Κεφάλαιο 5

Συντακτική ταξινόμηση προτύπων

5.1 Η δομή των συστημάτων συντακτικής ταξινόμησης

Στο κεφάλαιο αυτό περιγράφονται τα συντακτικά συστήματα ταξινόμησης προτύπων και δίνονται μερικές από τις σημαντικότερες εφαρμογές τους.

Κάθε γλώσσα αποτελείται από λέξεις που δομούν προτάσεις. Θα υποθέσουμε ότι λέξεις αποτελούν τα στοιχειώδη τμήματα του φαινομένου που μελετάμε. Θα υποθέσουμε επίσης ότι κάθε γλώσσα αποτελείται από ακολουθίες προτάσεων οι οποίες είναι συντακτικά ασυσχέτιστες μεταξύ τους.

5.2 Γλώσσα και γραμματική

Αρχικά δίνουμε κάποιους βασικούς ορισμούς από την θεωρία των συνόλων.

Ορισμός 4 *Αλφάβητο μιας γλώσσας A ονομάζεται το πεπερασμένο σύνολο των στοιχειωδών συμβόλων που την απαρτίζουν.*

Ορισμός 5 *Πρόταση του A είναι κάθε πεπερασμένη ακολουθία στοιχείων του A.*

Παράδειγμα 46 *Αν $A = \{\alpha, \beta, \gamma, \delta\}$ είναι το αλφάβητο μιας γλώσσας, η συμβολοσειρά ααβγγγβ είναι μία πρόταση του A.*

Ορισμός 6 *Μήκος πρότασης είναι ο αριθμός της σε σύμβολα και συμβολίζεται*
$|<$ *συμβολοσειρά* $>|$. *Το μήκος κάθε πρότασης είναι φυσικός αριθμός.*

Το μήκος της πρότασης του παραδείγματος 1 είναι $|$ ααβγγγβ $| = 8$.

Ορισμός 7 *Με λ συμβολίζεται η πρόταση με μηδενικό μήκος, δηλαδή η πρόταση που δεν περιέχει κανένα σύμβολο.*

Ορισμός 8 *Με A^* συμβολίζεται το σύνολο όλων των προτάσεων που παράγονται από το αλφάβητο A. Το σύνολο A^* περιέχει και την πρόταση μηδενικού μήκους.*

Ορισμός 9 *Με A^+ συμβολίζουμε το σύνολο των προτάσεων του A το οποίο δεν περιέχει την πρόταση μηδενικού μήκους. Ισχύει ότι: $A^+ = A^* - \{\lambda\}$.*

Ορισμός 10 *Γλώσσα L με αλφάβητο A ονομάζεται κάθε υποσύνολο του A^*.*

Αν αλφάβητο είναι το σύνολο $A = \{\alpha, \beta, \gamma, \delta\}$, γλώσσα L με αλφάβητο το A μπορεί να είναι το ακόλουθο σύνολο προτάσεων:

$$L = \{\alpha\beta, \alpha\beta\beta, \alpha\delta\gamma, \alpha\delta\delta, \gamma\delta\alpha\alpha, \beta\beta\beta\beta\beta\beta\beta\beta\beta\}.$$

Ορισμός 11 *Γραμματική G είναι κάθε διατεταγμένη τετράδα που αποτελείται από τα σύνολα $G = (V_N, V_T, S, P)$ και που το κάθε ένα περιγράφει τα ακόλουθα στοιχεία της γραμματικής:*

V_T είναι το σύνολο των τερματικών συμβόλων,

V_N είναι το σύνολο των μη τερματικών συμβόλων ή μεταβλητών. Το σύνολο αυτό δεν περιέχει τερματικά σύμβολα, δηλ. $V_T \cap V_N = \emptyset$.

S είναι η μεταβλητή έναρξης και,

P είναι ένα διατεταγμένο σύνολο από ζεύγη συμβολοσειρών (C, D) που αποτελούνται από τερματικά και μη τερματικά σύμβολα, δηλαδή $C, D \in (V_T \cup V_N)^*$ και το C περιέχει ένα τουλάχιστον μη τερματικό σύμβολο.

Τα διατεταγμένα ζεύγη (C, D) ονομάζονται και κανόνες επαναγραφής ή κανόνες παραγωγής και συνήθως παριστάνονται ως εξής: $C \to D$.

Ορισμός 12 *Η συμβολοσειρά μ παράγει άμεσα την συμβολοσειρά ν και συμβολίζουμε $\mu \underset{G}{\Rightarrow} \upsilon$ όταν ισχύει: $\mu = \alpha_1\alpha\alpha_2, \nu = \alpha_1\beta\alpha_2$ και ο κανόνας παραγωγής $\alpha \to \beta$ ανήκει στο σύνολο των κανόνων παραγωγής της γραμματικής G.*

Ορισμός 13 *Η συμβολοσειρά α_1 παράγει την συμβολοσειρά α_n ή $\alpha_1 \overset{*}{\underset{G}{\Rightarrow}} \alpha_n$ όταν το α_n παράγεται από το α_1 με διαδοχική εφαρμογή των κανόνων άμεσης παραγωγής συμβολοσειρών.*

Πρέπει δηλαδή να υπάρχουν $\alpha_i, i = 2, ..., n-1$ τέτοια ώστε $\alpha_i \underset{G}{\Rightarrow} \alpha_{i+1}$.

Ορισμός 14 *Γλώσσα μιας γραμματικής είναι το σύνολο των παραγόμενων προτάσεων από το σύμβολο εκκίνησης που αποτελούνται μόνο από τερματικά σύμβολα:*

$$L(G) = \{x \mid x \in V_T^*, S \overset{*}{\underset{G}{\Rightarrow}} x\} \tag{5.1}$$

Τα σύμβολα που βρίσκονται μέσα σε ανισότητες ή περιγράφονται με κεφαλαία γράμματα θα θεωρούμε ότι αναφέρονται σε μεταβλητές κάποιας γραμματικής. Τα πεζά σύμβολα θα θεωρούνται ότι είναι τερματικά σύμβολα.

Παράδειγμα 47 *Έστω η γραμματική $G = (\{S, A\}, \{a, b\}, S, \{S \to aA, A \to aA, A \to b\})$. Δείξε ότι η γραμματική παράγει την γλώσσα που αποτελείται από τις προτάσεις $L(G) = \{a^n b \mid n = 1, ...\}$.*

Η απόδειξη είναι πολύ εύκολη και δίνεται από την διαδοχική εφαρμογή των κανόνων παραγωγής:

$$S \Rightarrow aA \left\{ \begin{matrix} aaA \\ ab \end{matrix} \right. \Rightarrow \left\{ \begin{matrix} aaaA \\ aab \end{matrix} \right. \Rightarrow \left\{ \begin{matrix} aaaaA \\ aaab \end{matrix} \right. \cdots$$

Ορισμός 15 *Δύο γραμματικές G_1, G_2 λέγονται ισοδύναμες όταν οι αντίστοιχες γλώσσες τους είναι ίδιες, δηλαδή: $L(G_1) = L(G_2)$.*

Θεώρημα 7 *Σε κάθε γραμματική $G = (V_N, V_T, S, P)$ υπάρχει μία τουλάχιστον ισοδύναμη γραμματική $G' = (V_N', V_T, S, P')$ τέτοια ώστε στο αριστερό τμήμα των κανόνων παραγωγής της G' να μην υπάρχουν τερματικά σύμβολα.*

Η απόδειξη του θεωρήματος και η εύρεση της ισοδύναμης γραμματικής γίνεται με τον ακόλουθο τρόπο:

Δημιουργούμε μία νέα μεταβλητή για κάθε τερματικό σύμβολο που βρίσκεται στο αριστερό μέρος των κανόνων παραγωγής. Ετσι αν $a_i, i = 1, ..., N, a_i \in V_T$ είναι το σύνολο των τερματικών συμβόλων που υπάρχουν στο αριστερό τμήμα των κανόνων παραγωγής, τότε $A_i, i = 1, ..., N$, με $A_i \in V_N$ θα είναι το σύνολο των νέων μεταβλητών.

Προσθέτουμε τις νέες μεταβλητές, στο σύνολο V_N αντικαθιστούμε στους κανόνες παραγωγής τα τερματικά σύμβολα με τις νέες μεταβλητές και προσθέτουμε N νέους κανόνες παραγωγής $A_i \to a_i, i = 1, ..., N$.

Παράδειγμα 48 *Εστω η γραμματική $G = (\{A, B, S\}, \{x, y, z\}, S, P)$ με τους ακόλουθους κανόνες παραγωγής:*

$$\begin{array}{ccc} S \to xSAB & SA \to yyA & yA \to yzB \\ yS \to xB & Az \to xSAB & yyzB \to AB \\ S \to xB & A \to x & B \to xzy \end{array}$$

Βρείτε μία αντίστοιχη γραμματική G'στην οποία να μην υπάρχουν τερματικά σύμβολα στο αριστερό τμήμα των κανόνων παραγωγής.

Παρατηρούμε ότι στο αριστερό τμήμα των κανόνων παραγωγής βρίσκονται τα τερματικά σύμβολα y και z. Δημιουργούμε δύο νέες μεταβλητές τις Y, Z και ακολουθώντας τις οδηγίες του θεωρήματος 1 έχουμε την νέα ισοδύναμη γραμματική $G' = (\{A, B, X, Y, S\}, \{x, y, z\}, S, P')$. Οι νέοι κανόνες παραγωγής P' είναι:

$$\begin{array}{ccc} S \to xSAB & SA \to YYA & YA \to YZB \\ YS \to xB & AZ \to xSAB & YYZB \to AB \\ S \to xB & A \to x & B \to xZY \\ Y \to y & Z \to z & \end{array}$$

Ορισμός 16 *Η γραμματική G λέγεται μη μειούμενου μήκους όταν το μήκος της αριστερής συμβολοσειράς κάθε κανόνα παραγωγής είναι μικρότερο από το μήκος της αντίστοιχης δεξιάς συμβολοσειράς.*

Δηλαδή για κάθε κανόνα της γραμματικής $A \to B$ θα πρέπει να ισχύει: $| A | \leq | B |$.

Ο Chomsky όρισε τέσσερεις κατηγορίες γραμματικών που διακρίνονται ανάλογα με την μορφή των κανόνων παραγωγής τους. Οι κατηγορίες αυτές αναλύονται λεπτομερέστερα στις παραγράφους που ακολουθούν.

5.2.1 Γραμματική τύπου-0

Οι γραμματικές που δεν έχουν περιορισμούς στους κανόνες παραγωγής ονομάζονται γραμματικές χωρίς περιορισμούς ή γραμματικές τύπου-0.

Είναι σημαντικό να τονιστεί ότι οι γραμματικές χωρίς περιορισμούς μπορούν να περιγράψουν οποιαδήποτε γλώσσα αλλά μέχρι σήμερα δεν έχουν δημιουργηθεί αποτελεσματικοί μέθοδοι ανάλυσης

και επεξεργασίας αυτών των γραμματικών. Η αδυναμία αυτή είναι και το κυριότερο εμπόδιο για την χρήση αυτών των γραμματικών σε πραγματικές εφαρμογές.

Το μεγαλύτερο εμπόδιο στην χρήση γραμματικών τύπου-0 είναι το πρόβλημα της συνταχτικής ανάλυσης που μπορεί επιγραμματικά να περιγραφεί ως εξής: Δοσμένης μιας γραμματικής να βρεθεί αν μία συμβολοσειρά ανήκει στη γλώσσα της γραμματικής. Στο πρόβλημα αυτό δεν έχει δοθεί αναλυτική ή αλγοριθμική λύση αλλά υπάρχουν κάποιες γενικές κατευθύνσεις για τον μετασχηματισμό των κανόνων παραγωγής σε κάποιες τυποποιημένες μορφές έτσι ώστε να διευκολυνθεί η διαδικασία συνταχτικής ανάλυσης. Μερικές από αυτές τις κατευθύνσεις δίνονται στις επόμενες παραγράφους.

Θεώρημα 8 *Σε κάθε γραμματική G τύπου-0 υπάρχει μία τουλάχιστον ισοδύναμη γραμματική G' στην οποία όλοι οι κανόνες παραγωγής ανήκουν σε μία από τις ακόλουθες μορφές:*

$$\begin{array}{lll} S \to \lambda & A \to a & A \to B \\ A \to BC & AB \to AC & AB \to CB \\ AB \to B & & \end{array}$$

με $S, A, B, C \in V_N$, $a \in V_T$ και λ είναι η συμβολοσειρά μηδενικού μήκους.

Το θεώρημα μας λέει ότι σε οποιαδήποτε γραμματική μπορούμε να βρούμε μία τουλάχιστον ισοδύναμη γραμματική η οποία να έχει απλούς κανόνες δύο συμβόλων σε κάθε τμήμα των κανόνων παραγωγής, και μάλιστα στις μορφές που περιγράφει το θεώρημα.

Οι κανόνες αυτοί είναι όλοι μη μειούμενου μήκους εκτός από τους κανόνες της μορφής $S \to \lambda$ και τον $AB \to C$. Ο κανόνας $S \to \lambda$ τοποθετείται μόνο στην περίπτωση που η γλώσσα της γραμματικής περιέχει την πρόταση μηδενικού μήκους.

Για την ισοδύναμη γραμματική G' θα λέμε ότι βρίσκεται σε κανονική μορφή. Η κατασκευή μιας ισοδύναμης γραμματικής σε κανονική μορφή ακολουθεί τα εξής βήματα:

1. Βρίσκουμε μία ισοδύναμη γραμματική για την οποία όλοι οι όροι στο αριστερό τμήμα των κανόνων παραγωγής είναι μεταβλητές. Η μετατροπή της γραμματικής πραγματοποιείται με τους κανόνες του θεωρήματος 1.

2. Αντικαθιστούμε όλους τους κανόνες της μορφής $A \to \lambda$ με κανόνες της μορφής $AX \to X$ και $XA \to X$ για όλες τις μεταβλητές και τα τερματικά σύμβολα της γραμματικής. Με αυτή την αντικατάσταση έχουμε μία ισοδύναμη γραμματική για τις περιπτώσεις που η αντίστοιχη γλώσσα της γραμματικής δεν περιέχει την πρόταση λ. Αν στην γλώσσα περιέχεται και η πρόταση λ τότε προσθέτουμε και τον κανόνα $S \to \lambda$.

3. Αντικαθιστούμε όλους τους κανόνες που έχουν περισσότερα από δύο σύμβολα είτε στο αριστερό είτε στο δεξιό τους τμήμα ακολουθώντας την εξής τεχνική:

Εστω ο κανόνας παραγωγής $X_1X_2...X_m \to Y_1Y_2...Y_n$ με $m, n > 1$ και με τουλάχιστον ένα από τα m,n να είναι μεγαλύτερο του 2.

Δημιουργούμε τις νέες μεταβλητές: $U_1, U_2, ..., U_{n-1}, Z_1, ..., Z_{m-1}$ και αντικαθιστούμε τον κανόνα παραγωγής έτσι ώστε:

$$\begin{array}{ll} X_1U_1 \to Y_1Z_1 & Z_1 \to Y_2Z_2 \\ U_1 \to X_2U_2 & Z_2 \to Y_3Z_3 \\ \ldots & \ldots \\ U_{m-1} \to X_{m-1}X_m & Z_{n-1} \to Y_{n-1}Y_n \end{array}$$

4. Με την εφαρμογή και του τρίτου βήματος βλέπουμε ότι μερικοί από τους νέους κανόνες παραγωγής βρίσκονται στην μορφή $AB \to CD$. Κάθε κανόνας αυτής της μορφής μπορεί να αντικατασταθεί με κανόνες της μορφής $AB \to AC, AB \to CB$ με την βοήθεια δύο νέων μεταβλητών K,M:

$$AB \to AK \quad AK \to MK \quad MK \to CK \quad CK \to CD$$

Παράδειγμα 49 *Να βρεθεί μία ισοδύναμη γραμματική σε κανονική μορφή της γραμματικής του παραδείγματος 5.*

1. Οι κανόνες παραγωγής της γραμματικής μετά την αντικατάσταση των τερματικών συμβόλων στο αριστερό τμήμα των συμβόλων παραγωγής γίνονται ως εξής:

$$\begin{array}{ccc} S \to xSAB & SA \to YYA & YA \to YZB \\ YS \to xB & AZ \to xSAB & YYZB \to AB \\ S \to xB & A \to x & B \to xZY \\ Y \to y & Z \to z & \end{array}$$

2. Δεν υπάρχουν κανόνες της μορφής $A \to \lambda$.

3. Αντικαθιστούμε έξι κανόνες παραγωγής που έχουν περισότερα από δύο μέλη στο αριστερό ή στο δεξί τους τμήμα.

$$\begin{array}{ccc} S \to xR_1 & R_1 \to SR_2 & R_2 \to AB \\ SA \to YR_3 & R_3 \to YA & \\ YA \to YR_4 & R_4 \to ZB & \\ AZ \to xR_5 & R_5 \to SR_6 & R_6 \to AB \\ YR_7 \to AB & R_7 \to YR_8 & R_8 \to ZB \\ B \to xR_9 & R_9 \to ZY & \end{array}$$

4. Από τους παραπάνω κανόνες παραγωγής αυτοί που βρίσκονται στην μορφή $AB \to CD$ αντικαθίστανται από τους ακόλουθους κανόνες παραγωγής:

$$\begin{array}{lllll} SA \to YR_3 \Leftrightarrow & SA \to SR_{10} & SR_{10} \to R_{11}R_{10} & R_{11}R_{10} \to YR_{10} & YR_{10} \to YR_3 \\ AZ \to xR_5 \Leftrightarrow & AZ \to AR_{12} & AR_{12} \to R_{13}R_{12} & R_{13}R_{12} \to xR_{12} & xR_{12} \to xR_5 \\ YR_7 \to AB \Leftrightarrow & YR_7 \to YR_{14} & YR_{14} \to R_{15}R_{14} & R_{15}R_{14} \to AR_{14} & AR_{14} \to AB \end{array}$$

Η ισοδύναμη γραμματική σε κανονική μορφή θα έχει την εξής μορφή:

$$G' = (\{A, B, X, Y, Z, S, R_1, R_2, ..., R_{14}\}, \{x, y, z\}, S, P')$$

με 29 κανόνες παραγωγής για το P':

$$\begin{array}{cccc} Y \to y & Z \to z & S \to XB & A \to x \\ S \to XR_1 & R_1 \to SR_2 & R_2 \to AB & YR_{10} \to YR_3 \\ R_3 \to YA & YA \to YR_4 & R_4 \to ZB & XR_{12} \to XR_5 \\ R_5 \to SR_6 & SA \to SR_{10} & SR_{10} \to R_{11}R_{10} & R_{11}R_{10} \to YR_{10} \\ R_6 \to AB & AZ \to AR_{12} & AR_{12} \to R_{13}R_{12} & R_{13}R_{12} \to xR_{12} \\ YR_7 \to YR_{14} & YR_{14} \to R_{15}R_{14} & R_{15}R_{14} \to AR_{14} & AR_{14} \to AB \\ R_7 \to YR_8 & R_8 \to ZB & B \to XR_9 & R_9 \to ZY \\ X \to x & & & \end{array}$$

Μπορούμε να δούμε ότι όλοι οι κανόνες της ισοδύναμης γραμματικής είναι σε κανονική μορφή.

Η μετατροπή γραμματικών τύπου-0 σε κανονική μορφή είναι μία απαραίτητη προεπεξεργασία κατά την εφαρμογή μεθόδων συντακτικής ανάλυσης.

Θεώρημα 9 *Κάθε γραμματική τύπου-0 για την οποία ισχύει ότι η ισοδύναμη γραμματική δεν περιέχει κανόνα παραγωγής της μορφής $AB \to B$ και δεν περιέχει την πρόταση μηδενικού μήκους, είναι γραμματική μη μειούμενου μήκους.*

Η απόδειξη είναι προφανής διότι οι κανόνες κάθε γραμματικής τύπου-0 μπορούν να αναλυθούν σε απλούστερους κανόνες κανονικής μορφής.

Το θεώρημα είναι πολύ σημαντικό διότι μας δίνει τις προυποθέσεις που πρέπει να πληρεί μία συμβολοσειρά A που περιέχει μία τουλάχιστον μεταβλητή έτσι ώστε με την εφαρμογή των κανόνων παραγωγής της γραμματικής να είναι δυνατή η κατασκευή της συμβολοσειράς B. Με βάση αυτό το θεώρημα αν $\mid A \mid > \mid B \mid$ και η γραμματική είναι μη μειούμενη τότε δεν μπορεί να υπάρξει μέθοδος παραγωγής του B από το A.

Ευρεία χρήση του θεωρήματος αυτού γίνεται στους αλγόριθμους συντακτικής ανάλυσης.

5.2.2 Γραμματική τύπου-1

Οι γραμματικές τύπου-1 ονομάζονται και γραμματικές ευαίσθητης σύνταξης (context-sensitive).

Ορισμός 17 *Η γραμματική G ονομάζεται τύπου-1 ή ευαίσθητης σύνταξης όταν οι κανόνες παραγωγής της είναι του τύπου:*

$X_1AX_2 \rightarrow X_1BX_2$ *με* $X_1, B, X_2 \in (V_T \cup V_N)^*, A \in V_N$ *και* B *δεν είναι η πρόταση μηδενικού μήκους εκτός από την περίπτωση κατά την οποία η γλώσσα περιέχει και την πρόταση μηδενικού μήκους.*

Θεώρημα 10 *Κάθε γραμματική ευαίσθητης σύνταξης είναι μη μειούμενη γραμματική.*

Η απόδειξη είναι προφανής από το είδος των κανόνων της γραμματικής.

Θεώρημα 11 *Για κάθε μη μειούμενη γραμματική μπορούμε να βρούμε μία αντίστοιχη γραμματική ευαίσθητης σύνταξης.*

Η απόδειξη του θεωρήματος και η μέθοδος δημιουργίας της αντίστοιχης γραμματικής ευαίσθητης σύνταξης είναι η ακόλουθη:

Ολοι οι κανόνες μιας μη μειούμενης γραμματικής έχουν την μορφή:

$$X_1X_2...X_m \rightarrow Y_1Y_2...Y_n$$

, με $m > 0$, $n > 0$ και $m < n$.

Ο κανόνας αυτός μπορεί να αντικατασταθεί από το ακόλουθο σύνολο κανόνων παραγωγής:

$$\begin{gathered}
X_1X_2...X_m \rightarrow Z_1X_2...X_m \\
Z_1X_2...X_m \rightarrow Z_1Z_2...X_m \\
... \\
Z_1...Z_{m-1}X_m \rightarrow Z_1...Z_{m-1}X_mY_{m+1}...Y_n \\
Z_1...Z_{m-1}Z_mY_{m+1}...Y_n \rightarrow Y_1...Z_{m-1}Z_mY_{m+1}...Y_n \\
... \\
Y_1...Y_{m-1}Z_mY_{m+1}...Y_n \rightarrow Y_1...Y_{m-1}Y_mY_{m+1}...Y_n
\end{gathered}$$

Οπου Z_i, $i = 1, ..., m$ είναι νέες μεταβλητές. Μπορούμε να δούμε ότι όλοι οι ισοδύναμοι κανόνες είναι κανόνες ευαίσθητης σύνταξης.

Παράδειγμα 50 *Εστω η μη μειούμενη γραμματική*

$$G = (\{A, B, S, \}, \{x, y, z\}, S, P)$$

με P να περιέχει τους ακόλουθους κανόνες σύνταξης:

$$\begin{array}{ccc} S \to xSAB & SA \to yyA & yA \to yzB \\ yS \to xB & Az \to xSAB & yzB \to AByA \\ S \to xB & A \to x & B \to xzy \end{array}$$

Να βρεθεί η ισοδύναμη γραμματική ευαίσθητης σύνταξης.

Μερικοί από τους κανόνες παραγωγής βρίσκονται ήδη σε μορφή ευαίσθητης σύνταξης, οι υπόλοιποι μετασχηματίζονται ως εξής:

$$yS \to xB \Leftrightarrow \left\{ \begin{array}{cc} yS \to Z_1S & xZ_2 \to xB \\ Z_1S \to Z_1Z_2 & Z_1Z_2 \to xZ_2 \end{array} \right.$$

$$Az \to xSAB \Leftrightarrow \left\{ \begin{array}{cc} Az \to Z_3z & Z_3z \to Z_3SAB \\ Z_3SAB \to xSAB & \end{array} \right.$$

$$yzB \to AByA \Leftrightarrow \left\{ \begin{array}{cc} yzB \to Z_4zB & Z_4Z_5yA \to AZ_5yA \\ Z_4zB \to Z_4Z_5B & AZ_5yA \to AByA \\ Z_4Z_5B \to Z_4Z_5yA & \end{array} \right.$$

Η ισοδύναμη γραμματική ευαίσθητης σύνταξης είναι η ακόλουθη:
$G' = (\{A, B, Z_1, Z_2, Z_3, Z_4, Z_5, S\}, \{x, y, z\}, S, P')$ με κανόνες παραγωγής P' τους ακόλουθους:

$$\begin{array}{ccc} S \to xSAB & SA \to yyA & YA \to yzB \\ A \to x & B \to xzy & YS \to Z_1S \\ Z_1S \to Z_1Z_2 & Z_1Z_2 \to xZ_2 & Az \to Z_3z \\ YzB \to Z_4zB & Z_4Z_5YA \to AZ_5yA & Z_4zB \to Z_4Z_5B \\ Z_4Z_5B \to Z_4Z_5yA & Z_3z \to Z_3SAB & AZ_5YA \to ABYA \\ S \to xB & xZ_2 \to xB & Z_3SAB \to xSAB \\ Y \to y & & \end{array}$$

Θεώρημα 12 *Σε κάθε γραμματική μη μειούμενου μήκους ή ισοδύναμα για κάθε γραμματική ευαίσθητης σύνταξης μπορούμε πάντα να βρούμε αν μία τυχαία πρόταση P ανήκει στην γλώσσα της.*

Επειδή κάθε γραμματική ευαίσθητης σύνταξης είναι μη μειούμενου μήκους και επειδή το μήκος της πρότασης P είναι πεπερασμένο, το πλήθος των προτάσεων της γραμματικής που έχουν μήκος ίσο η μικρότερο της P είναι και αυτό πεπερασμένο.

Δημιουργώντας λοιπόν το σύνολο όλων των προτάσεων που παράγονται από την γραμματική και έχουν μήκος μικρότερο του μήκους της P μπορούμε να βρούμε αν η P μπορεί να παραχθεί από την δοσμένη γραμματική ευαίσθητης σύνταξης.

Θεώρημα 13 *Σε κάθε γραμματική μη μειούμενου μήκους μπορούμε να βρούμε μία τουλάχιστον ισοδύναμη γραμματική που να περιέχει κανόνες στην μορφή KURODA. Μορφή KURODA έχουν οι κανόνες παραγωγής που βρίσκονται σε μία από τις ακόλουθες τέσσερις μορφές:*

$$A \to a \quad A \to B \quad A \to BC \quad AB \to CD$$

με $A, B, C, D \in V_N$, και $a \in V_T$.

Η απόδειξη του θεωρήματος και η μέθοδος δημιουργίας της ισοδύναμης γραμματικής είναι η ακόλουθη:

Οι κανόνες κάθε μη μειούμενης γραμματικής έχουν την μορφή:

$$X_1X_2...X_m \rightarrow Y_1Y_2...Y_n$$

, με $m > 0$, $n > 0$ και $m < n$.

Η γενική μορφή των κανόνων παραγωγής μπορεί να αντικατασταθεί με ένα σύνολο ισοδύναμων κανόνων που βρίσκονται όλοι στην μορφή KURODA χρησιμοποιώντας m νέες μεταβλητές, ως εξής:

$$\begin{array}{ll} X_1X_2 \rightarrow Y_1Z_2 & Z_m \rightarrow Y_mZ_{m+1} \\ Z_2X_3 \rightarrow Y_2Z_3 & Z_{m+1} \rightarrow Y_{m+1}Z_{m+2} \\ \ldots & \ldots \\ Z_{m-1}X_m \rightarrow Y_{m-1}Z_m & Z_{n-1} \rightarrow Y_{n-1}Y_n \end{array}$$

Παράδειγμα 51 *Εστω η μη μειούμενη γραμματική* $G = (\{A, B, S\}, \{a, b\}, S, P)$, *με κανόνες παραγωγής:*

$$\begin{array}{llll} S \rightarrow Aa & Aa \rightarrow bBa & A \rightarrow Ba & Ba \rightarrow aaa \\ bBA \rightarrow bBaAA & B \rightarrow a & A \rightarrow b & AA \rightarrow BbA \end{array}$$

Να βρεθεί η αντίστοιχη γραμματική με κανόνες παραγωγής σε μορφή KURODA.

Μετασχηματίζουμε τους κανόνες παραγωγής που δεν βρίσκονται σε μορφή KURODA:

$$\begin{array}{rl} Aa \rightarrow bBa \Leftrightarrow & \left\{ \begin{array}{ll} Aa \rightarrow bZ_1 & Z_1 \rightarrow Ba \end{array} \right. \\ Ba \rightarrow aaa \Leftrightarrow & \left\{ \begin{array}{ll} Ba \rightarrow aZ_2 & Z_2 \rightarrow aa \end{array} \right. \\ bBA \rightarrow bBaAA \Leftrightarrow & \left\{ \begin{array}{ll} bZ_3 \rightarrow bZ_4 & Z_3 \rightarrow BA \\ Z_4 \rightarrow BZ_5 & Z_5 \rightarrow aZ_6 \\ Z_6 \rightarrow AA & \end{array} \right. \\ AA \rightarrow BbA \Leftrightarrow & \left\{ \begin{array}{ll} AA \rightarrow BZ_7 & Z_7 \rightarrow bA \end{array} \right. \end{array}$$

Με την χρήση δύο νέων μεταβλητών, των K,L με τις οποίες αντικαθιστούμε τα τερματικά σύμβολα a,b φτιάχνουμε μία ισοδύναμη γραμματική σε μορφή KURODA:

$G' = (\{A, B, Z_1, ..., Z_7, S\}, \{a, b\}, S, P')$, με κανόνες παραγωγής P':

$$\begin{array}{llll} S \rightarrow AK & B \rightarrow a & A \rightarrow BK & A \rightarrow b \\ AK \rightarrow BZ_1 & Z_1 \rightarrow BK & BK \rightarrow KZ_2 & Z_2 \rightarrow KK \\ LZ_3 \rightarrow LZ_4 & Z_3 \rightarrow BA & Z_4 \rightarrow BZ_5 & Z_5 \rightarrow KZ_6 \\ Z_6 \rightarrow AA & AA \rightarrow BZ_7 & Z_7 \rightarrow LA & K \rightarrow a \\ L \rightarrow b & & & \end{array}$$

5.2.3 Γραμματική τύπου-2

Η γραμματική τύπου-2 ή γραμματική ελεύθερης σύνταξης (context-free) είναι η πλέον διαδεδομένη γραμματική σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Ο σημαντικότερος λόγος χρήσης γραμματικών αυτού του τύπου είναι η ύπαρξη αλγορίθμων συντακτικής ανάλυσης που μπορούν να υλοποιηθούν στις περισσότερες των περπτώσεων σε πραγματικό χρόνο. Μειονέκτημα αποτελεί το γεγονός ότι στις περισσότερες περιπτώσεις η φυσική γλώσσα δεν μπορεί να μοντελοποιηθεί με ακρίβεια με μία γραμματική ελεύθερης σύνταξης.

Ορισμός 18 *Η γραμματική G ονομάζεται τύπου-2 ή γραμματική ελεύθερης σύνταξης όταν οι κανόνες παραγωγής της είναι στην μορφή* $A \rightarrow B$ *με* $B \in (V_T \cup V_N)^*, A \in V_N$.

Θεώρημα 14 *Σε κάθε γραμματική ελεύθερης σύνταξης μπορούμε να βρούμε αν μία πρόταση ανήκει στην γλώσσα της γραμματικής.*

Η απόδειξη είναι απλή διότι κάθε γραμματική ελεύθερης σύνταξης είναι και γραμματική μη μειούμενου μήκους. Γνωρίζουμε από προηγούμενο θεώρημα ότι για κάθε μη μειούμενη γραμματική μπορούμε να βρούμε με την βοήθεια αλγόριθμου που χρησιμοποιεί πεπερασμένο αριθμό υπολογισμών αν μία πρόταση ανήκει στην γλώσσα της γραμματικής.

Ορισμός 19 *Η γραμματική ελεύθερης σύνταξης G βρίσκεται στην μορφή Chomsky όταν όλοι οι κανόνες παραγωγής της είναι του τύπου $A \to a$ με $A \in V_N, a \in V_T$ ή $A \to BC$ με $A, B, C \in V_N$.*

Θεώρημα 15 *Σε κάθε γραμματική ελεύθερης σύνταξης μπορούμε πάντα να βρούμε μία τουλάχιστον ισοδύναμη γραμματική ελεύθερης σύνταξης σε μορφή Chomsky.*

Η απόδειξη του θεωρήματος και η μέθοδος δημιουργίας της ισοδύναμης γραμματικής είναι η ακόλουθη:

1. Με την βοήθεια νέων μεταβλητών αντικαθιστούμε τα τερματικά σύμβολα στους κανόνες παραγωγής που δεν βρίσκονται σε μορφή Chomsky.

2. Μετά τον πρώτο μετασχηματισμό οι κανόνες παραγωγής κάθε γραμματικής ελεύθερης σύνταξης έχουν την μορφή $X \to Y_1 Y_2 ... Y_n$, με $n > 2$. Ισοδύναμοι κανόνες μπορούν να δημιουργηθούν με την βοήθεια νέων μεταβλητών, ως εξής:

$$X \to Y_1 Z_1 \quad Z_1 \to Y_2 Z_2 \quad ... \quad Z_{n-2} \to Y_{n-1} Y_n$$

Αποδεικνύεται λοιπόν ότι με την βοήθεια n-2 καινούργιων μεταβλητών κάθε κανόνας μιας γραμματικής ελεύθερης σύνταξης μπορεί να έρθει στην μορφή Chomsky.

Παράδειγμα 52 *Εστω η γραμματική ελεύθερης σύνταξης*
$G = (\{A, B, S\}, \{a, b, c, +, (,)\}, S, P)$, *με κανόνες παραγωγής:*

$$\begin{array}{cccc} S \to S + A & A \to AB & B \to (S) & B \to b \\ S \to a & A \to a & B \to a & B \to c \end{array}$$

Βρείτε μία ισοδύναμη γραμματική σε μορφή Chomsky.

Η γραμματική του παραδείγματος μπορεί να χρησιμοποιηθεί για να παράγει απλές αλγεβρικές παραστάσεις των a,b,c.

Η εύρεση μιας ισοδύναμης γραμματικής σε μορφή Chomsky παραγματοποιείται με την ακόλουθη διαδικασία:

1. Με την βοήθεια νέων μεταβλητών αντικαθιστούμε τα τερματικά σύμβολα στους κανόνες παραγωγής που δεν βρίσκονται σε μορφή Chomsky. Στο παράδειγμα χρειάζεται να ορίσουμε τρεις νέες μεταβλητές, τις C,D,E. Οι κανόνες παραγωγής μετασχηματίζονται ως εξής:

$$\begin{array}{cccc} S \to SCA & A \to AB & B \to DSE & B \to b \\ S \to a & A \to B & B \to a & B \to c \\ C \to + & D \to (& E \to) & \end{array}$$

2. Μετασχηματίζουμε τους κανόνες παραγωγής που έχουν μήκος μεγαλύτερο από δύο σύμβολα ακολουθώντας την μέθοδο που αναφέρεται στο προηγούμενο θεώρημα. Οι κανόνες παραγωγής της ισοδύναμης γραμματικής σε μορφή Chomsky είναι οι ακόλουθοι:

$$\begin{array}{cccc} S \to SX & A \to AB & B \to DY & B \to b \\ S \to a & A \to B & B \to a & B \to c \\ C \to + & D \to (& E \to) & X \to CA \\ Y \to SE & & & \end{array}$$

5.2.4 Γραμματική τύπου-3

Η γραμματική τύπου-3 ή κανονική γραμματική (regular language) περιέχει τους πλέον αυστηρούς περιορισμούς στους κανόνες παραγωγής. Ονομάζεται επίσης και γραμματική πεπερασμένων καταστάσεων.

Ορισμός 20 *Η γραμματική G ονομάζεται τύπου-3 ή κανονική γραμματική όταν οι κανόνες παραγωγής της έχουν την μορφή:* $A \to aB, A \to a$ *με* $A, B \in V_N, a \in V_T$.

5.3 Συντακτική ταξινόμηση προτύπων

Η διαδικασία συντακτικής ταξινόμησης προτύπων ονομάζεται και συντακτική ανάλυση προτύπων. Το πρόβλημα της συντακτικής ανάλυσης μπορεί να περιγραφεί απλά σαν την μέθοδο που πρέπει να ακολουθήσει κάποιος για να βρεί αν μία πρόταση ανήκει στην γλώσσα μιας γραμματικής ή πιό απλά αν η πρόταση μπορεί να παραχθεί με διαδοχική εφαρμογή των κανόνων της γραμματικής.

Οι μέθοδοι συντακτικής ανάλυσης μπορούν να διακριθούν σε κατηγορίες με βάση το είδος της γραμματικής που μπορούν να αναλύσουν και την μέθοδο που το επιτυγχάνουν. Δύο είναι οι βασικές κατηγορίες μεθόδων συντακτικής ανάλυσης προτάσεων:

1. Οι μέθοδοι που αναζητούν την διαδοχή των κανόνων παραγωγής που οδηγούν από το μοναδικό σύμβολο εκκίνησης στην πρόταση και ονομάζονται μέθοδοι συντακτικής ανάλυσης προς την συμβολοσειρά (top-down parsing).

2. Οι μέθοδοι που υλοποιούν την αντίστροφη διαδικασία, μετασχηματίζοντας την δοσμένη πρόταση στο μοναδικό σύμβολο εκκίνησης. Ο μετασχηματισμός πραγματοποιείται με την διαδοχική εφαρμογή των κανόνων παραγωγής κατά την αντίστροφη φορά. Η διαδικασία αναζήτησης της διαδοχής των κανόνων που μπορούν να οδηγήσουν απο την πρόταση στο σύμβολο εκκίνησης ονομάζεται συντακτική ανάλυση προς το σύμβολο εκκίνησης (bottom-up parsing).

Οι τεχνικές που παρουσιάζονται σε αυτό το κεφάλαιο χρησιμοποιούνται κύρια από μεταφραστές (interpreters-compilers) συμβολικού κώδικα για τον έλεγχο ορθής σύνταξης του συμβολικού κώδικα εξετάζουν δηλαδή αν η σύνταξη του συμβολικού κώδικα ακολουθεί τους κανόνες της γλώσσας προγραμματισμού.

5.3.1 Συντακτική ανάλυση προς την συμβολοσειρά

Αναλυτική μέθοδος συντακτικής ανάλυσης προς την συμβολοσειρά για γραμματικές τύπου-0 δεν έχει ανακαλυφθεί μέχρι σήμερα. Για μη μειούμενες γραμματικές έχει ήδη αποδειχθεί (θεώρημα 6) ότι είναι δυνατόν να υπάρξει αλγόριθμος συντακτικής ανάλυσης με εφαρμογή ενός πεπερασμένου αριθμού κανόνων παραγωγής ο οποίος είναι σε θέση να παράγει όλες οι προτάσεις της γλώσσας με δοσμένο μήκος. Ετσι με απλή σύγκριση της ελεγχόμενης πρότασης με τις αντίστοιχες προτάσεις της γραμματικής που έχουν το ίδιο μήκος μπορεί να βρεθεί αν η ελεγχόμενη πρόταση μπορεί να παραχθεί από την γραμματική.

Αλγόριθμος 7 *Η συντακτική ανάλυση μη μειούμενων γραμματικών μπορεί να διευκολυνθεί όταν ακολουθήσουμε τις εξής οδηγίες:*

ΒΗΜΑ 1. Μετασχηματίζουμε τους κανόνες της γραμματικής που δεν βρίσκονται στην μορφή $X \to a, a \in V_T, X \in V_N$ σε απλούστερες μορφές (πιθανόν σε μορφή KURODA) είτε σε μία από τις ακόλουθες μορφές:

1η μορφή: Ανάλυση από αριστερά προς τα δεξιά:

$$X \to aY, a \in V_T, X, Y \in (V_N \cup V_T)^*$$

2η μορφή: Ανάλυση από δεξιά προς τα αριστερά:

$$X \to Ya, a \in V_T, X, Y \in (V_N \cup V_T)^*$$

Συνήθως διαλέγουμε εκείνη την μορφή στην οποία οι περισσότεροι κανόνες μπορούν να μετασχηματιστούν πιο εύκολα.

Η επιδίωξη αυτή στοχεύει στο να φέρουμε την γραμματική σε τέτοια μορφή έτσι ώστε να γίνεται ευκολότερη η ανάλυση από δεξιά προς τα αριστερά ή το αντίθετο. Η ύπαρξη τερματικού συμβόλου στα αριστερά ή στα δεξιά των κανόνων παραγωγής διευκολύνει την απόφασή μας για την χρησιμοποίηση του κανόνα στην προσπάθεια να βρούμε την πρόταση, ξεκινώντας από το σύμβολο εκκίνησης.

ΒΗΜΑ 2. Αρχίζοντας από το σύμβολο εκκίνησης και λαμβάνοντας υπόψη την μορφή των κανόνων της γραμματικής (μορφή 1 ή 2) αρχίζουμε να αντικαθιστούμε τις μεταβλητές που βρίσκονται στην γειτονιά των τερματικών συμβόλων εφαρμόζοντας τους κανόνες της γραμματικής. Σε περίπτωση που είναι δυνατές περισσότερες της μιας αντικατάστασης, όλες οι δυνατές λύσεις πρέπει να εξεταστούν. Κατά συνέπεια ο αλγόριθμος πρέπει να έχει αναδρομική ισχύ.

ΒΗΜΑ 3. Η απόρριψη λύσεων είναι δυνατή στις μη μειούμενες γραμματικές στις εξής περιπτώσεις:

α. Οταν η συμβολοσειρά των τερματικών συμβόλων (από το άκρο που έχει αρχίσει η ανάλυση) δεν συμφωνεί με τα τερματικά σύμβολα της ελεγχόμενης πρότασης.

β. Οταν το μήκος της συμβολοσειράς έχει ξεπεράσει το μήκος της ελεγχόμενης πρότασης.

Αλγόριθμος 8 *Η συντακτική ανάλυση γραμματικών ευαίσθητης σύνταξης πρέπει να ακολουθεί τους εξής κανόνες:*

ΒΗΜΑ 1. Βρίσκουμε μία ισοδύναμη γραμματική σε μορφή KURODA.

ΒΗΜΑ 2. Οι κανόνες παραγωγής μετασχηματιστίζονται σε κάποια από τις μορφές 1 ή 2 του αλγόριθμου 1. Κατόπιν εφαρμόζονται τα βήματα 2 και 3 του αλγόριθμου 1.

Αλγόριθμος 9 *Η συντακτική ανάλυση γραμματικών ελεύθερης σύνταξης ακολουθεί τους εξής κανόνες:*

ΒΗΜΑ 1. Οι κανόνες της γραμματικής μετατρέπονται σε μορφή Greibach.

ΒΗΜΑ 2. Ακολουθήσουν τα βήματα 2 και 3 του αλγόριθμου 1.

Παράδειγμα 53 *Εστω η γραμματική ελεύθερης σύνταξης*

$G = (\{A, S\}, \{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}, S, P)$ *με κανόνες παραγωγής P:*

$$P = \{S \to SA, S \to A, A \to 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9 \mid 0\}$$

όπου $D \to A \mid B \Leftrightarrow D \to A, D \to B$

Η πρόταση 345 ανήκει στην γλώσσα της γραμματικής;.

Με εφαρμογή των κανόνων ανάλυσης πρός την συμβολοσειρά έχουμε:

$$S \Rightarrow \begin{cases} SA \Rightarrow & \begin{cases} SAA \Rightarrow & \begin{cases} SAAA \Rightarrow \mid SAAA \mid > 3 \\ AAA \stackrel{*}{\Rightarrow} 345 \\ \ldots \\ AA \Rightarrow \mid AA \mid = 2 \\ S5 \stackrel{*}{\Rightarrow} 345 \end{cases} \end{cases} \\ A \Rightarrow \mid A \mid = 1 \end{cases}$$

Παρατηρούμε ότι η πρόταση μπορεί να παραχθεί με δύο διαφορετικούς τρόπους. Γιαυτό τον λόγο η πρόταση 345 είναι συνταχτικά ασαφής για την γραμματική G.

Παράδειγμα 54 *Εστω η μη μειούμενη γραμματική σε μορφή KURODA:*

$$G = (\{A, B, S\}, \{a, b\}, S, P)$$

με:

$$P = \left\{ \begin{array}{ccc} S \to SA & A \to AB & AA \to Aa \\ AB \to bB & BA \to Baa & bB \to bba \\ A \to a & B \to b & S \to a \end{array} \right\}$$

Η πρόταση aabb ανήκει στην γλώσσα της γραμματικής;.

Ας εξετάσουμε αν η πρόταση *aabb* μπορεί να δημιουργηθεί με διαδοχική εφαρμογή των κανόνων παραγωγής της γραμματικής G.

$$S \stackrel{S \to aA}{\Rightarrow} aA \stackrel{A \to AB}{\Rightarrow} aAB \stackrel{A \to AB}{\Rightarrow} aABB \stackrel{A \to a}{\Rightarrow} aaBB \stackrel{B \to b}{\Rightarrow} aabB \stackrel{B \to b}{\Rightarrow} aabb$$

Η πρόταση μπορεί να παραχθεί από το σύμβολο εκκίνησης. Συνεπώς πρόταση aabb ανήκει στην γλώσσα της γραμματικής G.

5.3.2 Συνταχτική ανάλυση προς το σύμβολο εκκίνησης

Οι μέθοδοι συνταχτικής ανάλυσης προς το σημείο εκκίνησης περιλαβάνουν όλες τις τεχνικές με τις οποίες προσπαθούμε να παράγουμε απο την πρόταση το σύμβολο εκκίνησης με την αντίστροφη εφαρμογή των κανόνων παραγωγής. Οι γενικές οδηγίες που ισχύουν για την συνταχτική ανάλυση προς την συμβολοσειρά (αλγόριθμος 1) έχουν ισχύ και για τους αλγορίθμους συνταχτικής ανάλυσης προς το σύμβολο εκκίνησης.

Αλγόριθμος 10 *Η συνταχτική ανάλυση προς το σύμβολο εκκίνησης για μη μειούμενων γραμματικών μπορεί να διευκολυνθεί αν ακολουθήσουμε τις εξής οδηγίες:*

ΒΗΜΑ 1. Ιδιο με το βημα 1 του αλγόριθμου 1.

ΒΗΜΑ 2. Αρχίζοντας από την συμβολοσειρά ελαττώνουμε τον αριθμό των συμβόλων της πρότασης με την αντίστροφη εφαρμογή των κανόνων παραγωγής, βρίσκοντας διαδοχικά το πλέον κατάλληλο

κανόνα απο τα αριστερά ή τα δεξιά της συμβολοσειράς, ανάλογα με την μορφή των κανόνων παραγωγής που δημιουργήθηκαν στο ΒΗΜΑ 1. Οι πλέον κατάλληλοι κανόνες είναι οι ακόλουθοι:

α) Βρίσκουμε ακολουθίες τερματικών συμβόλων και μεταβλητών που βρίσκονται στο αριστερό (ή δεξιό) τμήμα της συμβολοσειράς για το οποίο μπορεί να βρεθεί κανόνας που να τα περιέχει στο δεξί τμήμα του. Η κατεύθυνση της ανάλυσης έχει φορά αντίστροφη των κανόνων που έχουν παραχθεί από το πρώτο βήμα του αλγόριθμου. Αν οι κανόνες παραγωγής ανήκουν στην πρώτη κατηγορία αρχίζουμε τις αντικαταστάσεις συμβολοσειρών της πρότασης από τα αριστερά, ενώ αν οι κανόνες παραγωγής ανήκουν στην δεύτερη κατηγορία αρχίζουμε τις αντικαταστάσεις συμβολοσειρών της πρότασης από τα δεξιά.

β) Αντικαθιστούμε ακολουθίες μεταβλητών που βρίσκονται στο αριστερό (ή δεξιό) τμήμα της συμβολοσειράς.

γ) Αντικαθιστούμε την συμβολοσειρά με το σύμβολο εκκίνησης.

Οι αλγόριθμοι που αναφέραμε μέχρι τώρα παρουσιάζουν την δυνατότητα πολλαπλών αντικαταστάσεων σε κάθε βήμα των μετασχηματισμών της πρότασης. Σε αυτή την περίπτωση όλες οι εναλλακτικές λύσεις πρέπει να εξετάζονται. Δυστυχώς στις περισσότερες των πρακτικών εφαρμογών οι λύσεις αυξάνουν γεωμετρικά με το μήκος της ελεγχόμενης πρότασης, απαιτώντας πολλές φορές χρονοβόρους υπολογισμούς και δεσμεύοντας μεγάλη μνήμη ιδιαίτερα όταν αναλύονται προτάσεις μεγάλου μήκους.

Γενικά οι μέθοδοι συντακτικής ανάλυσης προς το σύμβολο εκκίνησης παρουσιάζουν λιγότερα προβλήματα απο τις μεθόδους συντακτικής ανάλυσης προς την συμβολοσειρά. Αυτό γίνεται φανερό στα παραδείγματα που ακολουθούν, στα οποία δίνεται η αντίστοιχη συντακτική ανάλυση προς το σύμβολο εκκίνησης των γραμματικών των παραδειγμάτων 13 και 14.

Παράδειγμα 55 *Εστω η γραμματική ελεύθερης σύνταξης G του παραδείγματος 13. Η πρόταση 345 ελέγχεται συντακτικά με την μέθοδο ανάλυσης προς το σύμβολο εκκίνησης ως εξής:*

$$345 \stackrel{*}{\Rightarrow} AAA \stackrel{S\to AA}{\Rightarrow} SAA \stackrel{S\to SA}{\Rightarrow} SA \stackrel{S\to SA}{\Rightarrow} S$$

Παράδειγμα 56 *Εστω η μη μειούμενη γραμματική του παραδείγματος 14. Η πρόταση aabb ανήκει στην γλώσσα της γραμματικής;.*

Με συντακτική ανάλυση προς το σύμβολο εκκίνησης έχουμε:

$$aabb \stackrel{S\to a}{\Rightarrow} Sabb \stackrel{A\to a}{\Rightarrow} SAbb \stackrel{B\to b}{\Rightarrow} SABB \stackrel{A\to AB}{\Rightarrow} SAB \stackrel{A\to AB}{\Rightarrow} SA \stackrel{S\to SA}{\Rightarrow} S$$

5.4 Ο Αλγόριθμος CYK

Οι αλγόριθμοι που περιγράφησαν παρουσιάζουν ένα σημαντικό μειονέκτημα. Ο χρόνος επεξεργασίας, που μπορεί να μετρηθεί από το πλήθος των κανόνων παραγωγής που εφαρμόζονται κατά την διαδικασία της συντακτικής ανάλυσης, είναι εκθετικά συνήθως αυξανόμενος με το μήκος της ελεγχόμενης πρότασης. Ο αλγόριθμος των Cocke-Younger-Kasami (CYK) έχει χρόνο επεξεργασίας που είναι, στην χειρότερη των περιπτώσεων, ανάλογος της τρίτης δύναμης του μήκους της πρότασης. Είναι αλγόριθμος συντακτικής ανάλυσης προτάσεων για γραμματικές ελεύθερης σύνταξης προς το σύμβολο εκκίνησης και βασίζεται στις τεχνικές του δυναμικού προγραμματισμού.

Αλγόριθμος 11 *Εστω η γραμματική ελεύθερης σύνταξης* $G = (V_N, V_T, S, P)$ *η οποία βρίσκεται στην μορφή Chomsky και η αντίστοιχη γλώσσα της γραμματικής δεν περιέχει την πρόταση μηδενικού μήκους. Αν* $a_1, a_2, ..., a_n$ *είναι μία πρόταση με σύμβολα από το αλφάβητο τερματικών συμβόλων της* G*, τότε: αν ο πίνακας* T *περιέχει στη χαμηλότερη θέση του το σύμβολο* S *τότε η πρόταση* $a_1, a_2, ..., a_n$ *ανήκει στην γλώσσα της γραμματικής* G*.*

Η κατασκευή του πίνακα T γίνεται με την τοποθέτηση συμβόλων στις θέσεις του t_{ji} ως εξής:

ΒΗΜΑ 1. Συμπληρώνω την πρώτη γραμμή με τις ακόλουθες μεταβλητές:

$$t_{1i} = \{A \mid A \to a_i \in P\}.$$

ΒΗΜΑ 2. Συμπληρώνω τις υπόλοιπες γραμμές κατά αύξουσα σειρά και τις στήλες από μικρότερες προς μεγαλύτερες τιμές:

$$t_{ji} = \{A \mid A \to BC \in P, B \in t_{ki}, C \in t_{j-k,i+k}, 0 < k < j\}.$$

ΒΗΜΑ 3. Αν $S \in t_{n1}$ τότε η πρόταση ανήκει στην γλώσσα της γραμματικής.

Παράδειγμα 57 *Εστω η γραμματική ελεύθερης σύνταξης*
$G = (\{A, S\}, \{a, b\}, S, P)$ *με κανόνες παραγωγής:*

$$\begin{array}{lll} S \to AA & S \to AS & S \to b \\ A \to SA & A \to AS & A \to a \end{array}$$

και η πρόταση abaab. Με την βοήθεια του αλγόριθμου CYK βρείτε αν η πρόταση ανήκει στην γλώσσα της γραμματικής.

Για την πρόταση *abaab*, ο πίνακας T είναι ο ακόλουθος:

Πίνακας 5.1: Πίνακας CYK του παραδείγματος 18

	a	b	a	a	b
1	A	S	A	A	S
2	A, S	A	S	A, S	
3	A, S	S	S, A		
4	A, S	A, S			
5	A, S				

Η πρόταση *abaab* ανήκει στην γλώσσα της γραμματικής διότι το σύμβολο S εμφανίζεται στο χαμηλότερο σημείο του πίνακα.

Παράδειγμα 58 *Εστω η γραμματική ελεύθερης σύνταξης*
$G = (\{A, B, C, S\}, \{a, b\}, S, P)$ *με κανόνες παραγωγής:*

$$\begin{array}{lll} S \to AB & S \to BC & A \to BA \\ A \to a & B \to CC & B \to b \\ C \to AB & C \to a & \end{array}$$

Με την βοήθεια του αλγόριθμου CYK βρείτε αν η πρόταση baaba ανήκει στην γλώσσα της γραμματικής.

Για την πρόταση αυτή ο πίνακας T είναι ο ακόλουθος:

Η πρόταση *baaba* ανήκει στην γλώσσα της γραμματικής διότι το σύμβολο S εμφανίζεται στο χαμηλότερο σημείο του πίνακα.

Πίνακας 5.2: Πίνακας CYK του παραδείγματος 19

	b	a	a	b	a
1	B	A, C	A, C	B	A, C
2	S, A	B	S, C	S, A	
3		B	B		
4		S, A, C			
5	S, A, C				

5.5 Η διαδικασία εκπαίδευσης

Ενα από τα δυσκολότερα προβλήματα που συναντώνται κατά την σχεδίαση συστημάτων συντακτικής ταξινόμησης προτύπων είναι η μέθοδος που πρέπει να ακολουθηθεί για την κατασκευή της γραμματικής.

Οι τεχνικές και και οι αλγόριθμοι που παρουσιάζονται μας δίνουν την δυνατότητα να κατασκευάζουμε μία γραμματική από δείγμα της γλώσσας είτε αυτό είναι θετικό (κάθε συμβολοσειρά του δείγματος ανήκει στην γλώσσα της γραμματικής) είτε είναι αρνητικό (κάθε συμβολοσειρά του δείγματος **δεν** ανήκει στην γλώσσα της γραμματικής).

5.5.1 Κατασκευή κανονικής γραμματικής

Εστω ότι έχουμε ένα δείγμα της γλώσσας L και θέλουμε να κατασκευάσουμε μία κανονική γραμματική που να παράγει την γλώσσα. Σε αυτή την περίπτωση αποδεικνύεται ότι η διαδικασία κατασκευής της γραμματικής είναι μία εύκολη υπόθεση. Αναλύουμε κάθε δείγμα της γλώσσας σαν μια διαδοχή κανόνων παραγωγής κατά την οποία παράγεται κάθε φορά και ένα τερματικό σύμβολο της συμβολοσειράς του δείγματος που διαθέτουμε.

Αλγόριθμος 12 *Εστω η γλώσσα* $L = \{X_i, i = 1, ..., N\}$. *Τα τερματικά σύμβολα της γλώσσας ορίζονται να είναι το σύνολο όλων των συμβόλων που βρίσκονται στα* X_i. *Οι κανόνες της γραμματικής βρίσκονται με την ανάλυση κάθε πρότασης ξεχωριστά:*

$$S \to x_1 x_2 ... x_m \Leftrightarrow \quad S \to x_1 Z_1 \quad Z_1 \to x_2 Z_2 \quad Z_2 \to x_3 Z_3, ..., \quad Z_{m-1} \to x_m$$

Παράδειγμα 59 *Εστω η γλώσσα* $L = \{ab, baa, bbb, aaba\}$. *Κατασκεύασε μία κανονική γραμματική που να παράγει την γλώσσα L.*

Τα τερματικά σύμβολα της γραμματικής είναι τα a,b. Για κάθε μία πρόταση του παραδείγματος εφαρμόζω την τεχνική του αλγόριθμου:

$$\begin{array}{llll} S \to ab \Leftrightarrow & S \to aZ_1 & Z_1 \to b & \\ S \to baa \Leftrightarrow & S \to bZ_2 & Z_2 \to aZ_3 & Z_3 \to a \\ S \to bbb \Leftrightarrow & S \to bZ_4 & Z_4 \to bZ_5 & Z_5 \to b \\ S \to aaba \Leftrightarrow & S \to aZ_6 & Z_6 \to aZ_7 & Z_7 \to bZ_8 \quad Z_8 \to a \end{array}$$

Η γραμματική που περιγράφει την γλώσσα του παραδείγματος είναι η ακόλουθη: $G = (\{S, Z_1, Z_2, ..., Z_8\}, \{a, b\}, P, S)$

Οι κανόνες P είναι το σύνολο των κανόνων παραγωγής που βρίσκονται στο δεξί τμήμα των προηγούμενων τεσσάρων ισοδυναμιών:

$$P = \left\{ \begin{array}{llll} S \to aZ_1 & Z_1 \to b & & \\ S \to bZ_2 & Z_2 \to aZ_3 & Z_3 \to a & \\ S \to bZ_4 & Z_4 \to bZ_5 & Z_5 \to b & \\ S \to aZ_6 & Z_6 \to aZ_7 & Z_7 \to bZ_8 & Z_8 \to a \end{array} \right\}$$

Οι κανόνες παραγωγής του παραδείγματος είναι πολλοί και θα ήταν σκόπιμο να τους μειώσουμε όσο το δυνατόν περισσότερο μία και η ταχύτητα συντακτικής ανάλυσης συνδέεται πολύ στενά με το μήκος της πρότασης που αναλύεται και τον αριθμό των κανόνων παραγωγής της γραμματικής.

Αλγόριθμος 13 *Οι κανόνες παραγωγής της κανονικής γραμματικής μπορούν να ελαττωθούν κατά τους ακόλουθους τρόπους:*

ΒΗΜΑ 1. Ενοποιούμε τις μεταβλητές στους κανόνες της μορφής $Z_i \to x$ για όλους τους κανόνες που καταλήγουν στο ίδιο τερματικό σύμβολο.

ΒΗΜΑ 2. Ενοποιούμε τις μεταβλητές στους κανόνες της μορφής $S \to xZ_i$ για όλους τους κανόνες που περιέχουν το ίδιο τερματικό σύμβολο.

Παράδειγμα 60 *Ελαττώστε τον αριθμό των κανόνων της γραμματικής του προηγούμενου παραδείγματος.*

ΒΗΜΑ 1. Ενοποιούμε τα σύμβολα (Z_1, Z_5) και τα (Z_3, Z_8).

$$\begin{array}{llll} S \to aZ_1 & Z_1 \to b & & \\ S \to bZ_2 & Z_2 \to aZ_3 & Z_3 \to a & \\ S \to bZ_4 & Z_4 \to bZ_1 & & \\ S \to aZ_6 & Z_6 \to aZ_7 & Z_7 \to bZ_3 & \end{array}$$

ΒΗΜΑ 2. Ενοποιούμε τα σύμβολα (Z_1, Z_6) και τα (Z_2, Z_4).

$$\begin{array}{llll} S \to aZ_1 & Z_1 \to b & & \\ S \to bZ_2 & Z_2 \to aZ_3 & Z_3 \to a & \\ & Z_2 \to bZ_1 & & \\ & Z_1 \to aZ_7 & Z_7 \to bZ_3 & \end{array}$$

Στην ισοδύναμη γραμματική οι κανόνες ελαττώθηκαν από 12 σε 8 και οι μεταβλητές από 8 σε 4.

Η παραπάνω ανάλυση έχει ένα μειονέκτημα. Η γραμματική που κατασκευάζεται μπορεί να δημιουργήσει την γλώσσα του δείγματος που έχουμε και όχι μία γλώσσα που να περιέχει περισσότερες προτάσεις που απλώς δεν έχουμε στην διάθεσή μας κατά την διαδικασία εκπαίδευσης.

Το ερώτημα που προκύπτει εύλογα είναι το ακόλουθο. Αφού δεν γνωρίζουμε τις προτάσεις τις γλώσσας, πως έχουμε απαίτηση να φτιάξουμε την γραμματική που να τις παράγει;

Η μέθοδος που ακολουθεί περιγράφει μία τεχνική δημιουργίας μιάς γραμματικής στην οποία το δείγμα που διαθέτουμε είναι ένα γνήσιο υποσύνολό της.

Ορισμός 21 *Το δείγμα της γλώσσας που διαθέτουμε είναι δομικά πλήρες όταν στο δείγμα αυτό έχουν χρησιμοποιηθεί όλοι οι κανόνες παραγωγής της γραμματικής.*

Στην πραγματικότητα ποτέ δεν είμαστε σίγουροι ότι το δείγμα της γλώσσας που έχουμε είναι δομικά πλήρες διότι απλούστατα δεν γνωρίζουμε τους κανόνες παραγωγής της. Είναι επίσης αλήθεια ότι οι γραμματικές που δημιουργούμε με τους αλγόριθμους που θα περιγραφούν αντιστοιχούν σε δομικά πλήρη δείγματα της γλώσσας.

Ορισμός 22 *Η γραμματική G_1 ονομάζεται απορρέουσα από την γραμματική G_2 όταν προκύπτει από την ενοποίηση μεταβλητών της γλώσσας G_2.*

Παράδειγμα 61 *Βρείτε την απορρέουσα γραμματική του προηγούμενου παραδείγματος που προκύπτει με την ενοποίηση των μεταβλητών (Z_1, Z_3, Z_5) σε Z_1, (Z_2, Z_6, Z_8) σε Z_2, και (Z_4, Z_7) σε Z_4.*

Η απορρέουσα γραμματική που περιγράφει την γλώσσα του παραδείγματος είναι η ακόλουθη:

$$G = (\{S, Z_1, Z_2, Z_4\}, \{a, b\}, P, S)$$

Οι κανόνες παραγωγής μετά την ενοποίηση και την απαλοιφή των ιδίων κανόνων γίνονται:

$$\begin{array}{lll} S \to aZ_1 & Z_1 \to b & S \to bZ_2 \\ Z_4 \to bZ_1 & Z_2 \to aZ_1 & Z_1 \to a \\ Z_2 \to aZ_1 & Z_1 \to bZ_4 & \end{array}$$

Η παραπάνω γραμματική περιγράφει ένα υπερσύνολο του δείγματος που διαθέτουμε. Δεν μπορούμε να απαντήσουμε με ακρίβεια στο ερώτημα αν η γραμματική αυτή μπορεί να περιγράψει πλήρως την γλώσσα που θέλουμε. Μπορεί να αποδειχθεί εύκολα ότι η απορρέουσα γραμματική μπορεί να παράγει όλες τις προτάσεις του δείγματος που διαθέτουμε.

Με την μέθοδο που περιγράψαμε μπορούμε να παράγουμε μεγάλο πλήθος γραμματικών. Η κατασκευή λοιπόν μιας κανονικής γραμματικής από δείγμα της γλώσσας της δεν έχει μονοσήμαντη λύση.

Στις περιπτώσεις που διαθέτουμε και αρνητικό δείγμα της γλώσσας (δηλαδή προτάσεις που δεν ανήκουν στην γλώσσα την γραμματικής την οποία προσπαθούμε να κατασκευάσουμε) τότε μετά την κατασκευή κάθε απορρέουσας γραμματικής αναλύουμε συντακτικά κάθε πρόταση του αρνητικού δείγματος έτσι ώστε να επιβεβαιώσουμε ότι κάθε αρνητικό δείγμα δεν ανήκει στην γλώσσα της γραμματικής. Αν βρεθεί έστω και μια πρόταση που ανήκει στην γλώσσα της γραμματικής που κατασκευάσαμε τότε πρέπει να αναζητήσουμε νέα απορρέουσα γραμματική.

5.5.2 Κατασκευή γραμματικής ελεύθερης σύνταξης

Η κατασκευή γραμματικής ελεύθερης σύνταξης είναι δυσκολότερο εγχείρημα από την κατασκευή μιας κανονικής γραμματικής γεγονός που οφείλεται στην μεγαλύτερη πολυπλοκότητα των κανόνων παραγωγής. Δεν έχει βρεθεί ακόμα αυτόματη μέθοδος κατασκευής γραμματικής ελεύθερης σύνταξης απο δείγμα της γλώσσας. Οι μέθοδοι που χρησιμοποιούνται είναι εμπειρικοί και βασίζονται σε μία διαδικασία επαναληπτικών δοκιμών. Γιαυτό τον λόγο η εμπειρία του σχεδιαστή παίζει σημαντικό ρόλο στην επιτυχία κατασκευής της γραμματικής.

Σε αυτή την παράγραφο παρουσιάζουμε γενικές οδηγίες κατασκευής μιας γραμματικής ελεύθερης σύνταξης από προτάσεις της γλώσσας.

Το θεώρημα που ακολουθεί μας βοηθά να αποφασίουμε **πότε** πρέπει να κατασκευάσουμε μία γραμματική ελεύθερης σύνταξης ελέγχοντας την δομή των προτάσεων. Το θεώρημα είναι γνωστό σαν pumping lemma.

Θεώρημα 16 *Σε κάθε γραμματική ελεύθερης σύνταξης $L(G)$ υπάρχουν ακέραιοι p και q τέτοιοι ώστε αν*

1. *Η συμβολοσειρά $s = abcde$ ανήκει στην γλώσσας της γραμματικής*
2. *Η συμβολοσειρά s έχει μήκος μεγαλύτερο του q,*
3. *το μήκος της bcd είναι μικρότερο του p,*
4. *το b, d δεν είναι συμβολοσειρές μηδενικού μήκους*

τότε κάθε συμβολοσειρά της μορφής $ab^icd^ie, i > 1$ ανήκει επίσης στην γλώσσα της γραμματικής.

Απο το παραπάνω θεώρημα βλέπουμε ότι έχουμε ελπίδες να αναγνωρίσουμε απο τις προτάσεις του δείγματος αν μία γλώσσα μπορεί να περιγραφεί από γραμματική ελεύθερης σύνταξης, βρίσκοντας ομάδες συμβολοσειρών που να μπορούν να περιγραφούν με βάση την σχέση που δίνεται στο θεώρημα.

Σε αυτή την περίπτωση το μέγεθος του δείγματος κρίνεται κρίσιμο διότι με αυτό τον τρόπο μπορούμε να βρούμε ευκολότερα τις σχέσεις που υπάρχουν στις συμβολοσειρές του δείγματος.

Παράδειγμα 62 *Εστω το δείγμα:*

$\Sigma = \{aba, abba, abbba, bab, baab, baaab, aabb\}$. *Βρές μια γραμματική ελεύθερης σύνταξης που να παράγει ένα υπερσύνολο του δείγματος.*

Μπορούμε να χωρίσουμε τα δείγματα σε τρείς ομάδες:

1. aba, abba, abbba
2. bab, baab, baaab
3. aabb

Οι κανόνες παραγωγής που μπορούν να παράγουν τις προτάσεις της πρώτης ομάδας είναι οι ακόλουθοι:

$$B \to BB \quad B \to b \quad S \to aBa$$

Οι κανόνες παραγωγής που παράγουν τις προτάσεις της δεύτερης ομάδας μπορούν να είναι και οι ακόλουθοι:

$$A \to AA \quad A \to a \quad S \to bAb$$

Η τρίτη πρόταση παράγεται από τον από κανόνα $S \to aabb$.

Η γραμματική που κατασκευάσαμε είναι:

$$G(L) = (\{S, A, B\}, \{a, b\}, P, S)$$

με κανόνες παραγωγής P:

$$\begin{array}{lll} B \to BB & B \to b & S \to aBa \\ A \to AA & A \to a & S \to bAb \\ S \to aabb & & \end{array}$$

5.6 Στοχαστικές γλώσσες και γραμματικές

Η συντακτική ταξινόμηση προτύπων που περιγράψαμε μπορεί να απαντήσει στο ερώτημα αν μία δοσμένη πρόταση μπορεί να παραχθεί από τους κανόνες της γραμματικής. Σε πραγματικές εφαρμογές πιο ρεαλιστικές, ευέλικτες και πιο αποδοτικές είναι συνήθως οι γλώσσες οι οποίες συνοδεύουν την παραγωγή κάθε πρότασης με ποσοτικό μέγεθος που δηλώνει την πιθανότητα με την οποία η γραμματική παράγει την συγκεκριμένη πρόταση. Οι γραμματικές αυτές ονομάζονται στοχαστικές γραμματικές.

Ορισμός 23 *Η γλώσσα L_s ονομάζεται στοχαστική όταν για κάθε της πρόταση υπάρχει αριθμός θετικός και μικρότερος της μονάδος ο οποίος δηλώνει την πιθανότητα εμφάνισης της πρότασης στην γλώσσα. Αν x είναι μία πρόταση στην στοχαστική γλώσσα L_s και $p(x)$ η αντίστοιχη πιθανότητα εμφάνισής της τότε ισχύει:*

$$\sum_{x \in L} p(x) = 1$$

Ορισμός 24 *Στοχαστική γραμματική G_s είναι κάθε διατεταγμένη τετράδα που αποτελείται από τα σύνολα $G_s = (V_N, V_T, S, P_s)$ και που το κάθε ένα από αυτά περιγράφει τα ακόλουθα στοιχεία της γραμματικής:*

V_T είναι το σύνολο των τερματικών συμβόλων,

V_N είναι το σύνολο των μη τερματικών συμβόλων ή μεταβλητών. Το σύνολο αυτό δεν περιέχει τερματικά σύμβολα, δηλ. $V_T \cap V_N = \emptyset$.

S είναι η μεταβλητή έναρξης και,

P είναι ένα διατεταγμένο σύνολο από ζεύγη συμβολοσειρών και έναν αριθμό που δηλώνει την πιθανότητα εφαρμογής του κανόνα παραγωγής κατά την διαδικασία παραγωγής των προτάσεων (C, D, p). Τα σύμβολα C, D αποτελούνται από τερματικά και μη τερματικά σύμβολα και το C περιέχει ένα τουλάχιστον μη τερματικό σύμβολο. Οι στοχαστικοί κανόνες παραγωγής παριστάνονται ως εξής: $C \xrightarrow{p} D$.

Ορισμός 25 *Γλώσσα μιας στοχαστικής γραμματικής είναι το σύνολο των παραγόμενων διατεταγμένων ομάδων (x, p_x) που αποτελούνται από μία συμβολοσειρά και την αντίστοιχη πιθανότητα παραγωγής της από το σύμβολο εκκίνησης:*

$$L(G_s) = \{(x, p(x)) \mid x \in V_T^*, S \overset{*}{\underset{p_j}{\Rightarrow}} x, j = 1, ..., k, p(x) = \prod_{j=1}^{k} p_j\}$$

Ορισμός 26 *Η στοχαστική γραμματική G_s ονομάζεται τύπου-1 ή ευαίσθητης σύνταξης όταν η αντίστοιχη γραμματική G που προκύπτει με την αφαίρεση των πιθανοτήτων από τους κανόνες παραγωγής είναι γραμματική ευαίσθητης σύνταξης.*

Ορισμός 27 *Η στοχαστική γραμματική G_s ονομάζεται τύπου-2 ή ελεύθερης σύνταξης όταν η αντίστοιχη γραμματική G που προκύπτει με την αφαίρεση των πιθανοτήτων από τους κανόνες παραγωγής είναι γραμματική ελεύθερης σύνταξης.*

Ορισμός 28 *Η στοχαστική γραμματική G_s ονομάζεται τύπου-3 ή στοχαστική κανονική γραμματική όταν η αντίστοιχη γραμματική G που προκύπτει με την αφαίρεση των πιθανοτήτων από τους κανόνες παραγωγής είναι κανονική γραμματική.*

5.7 Συνθήκες συνοχής

Για κάθε στοχαστική γραμματική ελεύθερης σύνταξης ισχύει ότι το άθροισμα των πιθανοτήτων παραγωγής για κάθε σύμβολο είναι ίσο με την μονάδα.

Εστω η στοχαστική γραμματική ελεύθερης σύνταξης $G_s = (V_N, V_T, S, P_s)$ για την οποία οι κανόνες παραγωγής για την τυχαία μεταβλητή A έχουν την μορφή:

$$A \xrightarrow{p_i^{(1)}} B_i C_i, i = 1, ..., I$$

$$A \xrightarrow{p_j^{(2)}} a_j, j = 1, ..., J$$

με $A, B_i, C_i, \in V_N, a_i \in V_T$.

Οι πιθανότητες της στοχαστικής γραμματικής πρέπει να υπακούουν στις ακόλουθες συνθήκες για κάθε μεταβλητή A:

$$\sum_{i=1}^{I} p_i^{(1)} + \sum_{j=1}^{J} p_j^{(2)} = 1$$

Οι κανόνες αυτοί ονομάζονται συνθήκες συνοχής της γραμματικής.

Προσοχή πρέπει να δοθεί στις προτάσεις οι οποίες μπορούν να παραχθούν με πολλαπλούς τρόπους από την στοχαστική γραμματική. Σε αυτές τις περιπτώσεις η πιθανότητα εμφάνισης της πρότασης ισούται με το άθροισμα των πιθανοτήτων παραγωγής της με όλους τους δυνατούς τρόπους.

Οι στοχαστικές γραμματικές διευκολύνουν σημαντικά την διαδικασία συντακτικής ανάλυσης διότι η εφαρμογή των κανόνων παραγωγής δεν είναι ισοπίθανη αλλά βεβαρυμένη με την πιθανότητα εφαρμογής των κανόνων. Χαρακτηριστικό παράδειγμα της χρησιμότητας που έχουν οι πιθανότητες εφαρμογής των κανόνων είναι η ιεράρχηση των κανόνων παραγωγής που θα χρησιμοποιήσουμε για την διαδικασία συνταχτικής ανάλυσης όταν έχουμε φτάσει σε κάποιο σημείο στο οποίο μπορούμε να επιτύχουμε περισσότερες από μία αντικαταστάσεις στις μεταβλητές ή τις σταθερές της πρότασης είτε χρησιμοποιούμε μεθόδους συντακτικής ανάλυσης προς την συμβολοσειρά είτε προς το σύμβολο εκκίνησης. Διατάσουμε λοιπόν κατά φθίνουσα σειρά πιθανότητας εφαρμογής τους κανόνες και κατόπιν να επιλέγουμε την χρήση των πλέον πιθανών έτσι ώστε να μεγιστοποιήσουμε την πιθανότητα να βρούμε την πρόταση σύντομα εάν βέβαια η πρόταση ανήκει στην γλώσσα της γραμματικής.

Επιπρόσθετα με την χρήση στοχαστικών γραμματικών έχουμε την δυνατότητα να εκτιμήσουμε και να διορθώσουμε συντακτικά λάθη σε προτάσεις.

Παράδειγμα 63 *Εστω η στοχαστική γραμματική ελεύθερης σύνταξης*

$$G_s = (\{A, B, S\}, \{a, b\}, S, P_s)$$

με κανόνες παραγωγής P_s*:*

$$S \xrightarrow{1} bA \quad A \xrightarrow{0.9} aB \quad A \xrightarrow{0.1} b$$
$$B \xrightarrow{0.4} a \quad B \xrightarrow{0.6} bS$$

η πρόταση baa μπορεί να παραχθεί από την γραμματική G_s*;.*

Αποδεικνύεται ότι η πρόταση baa μπορεί να παραχθεί από την γραμματική G_s ακολουθώντας την μέθοδο συνταχτικής ανάλυσης προς την συμβολοσειρά:

$$S \overset{S \xrightarrow{1} bA}{\Rightarrow} bA \overset{A \xrightarrow{0.9} aB}{\Rightarrow} baB \overset{B \xrightarrow{0.4} a}{\Rightarrow} baa$$

με πιθανότητα παραγωγής της πρότασης από την γραμματική: $p(baa) = 1 * 0.9 * 0.4 = 0.36$.

Είναι ευνόητο ότι όλοι οι υπόλοιποι ορισμοί και τα θεωρήματα που έχουν αναφερθεί στα προηγούμενα κεφάλαια για τις τυπικές γραμματικές ισχύουν και στην περίπτωση των στοχαστικών γραμματικών.

5.7.1 Εκπαίδευση στοχαστικών γραμματικών

Οταν γνωρίζουμε τους κανόνες παραγωγής της γραμματικής αλλά είναι άγνωστες οι αντίστοιχες πιθανότητές τους, οι συνθήκες συνοχής δεν επαρκούν συνήθως για να προσδιορίσουν τις αριθμητικές

τιμές των πιθανοτήτων. Χρειαζόμαστε επιπρόσθετη πληροφορία που προέρχεται συνήθως από πειραματικές μετρήσεις της συχνότητας παραγωγής προτάσεων. Με αυτές τις μετρήσεις με τις οποίες μπορούμε να υπολογίσουμε προσεγγιστικά την πιθανότητα παραγωγής των προτάσεων του δείγματος που διαθέτουμε. Η πληροφορία αυτή συμπληρώνει τον απαραίτητο αριθμό εξισώσεων που χρειάζεται για την λύση ενός συστήματος εξισώσεων.

Ενδεικτικό της μεθοδολογίας που ακολουθούμε κατά τον υπολογισμό των πιθανοτήτωντων κανόνων παραγωγγής είναι το παράδειγμα που ακολουθεί.

Παράδειγμα 64 *Εστω η στοχαστική γραμματική ελεύθερης σύνταξης*

$$G_s = (\{A, B, S\}, \{a, b\}, S, P_s)$$

με κανόνες παραγωγής P_s*:*

$$S \xrightarrow{p_s} bA \quad S \xrightarrow{1-p_s} AB \quad A \xrightarrow{p_a} aB$$
$$A \xrightarrow{1-p_a} a \quad B \xrightarrow{p_b} b \quad B \xrightarrow{1-p_b} bS$$

Τι επιπλέον πληροφορίες χρειάζεστε για να υπολογίσετε τις πιθανότητες των κανόνων παραγωγής;

Οι συνθήκες συνοχής έχουν ήδη χρησιμοποιηθεί για να ελαττώσουν τον αριθμό των άγνωστων μεταβλητών. Χρειαζόμαστε επιπλέον τρείς προτάσεις της γλώσσας με τις αντίστοιχες πιθανότητες εμφάνισης τους έτσι ώστε να μπορέσουμε να φτιάξουμε ένα σύστημα τριών εξισώσεων με τρείς αγνώστους. Απο πειράματα πληροφορούμεθα τρείς πιθανότητες εμφάνισης προτάσεων, τις ακόλουθες:

p(ab) = 0.004,

p(ba) = 0.1,

και p(bab) = 0.001.

Η συντακτική ανάλυση πρός την συμβολοσειρά για κάθε μία από τις προτάσεις μας δίνει τις ακόλουθες εξισώσεις:

$$S \overset{1-p_s}{\underset{S \to AB}{\Rightarrow}} AB \overset{p_b}{\underset{B \to b}{\Rightarrow}} Ab \overset{1-p_a}{\underset{A \to a}{\Rightarrow}} ab$$

Η πιθανότητα παραγωγής της πρότασης από την γραμματική μας δίνει την πρώτη εξίσωση: $2(1-p_s)p_b(1-p_a) = 0.004$

$$S \overset{p_s}{\underset{S \to bA}{\Rightarrow}} bA \overset{1-p_a}{\underset{A \to a}{\Rightarrow}} ba$$

Η πιθανότητα παραγωγής της πρότασης από την γραμματική μας δίνει την δεύτερη εξίσωση: $p_s(1-p_a) = 0.1$

$$S \overset{p_s}{\underset{S \to bA}{\Rightarrow}} bA \overset{p_a}{\underset{A \to aB}{\Rightarrow}} baB \overset{p_b}{\underset{B \to b}{\Rightarrow}} bab$$

Η πιθανότητα παραγωγής της πρότασης από την γραμματική μας δίνει την τρίτη εξίσωση: $p_s p_a p_b = 0.001$

Λύνουμε την δεύτερη εξίσωση ως προς p_s και το αντικαθιστούμε στην πρώτη και τρίτη εξίσωση:

$$(1 - \frac{1}{10(1-p_a)})p_b(1-p_a) = 0.002$$

$$p_s = \frac{1}{10(1-p_a)}$$

$$\frac{1}{10(1-p_a)p_a p_b} = 0.001$$

Λύνουμε την πρώτη εξίσωση ως προς p_b και αντικαθιστούμε στην τρίτη εξίσωση το p_b οπότε έχουμε:

$$\frac{0.002p_a}{(100p_a - 99)(p_a - 1)} = 0.001$$

Η εξίσωση έχει δύο λύσεις η μία των οποίων είναι εκτός του πεδίου τιμών ποσοτήτων πιθανότητας που είναι το κλειστό διάστημα $[0, 1]$ και γιαυτό τον λόγο απορρίπτεται.

Αντικαθιστώντας στις προηγούμενες εξισώσεις έχουμε μία λύση για τις πιθανότητες της στοχαστικής γραμματικής:

$$p_a = 0.86349 \quad p_b = 0.015809 \quad p_s = 0.0732547$$

Παράδειγμα 65 *Ο μηχανικός που αντικαθιστούμε μελετούσε ένα στοχαστικό φαινόμενο που παρήγαγε συμβολοσειρές. Στις σημειώσεις του βρήκαμε ότι μοντελοποίησε το φαινόμενο με στοχαστική γραμματική και οι μετρήσεις του για την συχνότητα εμφάνισης των κανόνων δίνονται στον πίνακα 5.1.*

Πίνακας 5.3: Πίνακας συχνότητας εφαρμογής κανόνων του παραδείγματος 30

Κανόνας	$S \to SA$	$S \to AB$	$A \to Sa$	$A \to a$	$B \to Ba$
Συχνότητα	10	50	30	10	10
Κανόνας	$B \to Ba$	$B \to b$			
Συχνότητα	30	4			

Συνεχίσαμε τις μετρήσεις του προκατόχου μας και πήραμε δύο νέες προτάσεις, τις: abaaa, aaba.

Μπορούμε άραγε να προσδιορίσουμε την πιθανότητα εμφάνισης της πρότασης aabaa ;

Παράδειγμα 66 *Φυσικό φαινόμενο προσομοιώθηκε με στοχαστική γραμματική ελεύθερης σύνταξης*

$$G_s = (\{A, B, S\}, \{a, b\}, S, P_s)$$

που περιέχει του ακόλουθους κανόνες παραγωγής:

$$S \xrightarrow{p_1} bA \quad S \xrightarrow{p_2} aB \quad S \xrightarrow{p_3} A$$
$$A \xrightarrow{p_4} AA \quad A \xrightarrow{p_5} a \quad B \xrightarrow{p_6} BB$$
$$B \xrightarrow{p_7} b$$

Υπολογίστε τις πιθανότητες των κανόνων παραγωγής όταν είναι γνωστές οι συχνότητες εμφάνισης προτάσεων που δίνονται στον πίνακα 5.2.

Πίνακας 5.4: Πίνακας μετρήσεων συχνότητας εμφάνισης προτάσεων για το παραδ. 32

Πρόταση	aa	ba	abb	abbb	baa	baaaa
Συχνότητα	30	76	4	1	22	2

Τι συμπεράσματα μπορείτε να βγάλετε για το μοντέλο που φτιάξατε;

Αρχικά μπορούμε να διαπιστώσουμε ότι από τις συνθήκες συνέχειας της στοχαστικής γραμματικής οι άγνωστες μεταβλητές δεν είναι επτά αλλά μονάχα τέσσερις:

$$S \xrightarrow{p_1} bA \quad S \xrightarrow{1-p_1-p_3} aB \quad S \xrightarrow{p_3} A$$
$$A \xrightarrow{1-p_5} AA \quad A \xrightarrow{p_5} a \quad B \xrightarrow{1-p_7} BB$$
$$B \xrightarrow{p_7} b$$

Οι μετρήσεις που έχουμε αφορούν μετρήσεις συχνότητας εμφάνισης έξι προτάσεων, γεγονός που μας προβληματίζει διότι αν ακολουθήσουμε την μεθοδολογία του παραδείγματος 64 θα έχουμε πλεονάζοντες εξισώσεις (οι εξισώσεις θα είναι περισσότερες από τις μεταβλητές).

Γιαυτό τον λόγο είναι προτιμότερο να χρησιμοποιήσουμε την μέθοδο του στατιστικού υπολογισμού των πιθανοτήτων παραγωγής των κανόνων από τον συχνότητα χρήσης τους στο πειραματικό δείγμα. Με αυτό τον τρόπο η πιθανότητα εφαρμογής του κανόνα $A_i \xrightarrow{p_{ij}} B_j$ προκύπτει από την σχέση:

$$p_{ij} \approx \frac{n_{ij}}{\sum_j n_{ij}}$$

n_{ij} είναι ο αριθμός εφαρμογής των κανόνων στις πειραματικές μετρήσεις. Ο παρονομαστής εκφράζει την συχνότητα εφαρμογής κανόνα με σύμβολο στο αριστερό του τμήμα το A_i.

Με βάση το παραπάνω σκεπτικό εκτελούμε συνταχτική ανάλυση για κάθε μία πρόταση έτσι ώστε να βρούμε πόσες φορές χρησιμοποιούμε κάθε κανόνα για να παράγουμε τις προτάσεις του δείγματος που διαθέτουμε.

Συμπληρώνουμε τον ακόλουθο πίνακα για κάθε μία από τις προτάσεις και για κάθε έναν από τους κανόνες παραγωγής:

Πίνακας 5.5: Χρήση κανόνων στις προτάσεις

Πρόταση	aa	ba	abb	abbb	baa	baaaa
$S \xrightarrow{p_1}$ b A	0	1	0	0	1	1
$S \xrightarrow{p_2}$ a B	0	0	1	1	0	0
$S \xrightarrow{p_3}$ A	1	0	0	0	0	0
$A \xrightarrow{p_4}$ A A	1	0	0	0	1	3
$A \xrightarrow{p_5}$ a	2	1	0	0	2	4
$B \xrightarrow{p_6}$ B B	0	0	1	2	0	0
$B \xrightarrow{p_7}$ b	0	0	2	3	0	0
Συχνότητα	30	76	4	1	22	2

Υπολογίζουμε:

1. Την συχνότητα εφαρμογής των κανόνων και
2. Τις αντίστοιχες συχνότητες εφαρμογής ομάδων κανόνων παραγωγής που έχουν στο αριστερό τους τμήμα την ίδια μεταβλητή για όλες τις προτάσεις του δείγματος.

Πίνακας 5.6: Επαναλήψεις χρήσης κανόνων στις προτάσεις του δείγματος

Πρόταση	Συχνότητα	Συχνότητα ομάδος	Πιθανότητα
$S \xrightarrow{p_1}$ b A	100	135	$p_1 = \frac{100}{135}$
$S \xrightarrow{p_2}$ a B	5	135	$p_2 = \frac{5}{135}$
$S \xrightarrow{p_3}$ A	30	135	$p_3 = \frac{30}{135}$
$A \xrightarrow{p_4}$ A A	58	246	$p_4 = \frac{58}{146}$
$A \xrightarrow{p_5}$ a	188	246	$p_5 = \frac{188}{246}$
$B \xrightarrow{p_6}$ B B	6	17	$p_6 = \frac{6}{17}$
$B \xrightarrow{p_7}$ b	11	17	$p_7 = \frac{11}{17}$

Παράδειγμα 67 *Δίνονται οι στοχαστικές γραμματικές*

$$G_{s1} = (\{A, B, S\}, \{a, b\}, S, P_{s1})$$

με κανόνες παραγωγής:

$$S \xrightarrow{0.8} bA \quad S \xrightarrow{0.1} aB \quad S \xrightarrow{0.1} A$$
$$A \xrightarrow{0.2} AA \quad A \xrightarrow{0.8} a \quad B \xrightarrow{0.1} BB$$
$$B \xrightarrow{0.9} b$$

$G_{s2} = (\{X, Y, S\}, \{a, b\}, S, P_{s2})$ *με κανόνες παραγωγής:*

$$S \xrightarrow{0.8} SX \quad S \xrightarrow{0.2} Y \quad X \xrightarrow{0.6} Xa$$
$$X \xrightarrow{0.4} a \quad Y \xrightarrow{1} b$$

$G_{s3} = (\{X, Y, S\}, \{a, b\}, S, P_{s3})$ *με κανόνες παραγωγής:*

$$S \xrightarrow{0.8} SX \quad S \xrightarrow{0.2} Y \quad X \xrightarrow{0.6} Xa$$
$$X \xrightarrow{0.4} a \quad Y \xrightarrow{0.4} b \quad Y \xrightarrow{0.6} ba$$

Ποιά γραμματική είναι πιο πιθανόν να παράγει την συμβολοσειρά baaaa;.

Οι στοχαστικές γραμματικές είναι ασαφείς και συνεπώς τα μεγέθη που θα υπολογισθούν δεν θα εκφράζουν μεγέθη πιθανοτήτων. Η γραμματική από την οποία μπορούμε να παράγουμε με την μεγαλύτερη "πιθανότητα" την πρόταση baaaa είναι και η πλέον πιθανή γραμματική.

Η πιθανότητα παραγωγής της πρότασης από την πρώτη στοχαστική γραμματική είναι:

$$S \overset{B \to bA}{\Rightarrow} bA \overset{A \to AA}{\Rightarrow} bAAAA \overset{A \to a}{\Rightarrow} baaaa$$

με $p(baaaa) = 0.8 * (6 * 0.2^3) * (24 * 0.8^4) = 0.33774$.

Η πρόταση baaaa είναι συνταχτικά ασαφής για την δεύτερη γραμματική διότι μπορεί να παραχθεί με δύο διαφορετικούς τρόπους.

$$S \overset{S \to SX}{\Rightarrow} SX \overset{X \to Xa}{\Rightarrow} SXaaa \overset{Y \to b, X \to a}{\Rightarrow} baaaa$$

με $p(baaaa) = 0.8 * (0.6^3) * 2 * 0.4 = 0.13824$.

$$S \overset{S \to SX}{\Rightarrow} SXXXX \overset{S \to Y}{\Rightarrow} YXXXX \overset{Y \to b, X \to a}{\Rightarrow} baaaa$$

με $p(baaaa) = 0.8^4 * 0.2 * (120 * 0.4^4) = 0.2516$.

Η πιθανότητα παραγωγής της πρότασης από την δεύτερη στοχαστική γραμματική είναι:

$$p(baaaa) = 0.13824 + 0.2516 = 0.38984.$$

Με τον ίδιο τρόπο υπολογίζεται και η πιθανότητα παραγωγής της πρότασης από την τρίτη στοχαστική γραμματική:

$$S \overset{S \to SX}{\Rightarrow} SXXXX \overset{S \to Y}{\Rightarrow} YXXXX \overset{Y \to b, X \to a}{\Rightarrow} baaaa$$

με $p(baaaa) = 0.8^4 * 0.2 * (120 * 0.6 * 0.4^4) = 0.15099$.

$$S \overset{S \to SX}{\Rightarrow} SXXX \overset{S \to Y}{\Rightarrow} YXXX \overset{Y \to ba, X \to a}{\Rightarrow} baaaa$$

με $p(baaaa) = 0.8^3 * 0.2 * (24 * 0.4 * 0.4^3) = 0.06291$.

Αθροιστικά η συνολική πιθανότητα παραγωγής της πρότασης από την τρίτη στοχαστική γραμματική είναι 0.21390.

Συνεπώς η πρόταση baaaa είναι πιθανότερο να παραχθεί από την δεύτερη στοχαστική γραμματική.

5.8 Αναγνώριση αλλοιωμένων προτάσεων

Στην ανάλυση που προηγήθηκε θεωρήσαμε ότι οι προτάσεις που αναγνωρίζαμε δεν είχαν προηγουμένως υποστεί παραμορφώσεις είτε διότι είχαν μεταφερθεί από δίαυλο ο οποίος δεν πρόσθεσε θόρυβο και η γραμματική που τελικά κατασκευάστηκε είναι πλήρης, παρόλο που σε πρακτικές εφαρμογές σπάνια μπορούμε να ικανοποιήσουμε αυτές τις δύο υποθέσεις.

Το γεγονός πάντως είναι ότι ένα από τα συνηθέστερα προβλήματα που έχουμε να αντιμετωπίσουμε σε συστήματα συντακτικής ταξινόμησης προτύπων είναι η εύρεση κάποιας μεθόδου με την οποία θα είμαστε σε θέση να μετρούμε την "διαφορά", "απόσταση" ή "ομοιότητα" οποιασδήποτε πρότασης από γραμματική. Με αυτό τον τρόπο αποφεύγουμε καταστάσεις στις οποίες **δεν μπορούμε** να ταξινομήσουμε ορισμένο αριθμό προτάσεων. Η ύπαρξη τέτοιων καταστάσεων αποδεικνύεται από το επόμενο παράδειγμα:

Παράδειγμα 68 *Εστω η γραμματική* $G_1 = (\{S\}, \{a, b\}, \{S \to SS, S \to ab\}, S)$ *και η γραμματική* $G_2 = (\{S\}, \{a, b\}, \{S \to SS, S \to ba\}, S)$

Σε ποιά γλώσσα ανήκει η πρόταση aba;

Εφαρμόζοντας τον αλγόριθμο CYK βρίσκουμε ότι ακόμα και αν οι γραμματικές που διαθέταμε ήταν στοχαστικές, πάλι δεν θα μπορούσαμε να ταξινομήσουμε την πρόταση σε κάποια από τις γλώσσες των γραμματικών διότι βρίσκουμε ότι η πρόταση aba δεν ανήκει σε καμμία από αυτές.

Ακόμα και και στην περίπτωση κατά την οποία γνωρίζουμε ότι η πρόταση που διαθέτουμε ανήκει πράγματι σε κάποια απο τις δύο γραμματικές η πρόταση δεν μπορεί να ταξινομηθεί διότι είναι το αποτέλεσμα μιας διαδικασίας παραμόρφωσης κάποιας άλλης πρότασης η οποία πραγματικά περιέχεται σε μία από τις δύο γλώσσες των γραμματικών.

Η κατασκευή συστήματος ταξινόμησης προτάσεων το οποίο θα είναι σε θέση να ταξινομεί οποιαδήποτε πρόταση απαιτεί την υιοθέτηση κάποιου μοντέλου που να περιγράφει τον τρόπο με τον οποίο παραμορφώνονται οι προτάσεις. Γενική λύση σε αυτή την περίπτωση δεν υπάρχει, διότι καθοριστικό ρόλο στην μοντελοποίηση του τρόπου παραμόρφωσης των προτύπων παίζει το φυσικό φαινόμενο που την προκαλεί.

Οι απλούστερες λύσεις αυτού του προβλήματος δίνονται σε αυτό το εδάφιο. Αρχικά ορίζουμε έναν τρόπο μέτρησης απόστασης προτάσεων και ακολούθως ορίζουμε το κριτήριο ταξινόμησης.

5.8.1 Απόσταση προτάσεων

Η συνάρτηση απόστασης προτάσεων που θα ορίσουμε πρέπει να ακολουθεί τις ισχυρές συνθήκες και να ικανοποιεί όσο γίνεται περισσότερες ασθενείς συνθήκες όπως αυτές έχουν ήδη περιγραφεί στο κεφάλαιο 2. Για να παρουσιάσουμε ένα σύστημα ταξινόμησης προτάσεων το οποίο παρουσιάζει σημαντική διάδοση σε πρακτικά συστήματα πρέπει να κάνουμε σε αυτό το σημείο μία υπόθεση. Θεωρούμε ότι η διαδικασία παραμόρφωσης των προτάσεων επιδρά σε κάθε ένα τερματικό σύμβολο ανεξάρτητα από τις τυχόν παραμορφώσεις στα γειτονικά σύμβολα, γεγονός που γενικά ισχύει σε μεγάλο αριθμό εφαρμογών.

Ορίζουμε την απόσταση των προτάσεων x, y σαν μία συνάρτηση που εξαρτάται από τον ελάχιστο αριθμό στοιχειωδών μετασηματισμών σε επίπεδο τερματικών συμβόλων με τους οποίους μπορούμε από την πρόταση x να λάβουμε την πρόταση y:

Ορισμός 29 *Απόσταση Levenshtein μεταξύ των προτάσεων* $x, y \in V_T^*$ *ορίζεται το μέγεθος* $d_T(x, y)$ *το οποίο υπολογίζεται σαν ο ελάχιστος αριθμός των μετασχηματισμών αντικατάστασης, απαλειφής και παραμβολής με τους οποίους λαμβάνουμε από την πρόταση x την πρόταση y.*

1. Μετασχηματισμός αντικατάστασης είναι κάθε αντικατάσταση τερματικού συμβόλου:

$$abc \rightarrow adc, \quad b, d \in V_T, \quad b \neq d, \quad a, c \in V_T{}^* \tag{5.2}$$

2. Μετασχηματισμός απαλειφής είναι η εξάλειψη τερματικού συμβόλου από την πρόταση:

$$abc \rightarrow ac, \quad b \in V_T, \quad a, c \in V_T{}^* \tag{5.3}$$

3. Μετασχηματισμός παρεμβολής είναι η προσθήκη τερματικού συμβόλου στην πρόταση:

$$ac \rightarrow adc, \quad d \in V_T, \quad a, c \in V_T{}^* \tag{5.4}$$

Παράδειγμα 69 *Βρείτε την απόσταση Levenshtein των προτάσεων abaabbbb και aabaaabba.*

Ο ελάχιστος αριθμός μετασχηματισμών που μπορούμε να εκτελέσουμε και να μεταβούμε απο την μία πρόταση στην αλλη είναι $d_T(abaabbbb, aabaaabba) = 3$.

Ορισμός 30 *Σταθμισμένη απόσταση Levenshtein μεταξύ των προτάσεων $x, y \in V_T{}^*$ ορίζεται το μέγεθος*

$$d_T(x, y) = \min_K (N_\alpha w_\alpha + N_\lambda w_\lambda + N_\pi w_\pi) \tag{5.5}$$

όπου $w_\alpha, w_\lambda, w_\pi$ είναι πραγματικοί αριθμοί που χαρακτηρίζουν την βαρύτητα των μετασχηματισμών αντικατάστασης, απαλειφής και παραμβολής αντίστοιχα, ενώ $N_\alpha, N_\lambda, N_\pi$ είναι ο αριθμός των αντίστοιχων μετασχηματισμών.

Πολλές φορές συνηθίζεται οι συντελεστές βαρύτητας να κανονικοποιούνται έτσι ώστε να ισχύει: $w_\alpha + w_\lambda + w_\pi = 1$.

Παράδειγμα 70 *Βρείτε την σταθμισμένη απόσταση Levenshtein των προτάσεων abaabbbb και aabaaabba, όταν οι συντελεστές βαρύτητας των μετασχηματισμών έχουν τιμές $w_\alpha = 0.1$, $w_\lambda = 0.6$ και $w_\pi = 0.3$.*

Η σταθμισμένη απόσταση αποτελεί το άθροισμα μιας απαλειφής συμβόλου και δύο αντικαταστάσεων. Συνεπώς $d_T(abaabbbb, aabaaabba) = 0.6 + 2 \times 0.1 = 0.8$.

Σε εφαρμογές συστημάτων οπτικής ταξινόμησης χαρακτήρων η λανθασμένη ταξινόμηση τυπογραφικών χαρακτήρων είναι ισχυρά ανομοιόμορφη. Τα λάθη αντικατάστασης του χαρακτήρα της τελείας με κόμμα, των τονούμενων φωνηέντων με τα αντίστοιχα φωνήεντα ή του κεφαλαίου γράμματος όμικρον με τον χαρακτήρα του αριθμού μηδέν είναι συντριπτικώς περισσότερα από τα υπόλοιπα λάθη αντικατάστασης.

Συνεπώς αν θα θέλαμε να κατασκευάσουμε ένα σύστημα αυτόματης διόρθωσης σφαλμάτων το οποίο θα αναγνώριζε παραμορφωμένες λέξεις από ένα λεξικό της Ελληνικής γλώσσας θα έπρεπε η συνάρτηση απόστασης η οποία θα μετρούσε την απόσταση της λέξης απο τις αντίστοιχες λέξεις του λεξικού να λαμβάνει υπόψη της το γεγονός ότι οι μετασχηματισμοί αντικατάστασης, απαλειφής και παραμβολής δεν έχουν την ίδια βαρύτητα αλλά εξαρτώνται και από το είδος του χαρακτήρα (το τερματικό σύμβολο) που ταξινομείται.

Ορισμός 31 *Σταθμισμένη απόσταση μεταξύ των προτάσεων $x, y \in V_T{}^*$ ορίζεται το μέγεθος $d_\sigma(x, y)$ το οποίο υπολογίζεται σαν το άθροισμα του κόστους ή της βαρύτητας των μετασχηματισμών με τους οποίους μεταβαίνουμε από την πρόταση x στην πρόταση y.*

Οταν χρησιμοποιήσουμε την συνάρτηση σταθμισμένης απόστασης προτάσεων τότε θα πρέπει να υπολογίσουμε κατά την διαδικασία εκπαίδευσης τα κόστη ή τους συντελεστές βαρύτητας των ακόλουθων μετασχηματισμών:

1. Μετασχηματισμός αντικατάστασης:

$$abc \overset{k_\alpha(b,d)}{\rightarrow} adc, \quad b, d \in V_T, \quad b \neq d, \quad a, c \in V_T{}^* \tag{5.6}$$

2. Μετασχηματισμός απαλειφής:

$$abc \overset{k_\lambda(b)}{\rightarrow} ac, \quad b \in V_T, \quad a, c \in V_T{}^* \tag{5.7}$$

3. Μετασχηματισμός παρεμβολής:

$$ac \overset{k_\pi(d)}{\rightarrow} adc, \quad d \in V_T, \quad a, c \in V_T{}^* \tag{5.8}$$

Ο αριθμός των συντελεστών κόστους των αντικαταστάσεων είναι ίσος με το τετράγωνο του αριθμού των τερματικών συμβόλων για τους μετασχηματισμούς αντικατάστασης και ίσος με τον αριθμό των τερματικών συμβόλων για τους μετασχηματισμούς απαλειφής και παρεμβολής.

Η επιλογή της συνάρτησης απόστασης προτάσεων που πρέπει να χρησιμοποιήσουμε στο σύστημα ταξινόμησης ή διόρθωσης προτάσεων εξαρτάται από το είδος του προβλήματος που αντιμετωπίζουμε. Αν θέλουμε να διορθώσουμε τυχαία σφάλματα που παραμορφώνουν προτάσεις που μεταδίδονται από κάποιο τηλεπικοινωνιακό κανάλι λόγω ύπαρξης λευκού θορύβου τότε η σταθμισμένη απόσταση Levenshtein καλύπτει συνήθως το είδος των σφαλμάτων που συναντώνται. Αν χρησιμοποιούμε σύστημα διόρθωσης σφαλμάτων σε συστήματα ταξινόμησης ομιλίας ή οπτικής ταξινόμησης τυπογραφικών λέξεων, τότε η σταθμισμένη απόσταση προτάσεων $d_\sigma(x, y)$ αποτελεί την προσφορότερη λύση.

Η μέθοδος που ακολουθούμε για να υπολογίσουμε την απόσταση των προτάσεων με το κριτήριο ελάχιστου κόστους βασίζεται στον δυναμικό προγραμματισμό. Η συνάρτηση συνολικής απόστασης των προτάσεων $\alpha(t) = a_1 a_2 ... a_t$ και $\beta(t) = b_1 b_2 ... b_t$ υπολογίζεται αναδρομικά ως εξής:

$$D(a_1 a_2 ... a_t, b_1 b_2 ... b_t) = \min \begin{cases} D(a_1 a_2 ... a_{t-1}, b_1 b_2 ... b_{t-1}) + k_\alpha(a_t, b_t) \\ D(a_1 a_2 ... a_t, b_1 b_2 ... b_{t-1}) + k_\pi(b_t) \\ D(a_1 a_2 ... a_{t-1}, b_1 b_2 ... b_t) + k_\lambda(a_t) \end{cases} \tag{5.9}$$

Το μέγεθος $D(a_1 a_2 ... a_t, b_1 b_2 ... b_t)$ ονομάζεται συνολική απόσταση της πρότασης $a_1 a_2 ... a_t$ από την πρόταση $b_1 b_2 ... b_t$ ενώ ο δεύτερος προσθετικός όρος είναι τα κόστη των μετασχηματισμών.

Η απόσταση των προτάσεων $a_1 a_2 ... a_N$ και $b_1 b_2 ... b_N$ δίνεται από την συνολική απόσταση $D(a_1 a_2 ... a_N, b_1 b_2 ..$

Παράδειγμα 71 *Υπολογείστε την απόσταση των λέξεων της Ελληνικής γλώσσας "ΚΑΘΕ" και "ΚΑΝΕΝΑ" λαμβάνοντας υπόψη ότι το κόστος μετασχηματισμού αντικατάστασης είναι ίσο με την αλφαβητική τους απόσταση ενώ οι μετασχηματισμοί απαλειφής και παραμβολής έχουν σταθερό κόστος 10.*

Πίνακας 5.7: Απόσταση των λέξεων "ΚΑΘΕ" και "ΚΑΝΕΝΑ"

	Κ	Α	Ν	Ε	Ν	Α
Κ	0	10	20	30	40	50
Α	10	0	30	24	34	40
Θ	20	10	5	15	25	35
Ε	30	20	15	5	15	25

Στον πίνακα 5.7 δίνεται ο πίνακας των συνολικών αποστάσεων για τις δύο λέξεις όπως αυτές υπολογίζονται αναδρομικά με την μέθοδο του δυναμικού προγραμματισμού κατά αύξουσα σειρά γραμμών και στηλών.

Η συνολική απόσταση των λέξεων "ΚΑΘΕ" και "ΚΑΝΕΝΑ" είναι 25.

Είναι φανερό ότι εφόσον ορίσαμε τον τρόπο μέτρησης της απόστασης μπορούμε να κατασκευάσουμε οποιοδήποτε δομικό σύστημα ταξινόμησης προτάσεων όπως αυτά έχουν ήδη περιγραφεί στο κεφάλαιο 2.

5.9 Λυμένα Προβλήματα

Πρόβλημα 7 *Εστω η γραμματική* $G = (\{S, A, B\}, \{0,1,2,3,4,5,6,7,8,9\}, P, S)$ *με κανόνες γραμματικής:*

$P = \{S \rightarrow AS|ABS|BA|BB, B \rightarrow BAB,\ A \rightarrow AA|AB|1|2|5|7|8|9, B \rightarrow 0|1|3|4|5|6|8\}$

Με $A \rightarrow B|C <=> A \rightarrow B, A \rightarrow C$.

Χρησιμοποιειώντας τον αλγόριθμο CYK βρείτε αν η αριθμοσειρά 235874 ανήκει στην γλώσσα της γραμματικής G.

Λύση:

1. Η γραμματική είναι ελεύθερης σύνταξης διότι κάθε κανόνας παραγωγής περιέχει μία μεταβλητή στο αριστερό τμήμα των κανόνων παραγωγής.

2. Η γραμματική G δεν βρίσκεται σε μορφή Chomsky. Αρχικά λοιπόν πρέπει να βρούμε μία ισοδύναμη γραμματική σε μορφή Chomsky.

Οι κανόνες που δεν βρίσκονται σε μορφή Chomsky είναι δύο:

$$S \rightarrow ABS, B \rightarrow BAB$$

με την βοήθεια νέων μεταβλητών φτιάχουμε ισοδύναμους κανόνες:

$$S \rightarrow ABS <=> S \rightarrow AK, K \rightarrow BS$$

$$B \rightarrow BAB <=> B \rightarrow BL, L \rightarrow AB$$

Μια ισοδύναμη γραμματική σε μορφή Chomsky είναι η ακόλουθη:

$G' = (\{S, A, B, K, L\}, \{0,1,2,3,4,5,6,7,8,9\}, P', S)$ με κανόνες γραμματικής

$$P' = \{S \to AS|AK|BA|BB, K \to BS, B \to BL, L \to AB,$$

$$A \to AA|AB|1|2|5|7|8|9, B \to BL|0|1|3|4|5|6|8\}$$

Πίνακας 5.8: Πίνακας CYK

	2	3	5	8	7	4
1	A	B	A,B	A,B	A	B
2	L,A	S	A,L,S	A,S	L,A	
3	S,A,L	S,B,K	A,S,K	A,B,S,L		
4	S,L,A	S,K	A,L,S,K,B			
5	S,A	S,B,K				
6	S,L,A					

Το σύμβολο εκκίνησης περιέχεται στην τελευταία θέση του πίνακα. Συνεπώς η αριθμοσειρά 235874 περιέχεται στην γλώσσα της γραμματικής.

Πρόβλημα 8 *Βρείτε μία κανονική γραμματική η οποία να περιέχει τους λιγότερους δυνατούς κανόνες παραγωγής και η γλώσσα της γραμματικής να περιέχει τουλάχιστον τις ακόλουθες προτάσεις:*

$$\{aaabb, abab, aabb, abb, ab\}$$

Επιπλέον η γλώσσα της γραμματικής **δεν** *πρέπει να περιέχει τις ακόλουθες προτάσεις:*

$$\{abba, aba, bbaa, ba\}$$

Λύση:

Για κάθε μία από τις προτάσεις του θετικού δείγματος φτιάχνουμε μια κανονική γραμματική η οποία να παράγει την συγκεκριμένη πρόταση:

$$aaabb : S \to aZ_1, Z_1 \to aZ_2, Z_2 \to aZ_3, Z_3 \to bZ_4, Z_4 \to b$$

$$abab : S \to aZ_5, Z_5 \to bZ_6, Z_6 \to aZ_7, Z_7 \to b$$

$$aabb : S \to aZ_8, Z_8 \to aZ_9, Z_9 \to bZ_{10}, Z_{10} \to b$$

$$abb : S \to aZ_{11}, Z_{11} \to bZ_{12}, Z_{12} \to b$$

$$ab : S \to aZ_{13}, Z_{13} \to b$$

$$ab : S \to aZ_{16}, Z_{17} \to b$$

Ενοποιούμε τις παραμέτρους $\{Z_1, Z_5, Z_8, Z_{11}, Z_{13},\}$ και $\{Z_4, Z_7, Z_{10}, Z_{12}, Z_{13}\}$. Τα δύο σύνολα περιέχουν την μεταβλητή το Z_{13} γεγονός που σημαίνει ότι μπορούμε να ενοποιήσουμε τα στοιχεία των δύο συνόλων κάτω από κοινό σύμβολο, έστω το Z.

Η κανονική γραμματική που παράγει το θετικό δείγμα των προτάσεων της άσκησης μετά την ενοποίηση των συμβόλων και την απαλοιφή των ίδιων κανόνων παραγωγής είναι η ακόλουθη:

$G = (\{S, Z, Z_2, Z_3, Z_4, Z_6, Z_9\}, \{a, b\}, P, S)$ με κανόνες παραγωγής P τους ακόλουθους:

$$\begin{array}{cccc} S \to aZ & Z \to aZ_2 & Z_2 \to aZ_3 & Z_3 \to bZ \\ Z \to b & Z \to bZ_6 & Z_6 \to aZ & Z \to aZ_9 \\ Z_9 \to bZ & Z \to bZ & & \end{array}$$

Για την περαιτέρω μείωση των κανόνων μπορούμε να δοκιμάσουμε των κατασκευή απορρεουσών γραμματικών οι οποίες όμως δεν θα πρέπει να περιέχουν στην αντίστοιχη γλώσσα τους προτάσεις από το αρνητικό δείγμα.

Δεν υπάρχει αλγόριθμος ο οποίος να μας εξασφαλίζει μία διαδικασία ενοποίησης μεταβλητών τέτοια ώστε η γλώσσα της απορρέουσας γραμματικής να μην περιέχει προτάσεις από το αρνητικό δείγμα. Συνεπώς η μεθοδολογία που πρέπει να ακολουθήσουμε είναι να ενοποιήσουμε αυθαίρετα μεταβλητές και κατόπιν να εξετάσουμε (με συντακτική ανάλυση) αν κάποιο δείγμα αρνητικής πρότασης ανήκει στην γλώσσα της γραμματικής.

Αφού δεν υπάρχει μέθοδος για να βρούμε ποιές μεταβλητές μπορούμε να ενοποιήσουμε, ας δούμε ποιές μεταβλητές **δεν** είναι σκόπιμο να ενοποιήσουμε.

Ανεξάρτητα από την επιλογή των συμβόλων που θα ενοποιήσουμε βλέπουμε ότι όλες οι προτάσεις τις απορρέουσας γραμματικής θα περιέχουν σαν τελικό σύμβολο το b αν δεν ενοποιήσουμε την μεταβλητή Z, γεγονός που σημαίνει ότι καμμία πρόταση από το αρνητικό δείγμα δεν θα περιέχεται στην γλώσσα της απορρέουσας γραμματικής.

Ας ενοποιήσουμε λοιπόν όλες τις μεταβλητές εκτός της Z έχουμε την ενοποίηση των $\{Z_2, Z_3, Z_6, Z_9, S\}$ σε S.

Η απορρέουσα γραμματική που θα προκύψει είναι:

$$G' = (\{S, Z,\}, \{a, b\}, P', S)$$

με κανόνες παραγωγής P':

$$\begin{array}{cccc} S \to aZ & S \to bZ & Z \to aS & Z \to bS \\ Z \to b & S \to aS & & \end{array}$$

Η γραμματική που φτιάξαμε έχει έξι κανόνες και ικανοποιεί τις προυποθέσεις που μας ζητήθηκαν.

Πρόβλημα 9 *Μελετάμε δύο άγνωστες κανονικές γραμματικές για τις οποίες διαθέτουμε μόνο μετρήσεις συχνότητας εμφάνισης προτάσεων όπως αυτές δίνονται στο ακόλουθο πίνακα:*

Κατασκευάστε αντίστοιχες απορρέουσες γραμματικές που να περιέχουν λιγότερους από δέκα κανόνες παραγωγής και με το κριτήριο ελάχιστου σφάλματος φτιάξτε σύστημα ταξινόμησης προτάσεων. Κατόπιν βρείτε σε ποιά γραμματική ανήκουν οι προτάσεις των παραδειγμάτων και υπολογίστε το ελάχιστο σφάλμα του συστήματος ταξινόμησης για το δείγμα που είναι διαθέσιμο.

Πίνακας 5.9: Συχνότητα εμφάνισης προτάσεων κανονικών γραμματικών

Προτάσεις	abb	ab	aabbb	baa	bbaa
Συχνότητα	10	22	8	37	15

Προτάσεις	ab	abab	aba	bab
Συχνότητα	30	8	10	22

Λύση:

1. *Κατασκευή της πρώτης στοχαστικής απορρέουσας γραμματικής.* Για κάθε μία από τις προτάσεις του θετικού δείγματος φτιάχνουμε μια κανονική γραμματική η οποία να παράγει την συγκεκριμένη πρόταση:

$$abb : S \to aZ_1, Z_1 \to bZ_2, Z_2 \to b$$

$$ab : S \to aZ_3, Z_3 \to b$$

$$aabbb : S \to aZ_4, Z_4 \to aZ_5, Z_5 \to bZ_6, Z_6 \to bZ_7, Z_7 \to b$$

$$baa : S \to bZ_8, Z_8 \to aZ_9, Z_9 \to a$$

$$bbaa : S \to bZ_{10}, Z_{10} \to bZ_{11}, Z_{11} \to aZ_{12}, Z_{12} \to a$$

Ενοποιούμε τις παραμέτρους $\{Z_1, Z_2, Z_3, Z_4, Z_7\}$ σε Z, τις μεταβλητές $\{Z_9, Z_{12}\}$ σε Z_9 και τις $\{Z_8, Z_{10}\}$ σε Z_8. Οι κανόνες μετασχηματίζονται σε:

$$S \to aZ, Z \to bZ, Z \to b$$

$$Z \to aZ_5, Z_5 \to bZ_6, Z_6 \to bZ$$

$$S \to bZ_8, Z_8 \to aZ_9, Z_9 \to a$$

$$Z_8 \to bZ_{11}, Z_{11} \to aZ_9$$

Απο τα παραδείγματα βλέπουμε ότι πιθανόν η κανονική γραμματική της γλώσσας να περιέχει προτάσεις της μορφής $a^n b^m$ ή $b^n a^m$. Συνεπώς θα πραγματοποιήσουμε ενοποιήσεις συμβόλων έτσι ώστε να προκύψουν κανόνες της μορφής $X \to aX, Y \to bY$.

Κατασκευάζουμε λοιπόν την απορρέουσα κανονική γραμματική η οποία διαθέτει λιγότερους από δέκα κανόνες παραγωγής ενοποιώντας τα σύμβολα $\{Z_5, Z_6\}$ σε Z_5 και $\{Z_{11}, Z_8, S\}$ σε S.

Η κανονική γραμματική του πρώτου δείγματος προτάσεων είναι η ακόλουθη:

$G_1 = (\{S, Z, Z_5, Z_9\}, \{a, b\}, P_1, S)$ με κανόνες παραγωγής P:

$$S \to aZ \quad Z \to bZ \quad Z \to b \quad Z \to aZ_5 \quad Z_5 \to bZ_5$$
$$Z_5 \to bZ \quad S \to bS \quad S \to aZ_9 \quad Z_9 \to a$$

Η δεύτερη ενέργεια που πρέπει να κάνουμε για τον πλήρη προσδιορισμό των κανόνων της στοχαστικής κανονικής γραμματικής είναι να υπολογίσουμε τις πιθανότητες των κανόνων παραγωγής. Για κάθε μία από τις προτάσεις του δείγματος υπολογίζουμε την πιθανότητα εφαρμογής κάθε κανόνα ξεχωριστά πραγματοποιώντας συντακτική ανάλυση.

$$S \overset{S\to aZ}{\Rightarrow} aZ \overset{Z\to bZ}{\Rightarrow} abZ \overset{Z\to b}{\Rightarrow} abb$$

$$S \overset{S\to aZ}{\Rightarrow} aZ \overset{Z\to b}{\Rightarrow} ab$$

$$S \overset{S\to aZ}{\Rightarrow} aZ \overset{Z\to aZ_5}{\Rightarrow} aaZ_5 \overset{Z_5\to bZ_5}{\Rightarrow} aabZ_5 \overset{Z_5\to bZ}{\Rightarrow} aabbZ \overset{Z\to b}{\Rightarrow} abbb$$

$$S \overset{S\to bS}{\Rightarrow} bS \overset{S\to aZ_9}{\Rightarrow} baZ_9 \overset{Z_9\to a}{\Rightarrow} baa$$

$$S \overset{S\to bS}{\Rightarrow} bS \overset{S\to bS}{\Rightarrow} bbS \overset{S\to aZ_9}{\Rightarrow} bbaZ_9 \overset{Z_9\to a}{\Rightarrow} bbaa$$

Ο πίνακας χρήσης των κανόνων στις προτάσεις του δείγματος ως και οι αντίστοιχες πιθανότητες των κανόνων υπολογίζονται στον πίνακα 5.10.

Πίνακας 5.10: Χρήση κανόνων στις προτάσεις

Πρόταση	abb	ab	aabbb	baa	bbaa	Συχ. κανόνα	Συχ. ομάδας	Πιθαν.
$S \overset{p_1}{\to} aZ$	1	1	1	0	0	40	159	0.2516
$S \overset{p_2}{\to} bS$	0	0	0	1	2	67	159	0.4214
$S \overset{p_3}{\to} aZ_9$	0	0	0	1	1	52	159	0.3270
$Z \overset{p_4}{\to} bZ$	1	0	0	0	0	10	58	0.1725
$Z \overset{p_5}{\to} b$	1	1	1	0	0	40	58	0.6896
$Z \overset{p_6}{\to} aZ_5$	0	0	1	0	0	8	58	0.1379
$Z_5 \overset{p_7}{\to} bZ_5$	0	0	1	0	0	8	16	0.5000
$Z_5 \overset{p_8}{\to} bZ$	0	0	1	0	0	8	16	0.5000
$Z_9 \overset{p_9}{\to} a$	0	0	0	1	1	52	52	1.0000
Συχνότητα	10	22	8	37	15			

Η πιθανότητα παραγωγής της πρότασης x από την γραμματική είναι $p(x)$, ενώ η στατιστική πιθανότητα από το δείγμα συμβολίζεται $P(x)$ και για κάθε μία από τις προτάσεις του δείγματος είναι αριθμητικά:

$$\begin{array}{ll} p(abb|G_1) = p_1 \times p_4 \times p_5 = 0.0299 & P(abb) = \frac{10}{92} = 0.1087 \\ p(ab|G_1) = p_1 \times p_5 = 0.1735 & P(abb) = \frac{22}{92} = 0.2391 \\ p(aabbb|G_1) = p_1 \times p_5 \times p_6 \times p_7 \times p_8 = 0.0119 & P(abb) = \frac{8}{92} = 0.0869 \\ p(baa|G_1) = p_2 \times p_3 \times p_9 = 0.1378 & P(abb) = \frac{37}{92} = 0.4021 \\ p(bbaa|G_1) = {p_2}^2 \times p_3 \times p_9 = 0.0580 & P(abb) = \frac{15}{92} = 0.1630 \end{array}$$

2. *Κατασκευή της δεύτερης στοχαστικής απορρέουσας γραμματικής.* Για κάθε μία από τις προτάσεις του θετικού δείγματος φτιάχνουμε μια κανονική γραμματική η οποία να παράγει την συγκεκριμένη πρόταση:

$$ab : S \to aZ_1, Z_1 \to b$$

$$abab : S \to aZ_2, Z_2 \to bZ_3, Z_3 \to aZ_4, Z_4 \to b$$

$$aba : S \to aZ_5, Z_5 \to bZ_6, Z_6 \to a$$

$$bab : S \to bZ_7, Z_7 \to aZ_8, Z_8 \to b$$

Ενοποιούμε τις παραμέτρους $\{Z_1, Z_2, Z_4, Z_5, Z_8\}$ σε Z. Στην περίπτωση αυτή δεν χρειάζεται να φτιάξουμε απορρέουσα γραμματική διότι μετά την ενοποίηση των μεταβλητών η κανονική γραμματική $G_2 = (\{S, Z, Z_3, Z_6, Z_7\}, \{a, b\}, P_2, S)$ έχει οκτώ κανόνες:

$$\begin{array}{llll} S \to aZ & Z \to b & Z \to bZ_3 & Z_3 \to aZ \\ Z \to bZ_6 & Z_6 \to a & S \to bZ_7 & Z_7 \to aZ \end{array}$$

Οι κανόνες της γραμματικής υπολογίζονται όπως και στην πρώτη περίπτωση. Για κάθε μία από τις προτάσεις του δείγματος πραγματοποιούμε την ακόλουθη συντακτική ανάλυση:

$$S \overset{S \to aZ}{\Rightarrow} aZ \overset{Z \to b}{\Rightarrow} ab$$

$$S \overset{S \to aZ}{\Rightarrow} aZ \overset{Z \to bZ_3}{\Rightarrow} abZ_3 \overset{Z_3 \to aZ}{\Rightarrow} abaZ \overset{Z \to b}{\Rightarrow} abab$$

$$S \overset{S \to aZ}{\Rightarrow} aZ \overset{Z \to bZ_6}{\Rightarrow} abZ_6 \overset{Z_6 \to a}{\Rightarrow} aba$$

Ο πίνακας χρήσης των κανόνων στις προτάσεις του δεύτερου δείγματος ως και οι αντίστοιχες πιθανότητες των κανόνων υπολογίζονται στον πίνακα 5.12.

Οι αντίστοιχες πιθανότητες παραγωγής της πρότασης από την γραμματική και οι στατιστικές πιθανότητες από το δείγμα είναι:

$$\begin{array}{ll} p(ab|G_2) = p_1 \times p_3 = 0.5274 & P(ab) = \frac{30}{70} = 0.4286 \\ p(abab|G_2) = p_1 \times p_3 \times p_4 \times p_6 = 0.0541 & P(abab) = \frac{8}{70} = 0.1143 \\ p(aba|G_2) = p_1 \times p_5 \times p_7 = 0.0879 & P(aba) = \frac{10}{70} = 0.1428 \\ p(bab|G_2) = p_2 \times p_3 \times p_8 = 0.2418 & P(abb) = \frac{22}{70} = 0.3143 \end{array}$$

Πίνακας 5.11: Χρήση κανόνων στις προτάσεις

Πρόταση	ab	abab	aba	bab	Συχ. κανόνα	Συχ. ομάδας	Πιθαν.
$S \stackrel{p_1}{\rightarrow} aZ$	1	1	1	0	48	70	0.6857
$S \stackrel{p_2}{\rightarrow} bZ_7$	0	0	0	1	22	70	0.3143
$Z \stackrel{p_3}{\rightarrow} b$	1	1	0	1	60	78	0.7692
$Z \stackrel{p_4}{\rightarrow} bZ_3$	0	1	0	0	8	78	0.1026
$Z \stackrel{p_5}{\rightarrow} bZ_6$	0	0	1	0	10	78	0.1282
$Z_3 \stackrel{p_6}{\rightarrow} aZ$	0	1	0	0	8	8	1.0000
$Z_6 \stackrel{p_7}{\rightarrow} a$	0	0	1	0	10	10	1.0000
$Z_7 \stackrel{p_8}{\rightarrow} aZ$	0	0	0	1	22	22	1.0000
Συχνότητα	30	8	10	22			

Η πιθανότητα λανθασμένης ταξινόμησης δίνεται από την εξίσωση:

$$\textit{Πιθανότητα Σφάλματος} = p(abb \in G_2|G_1)p(G_1)+p(ab \in G_2|G_1)p(G_1)+p(aabbb \in G_2|G_1)p(G_1)+p(baa \in ($$

Οι πιθανότητες εμφάνισης προτάσης από τις γραμματικές ($p(G_1)$ και $p(G_2)$) δεν δίνονται. Γιαυτό τον λόγο πρέπει να τις προσεγγίσουμε από τα παραδείγματα που έχουμε στην διάθεσή μας.

$$p(G_1) \simeq \frac{\textit{προτάσεις της } G_1}{\textit{συνολικές προτάσεις}} = \frac{92}{166} = 0.5542$$

$$p(G_2) \simeq \frac{\textit{προτάσεις της } G_2}{\textit{συνολικές προτάσεις}} = \frac{70}{166} = 0.4458$$

Για το υπολογισμό των υπόλοιπων πιθανοτήτων στην συνάρτηση πιθανότητας σφάλματος είναι απαραίτητο να πραγματοποιήσουμε την διαδικασία της ταξινόμησης για όλες τις προτάσεις των παραδειγμάτων.

Οι πρόταση bab μπορεί να παραχθεί και από την γραμματική G_1:

$$S \stackrel{S\rightarrow bA}{\Rightarrow} bS \stackrel{S\rightarrow aZ}{\Rightarrow} baZ \stackrel{Z\rightarrow b}{\Rightarrow} bab$$

με πιθανότητα παραγωγής $p(bab) = 0.0731$.

Αθροίζοντας τις πιθανότητες της τελευταίας στήλης έχουμε τη συνολική πιθανότητα σφάλματος:

$$\textit{σφάλμα} = p(ab|G_1)p(G_1) + p(bab|G_1)p(G_1) = 0.0962 + 0.0405 = 0.1327$$

Πρόβλημα 10 *Θέλουμε να κατασκευάσουμε σύστημα αυτόματης διόρθωσης σφαλμάτων σε κωδικοποιημένες δυαδικές προτάσεις που εκπέμπονται από τηλεπικοινωνιακό δορυφόρο. Κατά την αποστολή του σήματος στην γη παρουσιάζεται παρεμβολή λευκού θορύβου.*

Πίνακας 5.12: Αναγνώριση των προτάσεων του δείγματος

Πρόταση	$p(x\|G_1)$	$p(x\|G_2)$	$p(x\|G_1)p(G_1)$	$p(x\|G_2)p(G_2)$	Αναγν.	Σφάλμα
abb	0.0299	0.0000	0.0165	0.0000	G_1	0.0000
ab	0.1735	0.5274	0.0962	0.2351	G_2	0.0962
aabbb	0.0119	0.0000	0.0066	0.0000	G_1	0.0000
baa	0.1378	0.0000	0.0764	0.0000	G_1	0.0000
bbaa	0.0580	0.0000	0.0321	0.0000	G_1	0.0000
abab	0.0000	0.1143	0.0000	0.0501	G_2	0.0000
aba	0.0000	0.0879	0.0000	0.0392	G_2	0.0000
bab	0.0731	0.2418	0.0405	0.1078	G_2	0.0731

Στατιστικές μετρήσεις έδειξαν ότι ο θόρυβος επηρεάζει τα σύμβολα που μεταδίδονται κατά τον ακόλουθο τρόπο. Σε 500 παραμορφώσεις τερματικών συμβόλων παρατηρήθηκαν 400 αντικαταστάσεις του τερματικού συμβόλου, 70 παραμβολές και 30 απαλειφές τερματικών συμβόλων.

Οι προτάσεις που μεταδίδονται είναι τέσσερις "1100", "00011", "101010", "0101".

Μόλις λάβαμε το ακόλουθο σήμα "10011". Ποιά ήταν η πρόταση εκπομπής ;

Υποθέτοντας ότι το κόστος των μετασχηματισμών δεν εξαρτάται από τα σύμβολα θα χρησιμοποιήσουμε την σταθμισμένη απόσταση Levenshtein και η ταξινόμηση θα πραγματοποιηθεί με το κριτήριο ελάχιστης απόστασης διότι διαθέτουμε ένα μόνο προτότυπο για κάθε κατηγορία.

Αρχικά πρέπει να υπολογίσουμε τα μεγέθη $w_\alpha, w_\lambda, w_\pi$ των συντελεστών βαρύτητας για τους μετασχηματισμούς αντικαταστάσης, παραμβολής και απαλειφής.

Επειδή διαθέτουμε στατιστικές μετρήσεις θέτουμε σαν συντελεστές βαρύτητας το αντίστροφο της στατιστικής πιθανότητας, διότι όσο πιο μεγάλη πιθανότητα έχει να συμβεί ένας μετασχηματισμός τόσο μικρότερος πρέπει να είναι ο συντελεστής βαρύτητας του μετασχηματισμού.

$$w_\alpha = \frac{500}{400} = 1.25$$

$$w_\lambda = \frac{500}{70} = 7.143$$

$$w_\pi = \frac{500}{30} = 16.667$$

Κανονικοποιούμε τους συντελεστές βαρύτητας έτσι ώστε $w_\alpha + w_\lambda + w_\pi = 1$:

$$w_\alpha = \frac{1.25}{25.06} = 0.05$$

$$w_\lambda = \frac{1.143}{25.06} = 0.285$$

$$w_\pi = \frac{16.667}{25.06} = 0.665$$

Στον πίνακα 5.13 δίνεται ο πίνακας των συνολικών αποστάσεων για την πρόταση 10011 όπως αυτές υπολογίζονται αναδρομικά με την τεχνική του δυναμικού προγραμματισμού.

Πίνακας 5.13: Απόσταση της πρότασης 10011 από τις προτάσεις αποστολής

	1	0	0	1	1
1	0.000	0.285	0.570	0.855	1.140
1	0.665	0.050	0.290	0.570	0.855
0	1.330	0.670	0.050	0.340	0.620
0	1.995	1.330	0.670	0.100	0.385

	1	0	0	1	1
0	0.050	0.335	0.620	0.905	1.190
0	0.715	0.050	0.335	0.670	0.955
0	1.380	0.715	0.050	0.385	0.720
1	2.045	1.380	0.715	0.050	0.385
1	2.710	2.045	1.380	0.715	0.050

	1	0	0	1	1
1	0.000	0.285	0.570	0.855	1.140
0	0.665	0.000	0.285	0.620	0.905
1	1.330	0.665	0.050	0.285	0.620
0	1.995	1.330	0.665	0.100	0.335
1	2.660	1.995	1.330	0.665	0.100
0	3.325	2.660	1.995	1.330	0.715

	1	0	0	1	1
0	0.050	0.335	0.620	0.905	1.190
1	0.715	0.100	0.385	0.620	0.905
0	1.380	0.715	0.100	0.435	0.670
1	2.045	1.430	0.765	0.100	0.385

Το σύστημα διόρθωσης σφαλμάτων που κατασκευάσαμε αναγνωρίζει στην παραμορφωμένη πρόταση 10011 την πρόταση αποστολής 00011.

5.10 Αλυτα Προβλήματα

Ασκηση 20 *Μελετάμε δύο άγνωστες κανονικές γραμματικές για τις οποίες διαθέτουμε μόνο μετρήσεις συχνότητας εμφάνισης προτάσεων όπως αυτές δίνονται τον ακόλουθο πίνακα 5.14:*

1. Κατασκευάστε δύο κανονικές γραμματικές που να παράγουν τις προτάσεις. Υπολογίστε το σφάλμα εκτίμησης μεταξύ των πιθανοτήτων παραγωγής των προτάσεων απο τις γραμματικές ως και τις αντίστοιχες στατιστικές πιθανότητες που προκύπτουν από τα δείγματα που διαθέτουμε.

2. Κατασκευάστε απορρέουσα γραμματική για κάθε δείγμα η οποία να μην περιέχει περισσότερους από έξι κανόνες και βρείτε ξανά το σφάλμα εκτίμησης μεταξύ των πιθανοτήτων παραγωγής των προτάσεων απο τις γραμματικές ως και τις αντίστοιχες στατιστικές τους πιθανότητες.

Τι παρατηρείτε και πως μπορείτε να δικαιολογήσετε το αποτέλεσμα;

Πίνακας 5.14: Πίνακες εμφάνισης προτάσεων κανονικών γραμματικών

Προτάσεις	abbb	abb	aab	bba	bbaa
Συχνότητα	10	26	30	120	84

Προτάσεις	abb	abab	baba	abbb	aabb
Συχνότητα	4	45	80	2	35

Ασκηση 21 *Εστω η ακόλουθη κανονική γραμματική*
$G = (\{S\}, \{a, b, c\}, P, S)$ *με κανόνες παραγωγής P:*

$$S \to aS \quad S \to bS \quad S \to a$$
$$S \to b \quad S \to cS \quad S \to c$$

Σε σύνολο 1000 προτάσεων μετρήθηκαν οι συχνότητες εμφάνισης των προτάσεων του πίνακα 5.15.

Πίνακας 5.15: Πίνακας εμφάνισης προτάσεων

Προτάσεις	aaccac	aaacca	aabcbc	bbbcbc	cbccbc
Συχνότητα	20	30	40	20	20

1. *Υπολογίστε τις πιθανότητες των κανόνων παραγωγής.*
2. *Ποιά είναι η πιθανότητα παραγωγής της πρότασης abbabbcc;*

Ασκηση 22 *Σε σύνολο 2000 προτάσεων μετρήθηκαν οι συχνότητες εμφάνισης των προτάσεων μιας γραμματικής ελεύθερης σύνταξης 5.16.*

Πίνακας 5.16: Πίνακας εμφάνισης προτάσεων

Προτάσεις	ab	ababab	aabb	aaabbb	abab	ba	baba	bababa
Συχνότητα	250	20	80	15	120	210	90	5

1. *Κατασκευάστε τους κανόνες της γραμματικής και υπολογίστε τις πιθανότητες των.*
2. *Ποιές είναι οι προτάσεις της γραμματικής που κατασκευάσατε που έχουν μήκος μικρότερο των οκτώ συμβόλων.*
3. *Ποιά είναι η πιθανότητα παραγωγής πρόττασης από την γραμματική που έχει μήκος μικρότερο των οκτώ συμβόλων. Τι παρατηρείτε;*

Ασκηση 23 *Εστω τα δείγματα προτάσεων του πίνακα 5.17 που αποτελούν υποσύνολο δύο γλωσσών L_1 και L_2.*

1. *Κατασκεύασε κανονική γραμματική και ταξινόμησε την πρόταση abbcb με το κριτήριο της μέγιστης πιθανότητας.*

Πίνακας 5.17: Πίνακες εμφάνισης προτάσεων κανονικών γραμματικών

Προτάσεις	abbb	abbc	aaab	aabbac	bac
Συχνότητα	10	26	30	5	80

Προτάσεις	abc	cab	bcba	abcb	acbb
Συχνότητα	80	45	10	2	35

2. Με την μέθοδο διόρθωσης σφαλμάτων ελάχιστης απόστασης ταξινόμησε την ίδια πρόταση θεωρώντας ότι τα λάθη αντικατάστασης απαλειφής και παραμβολής έχουν το ίδιο βάρος.

3. Με την μέθοδο διόρθωσης σφαλμάτων και κριτήριο ταξινόμησης τα πέντε πλησιέστερα προτότυπα ταξινόμησε την πρόταση abcabc θεωρώντας ότι τα λάθη απαλειφής και παραμβολής έχουν διπλάσια βαρύτητα από τα λάθη αντικατάστασης.

Ασκηση 24 *Η συνάρτηση συνολικής απόστασης προτάσεων, η απόσταση Levenshtein και η σταθμισμένη Levenshtein ικανοποιούν όλες τις ασθενείς συνθήκες της συνάρτησης απόστασης;*

Ασκηση 25 *Υπολογείστε την απόσταση των λέξεων "ΑΝΔΡΑΣ" και "ΓΥΝΑΙΚΑ" λαμβάνοντας υπόψη ότι το κόστος μετασχηματισμού αντικατάστασης είναι ίσο με την αλφαβητική τους απόστασης και οι μετασχηματισμοί απαλειφής και παραμβολής έχουν κόστος ίσο με το λογάριθμο της απόλυτης αλφαβητικής θέσης των.*

Ασκηση 26 *Θέλουμε να κατασκευάσουμε σύστημα αυτόματης διόρθωσης σφαλμάτων σε προτάσεις που παραμορφώνονται από λευκό θόρυβο. Οι προτάσεις αποτελούνται από τρία τερματικά σύμβολα* $\{\alpha, \beta, \gamma\}$.

Στατιστικές μετρήσεις έδειξαν τους μετασχηματισμούς του πίνακα 5.18.

Πίνακας 5.18: Συχνότητες εμφάνισης μετασχηματισμών αντικατάστασης απαλοιφής και παρεμβολής

Τερματικό σύμβολο	α	β	γ
Αντικατάσταση του α	--	26	30
Αντικατάσταση του β	40	--	100
Αντικατάσταση του γ	10	80	--

Τερματικό σύμβολο	α	β	γ
Απαλειφή	13	2	20

Τερματικό σύμβολο	α	β	γ
Παρεμβολή	3	8	10

Οι προτάσεις που μεταδίδονται είναι τρείς "αβγαβγ", "ββααγ", "αααββ" και αντίστοιχες συχνότητες αποστολής είναι 100, 20, 500. Κατασκευάστε στοχαστικό μοντέλο και αναγνωρίστε την πρόταση "αγγγββ".