 \dots a_n,\ a_1, \dots, a_n \in \Sigma,\ \hat{\phi}(w) = \phi(a_1) \cdot \dots \cdot \phi(a_n)$, για $n \geq 2$
Ο $\hat{\phi}$ εξ ορισμού επεκτείνει την φ και είναι πράγματι μορφισμός:
Έστω $w = a_1 \dots a_n, u = b_1 \dots b_m \in \Sigma^*$. Τότε

$$\begin{aligned}\hat{\phi}(wu) &= \hat{\phi}(a_1 \dots a_n b_1 \dots b_m) \\ &= \phi(a_1) \cdot \dots \cdot \phi(a_n) \cdot \phi(b_1) \cdot \dots \cdot \phi(b_m) \\ &= \big(\phi(a_1) \cdot \dots \cdot \phi(a_n)\big) \cdot \big(\phi(b_1) \cdot \dots \cdot \phi(b_m)\big) \\ &= \hat{\phi}(w) \cdot \hat{\phi}(u)\end{aligned}$$

Επίσης, ο $\hat{\phi}$ είναι μοναδικός. Έστω ότι υπάρχει και άλλος μορφισμός ψ που να επεκτείνει την φ. Τότε για κάθε λέξη $w = a_1 \dots a_n \in \Sigma^*$ ισχύει

$$\begin{aligned}\psi(w) = \psi(a_1 \dots a_n) &= \psi(a_1) \cdot \dots \cdot \psi(a_n) = \\ &= \phi(a_1) \cdot \dots \cdot \phi(a_n) = \hat{\phi}(a_1) \cdot \dots \cdot \hat{\phi}(a_n) = \\ &= \hat{\phi}(a_1 \dots a_n) = \\ &= \hat{\phi}(w).\end{aligned}$$

Οπότε $\psi = \hat{\phi}$. □

Παρατηρούμε λοιπόν ότι για να ορίσουμε μορφισμό από το Σ^* σε κάποιο άλλο μονοειδές, αρκεί να προσδιορίσουμε τις εικόνες των γραμμάτων.

Ένα μονοειδές $(M,+,0)$ θα λέγεται *πλήρες* (complete) αν ορίζεται το (όχι απαραίτητα πεπερασμένο) άθροισμα $\sum_I : M^I \to M$ για κάθε σύνολο δεικτών I, έτσι ώστε:

$$\sum_{i\in\emptyset} a_i = 0,\ \sum_{i\in\{k\}} a_i = a_k,\ \sum_{i\in\{j,k\}} a_i = a_j + a_k \text{ για } j \neq k,$$

$$\sum_{j\in J}\left(\sum_{i\in I_j} a_i\right) = \sum_{i\in I} a_i \text{ για κάθε διαμέριση } \{I_j\}_{j\in J} \text{ του } I.$$

Ένα μονοειδές $(M,+,0)$ θα λέγεται *ταυτοδύναμο* (idempotent) αν $a+a=a$ για κάθε $a \in M$ και ένα πλήρες μονοειδές θα λέγεται *πλήρως ταυτοδύναμο* (completely idempotent) αν $\sum_I a = a$ για κάθε $a \in M$ και κάθε σύνολο δεικτών I.

Ορισμός 1.4. *Η πεντάδα $(M,+,\cdot,0,1)$ λέγεται ημιδακτύλιος (semiring) αν ικανοποιεί τις συνθήκες:*

1. *$(M,+,0)$ είναι αντιμεταθετικό μονοειδές.*
2. *$(M,\cdot,1)$ είναι μονοειδές.*
3. *Ο πολλαπλασιασμός επιμερίζεται ως προς την πρόσθεση.*
4. *Το 0 είναι απορροφητικό στοιχείο για τον πολλαπλασιασμό.*

Υπάρχουν ορισμένες περιπτώσεις που οι παραπάνω συνθήκες είναι πολύ αυστηρές, όπως για παράδειγμα στον υπολογισμό του μέσου κόστους των διαδικασιών ενός υπολογισμού.
Γι΄ αυτό το λόγο, εισάγεται η έννοια του μονοειδούς με εκτίμηση με μονάδα (unital valuation monoid).

Ορισμός 1.5. *Η πεντάδα (K,+,val,0,1) θα καλείται unital valuation monoid αν ικανοποιεί τις συνθήκες:*

1. *Η τριάδα (K,+,0) είναι αντιμεταθετικό μονοειδές.*
2. *Για την απεικόνιση $val : K^* \to K$ ισχύει $val(a) = a$ για κάθε $a \in K$ και $val(\varepsilon) = 1$.*
3. *$val(a_1,\ldots,a_n) = 0$ κάθε φορά που $a_i = 0$ για κάποιο $1 \leq i \leq n$.*

4. $val(a_1, \ldots, a_{i-1}, 1, a_{i+1}, \ldots, a_n) = val(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ *για* $1 \leq i \leq n$.

Παράδειγμα 1.1. *Κάθε ημιδακτύλιος είναι unital valuation monoid, καθώς οι συνθήκες που ικανοποιούνται για να είναι μια δομή ημιδακτύλιος είναι προφανώς περισσότερες από τις συνθήκες του Ορισμού 1.5. Μπορούμε για παράδειγμα να το επιβεβαιώσουμε με τον ημιδακτύλιο των φυσικών αριθμών* $\big(\mathbb{N}, +, \cdot, 0, 1\big)$.

1. *Η τριάδα* $\big(\mathbb{N}, +, 0\big)$ *είναι αντιμεταθετικό μονοειδές*

2. *Μπορούμε να επεκτείνουμε την πράξη του πολλαπλασιασμού σε πράξη* $\odot$ *η οποία να ορίζεται στο* $\mathbb{N}^*$ *θέτοντας*

$$\odot(\varepsilon) = 1,\ \odot(a) = a \text{ για κάθε } a \in \mathbb{N}$$

$$\odot(a_1 \ldots a_n) = a_1 \cdot \cdots \cdot a_n \text{ για κάθε } a_1 \ldots a_n \in \mathbb{N}^* \setminus (\mathbb{N} \cup \{\varepsilon\})$$

3. *Εξ ορισμού,* $\odot(a_1, \ldots, a_n) = 0$ *κάθε φορά που* $a_i = 0$ *για κάποιο* $1 \leq i \leq n$

4. *Επίσης,* $\odot(a_1, \ldots, a_{i-1}, 1, a_{i+1}, \ldots, a_n) = \odot(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ *για* $1 \leq i \leq n$

Άλλοι ημιδακτύλιοι που βρίσκουν εφαρμογή στη Θεωρητική Πληροφορική είναι ο ημιδακτύλιος του Boole $\big(\{0,1\}, \vee, \wedge, 0, 1\big)$ *και ο τροπικός ημιδακτύλιος* $(\mathbb{R}_+ \cup \{-\infty\}, max, +, -\infty, 0)$.

Παράδειγμα 1.2. *Θα δούμε στη συνέχεια ότι η υπόθεση της μονάδας δεν επιβάλλει περιορισμό της γενικότητας, οπότε τα γνωστά από τη βιβλιογραφία valuation monoids [4, 6] βρίσκουν εφαρμογή. Πράγματι, έστω* $(K, +, val, 0)$ *μονοειδές με εκτίμηση, χωρίς όμως μονάδα. Αν θέσουμε* $K' = K \cup \{1\}$ *και ορίσουμε* $(K', +', val', 0, 1)$ *έτσι ώστε οι* $+', val'$ *να επεκτείνουν τις* $+, val$ *αντίστοιχα, και συγκεκριμένα* $x +' 1 = 1 +' x = 1$ *για κάθε* $x \in K'$ *και* $val'(a_1, \ldots, a_n) = val(b_1, \ldots, b_m)$, *όπου* $b_1, \ldots, b_m$ *είναι η υπακολουθία των* $a_1, \ldots, a_n$ *από την οποία αφαιρέθηκαν οι μονάδες, η παραπάνω πεντάδα θα είναι unital valuation monoid.*

Παράδειγμα 1.3. *Η πεντάδα* $(\mathbb{R} \cup \{-\infty, \infty\}, sup, avg, -\infty, \infty)$ *όπου* avg *είναι η γνωστή πράξη του μέσου όρου* $avg(a_1, \ldots, a_n) = \frac{1}{n}\sum_{i=1}^{n} a_i$, *αγνοώντας κάθε εμφάνιση του* ∞ *στον υπολογισμό του, είναι unital valuation monoid, αλλά όχι ημιδακτύλιος, καθώς δεν ικανοποιεί τη συνθήκη της προσεταιριστικότητας του πολλαπλασιασμού.*

Παράδειγμα 1.4. *Η πεντάδα* $(\mathbb{N}\cup\{\infty\}, +, min, 0, \infty)$ *είναι unital valuation monoid, που επίσης δεν είναι ημιδακτύλιος, αν και ο "πολλαπλασιασμός" είναι προσεταιριστικός. Ωστόσο δεν ισχύει ο επιμερισμός ως προς την πρόσθεση:*

$$\min(a, b + c) \neq \min(a, b) + \min(a, c)$$

Για να διαπιστώσουμε αυτό το τελευταίο, αρκεί να θέσουμε $a = b = c = 1$.

Παράδειγμα 1.5. *Η πεντάδα* $\big([0, 1], \oplus, \cdot, 0, 1\big)$ *με το συνήθη πολλαπλασιασμό είναι unital valuation monoid εφοδιασμένη με οποιαδήποτε από τις παρακάτω προσθέσεις:*

$$a \oplus_1 b = a + b - a \cdot b$$

$$a \oplus_2 b = min\{a + b, 1\}.$$

1.2 Τυπικές Δυναμοσειρές

Οι τυπικές δυναμοσειρές αποτελούν γενίκευση των δυναμοσειρών που γνωρίζουμε από την Ανάλυση (και κατ΄ επέκταση των πολυωνύμων). Έστω Σ ένα αλφάβητο και K ένα unital valuation monoid.

Ορισμός 1.6. *Τυπική δυναμοσειρά ή απλά σειρά καλείται μια οποιαδήποτε απεικόνιση* $r : \Sigma^* \to K$.

Η τιμή της r για τη λέξη w συμβολίζεται με (r, w).

Η κλάση όλων των σειρών από το Σ στο K συμβολίζεται με $K\langle\langle\Sigma^*\rangle\rangle$.

Το *support* μιας σειράς r είναι το σύνολο $supp(r) = \{w \in \Sigma^* \mid (r, w) \neq 0\}$.

Μια σειρά r θα λέγεται *μονώνυμο* αν $supp(r) = \{w\}$ για κάποιο $w \in \Sigma^*$, και θα γράφουμε απλά $r = (r, w).w$. Με $K\langle\Sigma \cup \{\varepsilon\}\rangle$ θα συμβολίζουμε όλα τα μονώνυμα με support στο $\Sigma \cup \{\varepsilon\}$.

Η κλασσική θεωρία για τις σειρές έχει αναπτυχθεί πάνω από ημιδακτυλίους, όπου δοθέντων δύο σειρών $r, s \in K\langle\langle\Sigma^*\rangle\rangle$ ορίζονται το άθροισμα και το *Cauchy* γινόμενό τους:

•$(r + s, w) = (r, w) + (s, w)$

•$(r \cdot s, w) = \sum_{w_1 w_2 = w} (r, w_1) \cdot (s, w_2)$.

Μπορούμε ωστόσο να επεκτείνουμε τις πράξεις στην περίπτωση που έχουμε unital valuation monoid. Συγκεκριμένα, η πρόσθεση παραμένει ίδια, και για σειρές $s_1, \ldots, s_n \in K\langle\langle\Sigma^*\rangle\rangle$ ορίζουμε το $val - Cauchy$ γινόμενο τους να είναι η σειρά $\cdot_{val}(s_1, \ldots, s_n)$ για την οποία:

• $(\cdot_{val}(s_1, \ldots, s_n), w) = \sum_{w_1 \ldots w_n = w} val((s_1, w_1), \ldots, (s_n, w_n))$.

Αν συμβολίσουμε με 0 τη σειρά που αποδίδει σε κάθε λέξη την τιμή 0, είναι εύκολο να διαπιστώσουμε ότι η πεντάδα $(K\langle\langle\Sigma^*\rangle\rangle, +, \cdot_{val}, 0, 1.\varepsilon)$ είναι unital valuation monoid.

Εύλογα όμως γεννάται η απορία τι γίνεται όταν έχουμε ένα μη πεπερασμένο άθροισμα σειρών σε ένα μη πλήρες μονοειδές. Μια οικογένεια σειρών $(r_i \mid i \in I)$ λέγεται τοπικά πεπερασμένη, αν για κάθε $w \in \Sigma^*$ το σύνολο $I_w = \{i \in I \mid (r_i, w) \neq 0\}$ είναι πεπερασμένο. Οπότε αν το K είναι πλήρες ή η οικογένεια $(r_i \mid i \in I)$ είναι τοπικά πεπερασμένη, ορίζεται η σειρά $\sum_{i\in I} r_i$ θέτοντας $(\sum_{i\in I} r_i, w) = \sum_{i\in I_w}(r_i, w)$ για κάθε $w \in \Sigma^*$.

Συνδυάζοντας αρκετά από τα παραπάνω, μπορούμε να ορίσουμε την έννοια του αλφαβητικού μορφισμού. Δοθέντων αλφαβήτων Δ και Σ και απεικόνισης $h\colon \Delta \to K\langle\Sigma \cup \{\varepsilon\}\rangle$, ο *αλφαβητικός μορφισμός που επάγεται από την* h είναι η απεικόνιση $h'\colon \Delta^* \to K\langle\langle\Sigma^*\rangle\rangle$ έτσι ώστε για κάθε $n \geq 0,\ \delta_1, \ldots, \delta_n \in \Delta$ με $h(\delta_i) = a_i.y_i$ να έχουμε

$$h'(\delta_1 \ldots \delta_n) = val(a_1, \ldots, a_n).y_1 \ldots y_n \ .$$

Αξίζει να σημειωθεί ότι η σειρά $h'(w)$ είναι μονώνυμο για κάθε λέξη $w \in \Delta^*$, $h'(\delta) = h(\delta)$ για κάθε $\delta \in \Delta$ και $h'(\varepsilon) = 1.\varepsilon$. Δοθείσης μιας γλώσσας $L \subseteq \Delta^*$, αν το K είναι πλήρες ή η οικογένεια $(h'(w) \mid w \in L)$ είναι τοπικά πεπερασμένη, μπορούμε να ορίσουμε

$$h'(L) = \sum_{w\in L} h'(w) \ .$$

Στη συνέχεια δεν θα κάνουμε διάκριση μεταξύ των h' και h.

Κεφάλαιο 2

Αυτόματα

Ορισμός 2.1. *Ονομάζουμε πλήρες αυτόματο ή απλά αυτόματο κάθε πεντάδα της μορφής*

$$A = (Q, \Sigma, \delta, q_0, F)$$

όπου Q ένα σύνολο καταστάσεων, Σ το αλφάβητο, q_0 μια κατάσταση του Q που ονομάζουμε αρχική, $F \subseteq Q$ το σύνολο των τελικών καταστάσεων και τέλος $\delta\colon Q \times \Sigma \to Q$ μια απεικόνιση που αντιστοιχίζει σε κάθε ζεύγος κατάστασης και γράμματος μια κατάσταση.

Στην ουσία η δ αποτελεί ένα σύνολο *μεταβάσεων* και περιγράφει τη λειτουργία του αυτομάτου: αν το αυτόματο βρίσκεται στην κατάσταση $q \in Q$ και τροφοδοτηθεί με ένα γράμμα $x \in \Sigma$ τότε θα μεταβεί στην κατάσταση $\delta(q, x)$.

Η δ μπορεί εύκολα να επεκταθεί σε μια απεικόνιση

$$\delta^*\colon Q \times \Sigma^* \to Q$$

θέτοντας $\delta^*(q, \varepsilon) = q$ και $\delta^*(q, x_1 \dots x_n x_{n+1}) = \delta(\delta^*(q, x_1 \dots x_n), x_{n+1})$.

Έστω $w \in \Sigma^*$ και q_0 η αρχική κατάσταση του αυτομάτου. Αν $\delta^*(q_0, w) \in F$, δηλαδή αν η κατάσταση στην οποία θα φτάσει το αυτόματο αφού τροφοδοτηθεί με ολόκληρη τη λέξη w είναι τελική, τότε θα λέμε ότι το αυτόματο *αναγνωρίζει* τη λέξη w. Το σύνολο όλων των λέξεων που αναγνωρίζει το αυτόματο A ονομάζεται *συμπεριφορά* του αυτομάτου και συμβολίζεται με $|A|$, δηλαδή είναι

$$|A| = \{w \in \Sigma^* \mid \delta^*(q_0, w) \in F\}.$$

Μια γλώσσα θα λέγεται αναγνωρίσιμη αν είναι η συμπεριφορά ενός αυτομάτου.

Δύο αυτόματα θα λέγονται ισοδύναμα αν έχουν την ίδια συμπεριφορά, δηλαδή αν αναγνωρίζουν την ίδια γλώσσα.

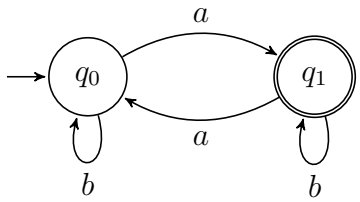
Είναι επίσης δυνατό η δ να είναι μια μερική απεικόνιση, δηλαδή να ορίζεται για *κάποια* μόνο ζεύγη κατάστασης-γράμματος, οπότε και θα έχουμε ένα *προσδιοριστό* αυτόματο. Εύκολα όμως αποδεικνύεται ότι κάθε προσδιοριστό αυτόματο είναι ισοδύναμο με ένα πλήρες: Αρκεί να προσθέσουμε στο Q μια ακόμη κατάσταση (μια κατάσταση "σκουπιδιών"), μη τελική, στην οποία θα καταλήγουν όλα τα ζεύγη που δεν έχουν εικόνα, και η οποία θα απεικονίζεται στον εαυτό της για κάθε γράμμα που τη συνοδεύει. Τα προσδιοριστά αυτόματα είναι συνήθως πιο "κομψά", καθώς δεν περιέχουν "περιττές" πληροφορίες. Υπάρχει επίσης η δυνατότητα η δ να είναι απλά μια σχέση μεταξύ των $Q \times \Sigma$ και Q, οπότε και θα έχουμε ένα *μη προσδιοριστό* αυτόματο. Και πάλι όμως αποδεικνύεται [1] ότι είναι ισοδύναμο με ένα πλήρες.

Μπορούμε να παραστήσουμε ένα αυτόματο με ένα γράφημα, όπου κάθε κατάσταση $q \in Q$ παριστάνεται με έναν κόμβο, κάθε μετάβαση $(q, x) \mapsto \delta(q, x)$ με ένα βελάκι από την κατάσταση q στην $\delta(q, x)$ το οποίο έχει πάνω του γραμμένο το γράμμα x. Η αρχική κατάσταση δηλώνεται με ένα βελάκι από αριστερά και οι τελικές δηλώνονται με έναν επιπλέον κύκλο.

Παράδειγμα 2.1. *Έστω το αυτόματο* $A = (Q, \Sigma, \delta, q_0, F)$ *όπου* $Q = \{q_0, q_1\}$, $\Sigma = \{a, b\}$, $F = \{q_1\}$ *και η απεικόνιση* δ *ορίζεται από τον πίνακα:*

δ	a	b
q_0	q_1	q_0
q_1	q_0	q_1

Τότε μπορούμε να παραστήσουμε το αυτόματο με το παρακάτω σχήμα:



Με απλές κατασκευές ([13],[1]) αποδεικνύεται ότι η ένωση, η τομή, η παράθεση δύο αναγνωρισίμων γλωσσών είναι επίσης αναγνωρίσιμη, όπως και ότι η θήκη και το συμπλήρωμα μιας αναγνωρίσιμης γλώσσας είναι επίσης αναγνωρίσιμη.

Κεφάλαιο 3

Γραμματικές χωρίς συμφραζόμενα

3.1 Εισαγωγή

Ήδη από το 300 π.Χ. ο Ινδός γλωσσολόγος Pāṇini [12] είχε περιγράψει τον τρόπο με τον οποίο δομούνται οι προτάσεις της Σανσκριτικής γλώσσας από μικρότερες φράσεις, οι οποίες με τη σειρά τους αποτελούνται από μέρη του λόγου και τελικά από λέξεις. Βασισμένος στη δουλειά του, ο Noam Chomsky συνέλαβε την ιδέα της γραμματικής χωρίς συμφραζόμενα με στόχο να περιγράψει το συντακτικό μιας φυσικής γλώσσας.

Αν και ο στόχος αυτός δεν έχει εκπληρωθεί ακόμα [9], οι γραμματικές χωρίς συμφραζόμενα έχουν βρει αρκετές και σημαντικές εφαρμογές στη Θεωρητική Πληροφορική, δύο από τις οποίες παραθέτουμε στη συνέχεια:
• Η κυριότερη ίσως χρήση εντοπίζεται στην περιγραφή γλωσσών προγραμματισμού. Κάθε πρόγραμμα αποτελείται από αντικείμενα, καθένα από τα οποία περιέχει συναρτήσεις. Κάθε συνάρτηση περιγράφεται από κάποιες ρουτίνες που κάνουν χρήση κάποιων μεταβλητών. Ο *parser*, το κομμάτι του *compiler* που αντιλαμβάνεται τη δομή του κώδικα, προέρχεται συνήθως από μια γραμματική χωρίς συμφραζόμενα. Αυτή είναι και μια από τις πρώτες εφαρμογές θεωρητικών αποτελεσμάτων στην Επιστήμη των Υπολογιστών.
• Η γλώσσα XML (Extensible Markup Language ή Επεκτάσιμη Γλώσσα Σήμανσης) αναμένεται, αν και έχει ήδη αρχίσει, να διευκολύνει την διεξαγωγή του ηλεκτρονικού εμπορίου, παρέχοντας στους συμμετέχοντες ένα κοινό

πλαίσιο περιγραφής προϊόντων, μορφής παραγγελιών και πολλών άλλων εγγράφων. Ίσως το σημαντικότερο κομμάτι της XML είναι η DTD (Document Type Definition) η οποία στην πραγματικότητα είναι μια γραμματική χωρίς συμφραζόμενα που περιγράφει τις επιτρεπτές ετικέτες και τους τρόπους που αυτές μπορούν να τοποθετηθούν. Οι ετικέτες αυτές είναι, όπως και στην HTML, λέξεις-κλειδιά που περιγράφουν το περιεχόμενο, όπως για παράδειγμα $\langle EM \rangle$ και $\langle /EM \rangle$ περικλείουν κείμενο στο οποίο πρέπει να δοθεί έμφαση. Ωστόσο, οι ετικέτες της XML δεν ασχολούνται μόνο με τη μορφοποίηση του κειμένου, αλλά κυρίως με το νόημα αυτού. Για παράδειγμα, μια ακολουθία χαρακτήρων που θέλουμε να δηλωθεί ως αριθμός τηλεφώνου μπορεί να περιέχεται από τις ετικέτες $\langle PHONE \rangle$ και $\langle /PHONE \rangle$.

Διαισθητικά, η γραμματική χωρίς συμφραζόμενα, κάνοντας χρήση συγκεκριμένων κανόνων, δομεί, ή όπως θα πούμε παρακάτω *παράγει* ακολουθίες συμβόλων. Σε γλωσσολογικό πλαίσιο, τα σύμβολα θα είναι οι λέξεις και οι ακολουθίες θα είναι προτάσεις. Στην Άλγεβρα τα σύμβολα είναι στοιχεία ενός αλφαβήτου και οι ακολουθίες ονομάζονται λέξεις. Στον προγραμματισμό το τελικό αποτέλεσμα θα είναι ένα πρόγραμμα.

3.2 Ορισμοί και παραδείγματα

Αρχικά δίνουμε αυστηρούς ορισμούς και συμβολισμό για τις γραμματικές χωρίς συμφραζόμενα κι έπειτα κάποια παραδείγματα.

Ορισμός 3.1. *Ονομάζουμε γραμματική χωρίς συμφραζόμενα κάθε τετράδα της μορφής*

$$G = (\Sigma, V, S, R)$$

όπου Σ το αλφάβητο (των τερματικών), V ένα σύνολο μεταβλητών, S μια συγκεκριμένη μεταβλητή που ονομάζουμε αξίωμα της γραμματικής και τέλος $R \subseteq V \times (V \cup \Sigma)^*$, ένα πεπερασμένο σύνολο κανόνων της μορφής $A \to w$ με $A \in V$ και $w \in (V \cup \Sigma)^*$.

Έστω $u_1, u_2 \in (V \cup \Sigma)^*$. Θα λέμε ότι η λέξη u_1 παράγει άμεσα την u_2 με εφαρμογή του κανόνα $A \to w$, αν υπάρχουν λέξεις $t_1, t_2 \in (V \cup \Sigma)^*$ έτσι ώστε $u_1 = t_1 A t_2$ και $u_2 = t_1 w t_2$. Το παραπάνω θα το συμβολίζουμε με $u_1 \underset{G}{\to} u_2$. Ακόμη, θα λέμε ότι η u_1 παράγει την u_2 σε k βήματα και θα συμβολίζουμε με

$u_1 \overset{k}{\underset{G}{\to}} u_2$ αν υπάρχουν λέξεις $w_0, w_1, \ldots, w_k \in (V \cup \Sigma)^*$ έτσι ώστε

$$u_1 = w_0 \underset{G}{\to} w_1 \underset{G}{\to} \ldots \underset{G}{\to} w_k = u_2.$$

Αν οι κανόνες που εφαρμόσαμε στη διαδικασία αυτή ήταν $\rho_1, \ldots, \rho_k$ αντίστοιχα, τότε θα λέμε ότι είχαμε την παραγωγή $d = \rho_1 \ldots \rho_k$.

Στο σημείο αυτό αξίζει να κάνουμε μια παρατήρηση. Αν έχουμε δύο κανόνες $\rho_1 = A \to t_1,\ \rho_2 = B \to t_2$ και μια λέξη $w = xAyBz,\ A, B \in V,\ x, y, z \in (V \cup \Sigma)^*$, τότε οι δύο παραγωγές

$$w = xAyBz \underset{G}{\to} xt_1yBz \underset{G}{\to} xt_1yt_2z$$

$$w = xAyBz \underset{G}{\to} xAyt_2z \underset{G}{\to} xt_1yt_2z$$

θα περιγράφονταν $d_1 = \rho_1\rho_2$ και $d_2 = \rho_2\rho_1$ αντίστοιχα. Είναι όμως φανερό ότι όχι μόνο έχουμε το ίδιο αποτέλεσμα, αλλά πρόκειται και για την ίδια ακριβώς διαδικασία. Για το λόγο αυτό, όταν θα κάνουμε λόγο για παραγωγή μιας λέξης, συνήθως θα αναφερόμαστε στην *αριστερότερη παραγωγή*, δηλαδή στην αντικατάσταση κάθε φορά της μεταβλητής που βρίσκεται πιο αριστερά στη λέξη. Οπότε, όταν θα λέμε ότι από την u_1 έχουμε την παραγωγή $d = \rho_1 \ldots \rho_k$ της u_2 θα εννοούμε ότι είχαμε αρχικά εφαρμογή του κανόνα ρ_1 για τη μεταβλητή που βρισκόταν πιο αριστερά στην $u_1 = w_0$ κι έτσι πήραμε τη λέξη w_1, στη συνέχεια του ρ_2 για την αριστερότερη μεταβλητή της w_1 κ.ο.κ. μέχρι και την εφαρμογή του ρ_k στην αριστερότερη μεταβλητή της w_{k-1} ώστε να προκύψει η λέξη $w_k = u_2$.

Τέλος, θα λέμε απλά ότι η u_1 παράγει την u_2 και θα συμβολίζουμε $u_1 \overset{*}{\underset{G}{\to}} u_2$ αν $u_1 = u_2$ ή υπάρχει $k > 0$ ώστε $u_1 \overset{k}{\underset{G}{\to}} u_2$. Στη συνέχεια, για λόγους απλότητας, αν δεν υπάρχει κίνδυνος σύγχυσης, δεν θα γίνεται διάκριση μεταξύ $\underset{G}{\to}$, $\overset{k}{\underset{G}{\to}}$ και $\overset{*}{\underset{G}{\to}}$. Όλα τα παραπάνω θα συμβολίζονται απλά $u_1 \to u_2$.

Μια *A-παραγωγή* της λέξης w είναι μια παραγωγή $d = \rho_1 \ldots \rho_k$ έτσι ώστε $A \overset{*}{\underset{G}{\to}} w$. Θα συμβολίζουμε $D(A, w)$ το σύνολο όλων των Α-παραγωγών της λέξης w και $D(w) = D(S, w)$ όλες τις παραγωγές της w ξεκινώντας από το αξίωμα της γραμματικής.

Η γλώσσα που παράγεται από τη γραμματική G συμβολίζεται με $L(G)$ και είναι $L(G) = \left\{w \in \Sigma^* \mid S \overset{*}{\underset{G}{\to}} w\right\} = \{w \in \Sigma^* \mid D(w) \neq \emptyset\}$.

Θα λέμε ότι η G είναι *ασαφής* αν υπάρχει λέξη $w \in L(G)$ τέτοια ώστε $|D(w)| \geq 2$. Διαφορετικά η γραμματική θα λέγεται *σαφής*.
Παρακάτω δίνουμε κάποια παραδείγματα γραμματικών χωρίς συμφραζόμενα και των γλωσσών που παράγουν.

Παράδειγμα 3.1. *Θεωρούμε τη γραμματική*

$$G = (\Sigma, V, S, R)$$

με $\Sigma = \{a, b\},\ V = \{S\},\ R = \{r_1 = S \to aSa,\ r_2 = S \to b\}$. *Για να βρούμε μια λέξη που να παράγεται από τη* G, *ξεκινάμε από το αξίωμα* S *και εφαρμόζουμε κανόνες από το* R *μέχρι να φτάσουμε σε μια λέξη που να αποτελείται μόνο από τερματικά γράμματα. Μπορούμε για παράδειγμα να έχουμε την παρακάτω παραγωγή:*

$$S \underset{G}{\to} aSa \underset{G}{\to} aaSaa \underset{G}{\to} aabaa$$

όπου εφαρμόσαμε αρχικά τον κανόνα r_1 *δύο φορές και στη συνέχεια τον* r_2 *μία φορά. Έχουμε με άλλα λόγια την παραγωγή* $d = r_1 r_1 r_2$. *Παρατηρούμε ότι ξεκινώντας και μετά από κάθε εφαρμογή του* r_1 *είχαμε λέξη με μία μεταβλητή, ενώ η εφαρμογή του δεύτερου κανόνα την εξαφάνισε και καταλήξαμε σε τερματική λέξη. Γενικότερα, οποιαδήποτε παραγωγή της συγκεκριμένης γραμματικής θα αποτελείται από εφαρμογές του* r_1 *και στη συνέχεια μία του* r_2, *θα είναι δηλαδή της μορφής* $d = r_1^n r_2$, *οπότε κάθε λέξη που παράγεται από τη γραμματική θα έχει τη μορφή* $a^n b a^n$ *για κάποιο* $n \in \mathbb{N}$. *Οπότε τελικά θα έχουμε*

$$L(G) = \{a^n b a^n \mid n \in \mathbb{N}\}.$$

Αξίζει να παρατηρήσουμε ότι κάθε λέξη της $L(G)$ *έχει μία και μόνο παραγωγή, οπότε η γραμματική είναι σαφής.*

Παράδειγμα 3.2. *Θεωρούμε τη γραμματική*

$$G = (\Sigma, V, S, R)$$

με αλφάβητο $\Sigma = \{a, b\}$, *μία μόνο μεταβλητή* $V = \{S\}$ *και σύνολο κανόνων* $R = \{r_1 = S \to aS,\ r_2 = S \to ba,\ r_3 = S \to aba,\ r_4 = S \to aaba\}$. *Εργαζόμενοι ανάλογα με πριν, είναι εύκολο να δούμε ότι η γλώσσα που παράγει η* G *είναι η* $L(G) = a^*ba$. *Συγκεκριμένα, η λέξη* $w_0 = ba$ *έχει μόνο μία παραγωγή, την* $d_0 = r_2$. *Ωστόσο, όλες οι υπόλοιπες λέξεις έχουν παραπάνω από μία παραγωγές. Η λέξη* $w_1 = aba$ *έχει δύο παραγωγές,* $d_{1,1} = r_1 r_2$ *και* $d_{1,2} = r_3$. *Όλες οι λέξεις της μορφής* $a^n ba$ *με* $n \geq 2$ *έχουν ακριβώς 3 παραγωγές,* $d_{n,1} = r_1^n r_2$, $d_{n,2} = r_1^{n-1} r_3$ *και* $d_{n,3} = r_1^{n-2} r_4$. *Έχουμε λοιπόν μια ασαφή γραμματική.*

Παράδειγμα 3.3. *Θεωρούμε τη γραμματική*

$$G = (\Sigma, V, S, R)$$

με $\Sigma = \{a\},\ V = \{S, A\},\ R = \{r_1 = S \to SA,\ r_2 = S \to a,\ r_3 = A \to \varepsilon\}$. *Μπορούμε να ξεκινήσουμε με οποιονδήποτε από τους κανόνες* r_1, r_2. *Ο δεύτερος τελειώνει την παραγωγή, ενώ ο πρώτος προσθέτει μία μεταβλητή* A, *η οποία θα εξαφανιστεί στη συνέχεια με μία εφαρμογή του κανόνα* r_3. *Βλέπουμε λοιπόν ότι η γραμματική παράγει μόνο μία λέξη,* $w = a$, *η οποία όμως έχει άπειρες παραγωγές, που είναι όλες της μορφής* $d_n = r_1^n r_2 r_3^n$.

Παράδειγμα 3.4. *(γλωσσολογικό) Εξετάζοντας το συντακτικό μιας γλώσσας, μπορούμε εύκολα να διατυπώσουμε κάποιους κανόνες. Μια Πρόταση (Π) αποτελείται από ένα Ονοματικό Σύνολο (ΟΣ) και ένα Ρηματικό Σύνολο (ΡΣ). Στην απλούστερη περίπτωση, το ΟΣ αποτελείται από ένα Άρθρο (Α) και ένα Ουσιαστικό (Ουσ) και το ΡΣ αποτελείται από το Ρήμα (Ρ) και το αντικείμενο, που είναι ένα ακόμη ΟΣ. Έχουμε λοιπόν τους εξής κανόνες:*

$$\begin{array}{rcl} \Pi & \to & O\Sigma\ P\Sigma \\ O\Sigma & \to & A\ Ous\sigma \\ P\Sigma & \to & P\ O\Sigma \end{array}$$

Αν προσθέσουμε επίσης τους κανόνες

$$\begin{array}{rcl} A & \to & the \\ Ous\sigma & \to & man \\ Ous\sigma & \to & dog \\ P & \to & saw \end{array}$$

έχουμε τις πρώτες παραγωγές μας:

$$\begin{array}{rl} \Pi \to & O\Sigma \quad P\Sigma \\ \to & A\ Ous\sigma \quad P\Sigma \\ \to & the\ Ous\sigma \quad P\Sigma \\ \to & the\ man \quad P\Sigma \\ \to & the\ man \quad P \quad O\Sigma \\ \to & the\ man\ saw \quad O\Sigma \\ \to & the\ man\ saw\ A\ Ous\sigma \\ \to & the\ man\ saw\ the\ Ous\sigma \\ \to & the\ man\ saw\ the\ dog \end{array}$$

Την παραπάνω παραγωγή μπορούμε να παραστήσουμε και με τη χρήση παρενθέσεων:

$$\Big[_{\Pi}\Big[_{O\Sigma}\Big[_{A}the\Big]\Big[_{Ov\sigma}man\Big]\Big]\Big[_{P\Sigma}\Big[_{P}saw\Big]\Big[_{O\Sigma}[_{A}the][_{Ov\sigma}dog]\Big]\Big]\Big]$$

Μπορούμε επίσης να έχουμε παραγωγή των προτάσεων "the man saw the man", "the dog saw the man", "the dog saw the dog". Για να έχουμε πιο πλούσιες προτάσεις πρέπει να προσθέσουμε κι άλλους κανόνες. Ένα πιο σύνθετο ΡΣ μπορεί να περιέχει, εκτός από το Ρ και το ΟΣ, έναν Εμπρόθετο Προσδιορισμό (ΕΠ). Το ίδιο ισχύει και για ένα ΟΣ: μπορεί να αποτελείται από Α, Ουσ και ΕΠ. Ο ΕΠ αποτελείται από μια πρόθεση (Προθ) και ένα ΟΣ. Έχουμε λοιπόν τους επιπλέον κανόνες:

$$\begin{array}{rcl} P\Sigma & \rightarrow & P\ O\Sigma\ E\Pi \\ O\Sigma & \rightarrow & A\ Ov\sigma\ E\Pi \\ E\Pi & \rightarrow & \Pi\rho o\theta\ O\Sigma \end{array}$$

Αν προσθέσουμε επίσης τους κανόνες:

$$\begin{array}{rcl} \Pi\rho o\theta & \rightarrow & with \\ Ov\sigma & \rightarrow & telescope \end{array}$$

έχουμε πλέον τη δυνατότητα να παράγουμε πιο περίπλοκες προτάσεις, όπως για παράδειγμα:[1]

$$\Big[_{\Pi}\Big[_{O\Sigma}\Big[_{A}the\Big]\Big[_{Ov\sigma}man\Big]\Big]\Big[_{P\Sigma}\Big[_{P}saw\Big]\Big[_{O\Sigma}[_{A}the][_{Ov\sigma}dog]\Big]\Big[_{E\Pi}[_{\Pi\rho o\theta}with][_{O\Sigma}(_{A}the)(_{Ov\sigma}telescope)]\Big]\Big]\Big]$$

ή και

$$\Big[_{\Pi}\Big[_{O\Sigma}\Big[_{A}the\Big]\Big[_{Ov\sigma}man\Big]\Big]\Big[_{P\Sigma}\Big[_{P}saw\Big]\Big[_{O\Sigma}[_{A}the][_{Ov\sigma}dog][_{E\Pi}[\Pi_{\rho o\theta}with][_{O\Sigma}(_{A}the)(_{Ov\sigma}telescope)]]\Big]\Big]\Big]$$

[1] Παρατηρούμε επίσης ότι λόγω του "κύκλου" ΟΣ→Α Ουσ ΕΠ, ΕΠ→Προθ ΟΣ η γραμματική μας μπορεί να παράγει οσοδήποτε μεγάλες προτάσεις, χωρίς απαραίτητα να έχουν νόημα, όπως για παράδειγμα "the man saw the dog with the telescope with the dog with the telescope". Αυτή η "δυνατότητα παρεκτροπής" είναι που δεν μας έχει επιτρέψει ακόμα να περιγράψουμε το συντακτικό μιας γλώσσας με χρήση γραμματικών χωρίς συμφραζόμενα: για να αντιμετωπιστεί το συγκεκριμένο πρόβλημα πρέπει να εισάγουμε περισσότερους και πιο εξειδικευμένους κανόνες, οπότε αφενός αυξάνεται η πολυπλοκότητα και αφετέρου δημιουργούνται νέα προβλήματα.

Παρατηρούμε ότι οι δύο παραγωγές, αν και διαφορετικές μεταξύ τους, έχουν το ίδιο αποτέλεσμα. Έχουμε λοιπόν μια ασαφή γραμματική. Μάλιστα, κάθε μία έχει και διαφορετικό νόημα: Στην πρώτη ο άνθρωπος είδε το σκύλο μέσα από το τηλεσκόπιο, ενώ στη δεύτερη είδε ένα σκύλο που είχε τηλεσκόπιο. [2]

Είδαμε ότι η χρήση παρενθέσεων μας βοηθά στο να παραστήσουμε πλήρως μια παραγωγή και να "ερμηνεύσουμε" το νόημα της παραγόμενης λέξης (πρότασης). Στην πραγματικότητα μάλιστα, οι "σωστά" τοποθετημένες παρενθέσεις αποτελούν μια γλώσσα χωρίς συμφραζόμενα, η οποία μάλιστα δομεί, υπό μία έννοια, όλες τις γλώσσες χωρίς συμφραζόμενα. Αυτό μας λέει το Θεώρημα Chomsky-Schützenberger, για την απόδειξη του οποίου όμως χρειαζόμαστε μερικές έννοιες ακόμα, με τις οποίες θα ασχοληθούμε στη συνέχεια.

3.3 Γλώσσες Παρενθέσεων

Αρχικά, θα ασχοληθούμε με τις γλώσσες από αλφάβητα που αποτελούνται από παρενθέσεις. Στην απλούστερη μορφή, μπορούμε να έχουμε ένα μόνο ζεύγος παρενθέσεων, δηλαδή το αλφάβητό μας να είναι το σύνολο $\Sigma = \Big\{ (\ ,\ ) \Big\}$, αλλά μπορεί να αποτελείται και από περισσότερα ζεύγη, όπως για παράδειγμα $\Big\{ (\ ,\ ), \{\ ,\ \}, [\ ,\ ] \Big\}$ ή ακόμα και $\Big\{ \underset{1}{(\ ,\ )}, \ \ldots\ , \underset{n}{(\ ,\ )} \Big\}$ κλπ.

Πώς όμως μπορούμε να ορίσουμε την έννοια "σωστά τοποθετημένες παρενθέσεις"; Καταρχάς θα πρέπει κάθε παρένθεση που ανοίγει κάποια στιγμή να κλείνει, καθώς και να μην κλείνει παρένθεση που δεν έχει ανοίξει προηγουμένως. Επίσης, στο εσωτερικό της να μην ανοίγει κάποια παρένθεση που να μην κλείνει, ακόμα και αν αυτό συμβαίνει αργότερα.

Για ευκολία, πολλές φορές συμβολίζουμε με κάποιο γράμμα a μια αριστερή παρένθεση και την αντίστοιχη δεξιά με $\bar{a}$. Ο συμβολισμός αυτός παραπέμπει

[2] Κάποιος θα μπορούσε να πει ότι ο ομιλητής είναι απίθανο να είχε στο μυαλό του τη δεύτερη πρόταση, καθώς οι σκύλοι συνήθως δεν κατέχουν τηλεσκόπια. Στο επιχείρημα αυτό, θα μπορούσαμε να απαντήσουμε απλώς αλλάζοντας τη λέξη *dog* με τη λέξη *girl*, οπότε και θα είχαμε δύο απόλυτα σωστές, τόσο συντακτικά όσο και εννοιολογικά, ίδιες προτάσεις με διαφορετικό νόημα.

Ο ίδιος ο Chomsky, για να διακρίνει την έννοια του συντακτικού από αυτή της σημασίας, είχε σχολιάσει σχετικά ότι μια πρόταση μπορεί να είναι απόλυτα σωστή γραμματικά (συντακτικά), αλλά να μην έχει μεγάλο εννοιολογικό ενδιαφέρον. Έδωσε μάλιστα το παράδειγμα, που έκτοτε έγινε κλασσικό στη βιβλιογραφία, "colorless green ideas sleep furiously"· μια πρόταση η οποία αν και έχει μια εύληπτη συντακτική δομή, ωστόσο δεν βγάζει νόημα και συνεπώς είναι τουλάχιστον απίθανο να την ξεστομίσει κάποιος.

στην Άλγεβρα, και, όπως θα δούμε παρακάτω, η έννοια των "σωστά τοποθετημένων παρενθέσεων" είναι πράγματι αλγεβρική.
Δεδομένου ενός αλφαβήτου παρενθέσεων Σ, το σύνολο όλων των "σωστά τοποθετημένων παρενθέσεων" αποτελεί την αντίστοιχη γλώσσα $Dyck$, και συμβολίζεται με D_Σ. Έστω η γραμματική $G = (\Sigma, V, S, R)$ όπου Σ ένα αλφάβητο παρενθέσεων, $V = \{S\}$ και $R = \{S \to aS\bar{a} \mid a \in \Sigma\} \cup \{S \to SS, S \to \varepsilon\}$. Είναι εύκολο να δούμε ότι η G παράγει μόνο σωστές ακολουθίες παρενθέσεων και ότι κάθε τέτοια μπορεί να παραχθεί από την G.

Η γλώσσα αυτή μπορεί όμως να παραχθεί και αλγεβρικά.
Έστω X ένα αλφάβητο. Θεωρούμε ένα δεύτερο αλφάβητο, $\widehat{X} = \{\hat{x} \mid x \in X\}$, και έστω $\Sigma = X \cup \widehat{X}$. Τα γράμματα του X θα παίζουν το ρόλο των αριστερών παρενθέσεων και τα γράμματα του $\widehat{X}$ των δεξιών. Έστω $u, v \in \Sigma^*$. Θα γράφουμε $u \sim v$ αν υπάρχουν $w_1, w_2 \in \Sigma^*$, $x \in X$ τέτοια ώστε $u = w_1 x \hat{x} w_2$ και $v = w_1 w_2$. Για κάθε $k \in \mathbb{N}$, θα γράφουμε $u \sim^k v$ αν υπάρχουν $u_1, \ldots, u_k \in \Sigma^*$ έτσι ώστε $u \sim u_1 \sim \cdots \sim u_k = v$. Τέλος, θα γράφουμε $u \sim^* v$ αν υπάρχει $k \in \mathbb{N}$ έτσι ώστε $u \sim^k v$.

Θεωρούμε τη σχέση $\mathcal{R}$ στο Σ^* η οποία ορίζεται ως εξής: $u\mathcal{R}v$ αν και μόνο αν υπάρχει $w \in \Sigma^*$ τέτοιο ώστε $u \sim^* w$ και $v \sim^* w$. Εύκολα αποδεικνύεται ότι η $\mathcal{R}$ είναι σχέση ισοδυναμίας, άρα μπορεί να οριστεί το σύνολο πηλίκο $\Sigma^*/\mathcal{R}$. Μελετώντας τον τρόπο κατασκευής της $\mathcal{R}$, η κλάση της κενής λέξης ε είναι ακριβώς η γλώσσα του $Dyck$ από το αλφάβητο Σ, δηλαδή η D_Σ.

3.4 Ειδικές μορφές γραμματικών

Πολλές φορές είναι χρήσιμο οι γραμματικές, και συνήθως οι κανόνες της, να έχουν μια συγκεκριμένη μορφή. Παρακάτω παρουσιάζουμε τα σημαντικότερα σχετικά αποτελέσματα, χωρίς αποδείξεις. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο [1].

3.4.1 Απαλοιφή κανόνων

Κατά τα συνήθη, έστω $G = (\Sigma, V, S, R)$ μια γραμματική. Μια μεταβλητή $A \in V$ θα λέγεται *χρήσιμη* αν διέρχεται από αυτήν μια επιτυχής παραγωγή.

Πρόταση 3.1. *Από κάθε γραμματική χωρίς συμφραζόμενα μπορούμε να κατασκευάσουμε μια ισοδύναμή της με όλες τις μεταβλητές χρήσιμες.*

Μπορούμε λοιπόν στο εξής, όταν αναφερόμαστε σε μια γραμματική να θεωρούμε ότι όλες οι μεταβλητές της είναι χρήσιμες.

Ένας κανόνας της μορφής $A \to \varepsilon$ λέγεται ε-κανόνας.

Πρόταση 3.2. *Για κάθε γραμματική χωρίς συμφραζόμενα G μπορούμε να κατασκευάσουμε μια άλλη, απαλλαγμένη από ε-κανόνες, που να παράγει τη γλώσσα $L(G) - \{\varepsilon\}$.*

Προφανώς αν $\varepsilon \notin L(G)$ τότε οι δύο γραμματικές είναι ισοδύναμες.
Ένας κανόνας της μορφής $A \to B$, $A, B \in V$ λέγεται μοναδιαίος κανόνας.

Πρόταση 3.3. *Για κάθε γραμματική χωρίς συμφραζόμενα G που στερείται ε-κανόνων, μπορούμε να κατασκευάσουμε μια άλλη ισοδύναμη, απαλλαγμένη από μοναδιαίους κανόνες.*

Συνοψίζοντας τα παραπάνω, έχουμε ότι κάθε γλώσσα χωρίς συμφραζόμενα που δεν περιέχει την κενή λέξη, μπορεί να παραχθεί από μια γραμματική χωρίς συμφραζόμενα η οποία στερείται άχρηστων μεταβλητών, ε- και μοναδιαίων κανόνων.

3.4.2 Κανονική μορφή Chomsky

Θα λέμε ότι η γραμματική G είναι σε κανονική μορφή $Chomsky$ αν οι κανόνες της είναι της μορφής

$$A \to BC \text{ ή } A \to a$$

όπου $A, B, C \in V,\ a \in \Sigma$.

Είναι γνωστό ότι κάθε γραμματική που δεν έχει ε-κανόνες είναι ισοδύναμη με μια άλλη σε κανονική μορφή $Chomsky$.

3.4.3 Κανονική μορφή ε-Greibach

Θα λέμε ότι η γραμματική G είναι σε κανονική μορφή $\varepsilon - Greibach$ αν οι κανόνες της είναι της μορφής

$$A \to aB$$

όπου $B \in V^*,\ a \in \Sigma \cup \{\varepsilon\}$.

Λήμμα 3.1. *Κάθε γραμματική είναι ισοδύναμη με μια άλλη σε κανονική μορφή $\varepsilon - Greibach$. Αν επιπλέον η αρχική γραμματική είναι σαφής, το ίδιο θα ισχύει και για την τελική.*

Απόδειξη. Έστω $G = (\Sigma, V, S, R)$ μια γραμματική. Για κάθε γράμμα $\sigma \in \Sigma$ θεωρούμε τη μεταβλητή A_σ και ορίζουμε $V' = V \cup \{A_\sigma \mid \sigma \in \Sigma\}$.
Για κάθε $\rho \in R$ ορίζουμε:

–Αν $\rho = A \to \varepsilon$, $\rho' = A \to \varepsilon$.

–Αν $\rho = A \to aW$ όπου $a \in \Sigma \cup \{\varepsilon\}$ και $W \in (V \cup \Sigma)^*$, $\rho' = aW'$ όπου το W' προκύπτει από το W αντικαθιστώντας κάθε γράμμα σ με την αντίστοιχη μεταβλητή A_σ.

Ορίζουμε επίσης τους κανόνες $\rho_\sigma = A_\sigma \to \sigma$, και στη συνέχεια θέτουμε $R' = \{\rho' \mid \rho \in R\} \cup \{\rho_\sigma \mid \sigma \in \Sigma\}$.

Θεωρούμε τη γραμματική $G' = (\Sigma, V', S, R')$ η οποία είναι σε $\varepsilon - Greibach$ κανονική μορφή. Για κάθε $w \in \Sigma^*$ κάθε παραγωγή της G για την w αντιστοιχεί μοναδικά σε μία παραγωγή της G' και αντίστροφα. Είναι λοιπόν προφανές ότι $L(G) = L(G')$ και η γραμματική που κατασκευάσαμε είναι ισοδύναμη με την αρχική. Επιπλέον, λόγω της αντιστοιχίας των αριστερότερων παραγωγών, αν η αρχική γραμματική είναι σαφής, το ίδιο θα ισχύει και για την τελική. □

3.5 Ιδιότητες γλωσσών χωρίς συμφραζόμενα

Με εύκολες κατασκευές αποδεικνύεται ότι η ένωση και η παράθεση δύο γλωσσών χωρίς συμφραζόμενα είναι γλώσσα χωρίς συμφραζόμενα. Το ίδιο ισχύει και για τη θήκη μιας γλώσσας χωρίς συμφραζόμενα. Ισχύουν επίσης οι παρακάτω προτάσεις:

Πρόταση 3.4. *Έστω X, Y αλφάβητα, $h : X^* \to Y^*$ μορφισμός, και L γλώσσα χωρίς συμφραζόμενα. Τότε η εικόνα της L μέσω του h, $h(L)$ είναι γλώσσα χωρίς συμφραζόμενα.*

Πρόταση 3.5. *Έστω L γλώσσα χωρίς συμφραζόμενα και R αναγνωρίσιμη γλώσσα. Τότε η τομή τους $L \cap R$ είναι γλώσσα χωρίς συμφραζόμενα.*

3.6 Αλγεβρικές Γραμματικές

Αν και ο ορισμός που έχει επικρατήσει για τις γραμματικές είναι αυτός που δόθηκε στην παράγραφο 3.2, όταν οι Chomsky και Schützenberger όριζαν τις γραμματικές χωρίς συμφραζόμενα [2], ο ορισμός που έδωσαν ήταν πιο αλγεβρικός. Συγκεκριμένα, οι κανόνες $r \in R$ μπορούν να ιδωθούν ως εξισώσεις, και η εύρεση της γλώσσας που ορίζεται από μία γραμματική ανάγεται στην επίλυση του συστήματος των εξισώσεων. Για το λόγο αυτό, οι γραμματικές χωρίς συμφραζόμενα συναντώνται συχνά στην ελληνική βιβλιογραφία με τον όρο *αλγεβρικές γραμματικές*.[3]

[3] Αν και στο [2] οι έννοιες algebraic και context-free γραμματικών διαχωρίζονται, στην ελληνική βιβλιογραφία η λέξη αλγεβρικές περιγράφει ακριβώς τις γραμματικές χωρίς συμφραζόμενα.

Κεφάλαιο 4

Θεώρημα Chomsky-Schützenberger

Έχοντας πλέον όλα τα εργαλεία που χρειαζόμαστε, διατυπώνουμε το περίφημο θεώρημα και στη συνέχεια παραθέτουμε μια απόδειξή του. Η πορεία της απόδειξης ακολουθεί αυτή που δίνεται στο [10].

Θεώρημα 4.1. *Για κάθε γλώσσα χωρίς συμφραζόμενα L υπάρχουν αλφάβητο Γ, μια αναγνωρίσιμη γλώσσα R με γράμματα από το Γ και μορφισμός h έτσι ώστε $L = h(D_\Gamma \cap R)$.*

Απόδειξη. Έστω $G = (\Sigma, V, S, R)$ μια γραμματική σε κανονική μορφή Chomsky που να παράγει την L με $R = \{\pi_1, \ldots, \pi_k\}$. Για κάθε $\pi \in R$ ορίζουμε

$$\pi' = \begin{cases} A \to \overset{1}{\underset{\pi}{(}} B \overset{1}{\underset{\pi}{)}} \overset{2}{\underset{\pi}{(}} C \overset{2}{\underset{\pi}{)}}, & \text{αν } \pi = A \to BC \\ A \to \overset{1}{\underset{\pi}{(}} \overset{1}{\underset{\pi}{)}} \overset{2}{\underset{\pi}{(}} \overset{2}{\underset{\pi}{)}}, & \text{αν } \pi = A \to \alpha \end{cases}$$

και θεωρούμε τη γραμματική $G' = (\Gamma, V, S, R')$,
όπου $\Gamma = \left\{ \overset{1}{\underset{\pi}{(}}, \overset{1}{\underset{\pi}{)}}, \overset{2}{\underset{\pi}{(}}, \overset{2}{\underset{\pi}{)}} \mid \pi \in R \right\}$ και $R' = \{\pi' \mid \pi \in R\}$.

Προφανώς $L(G') \subseteq D_\Gamma$, καθώς κάθε λέξη που παράγεται από την G' αποτελείται από "σωστά" τοποθετημένες παρενθέσεις. Επίσης, κάθε λέξη που παράγεται από την G' αντιστοιχεί σε μια λέξη της G, περιλαμβάνοντας και τον τρόπο που σχηματίστηκε. Μπορούμε λοιπόν εύκολα να "γυρίσουμε" στην L: Θεωρούμε το μορφισμό $h\colon \Gamma^* \to \Sigma^*$ που ορίζεται ως εξής:

για κάθε κανόνα $\pi \in R$ της μορφής $A \to BC$ θέτουμε

$$h\left(\overset{1}{\underset{\pi}{(}}\right) = h\left(\overset{1}{\underset{\pi}{)}}\right) = h\left(\overset{2}{\underset{\pi}{(}}\right) = h\left(\overset{2}{\underset{\pi}{)}}\right) = \varepsilon$$

για κάθε κανόνα $\pi \in R$ της μορφής $A \to \alpha$ θέτουμε

$$h\left(\overset{1}{\underset{\pi}{)}}\right) = h\left(\overset{2}{\underset{\pi}{(}}\right) = h\left(\overset{2}{\underset{\pi}{)}}\right) = \varepsilon \text{ και } h\left(\overset{1}{\underset{\pi}{(}}\right) = \alpha$$

Οπότε έχουμε ότι $h\left(L(G')\right) = L(G) = L$.
Μένει μόνο να δείξουμε ότι υπάρχει R αναγνωρίσιμη ώστε $L(G') = D_\Gamma \cap R$.
Θεωρούμε τις παρακάτω αναγνωρίσιμες γλώσσες με λέξεις από το Γ^*:

(i) Κάθε $\overset{1}{\underset{\pi}{)}}$ ακολουθείται από ένα $\overset{2}{\underset{\pi}{(}}$.
Ένα αυτόματο που αναγνωρίζει ακριβώς την παραπάνω γλώσσα είναι το $\mathcal{A} = (Q, \Gamma, \delta, q_0, \{q_0\})$ με $Q = \{q_0, q_1, \ldots, q_k\}$ και

$$\delta = \left\{ (q_0, a, q_0) \mid a \in \left\{ \overset{1}{\underset{\pi_i}{(}}, \overset{2}{\underset{\pi_i}{(}}, \overset{2}{\underset{\pi_i}{)}} \mid i = 1, 2, \ldots, k \right\} \right\}$$

$$\cup \left\{ \left(q_0, \overset{1}{\underset{\pi_i}{)}}, q_i \right) \mid i = 1, \ldots, k \right\} \cup \left\{ \left(q_i, \overset{2}{\underset{\pi_i}{(}}, q_0 \right) \mid i = 1, \ldots, k \right\}.$$

(ii) Κανένα $\overset{2}{\underset{\pi}{)}}$ δεν ακολουθείται από αριστερή παρένθεση.
Ένα αυτόματο που αναγνωρίζει αυτή τη γλώσσα είναι το $\mathcal{A} = (Q, \Gamma, \delta, q_0, \{q_0, q_1\})$ όπου $Q = \{q_0, q_1\}$ και

$$\delta = \left\{ (q_0, a, q_0) \mid a \in \left\{ \overset{1}{\underset{\pi_i}{(}}, \overset{1}{\underset{\pi_i}{)}}, \overset{2}{\underset{\pi_i}{(}} \mid i = 1, 2, \ldots, k \right\} \right\}$$

$$\cup \left\{ \left(q_0, \overset{2}{\underset{\pi_i}{)}}, q_1 \right), \left(q_1, \overset{1}{\underset{\pi_i}{)}}, q_0 \right), \left(q_1, \overset{2}{\underset{\pi_i}{)}}, q_1 \right) \mid i = 1, \ldots, k \right\}.$$

(iii) Αν $\pi = A \to BC$ τότε κάθε $\overset{1}{\underset{\pi}{(}}$ ακολουθείται από ένα $\overset{1}{\underset{\rho}{(}}$ για κάποιο $\rho \in R$ με B στην αριστερή μεριά και κάθε $\overset{2}{\underset{\pi}{(}}$ ακολουθείται από ένα $\overset{1}{\underset{\sigma}{(}}$ για κάποιο $\sigma \in R$ με C στην αριστερή μεριά, ενώ αν $\pi = A \to \alpha$ κάθε $\overset{1}{\underset{\pi}{(}}$ ακολουθείται από $\overset{1}{\underset{\pi}{)}}$ και κάθε $\overset{2}{\underset{\pi}{(}}$ από ένα $\overset{2}{\underset{\pi}{)}}$.

Αυτή είναι η πιο απαιτητική και περίπλοκη ιδιότητα, οπότε είναι αναμενόμενο να απαιτεί ένα πιο περίπλοκο αυτόματο. Μπορούμε να ορίσουμε $\mathcal{A} = (Q, \Gamma, \delta, q_0, \{q_0\})$ όπου $Q = \{q_0\} \cup \{q_{i,1}, q_{i,2} \mid i = 1, \ldots, k\}$

$$\delta = \left\{ (q_0, a, q_0) \mid a \in \left\{ \overset{1}{\underset{\pi}{)}}, \overset{2}{\underset{\pi}{)}} \mid \ \pi \in R \right\} \right\}$$
$$\cup \left\{ \left(q_0, \overset{1}{\underset{\pi_i}{(}}, q_{i,1} \right), \left(q_{i,1}, \overset{1}{\underset{\pi_j}{(}}, q_{j,1} \right) \mid \pi_i = A \to BC, \pi_j = B \to DE \text{ ή } \pi_j = B \to a \right\}$$
$$\cup \left\{ \left(q_0, \overset{2}{\underset{\pi_i}{(}}, q_{i,2} \right), \left(q_{i,2}, \overset{1}{\underset{\pi_j}{(}}, q_{j,1} \right) \mid \pi_i = A \to BC, \pi_j = C \to DE \text{ ή } \pi_j = C \to a \right\}$$
$$\cup \left\{ \left(q_0, \overset{k}{\underset{\pi_i}{(}}, q_{i,k} \right), \left(q_{i,k}, \overset{k}{\underset{\pi_i}{)}}, q_0 \right) \mid \pi_i = A \to a, \ k = 1, 2 \right\}.$$

(iv$_A$) Η λέξη ξεκινά με $\overset{1}{\underset{\pi}{(}}$ για κάποιο $\pi = A \to BC$ ή $\pi = A \to \alpha$. Σύμφωνα με τα προηγούμενα, αρκεί να ορίσουμε $\mathcal{A} = (\{q_0, q_1\}, \Gamma, \delta_A, q_0, \{q_1\})$ όπου

$$\delta_A = \left\{ \left(q_0, \overset{1}{\underset{\pi}{(}}, q_1 \right) \mid \pi = A \to BC \text{ ή } \pi = A \to \alpha \right\} \cup \{(q_1, a, q_1) \mid a \in \Gamma\}.$$

Θέτουμε R_A τη γλώσσα που ικανοποιεί όλες τις παραπάνω ιδιότητες, δηλαδή την τομή όλων των παραπάνω γλωσσών. Είναι γνωστό ότι η τομή αναγνωρισίμων γλωσσών είναι αναγνωρίσιμη. Άρα και η R_A είναι αναγνωρίσιμη.[1]

[1] Στο σημείο αυτό αξίζει να σημειωθεί το εξής: Τα παραπάνω αυτόματα που περιγράφηκαν για κάθε ιδιότητα έχουν $k+1$, 2, $2k+1$ και 2 καταστάσεις αντίστοιχα. Για την κατασκευή του αυτομάτου που αναγνωρίζει την τομή τους πρέπει αρχικά να τα κάνουμε πλήρη και στη συνέχεια να πάρουμε το καρτεσιανό γινόμενό τους. Για τα αυτόματα (i), (ii) και (iv$_A$) αρκεί να προσθέσουμε μια κατάσταση ακόμη, καθώς είναι προσδιοριστά. Το αυτόματο (iii) όμως είναι εν γένει μη προσδιοριστό, οπότε το αντίστοιχο πλήρες θα αποτελείται από 2^{2k+1} καταστάσεις. Συνεπώς το τελικό αυτόματο θα έχει $(k+2) \cdot 3 \cdot 2^{2k+1} \cdot 3 = 18 \cdot (k+2) \cdot 4^k$ καταστάσεις.

Δείχνουμε ότι $L(G') = D_\Gamma \cap R_S$.

- Έστω $w \in L(G')$. Τότε $w \in D_\Gamma$ και ικανοποιεί καθεμιά από τις ιδιότητες $(i) - (iv_S)$, άρα $w \in R_S$. Οπότε $w \in D_\Gamma \cap R_S$ και άρα $L(G') \subseteq D_\Gamma \cap R_S$.

- Θα δείξω αρχικά ότι $w \in D_\Gamma \cap R_A \Rightarrow A \underset{G'}{\overset{*}{\to}} w$. Δουλεύουμε με επαγωγή στο μήκος της w: Από την (iv_A) η w αρχίζει με $\underset{\pi}{\overset{1}{(}}$ για κάποιο $\pi \in R$ με Α στην αριστερή μεριά. Καθώς $w \in D_\Gamma$, θα εμφανίζεται ένα $\underset{\pi}{\overset{1}{)}}$, το οποίο σύμφωνα με την (i) θα ακολουθείται από ένα $\underset{\pi}{\overset{2}{(}}$. Κάπου στη συνέχεια θα εμφανίζεται ένα $\underset{\pi}{\overset{2}{)}}$. Από την (ii) δε γίνεται να υπάρχει αριστερή παρένθεση μετά το $\underset{\pi}{\overset{2}{)}}$, αλλά ούτε και δεξιά, αφού τότε δεν θα ήταν "σωστά" τοποθετημένες οι παρενθέσεις.
  Οπότε θα είναι $w = \underset{\pi}{\overset{1}{(}} y \underset{\pi}{\overset{1}{)}} \underset{\pi}{\overset{2}{(}} z \underset{\pi}{\overset{2}{)}}$ για $y, z \in \Gamma^*$. Αν $\pi = A \to BC$ τότε από την (iii) το y θα ξεκινάει με $\underset{\rho}{\overset{1}{(}}$ για κάποιο $\rho \in R$ με Β στην αριστερή μεριά και άρα θα ικανοποιεί την (iv_B) και ομοίως το z την (iv_C). Επίσης, τα y, z ικανοποιούν τις $(i), (ii), (iii)$ και αποτελούνται από "σωστά" τοποθετημένες παρενθέσεις. Οπότε $y \in D_\Gamma \cap R_B$ και $z \in D_\Gamma \cap R_C$. Από την υπόθεση της επαγωγής, $B \underset{G'}{\overset{*}{\to}} y$ και $C \underset{G'}{\overset{*}{\to}} z$. Οπότε

$$A \underset{G'}{\to} \underset{\pi}{\overset{1}{(}} B \underset{\pi}{\overset{1}{)}} \underset{\pi}{\overset{2}{(}} C \underset{\pi}{\overset{2}{)}} \underset{G'}{\overset{*}{\to}} \underset{\pi}{\overset{1}{(}} y \underset{\pi}{\overset{1}{)}} \underset{\pi}{\overset{2}{(}} z \underset{\pi}{\overset{2}{)}} = w.$$

  Αν $\pi = A \to \alpha$, τότε από την ιδιότητα (iii) θα είναι $y = z = \varepsilon$ και άρα $A \underset{G'}{\to} \underset{\pi}{\overset{1}{(}} \underset{\pi}{\overset{1}{)}} \underset{\pi}{\overset{2}{(}} \underset{\pi}{\overset{2}{)}} = w$.

Καθώς όμως όλες οι ιδιότητες είναι "τοπικές", είναι δυνατό να κατασκευάσουμε ένα αυτόματο με μόλις $4k + 1$ καταστάσεις! Πράγματι, τα τρία πρώτα αυτόματα μας δίνουν περιορισμούς σχετικά με τη συμπεριφορά μετά την ανάγνωση ενός γράμματος, ενώ το τέταρτο τη συνθήκη για έναρξη της λέξης. Συνεπώς αρκεί μια κατάσταση για κάθε γράμμα $(4k)$ και ακόμη μία για την έναρξη.

Δείξαμε λοιπόν ότι

$$w \in D_\Gamma \cap R_A \Rightarrow A \xrightarrow[G']{*} w.$$

Άρα

$$w \in D_\Gamma \cap R_S \Rightarrow S \xrightarrow[G']{*} w \Rightarrow w \in L(G')$$

και άρα

$$D_\Gamma \cap R_S \subseteq L(G').$$

Τελικά έχουμε ότι

$$L(G') = D_\Gamma \cap R_S$$

και άρα

$$h(D_\Gamma \cap R_S) = h(L(G')) = L(G) = L.$$

□

Το αντίστροφο του θεωρήματος προκύπτει εύκολα από τις ιδιότητες κλειστότητας της κλάσης των γλωσσών χωρίς συμφραζόμενα. Ισχύει δηλαδή:

Πρόταση 4.1. *Έστω αλφάβητο παρενθέσεων Γ, αναγνωρίσιμη γλώσσα R και μορφισμός $h\colon \Gamma^* \to \Sigma^*$. Τότε η γλώσσα $L = h(D_\Gamma \cap R)$ είναι χωρίς συμφραζόμενα.*

Κεφάλαιο 5

Γραμματικές με βάρη

5.1 Εισαγωγή

Οι γραμματικές χωρίς συμφραζόμενα μας δίνουν ένα μοντέλο παραγωγής λέξεων με περισσότερες δυνατότητες σε σχέση με τα αυτόματα. Θα θέλαμε ωστόσο να μπορούμε σε κάθε λέξη να αντιστοιχούμε ένα βάρος, το οποίο να συμβολίζει χρονικό κόστος, απαιτούμενους πόρους, κέρδος, πιθανότητα, βαθμό αλήθειας κλπ. Στο κεφάλαιο αυτό μελετάμε τέτοιες δομές με βάρη από ένα unital valuation monoid K, παρουσιάζοντας κάποια σχετικά παραδείγματα και αναφέροντας κάποια αποτελέσματα. Αξίζει να σημειωθεί ότι αν θεωρήσουμε ως K τον ημιδακτύλιο του Boole, έχουμε ακριβώς τις κλασσικές γραμματικές χωρίς συμφραζόμενα. Συνεπώς, όσα συζητιούνται εδώ αποτελούν μια γενίκευση της κλασσικής θεωρίας.

5.2 Ορισμοί και παραδείγματα

Στα επόμενα θα θεωρούμε ότι το K είναι unital valuation monoid.

Ορισμός 5.1. *Αλγεβρική γραμματική με βάρη ή γραμματική χωρίς συμφραζόμενα με βάρη (WCFG) ονομάζουμε κάθε πεντάδα της μορφής*

$$G = (\Sigma, V, S, R, wt)$$

για την οποία ισχύουν τα εξής:

- *Η τετράδα $G = (\Sigma, V, S, R)$ είναι CFG*

- *Η* $wt\colon R \to K$ *είναι απεικόνιση (βάρους)*
- *Αν το* K *δεν είναι πλήρες, τότε* $D(w)$ *είναι πεπερασμένο για κάθε λέξη* $w \in L(G)$.

Η απεικόνιση wt επεκτείνεται με τον αναμενόμενο τρόπο στις παραγωγές: το *βάρος* μιας παραγωγής $d = r_1 \dots r_n$ ορίζεται να είναι

$$wt(d) = val(wt(r_1), \dots, wt(r_n)).$$

Η συμπεριφορά $\|G\|$ μιας WCFG G ορίζεται να είναι η σειρά για την οποία ισχύει $(\|G\|, w) = \sum_{d \in D(w)} wt(d)$.
Φαίνεται πλέον η χρησιμότητα της τρίτης συνθήκης του Ορισμού 5.1 :
Αν το K δεν είναι πλήρες και μια λέξη w έχει άπειρες παραγωγές, τότε η $\|G\|$ δεν θα μπορεί να οριστεί πάνω στη w.

Μια σειρά $s \in K\langle\langle \Sigma^* \rangle\rangle$ θα λέγεται *ποσοτική γλώσσα χωρίς συμφραζόμενα* (quantitative context-free language) ή απλούστερα CF *σειρά* αν υπάρχει μια WCFG G έτσι ώστε $s = \|G\|$.

Στη συνέχεια δίνουμε παραδείγματα που επεκτείνουν τα αντίστοιχα που δόθηκαν στο Κεφάλαιο 3 σε διάφορα unital valuation monoids.

Παράδειγμα 5.1. *Στο Παράδειγμα 3.1 είχαμε τη γραμματική*

$$G = (\Sigma, V, S, R)$$

με $\Sigma = \{a, b\},\ V = \{S\},\ R = \{r_1 = S \to aSa,\ r_2 = S \to b\}$. *Θεωρώντας το unital valuation monoid* $(\mathbb{R} \cup \{-\infty, \infty\}, sup, avg, -\infty, \infty)$. *μπορούμε να την επεκτείνουμε σε μια WCFG προσθέτοντας μια συνάρτηση βάρους. Μπορούμε για παράδειγμα να θέσουμε* $wt(r_1) = 2,\ wt(r_2) = 1$. *Πράγματι, οι 3 συνθήκες του Ορισμού 5.1 ικανοποιούνται, καθώς όπως είδαμε στο απλό παράδειγμα, η γραμματική είναι σαφής, και άρα* $D(w) \leq 1$ *για κάθε* $w \in \Sigma^*$. *Στην παραγωγή* $d = r_1 r_1 r_2$ *της λέξης aabaa που είχαμε δει στο παράδειγμα, αντιστοιχεί το βάρος*

$$wt(d) = val(wt(r_1), wt(r_1), wt(r_2)) = avg(2, 2, 1) = \frac{2+2+1}{3} = \frac{5}{3}.$$

Καθώς αυτή είναι η μοναδική παραγωγή της λέξης $w = aabaa$, *θα έχουμε και*

$$(\|G\|, w) = \sum_{d \in D(w)} wt(d) = \sup_{d \in D(w)} wt(d) = wt(d) = \frac{5}{3}.$$

Στο συγκεκριμένο παράδειγμα μάλιστα, μπορούμε εύκολα να μιλήσουμε γενικότερα: Είδαμε ήδη ότι κάθε λέξη που παράγεται από τη γραμματική είναι της μορφής $a^n b a^n$, και μάλιστα έχει μοναδική παραγωγή της μορφής $d = r_1^n r_2$, οπότε για τη λέξη $w = a^n b a^n$ θα έχουμε:

$$\begin{aligned}(\|G\|, w) &= \sup_{d \in D(w)} wt(d) = wt(d) \\ &= val\big(wt(r_1), \ldots, wt(r_1), wt(r_2)\big) \\ &= avg(2, \ldots, 2, 1) = \frac{2n+1}{n+1}.\end{aligned}$$

Παράδειγμα 5.2. *Θεωρούμε τη γραμματική του Παραδείγματος 3.2 με κανόνες $R = \{r_1 = S \to aS,\ r_2 = S \to ba,\ r_3 = S \to aba,\ r_4 = S \to aaba\}$, το unital valuation monoid $\big([0,1], \oplus, \cdot, 0, 1\big)$ του Παραδείγματος 1.5, όπου $a \oplus b = a + b - a \cdot b$ και τη συνάρτηση βάρους $wt(r_1) = 0.8,\ wt(r_2) = 0.5,\ wt(r_3) = 0.4,\ wt(r_4) = 0.6$. Αφού για κάθε λέξη w που παράγεται από την G ισχύει $D(w) \leq 3$, η $G' = (\Sigma, V, S, R, wt)$ θα είναι πράγματι μια $WCFG$. Μάλιστα θα είναι:*

$$\begin{aligned}wt(d_0) &= wt(r_2) = 0.5 \\ wt(d_{1,1}) &= wt(r_1) \cdot wt(r_2) = 0.8 \cdot 0.5 = 0.4 \\ wt(d_{1,2}) &= wt(r_3) = 0.4\end{aligned}$$

και για $n \geq 2$

$$\begin{aligned}wt(d_{n,1}) &= wt(r_1)^n \cdot wt(r_2) = 0.8^n \cdot 0.5 \\ wt(d_{n,2}) &= wt(r_1)^{n-1} \cdot wt(r_3) = 0.8^{n-1} \cdot 0.4 \\ wt(d_{n,3}) &= wt(r_1)^{n-2} \cdot wt(r_4) = 0.8^{n-2} \cdot 0.6\end{aligned}$$

Τελικά παίρνουμε ότι

$$
\begin{aligned}
(\|G'\|, w_0) &= \bigoplus_{d \in D(w_0)} wt(d) = wt(d_0) = 0.5 \\
(\|G'\|, w_1) &= \bigoplus_{d \in D(w_1)} wt(d) = wt(d_{1,1}) \oplus wt(d_{1,2}) \\
&= 0.4 \oplus 0.4 = 0.64 \\
(\|G'\|, w_n) &= \bigoplus_{d \in D(w_n)} wt(d) = wt(d_{n,1}) \oplus wt(d_{n,2}) \oplus wt(d_{n,3}) = \\
&= 0.8^{n-2} \cdot 1.24 - 0.8^{2n-3} \cdot 0.608 + 0.8^{3n-3} \cdot 0.12
\end{aligned}
$$

Οπότε για παράδειγμα $(\|G'\|, w_2) = 0.81504$.

Παράδειγμα 5.3. *Θεωρούμε το unital valuation monoid* $(\mathbb{N} \cup \{\infty\}, min, +, 0, \infty)$ *και τη γραμματική του Παραδείγματος 3.3,*

$$G = (\Sigma, V, S, R, wt)$$

με $V = \{S, A\}$, $R = \{r_1 = S \to SA,\ r_2 = S \to a,\ r_3 = A \to \varepsilon\}$ *και* $wt(r_1) = 2,\ wt(r_2) = 3,\ wt(r_3) = 1$. *Είδαμε ότι η* G *παράγει μόνο τη λέξη* $w = a$, *και μάλιστα με άπειρες παραγωγές της μορφής* $d_n = r_1^n r_2 r_3^n$. *Για καθεμία θα έχουμε*

$$
\begin{aligned}
wt(d_n) &= val\big(\underbrace{wt(r_1), \ldots, wt(r_1)}_{n \text{ φορές}}, wt(r_2), \underbrace{wt(r_3), \ldots, wt(r_3)}_{n \text{ φορές}}\big) \\
&= n \cdot 2 + 3 + n \cdot 1 = 3n + 3
\end{aligned}
$$

Οι άπειρες παραγωγές μπορεί να δημιουργούσαν πρόβλημα σε κάποιο άλλο unital valuation monoid, αλλά το συγκεκριμένο είναι πλήρες, οπότε ορίζεται:

$$(\|G\|, w) = \min_{d \in D(w)} wt(d) = \min_{n \in \mathbb{N}} wt(d_n) = \min_{n \in \mathbb{N}} (3n + 3) = 3$$

Αν βρισκόμασταν σε άλλο χώρο, όπως για παράδειγμα στον ημιδακτύλιο των φυσικών αριθμών, η G *δεν θα ήταν* $WCFG$, *καθώς ούτε ο χώρος θα ήταν πλήρης, ούτε το* $D(w)$ *θα ήταν πεπερασμένο για τη λέξη* $w = a$.

5.3 Ειδικές μορφές

Οι κατασκευές που χρησιμοποιούνται στην απλή περίπτωση, δεν είναι δυνατόν να εφαρμοστούν πάντα στις γραμματικές με βάρη από ένα unital valuation monoid.

Αν η πράξη val είναι προσεταιριστική, και πολύ περισσότερο αν το K είναι ημιδακτύλιος, τότε οι κατασκευές εύκολα επεκτείνονται. Στη γενική περίπτωση όμως, όταν δηλαδή η προσεταιριστικότητα απουσιάζει, δεν μπορούμε πάντα να απαλείψουμε μοναδιαίους και ε-κανόνες, αφού τότε οι παραγωγές αποκτούν διαφορετικά βάρη. Συνεπώς, μια γραμματική με βάρη δεν έχει πάντα μια ισοδύναμη σε κανονική μορφή $Chomsky$.

Το ίδιο πάντως δεν ισχύει για την κανονική μορφή $\varepsilon - Greibach$. Πράγματι, μπορούμε να συμπληρώσουμε την απόδειξη που παραθέσαμε στο Λήμμα 3.1:

Δοθείσης μιας $WCFG$ $G = (\Sigma, V, S, R, wt)$, αρχικά κάνουμε την κατασκευή του Λήμματος. Στη συνέχεια, αρκεί να θεωρήσουμε $wt(\rho') = wt(\rho)$ για κάθε $\rho \in R$ και $wt(\rho_\sigma) = 1$ για κάθε $\sigma \in \Sigma$. Σε κάθε παραγωγή $d = \rho_1 \dots \rho_n$ της αρχικής γραμματικής G αντιστοιχεί μια d' της G', η οποία διαφέρει μόνο στο ότι περιέχει τους κανόνες ρ'_i αντί για τους ρ_i και μάλιστα στην ίδια (σχετική) σειρά, καθώς επίσης και (πιθανόν) κάποιους έξτρα κανόνες ρ_σ, οι οποίοι όμως έχουν βάρος 1. Από τον ορισμό της συνάρτησης val, όλα τα βάρη που είναι ίσα με 1 εξαιρούνται από τον υπολογισμό, οπότε τελικά έχουμε:

$$wt(d') = val(wt(\rho'_1), \dots, wt(\rho'_n)) = val(wt(\rho_1), \dots, wt(\rho_n)) = wt(d)$$

και άρα η γραμματική που κατασκευάσαμε είναι πράγματι μια $WCFG$ ισοδύναμη με την αρχική.

5.4 Στοχαστικές Γραμματικές

Μια κλάση WCFG με ιδιαίτερο ενδιαφέρον είναι οι *στοχαστικές γραμματικές*. Στις στοχαστικές γραμματικές τα βάρη των μεταβάσεων αντιστοιχούν σε πιθανότητες κι έτσι υπολογίζεται πόσο πιθανή είναι μια παραγωγή. Με αυτό τον τρόπο, μπορούμε σε μια ασαφή γραμματική όπως στο Παράδειγμα 3.4 να βρούμε ποια παραγωγή (και άρα ποιο νόημα!) είναι πιθανότερο να έχει συμβεί για μια συγκεκριμένη λέξη. Η κλάση αυτή αποτελεί την πιο πολυμελετημένη κλάση WCFG, καθώς έχει άμεσες εφαρμογές μεταξύ άλλων στην ανάλυση φυσικών γλωσσών[11] και στη μελέτη κλώνων RNA και άλλων ακολουθιών στη βιολο-

γία [8]. Παρακάτω δίνουμε τον αυστηρό ορισμό, καθώς επίσης επεικτείνουμε το γλωσσολογικό μας παράδειγμα.

Ορισμός 5.2. *Στοχαστική γραμματική ονομάζεται μια WCFG πάνω από το unital valuation monoid* $(\mathbb{R}, +, \cdot, 0, 1)$ *(το οποίο βέβαια είναι σώμα) για την οποία ισχύει επίσης ότι το βάρος κάθε κανόνα αντιστοιχεί σε πιθανότητα (οπότε έχει τιμή μεταξύ 0 και 1) και ακόμα τα βάρη όλων των κανόνων με την ίδια μεταβλητή στα αριστερά αθροίζουν στη μονάδα. Αν συμβολίσουμε τους κανόνες της γραμματικής με* $r_{ij} = A_i \to w_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, s_i$ *, οι περιορισμοί γράφονται:*

- $0 \leq wt\,(r_{ij}) \leq 1$, *για κάθε* $i = 1, \ldots, k,\ j = 1, \ldots, s_i$
- $\sum_{j=1}^{s_i} wt(r_{ij}) = 1$, *για κάθε* $i = 1, \ldots, k$.

Παράδειγμα 5.4. *Στο γλωσσολογικό παράδειγμα που δώσαμε παραπάνω, μπορούμε να δώσουμε πιθανότητες σε κάθε κανόνα, που να συμφωνούν με τους περιορισμούς που δόθηκαν μόλις. Οπότε μπορούμε να έχουμε:*

$r_{11} =$	Π	$\to$	$O\Sigma\ P\Sigma$	*με βάρος*	1
$r_{21} =$	$O\Sigma$	$\to$	$A\ Ov\sigma$	*με βάρος*	$^2/_3$
$r_{22} =$	$O\Sigma$	$\to$	$A\ Ov\sigma\ E\Pi$	*με βάρος*	$^1/_3$
$r_{31} =$	$P\Sigma$	$\to$	$P\ O\Sigma$	*με βάρος*	$^1/_2$
$r_{32} =$	$P\Sigma$	$\to$	$P\ O\Sigma\ E\Pi$	*με βάρος*	$^1/_2$
$r_{41} =$	A	$\to$	the	*με βάρος*	1
$r_{51} =$	$E\Pi$	$\to$	$\Pi ro\theta\ O\Sigma$	*με βάρος*	1
$r_{61} =$	$Ov\sigma$	$\to$	man	*με βάρος*	$^1/_3$
$r_{62} =$	$Ov\sigma$	$\to$	dog	*με βάρος*	$^1/_3$
$r_{63} =$	$Ov\sigma$	$\to$	$telescope$	*με βάρος*	$^1/_3$
$r_{71} =$	P	$\to$	saw	*με βάρος*	1
$r_{81} =$	$\Pi ro\theta$	$\to$	$with$	*με βάρος*	1

Συνεπώς, για τις δύο παραγωγές για τη φράση "the man saw the dog with the telescope" που δόθηκαν στο παράδειγμα, στην πρώτη αντιστοιχεί πιθανότητα $^4/_{162} \simeq 0.025$ *ενώ στη δεύτερη* $^2/_{162} \simeq 0.0125$. *Αν θεωρήσουμε λοιπόν ότι τα βάρη που ορίσαμε έχουν κάποια σχέση με την πραγματικότητα,*[1] *έχουμε το αποτέλεσμα που μας έλεγε η διαίσθησή μας· είναι πιο πιθανό ο ομιλητής όταν έλεγε την πρόταση να είχε στο μυαλό του την πρώτη παραγωγή.*

[1] Περαιτέρω αναφορά υπερβαίνει τους σκοπούς της εργασίας. Υπάρχει μεγάλη βιβλιογραφία σχετικά με το πώς υπολογίζονται οι πιθανότητες κάθε κανόνα από ένα σύνολο δεδομένων. Ενδεικτικά αναφέρουμε τα [11, 3].

Κεφάλαιο 6

Θεώρημα Chomsky-Schützenberger για WCFG

Στην παρούσα ενότητα θα παρουσιάσουμε μια εκδοχή του Θεωρήματος των Chomsky και Schützenberger για γραμματικές με βάρη.

Θεώρημα 6.1. *Για κάθε CF σειρά s υπάρχει αλφάβητο Γ, αναγνωρίσιμη γλώσσα R και αλφαβητικός μορφισμός h έτσι ώστε*

$$s = h(D_\Gamma \cap R).$$

Στην απλή περίπτωση αρκούσε να "μετατρέψουμε" τους κανόνες και να βρούμε κατάλληλο μορφισμό. Εδώ, επειδή μας ενδιαφέρει το βάρος κάθε παραγωγής πρέπει να είμαστε πιο προσεκτικοί. Το κυριότερο πρόβλημα είναι ότι δεν μπορούμε να φέρουμε τη γραμματική σε κανονική μορφή *Chomsky*, συνεπώς καλούμαστε να ξεπεράσουμε το πρόβλημα αυτό.

Σκιαγραφόντας την απόδειξη, θα δείξουμε αρχικά ότι υπάρχει CF γλώσσα L και μορφισμός h_1 έτσι ώστε

$$s = h_1(L)$$

και στη συνέχεια από το κλασικό θεώρημα θα έχουμε ότι

$$L = h_2(D_\Gamma \cap R)$$

οπότε θα έχουμε

$$s = h_1(h_2(D_\Gamma \cap R)) = h_1 \circ h_2(D_\Gamma \cap R) = h(D_\Gamma \cap R).$$

Λήμμα 6.1. *Έστω CF σειρά s. Τότε υπάρχει αλφάβητο Δ, σαφής CFG G και αλφαβητικός μορφισμός $h\colon \Delta^* \to K\langle\langle\Sigma^*\rangle\rangle$ έτσι ώστε $s = h(L(G))$.*

Απόδειξη. Έστω $s = \|H\|$ για κάποια WCFG $H = (\Sigma, V, S, R, wt)$. Από το Λήμμα 3.1 μπορούμε να υποθέσουμε ότι η H βρίσκεται σε κανονική μορφή ε-Greibach, δηλαδή όλοι οι κανόνες της είναι της μορφής $\rho = A \to aB_1 \dots B_n$.

Για κάθε $\rho \in R$ ορίζουμε $\rho' = A \to \rho B_1 \dots B_n$ και θεωρούμε τη γραμματική $G = (R, V, S, R')$ όπου $R' = \{\rho' \mid \rho \in R\}$.

Θεωρούμε επίσης το μορφισμό $h_1\colon R^* \to K\langle\langle\Sigma^*\rangle\rangle$ που ορίζεται από τη σχέση $h(\rho) = wt(\rho).a$.

Η G είναι προφανώς σαφής. Πράγματι, για κάθε $w \in L(G)$ η ίδια η λέξη προσδιορίζει ποιοι κανόνες χρησιμοποιήθηκαν στην παραγωγή της. Διαφορετικοί κανόνες θα οδηγήσουν σε διαφορετική λέξη.

Έστω $w \in \Sigma^*$. Κάθε παραγωγή της $d = \rho_1 \dots \rho_n \in D_H(w)$ μπορούμε να τη θεωρήσουμε λέξη του R^*. Οπότε από τον ορισμό του h_1 έχουμε

$$h_1(d) = val(wt(\rho_1), \dots, wt(\rho_n)).w = wt(d).w$$

Άρα έχουμε $wt(d) = (h_1(d), w)$. Το $L(G)$ θα αποτελείται από όλες τις παραγωγές λέξεων της H. Οπότε $L(G) = \bigcup_{w\in\Sigma^*} D_H(w)$. Προφανώς θα ισχύει $D_H(w) \cap D_H(w') = \emptyset$ για $w \neq w'$, αφού δεν είναι δυνατόν μια παραγωγή να έχει δύο διαφορετικά αποτελέσματα. Οπότε για κάθε $w \in \Sigma^*$ έχουμε

$$\{d \in L(G) \mid (h(d), w) \neq 0\} \subseteq D_H(w).$$

Από τον ορισμό της WGFG, είτε το K θα είναι πλήρες, είτε το $D_H(w)$ θα είναι πεπερασμένο, άρα η οικογένεια $\{h(d) \mid d \in L(G)\}$ είναι τοπικά πεπερασμένη. Οπότε για κάθε $w \in \Sigma^*$ έχουμε

$$(\|H\|, w) = \sum_{d\in D_H(w)} wt(d) = \sum_{d\in D_H(w)} (h_1(d), w) = {}^1 \sum_{d\in L(G)} (h_1(d), w) =$$

$$= \left(\sum_{d\in L(G)} h_1(d), w\right) = (h_1(L(G), w)$$

Άρα $s = h_1(L(G))$. □

[1] Έστω $d \in L(G)\backslash D_H(w)$. Τότε $d \notin D_H(w)$ και άρα $(h_1(d), w) = 0$

Απόδειξη (του Θεωρήματος). Δεδομένων των αποτελεσμάτων του Λήμματος 6.1, από το κλασσικό θεώρημα (Θεώρημα 4.1) υπάρχουν αλφάβητο Γ, αναγνωρίσιμη γλώσσα R, μορφισμός $h_2\colon \Gamma^* \to R^*$ και γραμματική χωρίς συμφραζόμενα G' έτσι ώστε $L(G) = h_2(L(G')) = h_2(D_\Gamma \cap R)$. Έστω $v \in D_G(d)$ για $d \in L(G)$. Αφού η G είναι σαφής, θα είναι $|D_G(d)| = 1$, και άρα από παρατήρηση στην κλασσική κατασκευή θα είναι $|h_2^{-1}(d) \cap L(G')| = 1$. Είδαμε πριν ότι το σύνολο $I_w = \{d \in L(G) \mid (h_1(d), w) \neq 0\}$ είναι πεπερασμένο για κάθε $w \in \Sigma^*$. Αφού το $h_2^{-1}(d) \cap L(G')$ είναι μονοσύνολο για κάθε $d \in L(G)$, το σύνολο $\{v \in L(G') \mid (h_1(h_2(v)), w) \neq 0\}$ είναι πεπερασμένο, κι έτσι η οικογένεια $\{(h_1 \circ h_2)(v) \mid v \in L(G')\}$ είναι τοπικά πεπερασμένη. Οπότε το $(h_1 \circ h_2)(D_\Gamma \cap R)$ είναι καλά ορισμένο, και μάλιστα $s = (h_1 \circ h_2)(D_\Gamma \cap R)$. □

Παράδειγμα 6.1. *Έστω $K = (\mathbb{R}, +, \cdot, 0, 1)$ και η CF σειρά που παράγεται από τη γραμματική με βάρη $H = (\Sigma, V, S, R, wt)$ με $\Sigma = \{a, b\},\ V = \{S\},\ R = \{\rho_1, \rho_2\}$ όπου $\rho_1 = S \to aSa,\ \rho_2 = S \to b$ και $wt(\rho_1) = \frac{1}{4},\ wt(\rho_2) = \frac{3}{4}$.*

Θα εφαρμόσουμε την κατασκευή του Θεωρήματος.

Αρχικά φέρνουμε την H σε κανονική μορφή ε-Greibach, οπότε και έχουμε τους κανόνες:

$$\pi_1 = S \to aSA_a$$

$$\pi_2 = A_a \to a$$

$$\pi_3 = S \to b$$

με $wt(\pi_1) = \frac{1}{4},\ wt(\pi_2) = 1,\ wt(\pi_3) = \frac{3}{4}$.

Ορίζουμε τους κανόνες

$$\pi_1' = S \to \pi_1 S A_a$$

$$\pi_2' = A_a \to \pi_2$$

$$\pi_3' = S \to \pi_3$$

και το μορφισμό h_1 ο οποίος ορίζεται από

$$h_1(\pi_1) = \frac{1}{4}.a$$

$$h_1(\pi_2) = 1.a$$

$$h_1(\pi_3) = \frac{3}{4}.b$$

Έχουμε λοιπόν την CF γραμματική $G = (\{\pi_1, \pi_2, \pi_3\}, \{S, A_a\}, S, \{\pi'_1, \pi'_2, \pi'_3\})$ *Για να εφαρμόσουμε το κλασσικό θεώρημα πρέπει να φέρουμε τη* G *σε κανονική μορφή* $Chomsky$. *Κάνοντας την αντίστοιχη κατασκευή, έχουμε τους κανόνες:*

$$\sigma_1 = S \to X_1 X_2 \qquad \sigma_2 = X_1 \to \pi_1$$

$$\sigma_3 = X_2 \to S A_a \qquad \sigma_4 = A_a \to \pi_2$$

$$\sigma_5 = S \to \pi_3$$

Εφαρμόζοντας την κλασσική κατασκευή, έχουμε το αλφάβητο

$$\Gamma = \left\{ \underset{\sigma_i}{\overset{1}{(}}, \underset{\sigma_i}{\overset{1}{)}}, \underset{\sigma_i}{\overset{2}{(}}, \underset{\sigma_i}{\overset{2}{)}} \mid i = 1, 2, 3, 4, 5 \right\}$$

τους κανόνες

$$\sigma'_1 = S \to \underset{\sigma_1}{\overset{1}{(}} X_1 \underset{\sigma_1\sigma_1}{\overset{1}{)}} \underset{\sigma_1}{\overset{2}{(}} X_2 \underset{\sigma_1}{\overset{2}{)}} \qquad \sigma'_2 = X_1 \to \underset{\sigma_2\sigma_2\sigma_2\sigma_2}{\overset{1\,1\,2\,2}{(\,)\,(\,)}}$$

$$\sigma'_3 = X_2 \to \underset{\sigma_3}{\overset{1}{(}} S \underset{\sigma_3\sigma_3}{\overset{1}{)}} \underset{\sigma_3}{\overset{2}{(}} A_a \underset{\sigma_3}{\overset{2}{)}} \qquad \sigma'_4 = A_a \to \underset{\sigma_4\sigma_4\sigma_4\sigma_4}{\overset{1\,1\,2\,2}{(\,)\,(\,)}}$$

$$\sigma'_5 = S \to \underset{\sigma_5\sigma_5\sigma_5\sigma_5}{\overset{1\,1\,2\,2}{(\,)\,(\,)}}$$

και το μορφισμό h_2 *που ορίζεται από*

$$h_2(\underset{\sigma_2}{\overset{1}{(}}\,) = \pi_1$$

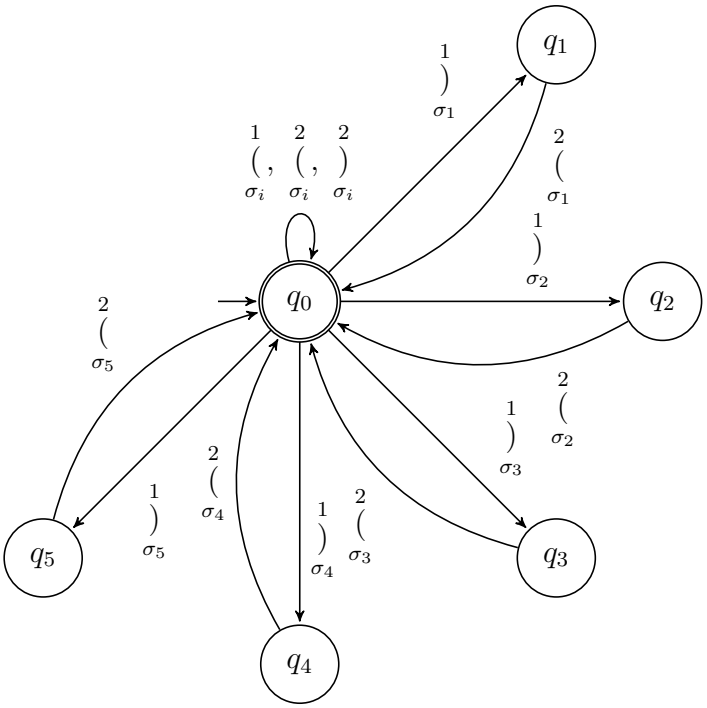
$$h_2(\underset{\sigma 4}{\overset{1}{(}}\,) = \pi_2$$

$$h_2(\underset{\sigma_5}{\overset{1}{(}}\,) = \pi_3$$

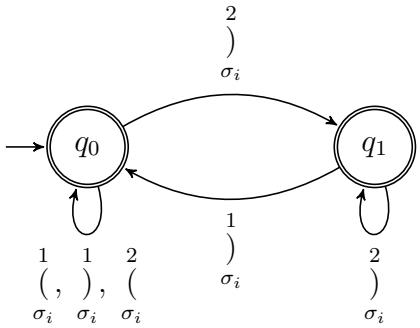
και $h_2(\gamma) = \varepsilon$ *για κάθε άλλο* $\gamma \in \Gamma$.

Σχετικά με τη γλώσσα R, *σύμφωνα με όσα συζητήθηκαν στην απόδειξη του κλασσικού Θεωρήματος, κατασκευάζουμε αυτόματα που να ικανοποιούν κάθε ιδιότητα:*

i) *Κάθε* $\overset{1}{\underset{\sigma_i}{)}}$ *ακολουθείται από* $\overset{2}{\underset{\sigma_i}{(}}$.

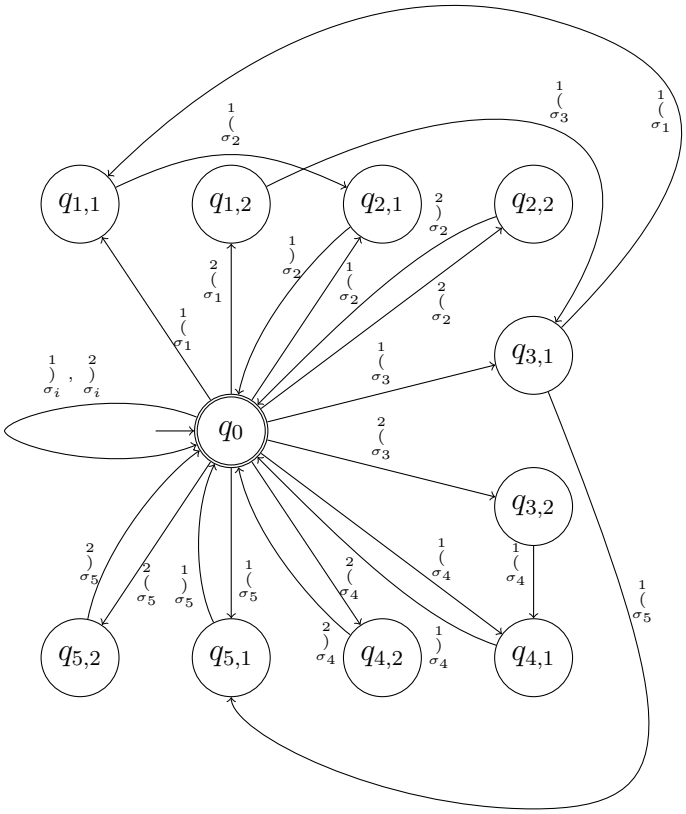


ii) *Μετά από* $\overset{2}{\underset{\sigma_i}{)}}$ *δεν υπάρχει αριστερή παρένθεση.*

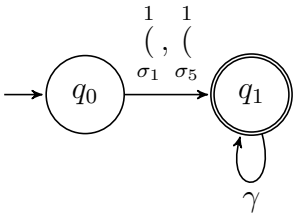


iii) *Κάθε* $\overset{1}{\underset{\sigma_1}{(}}$ *ακολουθείται από* $\overset{1}{\underset{\sigma_2}{(}}$ *και κάθε* $\overset{2}{\underset{\sigma_1}{(}}$ *ακολουθείται από* $\overset{1}{\underset{\sigma_3}{(}}$.
Κάθε $\overset{1}{\underset{\sigma_2}{(}}$ *ακολουθείται από* $\overset{1}{\underset{\sigma_2}{)}}$ *και κάθε* $\overset{2}{\underset{\sigma_2}{(}}$ *ακολουθείται από* $\overset{2}{\underset{\sigma_2}{)}}$.

$\mathit{Κάθε}\ \overset{1}{\underset{\sigma_3}{(}}\ \mathit{ακολουθείται\ από}\ \overset{1}{\underset{\sigma_1}{(}}\ \mathit{ή\ από}\ \overset{1}{\underset{\sigma_5}{(}}\ \mathit{και\ κάθε}\ \overset{2}{\underset{\sigma_3}{(}}\ \mathit{ακολουθείται\ από}\ \overset{1}{\underset{\sigma_4}{(}}.$
$\mathit{Κάθε}\ \overset{1}{\underset{\sigma_4}{(}}\ \mathit{ακολουθείται\ από}\ \overset{1}{\underset{\sigma_4}{)}}\ \mathit{και\ κάθε}\ \overset{2}{\underset{\sigma_4}{(}}\ \mathit{ακολουθείται\ από}\ \overset{2}{\underset{\sigma_4}{)}}.$
$\mathit{Κάθε}\ \overset{1}{\underset{\sigma_5}{(}}\ \mathit{ακολουθείται\ από}\ \overset{1}{\underset{\sigma_5}{)}}\ \mathit{και\ κάθε}\ \overset{2}{\underset{\sigma_5}{(}}\ \mathit{ακολουθείται\ από}\ \overset{2}{\underset{\sigma_5}{)}}.$



iv$_S$) *Η λέξη ξεκινά με* $\overset{1}{\underset{\sigma_1}{(}}$ *ή με* $\overset{1}{\underset{\sigma_5}{(}}$.



Παρατηρούμε ότι τα παραπάνω αυτόματα έχουν 6,3,11 και 2 καταστάσεις αντίστοιχα. Σύμφωνα με την κατασκευή του αυτομάτου που αναγνωρίζει την τομή δύο αυτομάτων, και άρα τη γλώσσα R που θέλουμε, χρειαζόμαστε $18 \cdot (5+2) \cdot 4^5 = 129024$ καταστάσεις! Σύμφωνα όμως με την παρατήρηση που έγινε στην κλασσικό θεώρημα, όλες οι ιδιότητες είναι δυνατό να ικανοποιηθούν από ένα αυτόματο με μόλις 21 καταστάσεις!

Τελικά έχουμε το αλφάβητο Γ, την αναγνωρίσιμη γλώσσα R και τον αλφαβητικό μορφισμό $h = h_1 \circ h_2$ που ορίζεται από

$$h\binom{1}{(}_{\sigma_2} = \frac{1}{4}.a \qquad h\binom{1}{(}_{\sigma_4} = 1.a \qquad h\binom{1}{(}_{\sigma_5} = \frac{3}{4}.b$$

$$\textit{και } h(\gamma) = 1.\varepsilon \textit{ για κάθε άλλο } \gamma \in \Gamma.$$

Το αντίστροφο του θεωρήματος είναι λίγο δυσκολότερο να αποδειχθεί απ΄ ότι στην κλασσική περίπτωση, συνεπώς παραθέτουμε την απόδειξή του. Θα αποδείξουμε μάλιστα κάτι πιο γενικό:

Θεώρημα 6.2. *Έστω L μια γλώσσα χωρίς συμφραζόμενα από το αλφάβητο Δ και αλφαβητικός μορφισμός $h\colon \Delta^* \to K\langle\langle\Sigma^*\rangle\rangle$, έτσι ώστε η οικογένεια $(h(v) \mid v \in L)$ να είναι τοπικά πεπερασμένη σε περίπτωση που το K δεν είναι πλήρες. Αν η L παράγεται από μια σαφή CFG ή το K είναι πλήρες και πλήρως ταυτοδύναμο, τότε η $h(L)$ είναι CF σειρά.*

Απόδειξη. Έστω ότι η L παράγεται από μια CF γραμματική $G = (\Sigma, V, S, R)$. Από το Λήμμα 3.1 μπορούμε να υποθέσουμε ότι η G βρίσκεται σε κανονική μορφή ε-Greibach, δηλαδή όλοι οι κανόνες της γραμματικής είναι της μορφής $\rho = A \to aB_1 \dots B_n,\ a \in \Delta \cup \{\varepsilon\}$. Για κάθε $\rho \in R$ ορίζουμε $\rho' = A \to \beta B_1 \dots B_n$ και $wt(\rho') = k$, αν $h(a) = k.\beta$. Προφανώς, αν $a = \varepsilon$ τότε θα είναι και $\beta = \varepsilon$ και $wt(\rho') = 1$. Θεωρούμε τη γραμματική $G' = (\Sigma, V, S, R', wt)$ όπου $R' = \{\rho' \mid \rho \in R\}$ και την απεικόνιση wt όπως ορίστηκε παραπάνω, και θα δείξουμε ότι είναι WCFG.

Το μόνο που έχουμε να δείξουμε είναι ότι αν το K δεν είναι πλήρες, τότε το $D_{G'}(w)$ είναι πεπερασμένο για κάθε $w \in \Sigma^*$. Αν το K δεν είναι πλήρες (άρα δεν θα είναι ούτε πλήρες και πλήρως ταυτοδύναμο), από το δεύτερο μέρος υποθέσεων απαιτούμε η L να παράγεται από μια σαφή γραμματική. Οι παραγωγές της λέξης $w \in \Sigma^*$ είναι ακριβώς όσες αντιστοιχούν στις λέξεις του συνόλου $A_w = \{v \in L \mid supp(h(v)) = \{w\}\}$. Αφού η οικογένεια $(h(v) \mid v \in L)$ είναι τοπικά πεπερασμένη και κάθε $h(v)$ είναι μονώνυμο, τότε για κάθε $w \in \Sigma^*$ το

σύνολο A_w είναι πεπερασμένο, και αφού η L παράγεται από μια σαφή CFG, οι παραγωγές της w είναι πεπερασμένες, δηλαδή το $D(w)$ είναι πεπερασμένο. Τελικά η G' είναι πράγματι μια WCFG.

Μένει λοιπόν να δειχθεί ότι η CF σειρά που παράγεται από την G' είναι ακριβώς η $h(L)$. Μας αρκεί η παρατήρηση που κάναμε παραπάνω, ότι δηλαδή κάθε παραγωγή της w από την G' αντιστοιχεί σε μια παραγωγή μιας λέξης $v \in A_w$ της G και προσδίδει στην w βάρος ίσο με $(h(v), w)$. Συνεπώς έχουμε:

$$(\|G'\|, w) = \sum_{d' \in D_{G'}(w)} wt(d') = \sum_{v \in A_w} \sum_{d \in D_G(v)} (h(v), w) \qquad (*)$$

Αν η L μπορεί να παραχθεί από μια σαφή CFG, τότε

$$(*) = \sum_{v \in A_w} \sum_{d \in \{d\}} (h(v), w) = \sum_{v \in A_w} (h(v), w).$$

Διαφορετικά, αν το K είναι πλήρες και πλήρως ταυτοδύναμο, τότε

$$\sum_{d \in D_G(v)} (h(v), w) = (h(v), w).$$

Σε κάθε περίπτωση δηλαδή, αν λάβουμε υπόψη ότι το A_w περιέχει ακριβώς τις λέξεις εκείνες της $L(G)$ που δίνουν τιμή διαφορετική του 0 στη w έχουμε:

$$\begin{aligned}(*) &= \sum_{v \in A_w} (h(v), w) = \sum_{v \in L(G)} (h(v), w) \\ &= \left(\sum_{v \in L(G)} h(v), w \right) = \big(h(L(G)), w\big) \\ &= (h(L), w).\end{aligned}$$

Δηλαδή τελικά έχουμε

$$\|G'\| = h(L).$$

□

Κεφάλαιο 7

Επίλογος

Στην εργασία αυτή παρουσιάσαμε το κλασσικό θεώρημα των Chomsky και Schützenberger, που συσχετίζει τις γλώσσες χωρίς συμφραζόμενα με τις γλώσσες του Dyck και τις αναγνωρίσιμες γλώσσες. Στη συνέχεια έγινε μια εισαγωγή στις αλγεβρικές γραμματικές με βάρη από μια γενική αλγεβρική δομή, το unital valuation monoid, και μελετήσαμε τις ποσοτικές γλώσσες χωρίς συμφραζόμενα που αυτές παράγουν, που δεν είναι τίποτα άλλο παρά τυπικές δυναμοσειρές. Τέλος αποδείχθηκε μια εκδοχή του παραπάνω θεωρήματος για τις γλώσσες αυτές.

Στην πορεία της εργασίας ανέκυψαν διάφορα ενδιαφέροντα ερωτήματα τα οποία ενδείκνυνται για μελλοντική μελέτη. Καταρχάς, η έννοια του αλφαβητικού μορφισμού που ορίστηκε για την εκδοχή του θεωρήματος με βάρη παρουσιάζει κάποιες αποκλίσεις από τον κλασσικό αλγεβρικό ορισμό του μορφισμού. Θα είχε νόημα λοιπόν να αναζητηθεί μήπως το θεώρημα μπορεί να διατυπωθεί με πιο κλασσικούς αλγεβρικούς όρους. Μεγάλη σημασία θα είχε επίσης αν θα μπορούσαμε να διατυπώσουμε μια εκδοχή του θεωρήματος για στοχαστικές γραμματικές, λόγω των πολλών τους εφαρμογών.

Βιβλιογραφία

[1] Μποζαπαλίδης, Σ. (1997). *Αυτόματα Γλώσσες Γραμματικές*. Leader Books A.E.

[2] Chomsky, N., Schützenberger, M.P. (1963). The algebraic theory of context-free languages, In: *Computer Programming and Formal Systems*, by Brafford, P., Hirschberg, D. (eds), North-Holland, pp. 118-161.

[3] Collins, M. *Probabilistic context-free grammars (PCFGs)*. Available from: <http://www.cs.columbia.edu/∼mcollins/courses/nlp2011/notes/pcfgs.pdf>. [4 November 2014].

[4] Droste, M., Meinecke, I. (2010). Describing average- and longtime-behavior by weighted MSO logics. In: *Mathematical Foundations of Computer Science 2010, Lecture Notes in Computer Science 6281*, by Hliněný, P., Kučera, A. (eds), Springer, pp. 537-548.

[5] Droste, M., Stüber, T., Vogler, H. (2010). Weighted finite automata over strong bimonoids, *Information Sciences* **180**(1) pp. 156-166.

[6] Droste, M., Meinecke, I. (2011). Weighted automata and regular expressions over valuation monoids, *International Journal of Foundations of Computer Science* **22**(8) pp. 1829-1844.

[7] Droste, M., Vogler, H. (2013). The Chomsky-Schützenberger theorem for quantitative context-free languages. In: *Developments in Language Theory, Lecture Notes in Computer Science 7907*, by Beal, M, Carton, O. (eds), Springer, pp. 203-214.

[8] Durbin, R., Eddy, S., Krogh, A., Mitchinson, G. (2006). *Biological sequence analysis*. Cambridge University Press.

[9] Hopcroft, J., Motwani, R., Ullman, J. (2006). *Introduction to automata theory, languages and computation.* Addison-Wesley Longman Publishing Co.

[10] Kozen, D. C. (1997). *Automata and computability*. Springer.

[11] Manning, C., Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

[12] Meyer, W. S., Burn R., Cotton J. S., Risley H. H. (1909). Sanskrit literature. In: *Imperial Gazetteer of India 2 (VI)*, Clarendon Press, pp. 206-269.

[13] Yu, S. (1997). Regular languages. In: *Handbook of Formal Languages 1*, by Rozenberg, G., Salomaa, A. (eds.), Springer, pp. 41-110.