§ 6. LR(k)-автомат. LR(k)-грамматики. LR-язык.

Напомним определение LR(1)-пункта.

Опр. **LR(1)-пунктом** расширенной грамматики G называется набор (A, β_1, β_2, a), где $A \to \beta_1\beta_2$ – правило грамматики, $a \in \Sigma \cup \{\dashv\}$.
Обозначение: $[A \to \beta_1 \bullet \beta_2, a]$.

Введем опред. LR(k)-пункта – обобщение определения LR(1)-пункта.

Опр. **LR(k)-пунктом грамматики G** называется набор $(A, \beta_1, \beta_2, \mathrm{v})$, где $A \to \beta_1\beta_2$ – правило грамматики, v – цепочка терминалов длины равной k, либо цепочка терминалов длины меньше k, дополненная в конце символом $\dashv$.
Обозначение: $[A \to \beta_1 \bullet \beta_2, \mathrm{v}]$

Напомним определение допустимого LR(1)-пункта $[A \to \beta_1 \bullet \beta_2, a]$.

Опр. Пункт $[A \to \beta_1 \bullet \beta_2]$ называется **допустимым** для активного префикса $\gamma = \gamma' \beta_1$ некоторой *r*-формы $\gamma\alpha$, если существует правый вывод

$$S' \Rightarrow^* \gamma' A \mathrm{w} \Rightarrow \gamma' \beta_1 \beta_2 \mathrm{w} \Rightarrow^* \mathrm{uw}$$

и *a* – первый символ слова w.

Дадим опр. допуст. LR(k)-пункта – обобщение опр. допуст. LR(1)-пункта.

Опр. LR(k)-пункт $[A \to \beta_1 \bullet \beta_2, \mathrm{v}]$ называется **допустимым** для активного префикса $\gamma = \gamma' \beta_1$ некоторой r-формы, если существует вывод

$S' \Rightarrow^* \gamma' A \mathrm{w} \Rightarrow \gamma' \beta_1 \beta_2 \mathrm{w} \Rightarrow^* \mathrm{uw}$, и цепочка v является префиксом цепочки $\mathrm{w}\dashv$.

Напомним определение автомата LR(1) пунктов.

Опр. **Автомат LR(1)-пунктов** содержит переходы двух видов:

1) $\delta([A \to \beta_1 \bullet X \beta_3, a], X) = [A \to \beta_1 X \bullet \beta_3, a]$;

2) $\delta([A \to \beta_1 \bullet B \beta_3, a], \varepsilon) = [B \to \bullet \beta, b]$, где $b \in$FIRST($\beta_3 a$).

Начальное состояние автомата: LR(1)-пункт $[S' \to \bullet\, S, \dashv]$.
Заключительные – все состояния.

Комментарий: из основной теоремы LR-анализа следует, что процессе построения переходов такого автомата появляются только допустимые LR(1)-пункты.

Напомним определение множества $FIRST_k$.

Опр. **$FIRST_k(\alpha)$** – это подмножество множества $\sum^*$, состоящее из всех цепочек u из $\sum^*$ таких, что

- либо $|u|=k$ и $\alpha \Rightarrow u\beta$ для некоторого β
- либо $|u|<k$ и $\alpha \Rightarrow u$

Очевидно, $FIRST_1(\alpha)= FIRST(\alpha)$.

Очевидно, если $\alpha \in \Sigma^*$, то $FIRST_k(\alpha)$ – это префикс длины k слова α, если $|\alpha| \geq k$ и слово α – $|$, если $|\alpha|< k$.

Опр. **Автомат LR(k)-пунктов** содержит переходы двух видов:

1) $\delta([A \to \beta_1 \bullet X \beta_3, \mathrm{v}], X) = [A \to \beta_1 X \bullet \beta_3, \mathrm{v}]$;

2) $\delta([A \to \beta_1 \bullet B \beta_3, \mathrm{v}], \varepsilon) = [B \to \bullet \beta, \mathrm{w}]$, где w $\in \mathrm{FIRST_k}(\beta_3 \mathrm{v})$.

Начальное состояние автомата: LR(k)-пункт $[S' \to \bullet\ S, \dashv]$.
Заключительные – все состояния.

Приведем обобщение определения автомата LR(1) – определение автомата LR(k).

Опр. **LR(k)-автоматом** грамматики G называется детерминированный конечный автомат, эквивалентный ε-недетерминированному конечному автомату LR(k)-пунктов.

Имеет место (приведем без доказательства)

Обобщение основной теоремы LR-анализа:
Автомат LR(k)-пунктов состоит только из допустимых пунктов и распознает множество всех активных префиксов грамматики G.

Следствие 1. LR(k)-автомат распознает множество всех активных префиксов грамматики G.

LR(k)-анализатор строится аналогично LR(1)-анализатору с той лишь разницей, что столбцы таблицы ACTION строятся не только для терминалов $a \in \Sigma \cup \{-|\}$, но и для всех слов $v \in \Sigma^l \cup \{-|\}$ таких, что $l \leq k$. Пустые столбцы при этом удаляются.

Напомним определение LR(1)-грамматики.

Опр. Грамматика G является **LR(1)-грамматикой**, если LR(1)-анализатор не имеет конфликтов.

Опр. Грамматика G является **LR(k)-грамматикой**, если LR(k)-анализатор не имеет конфликтов.

Напомним свойство LR(1)-грамматики.

Замечание 1. Грамматика G является LR(1)-грамматикой, если из того, что при правостороннем выводе некоторой r-формы $\alpha\beta$u последним применялось правило $A \to \beta$, следует, что это же правило применялось последним при выводе любой r-формы $\alpha\beta$v, такой, что $\mathrm{FIRST}(u)=\mathrm{FIRST}(v)$. (Напоминание: $u,v \in \Sigma^*$).

Справедливо обобщение этого свойства на LR(k)-грамматику.

Замечание 2. Грамматика G является LR(k)-грамматикой, если из того, что при правостороннем выводе некоторой r-формы $\alpha\beta$u последним применялось правило $A \to \beta$, следует, что это же правило применялось последним при выводе любой r-формы $\alpha\beta$v, такой, что $\mathrm{FIRST}_k(v) = \mathrm{FIRST}_k(u)$. (Напоминание: $u, v \in \Sigma^*$)

Опр. Грамматика G называется **LR-грамматикой**, если она является LR(k)-грамматикой для некоторого $k \geq 0$.

Теорема 1. $LR(0) \subseteq LR(1) \subseteq LR(2) \subseteq ... \subseteq LR(k) \subseteq \ldots \subseteq LR$.

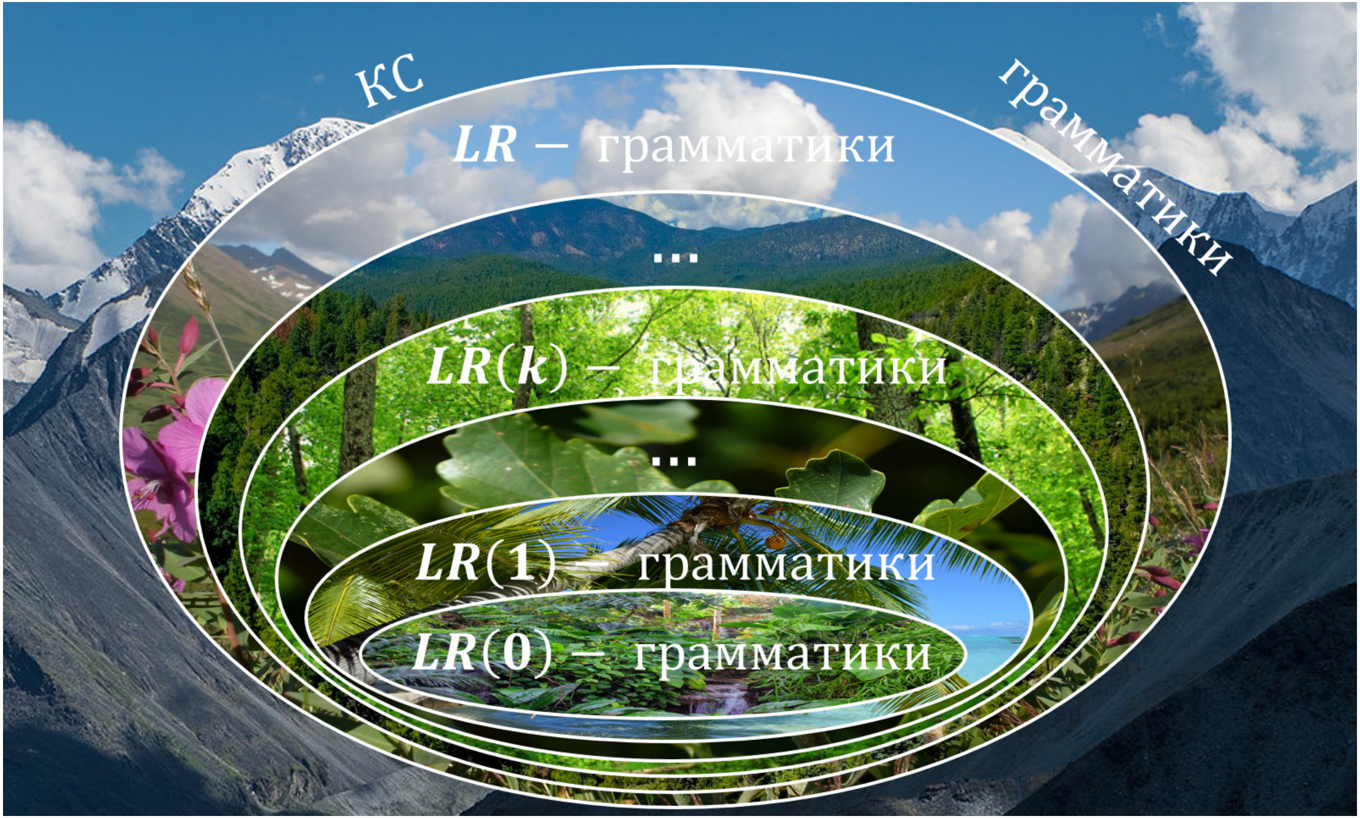
Доказательство следует из определения LR(k-1) грамматик и того простого факта, что из равенства $FIRST_k(\alpha)= FIRST_k(\beta)$ следует равенство $FIRST_{k-1}(\alpha)= FIRST_{k-1}(\beta)$.

Теорема 2. Для любого k выполняется строгое включение $LR(k) \subset LR(k+1)$.

Для **доказательства** используется похожий пример, что и для доказательства того, что $LL(k) \subset LL(k+1)$.

Пусть $G = \{S \to Ba^k b | Ca^k c, B \to d, C \to d\}$.
При правом выводе цепочки $da^k b$ последним может применяться как правило $B \to d \ (\alpha = \varepsilon, u = a^k b)$,
так и правило $C \to d(\beta = \varepsilon, v = a^k c)$,
причем $FIRST_k(a^k b) = FIRST_k(a^k c) = a^k$. Здесь β=*d*.
Таким образом, $G \notin LR(k)$.
В то же время непосредственно проверяется, что $G \in LR(k+1)$.

Следствие 2. $LR(0) \subset LR(1) \subset LR(2) \subset \ldots \subset LR(k) \subset \ldots \subset LR$.


КС грамматики
LR – грамматики
...
$LR(k)$ – грамматики
...
$LR(1)$ – грамматики
$LR(0)$ – грамматики

Замечание 3: существуют грамматики, не являющиеся LR(k)-грамматиками.

Пример 1. $G = \{S' \to S,\ S \to Bb \mid Cc,\ B \to Ba \mid \varepsilon,\ C \to Ca \mid \varepsilon\}$.

Вопрос: $L(G) = ?$

Замечание 3: существуют грамматики, не являющиеся LR(k)-грамматикой.

Пример 1. $G = \{S' \to S,\ S \to Bb \mid Cc,\ B \to Ba \mid \varepsilon,\ C \to Ca \mid \varepsilon\}$.

$L(G) = \{a^*b, a^*c\}$.

Покажем, что ни для какого k G не является LR(k)-грамматикой.

<u>От противного:</u> предположим, что G является LR(k)-грамматикой для некоторого k.

Рассмотрим выводы цепочек $a^k b$, $a^k c$.

$$S' \Rightarrow^* Ba^{k-1}b \Rightarrow Ba^k b \Rightarrow \underbrace{a^k b}_{\varepsilon u}, \quad (B \to \varepsilon - \text{последнее правило})$$

$$S' \Rightarrow^* Ca^{k-1}c \Rightarrow Ca^k c \Rightarrow \underbrace{a^k c}_{\varepsilon v}, \quad (C \to \varepsilon - \text{последнее правило})$$

Определение LR(k)-грамматики нарушено:
$FIRST_k(a^k b) = FIRST_k(a^k c) = a^k$. То есть $\mathrm{FIRST}_k(u)=\mathrm{FIRST}_k(v)$.

Опр. Язык называется **LR(k)-языком**, если существует LR(k)-грамматика, порождающая этот язык.

Приведем список основных результатов без доказательства.

Теорема 3. Любой язык, порождаемый LR-грамматикой, порождается LR(1)-грамматикой.

Теорема 4. Любой язык, порождаемый LR-грамматикой, порождается SLR(1)-грамматикой.

Теорема 5. Класс LR-языков совпадает с классом языков, распознаваемых ДАМП.

Следствие 3. Класс LL-языков включается в класс LR-языков.