VI Lingwistyka i podstawy teorii obliczeń

4. Gramatyki bezkontekstowe, notacja Backusa-Naura

notacja Backusa-Naura -(BNF)

W pierwszym przybliżeniu BNF jest formalizmem przypominającym definicje nazwanych wzorców ze specyfikacji dla Lex'a. Główna różnica polega na tym, że zależności między poszczególnymi nazwami mogą być rekurencyjne.

Specyfikacja w BNF-ie określa składnię szeregu konstrukcji składniowych, nazywanych też nieterminalami. Każda taka konstrukcja składniowa ma swoją nazwę. Specyfikacja składa się z szeregu reguł (nazywanych produkcjami), które określają jaką postać mogą przybierać konstrukcje składniowe. Nazwy konstrukcji składniowych ujmujemy w trójkątne nawiasy $\langle \ldots \rangle$. Produkcje mają postać:

`<definiowana konstrukcja>` `::=` wyrażenie opisujące definiowaną konstrukcję
Wyrażenie opisujące definiowaną konstrukcję może zawierać:

- $\langle$konstrukcja$\rangle$ -- nazwy konstrukcji, w tym również definiowanej, rekurencyjne zależności są dozwolone,
- tekst, ujęty w cudzysłów, który pojawia się dosłownie w danej konstrukcji,
- $\ldots \mid \ldots$ -- oddziela różne alternatywne postaci, jakie może mieć dana konstrukcja,
- $[\ldots]$ -- fragment ujęty w kwadratowe nawiasy jest opcjonalny.

Produkcje traktujemy jak możliwe reguły przepisywania -- nazwę konstrukcji stojącej po lewej stronie produkcji możemy zastąpić dowolnym napisem pasującym do wyrażenia podanego po prawej stronie. Każdej konstrukcji składniowej odpowiada pewien język. Składa sie on z wszystkich tych napisów, które można uzyskać zaczynając od nazwy konstrukcji i stosując produkcje, aż do momentu uzyskania napisu, który nie zawiera już żadnych nazw konstrukcji.

Gramatyki bezkontekstowe

Gramatyki bezkontekstowe są uproszczoną wersją notacji BNF. Konstrukcje składniowe nazywamy tu nieterminalami lub symbolami nieterminalnymi i będziemy je oznaczać wielkimi literami alfabetu łacińskiego: *A, B, ..., X, Y, Z*. Wśród nieterminali jest jeden wyróżniony symbol, nazywany aksjomatem. To właśnie język związany z tym nieterminalem opisuje gramatyka.

Oprócz nieterminali używamy również terminali (symboli terminalnych) -- są to znaki stanowiące alfabet opisywanego przez gramatykę języka. Będziemy je oznaczać małymi literami alfabetu łacińskiego: *a, b, ...*

Produkcje w gramatykach bezkontekstowych mają bardzo prostą postać: $A \rightarrow \alpha$, gdzie *A* to nieterminal, a α to słowo, które może zawierać terminale i nieterminale. Produkcja taka oznacza, że nieterminal *A* możemy zastąpić napisem α. Dla jednego nieterminala możemy mieć wiele produkcji. Oznacza to, że możemy je dowolnie stosować. Zwykle zamiast pisać:

$$\begin{aligned} A &\rightarrow \alpha \\ A &\rightarrow \beta \\ &\vdots \\ A &\rightarrow \gamma \end{aligned}$$

Będziemy pisać krócej:

$$A \rightarrow \alpha \mid \beta \mid \ldots \mid \gamma$$

Przykład

Zanim formalnie zdefiniujemy gramatyki bezkontekstowe i opisywane przez nie języki, spójrzmy na gramatykę opisującą język: $\{a^n b^n : n \geq 0\}$.

$$A \rightarrow aAb \mid \varepsilon$$

Oto sekwencja zastosować produkcji, która prowadzi do uzyskania słowa *aaabbb*:

$$A \rightarrow aAb \rightarrow aaAbb \rightarrow aaaAbbb \rightarrow aaabbb$$

Trzykrotnie stosowaliśmy tutaj produkcję $A \rightarrow aAb$, aby na koniec zastosować $A \rightarrow \varepsilon$.

Przykład ten pokazuje, że gramatyki bezkontekstowe mogą opisywać języki, które nie są regularne.

Definicja

Gramatyka bezkontekstowa, to dowolna taka czwórka $G = \langle N, \Sigma, P, S \rangle$, gdzie:

- *N* to (skończony) alfabet symboli nieterminalnych,
- Σ to (skończony) alfabet symboli terminalnych, $\Sigma \cap N = \emptyset$,
- *P* to skończony zbiór produkcji, reguł postaci $A \rightarrow \alpha$ dla $A \in N$, $\alpha \in (N \cup \Sigma)^*$,
- $S \in N$ to aksjomat wyróżniony symbol nieterminalny.

Zwykle nie będziemy podawać gramatyki jako formalnej czwórki. Zamiast tego będziemy podawać zbiór produkcji. Zwykle, w sposób niejawny z produkcji wynika zbiór nieterminali i terminali. Jeżeli nie będzie oczywiste, który nieterminal jest aksjomatem, będziemy to dodatkowo zaznaczać.

Definicja

Niech $G = \langle N, \Sigma, P, S \rangle$ będzie ustaloną gramatyką bezkontekstową. Przez $\longrightarrow$ będziemy oznaczać zbiór produkcji *P* traktowany jako relacja, $\longrightarrow \subseteq N \times (N \cup \Sigma)^*$. Przez $\longrightarrow^*$ będziemy oznaczać zwrotno-przechodnie domknięcie $\longrightarrow$, $\longrightarrow^* \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$.

Relacja $\longrightarrow$ opisuje pojedyncze zastosowanie produkcji, jako relacja między nieterminalem, a zastępującym go słowem. Relacje $\longrightarrow^*$ opisuje co można zrobić stosując dowolną liczbę (łącznie z zero) dowolnych produkcji. Możemy też przedstawić sobie relację $\longrightarrow^*$ jako skrót do wielokrotnego iterowania relacji $\longrightarrow$:

Fakt 11. *Jeżeli* $x \longrightarrow^* y$, *to istnieje ciąg słów* (z_i) *taki, że:*

$$x = z_0 \longrightarrow z_1 \longrightarrow \cdots \longrightarrow z_k = y$$

Gramatyka opisuje język złożony z tych wszystkich słów (nad alfabetem terminalnym), które możemy uzyskać z aksjomatu stosując produkcje.

Definicja

Niech $G = \langle N, \Sigma, P, S \rangle$ będzie ustaloną gramatyką bezkontekstową. Język generowany (lub opisywany) przez gramatykę *G*, to:

$$L(G) = \{x \in \Sigma^* : S \longrightarrow^* x\}$$

Alternatywnie, *L*(*G*) to zbiór takich słów $x \in \Sigma^*$, dla których istnieją ciągi słów (z_i), $z_i \in (N \cup \Sigma)^*$ takie, że:

$$S = z_0 \longrightarrow z_1 \longrightarrow \cdots \longrightarrow z_k = x$$

Ciąg (z_i) nazywamy wyprowadzeniem słowa *x* (w gramatyce *G*).

Definicja

Powiemy, że język jest bezkontekstowy, jeżeli istnieje generująca go gramatyka.

Wyprowadzenie stanowi coś w rodzaju dowodu, że dane słowo faktycznie należy do języka generowanego przez gramatykę. Istnieje jeszcze inny, bardziej strukturalny, sposób ilustrowania jak dane słowo można uzyskać w danej gramatyce. Jest to drzewo wyprowadzenia.

Definicja

Drzewo wyprowadzenia, to sposób ilustrowania jak dane słowo może być wyprowadzone w danej gramatyce, w postaci drzewa. Drzewo wyprowadzenia musi spełniać następujące warunki:

- W korzeniu drzewa znajduje się aksjomat.

- W węzłach wewnętrznych drzewa znajdują się nieterminale, a w liściach terminale lub słowa puste ε.
- Jeśli w danym węźle mamy nieterminal *X*, a w jego kolejnych synach mamy $x_1, x_2, \ldots, x_k$, to musi zachodzić $X \to x_1 x_2 \ldots x_k$.
- Terminale umieszczone w liściach, czytane od lewej do prawej, tworzą wyprowadzone słowo.

Przyjrzyjmy się zdefiniowanym powyżej pojęciom na przykładach:

Przykład

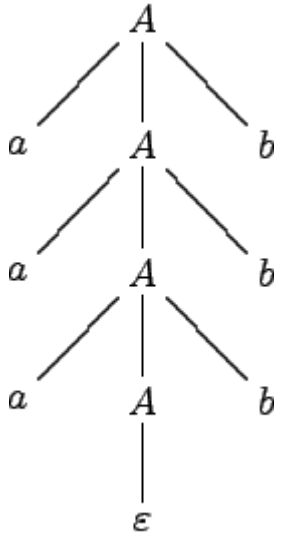
Weźmy gramatykę generującą język $\{a^n b^n : n \geq 0\}$:

$$A \to aAb \mid \varepsilon$$

Wyprowadzenie słów *aaabbb* ma postać:

$$A \to aAb \to aaAbb \to aaaAbbb \to aaabbb$$

a jego drzewo wyprowadzenia wygląda następująco:



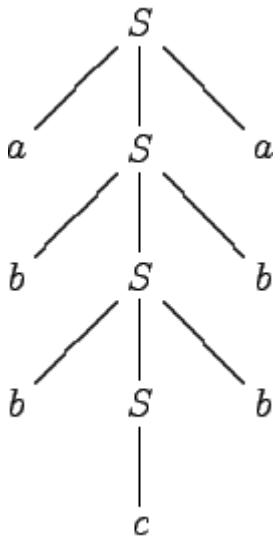
Przykład

Oto gramatyka generująca język palindromów nad alfabetem {a,*b*,*c*}:

$$S \to aSa \mid bSb \mid cSc \mid a \mid b \mid c \mid \varepsilon$$

Weźmy słowo *abbcbba*. Oto jego wyprowadzenie i drzewo wyprowadzenia:

$$S \longrightarrow aSa \longrightarrow abSba \longrightarrow abbSbba \longrightarrow abbcbba$$



Nie wiem czy to o to chodzi? Bo niby skąd mam wiedzieć :P