УДК 519.766.23

НОРМАЛЬНЫЕ ФОРМЫ И АВТОМАТЫ ДЛЯ КАТЕГОРИАЛЬНЫХ ГРАММАТИК ЗАВИСИМОСТЕЙ

Карлов Б.Н.
Кафедра информатики

Поступила в редакцию 27.11.2008, после переработки 03.12.2008.

В работе изучаются свойства обобщенных категориальных грамматик зависимостей (оКГЗ), введенных в работах [5, 6]. Для них определяются нормальные формы, аналогичные нормальной форме Грейбах для кс-грамматик. Доказывается, что каждый оКГЗ-язык можно получить с помощью гомоморфизма из пересечения кс-языка и скобочного языка. Определен класс магазинных автоматов со счетчиками, которые допускают оКГЗ-языки.

We establish some properties of generalized Categorial Dependency Grammars (gCDG) introduced in the papers [5, 6]. A normal form of gCDGs is defined which is similar to Greibach normal form for cf-grammars. It is proved that each gCDG-language can be obtained via homomorphism of the intersection of cf-language and some kind of multibracket language. A class of push-down automata with counters is defined which accept all gCDG-languages.

Ключевые слова: формальные грамматики, категориальные грамматики, структура зависимостей, нормальные формы грамматик, магазинные автоматы со счетчиками.
Keywords: formal grammars, categorial grammars, dependency structure, normal forms of grammars, push-down automata with counters.

1. Введение

Формальные способы описания синтаксической структуры предложения имеют первостепенную важность для большинства задач информатики, связанных с обработкой информации на естественном языке (см. [3]). После основополагающих работ Н. Хомского, определившего четыре базовых класса порождающих грамматик, был определен еще целый ряд типов грамматик, позволяющих получить ту или иную информацию о синтаксической структуре предложения. В частности, специальный тип грамматик, *грамматики зависимостей*, присваивают структуры зависимостей (структуры подчинения) предложениям языка, который они определяют (см. [3]). Структура зависимостей (СЗ) предложения – это ориентированный граф, вершинами которого являются слова предложения, а ребра помечены именами зависимостей. Таким образом, структура предложения задается в терминах

некоторых бинарных отношений на словах. Если два слова v_1 и v_2 связаны зависимостью d (обозначение: $v_1 \longrightarrow^d v_2$), то v_1 является *хозяином*, а v_2 – *подчиненным*. Скажем, что в СЗ слово w зависит от слова v, если существует путь из v в w. Для естественных языков СЗ являются деревьями (иногда их называют деревьями подчинения). Для большинства «типичных» предложений их деревья зависимостей являются *проективными*: для любых трех слов v, w и u из того, что $v \to u$ и w лежит между v и u, следует, что w зависит от v. Вот пример такого дерева:

С другой стороны, в естественных языках достаточно распространены (особенно в художественной литературе) и непроективные конструкции, когда подчинённое слово стоит вне синтаксической группы своего хозяина. В качестве примера приведём оригинальную строчку А. С. Пушкина:

Непроективность всегда связана с *разрывными* зависимостями, в которых хозяин v разделен со подчиненным словом u некоторым словом w, которое не зависит от v. На рисунке эта связь показана штриховой линией.

Классические кс-грамматики (см. [1, 2, 3]), как и обычные категориальные грамматики (см. [2, 4]) неспособны обнаруживать в подобных предложениях разрывные составляющие. М.И. Дехтярь и А.Я. Диковский в работах [5, 6] определили новые виды грамматик — *категориальные грамматики зависимостей (КГЗ)* и *обощенные категориальные грамматики зависимостей (оКГЗ)*. Эти грамматики работают как грамматики-распознаватели и структуру предложения раскрывают подобно грамматикам зависимостей. Их особенностью является то, что они способны обнаруживать дальние связи (как в приведённом примере из Пушкина). Подобные ситуации обрабатываются в них с помощью так называемых поляризованных валентностей: положительных, отвечающих за слово-хозяин, и отрицательных, отвечающих за подчиненное слово. В указанных работах получены результаты о выразительной силе КГЗ и оКГЗ, предложены алгоритмы их анализа.

Настоящая работа продолжает исследование свойств оКГЗ. В разделе 2 мы приводим основные определения, связанные с оКГЗ. В разделе 3 определяется нормальная форма оКГЗ, аналогичная нормальной форме Грейбах (см. [1, 2]) для кс-грамматик. Показано, как для любой оКГЗ можно построить эквивалентную оКГЗ в нормальной форме. Затем в разделе 4 этот результат используется для получения нового доказательства замкнутости класса оКГЗ-языков относительно обращений гомомоморфизмов. В разделе 5 устанавливается теорема о пред-

ставлении каждого оКГЗ-языка как результата гоморфизма пересечения некоторого кс-языка со скобочным языком. Эта теорема аналогична известной теореме Хомского-Шютценберже (см. [2]) для кс-языков. В разделе 6 мы вводим понятие счётчиковых автоматов с магазинной памятью и доказываем, что эти автоматы распознают в точности класс оКГЗ-языков.

2. Основные определения

Для формализации лингвистического понятия синтаксического типа мы будем использовать понятие категории. Пусть $\mathbf{C}$ — непустое конечное множество элементарных категорий (например, сказуемое, определение, дополнение и т. д.). Элементарные категории могут быть итерированы: для $C \in \mathbf{C}$ C^* означает соответствующую итеративную категорию. Множество всех итеративных категорий будем обозначать $\mathbf{C}^*$. Элементарные и итеративные категории соединяются конструкторами $\backslash$ и $/$в локальные категории.

Определение 1. *Множество локальных категорий $LCat(\mathbf{C})$ — это минимальное множество такое, что:*
1) $\mathbf{C} \cup \{\varepsilon\} \subseteq LCat(\mathbf{C})$, где ε — пустой символ;
2) если $\alpha \in LCat(\mathbf{C}), A \in \mathbf{C} \cup \mathbf{C}^*$, то $[A\backslash\alpha], [\alpha/A] \in LCat(\mathbf{C})$.

Полагаем, что конструкторы $\backslash$ и $/$ассоциативны. Поэтому любая локальная категория γ может быть представлена в виде $\gamma = [L_k\backslash \ldots \backslash L_1 \backslash C / R_1 / \ldots / R_m]$.

Для описания дальних связей между словами в предложении в [5] вводятся понятия полярности и поляризованной валентности. Полярностью v называется элемент множества $V = \{\searrow, \swarrow, \nwarrow, \nearrow\}$. Для каждой полярности v существует двойственная полярность $\breve{v}$:

$$\breve{\nearrow} = \searrow, \breve{\nwarrow} = \swarrow, \breve{\swarrow} = \nwarrow, \breve{\searrow} = \nearrow$$

Поляризованная валентность β — это выражение вида vC, где $v \in V, C \in \mathbf{C}$. Множество всех поляризованных валентностей будем обозначать $V(C)$. В нём, в соответствии с типом полярности, можно выделить подмножества:
$\nwarrow\mathbf{C} = \{\nwarrow C \mid C \in \mathbf{C}\}$, $\nearrow\mathbf{C} = \{\nearrow C \mid C \in \mathbf{C}\}$,
$\searrow\mathbf{C} = \{\searrow C \mid C \in \mathbf{C}\}$, $\nwarrow\mathbf{C} = \{\nwarrow C \mid C \in \mathbf{C}\}$,
$V^-(\mathbf{C}) = \searrow\mathbf{C} \cup \swarrow\mathbf{C}$, $V^+(\mathbf{C}) = \nwarrow\mathbf{C} \cup \nearrow\mathbf{C}$,
$V^l(\mathbf{C}) = \nearrow\mathbf{C} \cup \swarrow\mathbf{C}$, $V^r(\mathbf{C}) = \nwarrow\mathbf{C} \cup \searrow\mathbf{C}$.

Последовательность поляризованных валентностей называется потенциалом. Будем говорить, что потенциал θ сбалансирован, если всякая его проекция на $\{v, \breve{v}\}$, где $v \in V^+(\mathbf{C})$, представляет собой правильную скобочную последовательность. Примером сбалансированного потенциала может служить $\nearrow A \nearrow B \searrow A \searrow B$. Множество всех возможных потенциалов обозначается $Pot(\mathbf{C})$.

Из локальных категорий и потенциалов строятся оКГЗ-категории.

Определение 2. *Категорией γ называется выражение вида α^θ, где*

$$\alpha \in LCat(\mathbf{C}), \theta \in Pot(\mathbf{C}).$$

Множество всех категорий обозначается $Cat(\mathbf{C})$.

На множестве оКГЗ-категорий определяется исчисление зависимостей, задаваемое следующим набором правил.

Определение 3. *Пусть* Γ_1, Γ_2 *— строки категорий из* $Cat(\mathbf{C})^*$, $\theta, \theta_1, \theta_2, \theta_3$ *— потенциалы,* α *— локальная категория из* $LCat(\mathbf{C})$.

Правила локальной зависимости:
$L^l : \Gamma_1[C]^{\theta_1}[C\backslash\alpha]^{\theta_2}\Gamma_2 \vdash \Gamma_1[\alpha]^{\theta_1\theta_2}\Gamma_2$
$L^r : \Gamma_1[\alpha/C]^{\theta_1}[C]^{\theta_2}\Gamma_2 \vdash \Gamma_1[\alpha]^{\theta_1\theta_2}\Gamma_2$,
где $C \in \mathbf{C} \cup \{\varepsilon\}$

Правила итеративной зависимости:
$I^l : \Gamma_1[C]^{\theta_1}[C^*\backslash\alpha]^{\theta_2}\Gamma_2 \vdash \Gamma_1[C^*\backslash\alpha]^{\theta_1\theta_2}\Gamma_2$
$I^l_0 : \Gamma_1[C^*\backslash\alpha]^{\theta}\Gamma_2 \vdash \Gamma_1[\alpha]^{\theta}\Gamma_2$
$I^r : \Gamma_1[\alpha/C^*]^{\theta_1}[C]^{\theta_2}\Gamma_2 \vdash \Gamma_1[\alpha/C^*]^{\theta_1\theta_2}\Gamma_2$
$I^r_0 : \Gamma_1[\alpha/C^*]^{\theta}\Gamma_2 \vdash \Gamma_1[\alpha]^{\theta}\Gamma_2$,
где $C \in \mathbf{C} \cup \{\varepsilon\}$

Правило дальней зависимости:
$D : \Gamma_1\alpha^{\theta_1\beta\theta_2\breve{\beta}\theta_3}\Gamma_2 \vdash \Gamma_1\alpha^{\theta_1\theta_2\theta_3}\Gamma_2$,
где валентности $(\beta, \breve{\beta})$ *образуют правильную пару и* θ_2 *не содержит* $\beta, \breve{\beta}$.

При выполнении сокращения по каждому из правил в структуру зависимостей добавляется дуга графа, идущая из слова-хозяина в зависимую категорию с меткой-именем сократившейся категории.

Обозначим отношение выводимости за один шаг на множестве $Cat(\mathbf{C})^*$ в этом исчислении через $\vdash^R$, где R — одно из вышеприведенных правил, или просто $\vdash$, если имя правила для нас несущественно. Если Γ_2 получается из Γ_1 за n шагов, то будем писать $\Gamma_1 \vdash^n \Gamma_2$. Через $\vdash^*$ обозначим транзитивное замыкание отношения $\vdash$ на множестве $Cat(\mathbf{C})^*$.

Теперь дадим определение обобщённой категориальной грамматики зависимостей.

Определение 4. *Обобщённой категориальной грамматикой зависимостей (оКГЗ) называется система* $G =< W, \mathbf{C}, S, \delta >$, *где:*
W *— конечное множество слов,*
$\mathbf{C}$ *— конечное множество элементарных категорий,*
S *— выделенная в* $\mathbf{C}$ *главная категория,*
δ *— лексикон, функция на* W, *сопоставляющая каждому слову* $w \in W$ *конечное множество* $\delta(w) \subseteq Cat(\mathbf{C})$ *его возможных категорий.*

Каждая оКГЗ определяет некоторое множество предложений из W^*, словам которых можно корректно сопоставить их категории, определяемые лексиконом, так, что из них можно вывести главную категорию S. Пусть $s = w_1 w_2 \dots w_n \in W^*$ — предложение. Положим $\delta(s) = \delta(w_1)\delta(w_2)\dots\delta(w_n)$.

Определение 5. *оКГЗ* G *порождает язык* $L(G)$, *состоящий из всех предложений* $s \in W^*$, *для которых существует строка категорий* $\Gamma \in \delta(s)$ *такая, что* $\Gamma \vdash^* S$.

Пример 1. В качестве примера рассмотрим язык $L = \{ w_1^n w_2^n w_3^n \mid n = 0, 1, \dots \}$. Известно, что этот язык не является контекстно-свободным (см. [2]). Однако, как показано в работе [6], он порождается следующей оКГЗ:

$w_1 \mapsto [S/A_1]^{\nearrow A_2}, [A_1/A_1]^{\nearrow A_2}, [A_1/A_2]^{\nearrow A_2}$
$w_2 \mapsto [A_2/A_2]^{\searrow A_2 \nearrow A_3}, [A_2/A_3]^{\searrow A_2 \nearrow A_3}$
$w_3 \mapsto [A_3/A_3]^{\searrow A_3}, [A_3]^{\searrow A_3}$

В качестве примера рассмотрим предложение $w_1 w_1 w_2 w_2 w_3 w_3$. Сопоставим его словам следующие категории:

$$\overset{w_1}{[\,S\,/A_1]^{\nearrow A_2}} \quad \overset{w_1}{[\,A_1\,/A_2]^{\nearrow A_2}} \quad \overset{w_2}{[\,A_2\,/A_2]^{\searrow A_2 \nearrow A_3}} \quad \overset{w_2}{[\,A_2\,/A_3]^{\searrow A_2 \nearrow A_3}}$$

$$\overset{w_3}{[\,A_3\,/A_3]^{\searrow A_3}} \quad \overset{w_3}{[\,A_3\,]^{\searrow A_3}}$$

Эту строку категорий можно сократить до S следующим образом :

$[\,S\,/A_1]^{\nearrow A_2}[\,A_1\,/A_2]^{\nearrow A_2}[\,A_2\,/A_2]^{\searrow A_2 \nearrow A_3}[\,A_2\,/A_3]^{\searrow A_2 \nearrow A_3}[\,A_3\,/\mathbf{A_3}]^{\searrow A_3}[\,\mathbf{A_3}\,]^{\searrow A_3} \vdash^{L^r}$

$[\,S\,/A_1]^{\nearrow A_2}[\,A_1\,/A_2]^{\nearrow A_2}[\,A_2\,/A_2]^{\searrow A_2 \nearrow A_3}[\,A_2\,/\mathbf{A_3}]^{\searrow A_2 \nearrow A_3}[\,\mathbf{A_3}\,]^{\searrow A_3 \searrow A_3} \vdash^{L^r}$

$[\,S\,/A_1]^{\nearrow A_2}[\,A_1\,/A_2]^{\nearrow A_2}[\,A_2\,/\mathbf{A_2}]^{\searrow A_2 \nearrow A_3}[\,\mathbf{A_2}\,]^{\searrow A_2 \nearrow A_3 \searrow A_3 \searrow A_3} \vdash^{L^r}$

$[\,S\,/A_1]^{\nearrow A_2}[\,A_1\,/\mathbf{A_2}]^{\nearrow A_2}[\,\mathbf{A_2}\,]^{\searrow A_2 \nearrow A_3 \searrow A_2 \nearrow A_3 \searrow A_3 \searrow A_3} \vdash^{L^r}$

$[\,S\,/\mathbf{A_1}]^{\nearrow A_2}[\,\mathbf{A_1}\,]^{\nearrow A_2 \searrow A_2 \nearrow A_3 \searrow A_2 \nearrow A_3 \searrow A_3 \searrow A_3} \vdash^{L^r}$

$[\,S\,]^{\nearrow A_2 \nearrow \mathbf{A_2} \searrow \mathbf{A_2} \nearrow A_3 \searrow A_2 \nearrow A_3 \searrow A_3 \searrow A_3} \vdash^{D}$

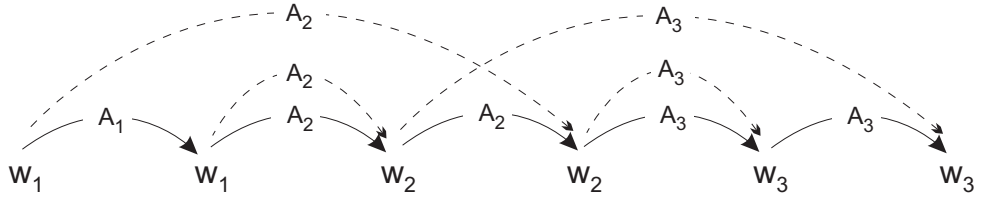
$[\,S\,]^{\nearrow \mathbf{A_2} \nearrow A_3 \searrow \mathbf{A_2} \nearrow A_3 \searrow A_3 \searrow A_3} \vdash^{D}$

$[\,S\,]^{\nearrow A_3 \nearrow \mathbf{A_3} \searrow \mathbf{A_3} \searrow A_3} \vdash^{D}$

$[\,S\,]^{\nearrow \mathbf{A_3} \searrow \mathbf{A_3}} \vdash^{D}$

$[\,S\,]$

Категория S является главной, поэтому $w_1^2 w_2^2 w_3^3 \in L(G_3)$. Соответствующая выводу структура зависимостей выглядит следующим образом:



Несложно видеть, что локальные правила сокращения $L^l, L^r, I^l, I_0^l, I^r, I_0^r$ действуют только на локальную часть категории, тогда как правило сокращения потенциала D затрагивает только потенциал. Это даёт возможность разделить анализ предложения в оКГЗ на два независимых теста: первый — в терминах локальных категорий и локальных правил сокращения, а второй — в терминах сбалансированности потенциала. В [6] доказана теорема об анализе для оКГЗ.

Теорема 1. *Предложение $s = w_1 w_2 \dots w_n$ принадлежит языку $L(G)$, задаваемому оКГЗ $G =< W, \mathbf{C}, S, \delta >$, тогда и только тогда, когда существует строка категорий $\Gamma = \alpha_1^{\theta_1} \dots \alpha_n^{\theta_n}$, где $\alpha_i^{\theta_i} \in \delta(w_i)$, такая, что:*
1. $\alpha_1 \dots \alpha_n \vdash^* S$,
2. $\theta_1 \dots \theta_n$ сбалансирован.

3. Нормальная форма оКГЗ

Известно, что для кс-грамматик существуют различные специальные формы (нормальная форма Хомского, нормальная форма Грейбах (см. [2])), упрощающие анализ порождаемых ими языков. В этом разделе мы определяем нормальную и сильную нормальную формы для оКГЗ, аналогичные нормальной форме Грейбах, и показываем, как для произвольной оКГЗ построить эквивалентные, т.е. порождающие тот же язык, оКГЗ в этих нормальных формах.

Определение 6. (Нормальная форма Грейбах) *Кс-грамматика* $G =< \Sigma, N, S, R >$ *находится в сильной нормальной форме Грейбах, если* G *есть грамматика без* ε*-правил и каждое правило из* R *имеет один из следующих видов:*
1) $S \to \varepsilon$,
2) $A \to x\alpha$, *где* $x \in \Sigma$, $A \in N$, $\alpha \in N^*$, $|\alpha| \leqslant 2$.

Известно, что для любой кс-грамматики можно построить эквивалентную ей кс-грамматику в сильной нормальной форме Грейбах. Определим нормальную форму для оКГЗ следующим образом.

Определение 7. *Будем называть оКГЗ* $G =< W, \mathbf{C}, S, \delta >$ *грамматикой в нормальной форме, если все её категории имеют один из следующих видов:*
1) $[X]^\theta$
2) $[X/Y]^\theta$
3) $[X/Y/Z]^\theta$,
где X, Y, Z *— элементарные категории,* θ *— потенциал.*

Нетрудно понять, что категориальная грамматика, построенная по кс-грамматике в сильной нормальной форме Грейбах, удовлетворяет определению 7, в котором опущены потенциалы (см. определение 11). В определении 7 не накладываются никакие ограничения на вид потенциала. Можно рассматривать грамматики, в которых категории имеют ещё более простой вид.

Определение 8. *Будем называть оКГЗ* $G =< W, \mathbf{C}, S, \delta >$ *грамматикой в сильной нормальной форме, если она находится в нормальной форме и все потенциалы имеют вид* v^n, *где* $v \in \nearrow\mathbf{C} \cup \searrow\mathbf{C}$.

Для перехода от произвольной оКГЗ к оКГЗ в нормальной форме мы определим вспомогательную кс-грамматику и докажем, что сокращения, выполняемые оКГЗ, можно промоделировать в этой грамматике.

Определение 9. *Пусть* $G =< W, \mathbf{C}, S, \delta >$. *Через CF(G) обозначим кс-грамматику*
$G' =< \Sigma, N, S, R >$, *где:*
$\Sigma = \{\, w^\theta \mid w \mapsto [\alpha]^\theta \in \delta$ *для некоторого* $\alpha \,\}$;
N *— множество всех локальных подкатегорий из* δ ;
R *определяется следующим образом:*
$[\alpha] \to w^\theta \in R \Leftrightarrow w \mapsto [\alpha]^\theta \in \delta$
$[\alpha] \to [A][A\backslash\alpha] \in R \Leftrightarrow [A\backslash\alpha] \in N$
$[\alpha] \to [\alpha/A][A] \in R \Leftrightarrow [\alpha/A] \in N$
$[\alpha] \to [A^*\backslash\alpha] \in R \Leftrightarrow [A^*\backslash\alpha] \in N$

$[A^*\backslash\alpha] \to [A][A^*\backslash\alpha] \in R \Leftrightarrow [A^*\backslash\alpha] \in N$
$[\alpha] \to [\alpha/A^*] \in R \Leftrightarrow [\alpha/A^*] \in N$
$[\alpha/A^*] \to [\alpha/A^*][A] \in R \Leftrightarrow [\alpha/A^*] \in N$

Слова в новом алфавите Σ можно разделить на две части: слово в исходном алфавите W и потенциал.

Определение 10. *Пусть* $u = w_1^{\theta_1} \dots w_n^{\theta_n} \in \Sigma^*$. *Тогда*

$$word(u) = w_1 \dots w_n, pot(u) = \theta_1 \dots \theta_n.$$

Связь исходной оКГЗ с построенной по ней кс-грамматикой выражается следующей леммой.

Лемма 1. *Пусть* G *— оКГЗ,* $G' = CF(G)$, $[\alpha] \in N$. *Тогда* $[\alpha] \Rightarrow^*_{G'} u \in \Sigma^*$ *тогда и только тогда, когда существует строка категорий* $\Gamma \in \delta(word(u))$ *такая, что* $\Gamma \vdash^*_G [\alpha]^{pot(u)}$.

Доказательство. [$\Rightarrow$] Индукция по длине вывода u в G'.
Базис индукции.
Длина вывода равна 1, вывод имеет вид $\alpha \Rightarrow w_1^{\theta_1}$. По построению, $[\alpha]^{\theta_1} \in \delta(w_1)$. В качестве Γ возьмём $[\alpha]^{\theta_1}$.
Индукционный шаг.
Пусть для выводов длины меньше m утверждение верно. Пусть $[\alpha] \Rightarrow^* u = w_1^{\theta_1} \dots w_n^{\theta_n}$.
1) Вывод имеет вид $[\alpha] \Rightarrow [\alpha/A][A] \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$. Тогда $[\alpha/A] \Rightarrow^* w_1^{\theta_1} \dots w_k^{\theta_k}$, $[A] \Rightarrow^* w_{k+1}^{\theta_{k+1}} \dots w_n^{\theta_n}$ для некоторого k. По индукционному предположению существуют строки категорий $\Gamma_1 \in \delta(w_1 \dots w_k)$ и $\Gamma_2 \in \delta(w_{k+1} \dots w_n)$ такие, что $\Gamma_1 \vdash^*_G [\alpha/A]^{\theta_1 \dots \theta_k}$, $\Gamma_2 \vdash^*_G [A]^{\theta_{k+1} \dots \theta_n}$. Тогда $\Gamma_1\Gamma_2 \vdash^*_G [\alpha]^{\theta_1 \dots \theta_n}$ и в качестве Γ можно взять $\Gamma_1\Gamma_2$.
2) Вывод имеет вид $[\beta/A^*] \Rightarrow [\beta/A^*][A]$, где $[\beta/A^*]$ есть $[\alpha]$. Тогда $[\beta/A^*] \Rightarrow^* w_1^{\theta_1} \dots w_k^{\theta_k}$, $[A] \Rightarrow^* w_{k+1}^{\theta_{k+1}} \dots w_n^{\theta_n}$ для некоторого k. По индукционному предположению существуют строки категорий $\Gamma_1 \in \delta(w_1 \dots w_k)$ и $\Gamma_2 \in \delta(w_{k+1} \dots w_n)$ такие, что $\Gamma_1 \vdash^*_G [\beta/A^*]^{\theta_1 \dots \theta_k}$, $\Gamma_2 \vdash^*_G [A]^{\theta_{k+1} \dots \theta_n}$. Тогда $\Gamma_1\Gamma_2 \vdash^*_G [\alpha]^{\theta_1 \dots \theta_n}$ и в качестве Γ можно взять $\Gamma_1\Gamma_2$.
3) Вывод имеет вид $[\alpha] \Rightarrow [\alpha/A^*] \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$. По индукционному предположению существует строка категорий $\Gamma_1 \in \delta(w_1 \dots w_n)$ такая, что $\Gamma_1 \vdash^*_G [\alpha/A^*]^{\theta_1 \dots \theta_n}$ и в качестве Γ можно взять Γ_1.
Случаи $[\alpha] \Rightarrow [A][A\backslash\alpha]$, $[A^*\backslash\beta] \Rightarrow [A][A^*\backslash\beta]$ и $[\alpha] \Rightarrow [A^*\backslash\alpha]$ рассматриваются аналогично.
[$\Leftarrow$] Индукция по длине вывода для u в G.
Базис индукции.
Длина вывода равна 1. Тогда $u = w_1^{\theta_1}$. Строка категорий γ имеет вид $[\alpha]^{\theta_1}$. По построению $[\alpha] \to w_1^{\theta_1} \in R$ и искомый вывод имеет вид $[\alpha] \Rightarrow w_1^{\theta_1}$.
Индукционный шаг.
Пусть для выводов длины меньше n утверждение верно. Пусть $u = w_1^{\theta_1} \dots w_n^{\theta_n}$ и $\Gamma \vdash^*_G [\alpha]^{\theta_1 \dots \theta_n}$.
1) Последним было выполнено сокращение $[\alpha/A]^{\theta'}[A]^{\theta''} \vdash [\alpha]^{\theta}$, где $\theta = \theta'\theta''$. Это

значит, что $\Gamma = \Gamma'\Gamma''$, так что $\Gamma' \in \delta(w_1 \dots w_k)$ и $\Gamma'' \in \delta(w_{k+1} \dots w_n)$ для некоторого k и $\Gamma' \vdash_G^* [\alpha/A]^{\theta'}$, $\Gamma'' \vdash_G^* [A]^{\theta''}$. По индукционному предположению $[\alpha/A] \Rightarrow^* w_1^{\theta_1} \dots w_k^{\theta_k}$, $[A] \Rightarrow^* w_{k+1}^{\theta_{k+1}} \dots w_n^{\theta_n}$. Следовательно, $[\alpha] \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$.
Случаи $[\alpha/A^*]^{\theta'}[A]^{\theta''} \vdash [\alpha/A^*]^{\theta'\theta''}$, $[A]^{\theta'}[A\backslash\alpha]^{\theta''} \vdash [\alpha]^{\theta'\theta''}$ и $[A]^{\theta'}[A^*\backslash\alpha]^{\theta''} \vdash [A^*\backslash\alpha]^{\theta'\theta''}$ рассматриваются аналогично.
2) Последним было выполнено сокращение $[\alpha/A^*]^{\theta} \vdash [\alpha]^{\theta}$. По индукционному предположению $[\alpha/A^*] \vdash^* w_1^{\theta_1} \dots w_n^{\theta_n}$. Следовательно, $[\alpha] \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$.
Случай $[A^*\backslash\alpha]^{\theta} \vdash [\alpha]^{\theta}$ рассматривается аналогично. □

Следствие 1. *Пусть G — оКГЗ, $G' = CF(G)$. Тогда $w_1 \dots w_n \in L(G)$ тогда и только тогда, когда $w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G')$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован.*

Доказательство. Пусть слово $u \in \Sigma^*$ таково, что $pot(u)$ сбалансирован. По лемме 1 $S \Rightarrow_{G'}^* u$ тогда и только тогда, когда существует строка категорий $\Gamma \in \delta(word(u))$ такая, что $\Gamma \vdash_G^* [S]^{pot(u)}$. Последнее по теореме об анализе означает, что слово $word(u)$ принадлежит языку, порождаемому грамматикой G. Таким образом, слово w принадлежит языку $L(G)$ тогда и только тогда, когда оно выводимо в G' со сбалансированным потенциалом. □

Теперь опишем «обратную» процедуру, которая по кс-грамматике строит оКГЗ.

Определение 11. *Пусть $G' =< \Sigma, N, S, R >$ —кс-грамматика в сильной нормальной форме Грейбах, где элементы Σ имеют вид w^{θ}. Через $gCDG(G')$ обозначим оКГЗ $G =< W, N, S, \delta >$, где:*
$W = \{ w \mid w^{\theta} \in \Sigma$ *для некоторого* $\theta \}$;
δ определяется следующим образом:
$w \mapsto [X]^{\theta} \in \delta \Leftrightarrow X \to w^{\theta} \in R$
$w \mapsto [X/Y]^{\theta} \in \delta \Leftrightarrow X \to w^{\theta} Y \in R$
$w \mapsto [X/Z/Y]^{\theta} \in \delta \Leftrightarrow X \to w^{\theta} Y Z \in R$

Имеет место свойство аналогичное лемме 1.

Лемма 2. *Пусть $G = gCDG(G')$. Тогда $X \Rightarrow_{G'}^* u \in \Sigma^*$ тогда и только тогда, когда существует строка категорий $\Gamma \in \delta(word(u))$ такая, что $\Gamma \vdash_G^* [X]^{pot(u)}$.*

Доказательство. [⇒] Индукция по длине u.
Базис индукции.
$|u| = 1, u = w^{\theta}$
Вывод u в G' имеет вид $X \Rightarrow w^{\theta}$. По построению $w \mapsto [X]^{\theta} \in \delta$ и в качестве Γ возьмём $[X]^{\theta}$.
Индукционный шаг.
Пусть для слов длины меньше n утверждение верно, $u = w_1^{\theta_1} \dots w_n^{\theta_n}$, $X \Rightarrow^* u$.
1) Пусть первым было применено правило $X \to w_1^{\theta_1} Y$. Тогда $Y \Rightarrow^* w_2^{\theta_2} \dots w_n^{\theta_n}$. По индукционному предположению существует строка категорий $\Gamma_1 \in \delta(w_2 \dots w_n)$ такая, что $\Gamma_1 \vdash_G^* [Y]^{\theta_2 \dots \theta_n}$. По построению $w_1 \mapsto [X/Y]^{\theta_1} \in \delta$. Следовательно, можно взять $\Gamma = [X/Y]^{\theta_1}\Gamma_1$.
2) Пусть первым было применено правило $X \to w_1^{\theta_1} Y Z$. Тогда $Y \Rightarrow^* w_2^{\theta_2} \dots w_k^{\theta_k}$, $Z \Rightarrow^* w_{k+1}^{\theta_{k+1}} \dots w_n^{\theta_n}$ для некоторого k. По индукционному предположению существуют строки категорий $\Gamma_1 \in \delta(w_2 \dots w_k)$, $\Gamma_2 \in \delta(w_{k+1} \dots w_n)$ такие, что $\Gamma_1 \vdash_G^*$

$[Y]^{\theta_2 \dots \theta_k}$, $\Gamma_2 \vdash_G^* [Z]^{\theta_{k+1} \dots \theta_n}$. По построению $w_1 \mapsto [X/Z/Y]^{\theta_1} \in \delta$. Следовательно, можно взять $\Gamma = [X/Z/Y]^{\theta_1}\Gamma_1\Gamma_2$.
[$\Leftarrow$] Индукция по длине u.
Базис индукции.
$|u| = 1, u = w^\theta$ В этом случае $\Gamma = [X]^\theta$. Поэтому $X \to w^\theta \in R$ и искомым выводом будет $X \Rightarrow w^\theta$.
Индукционный шаг.
Пусть для слов длины меньше n утверждение верно, $u = w_1^{\theta_1} \dots w_n^{\theta_n}$ и существует строка категорий $\Gamma \in \delta(w_1 \dots w_n)$ такая, что $\Gamma \vdash_G^* [X]^{\theta_1 \dots \theta_n}$.
1) w_1 приписана категория $[X/Y]^{\theta_1}$. Значит, существует строка категорий $\Gamma_1 \in \delta(w_2 \dots w_n)$ такая, что $\Gamma_1 \vdash_G^* [Y]^{\theta_2 \dots \theta_n}$. По индукционному предположению $Y \Rightarrow^* w_2^{\theta_2} \dots w_n^{\theta_n}$. Следовательно, $X \Rightarrow w_1^{\theta_1} Y \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$.
2) w_1 приписана категория $[X/Z/Y]^{\theta_1}$. Значит, существуют строки категорий $\Gamma_1 \in \delta(w_2 \dots w_k)$, $\Gamma_2 \in \delta(w_{k+1} \dots w_n)$ для некоторого k такие, что $\Gamma_1 \vdash_G^* [Y]^{\theta_2 \dots \theta_k}$, $\Gamma_2 \vdash_G^* [Z]^{\theta_{k+1} \dots \theta_n}$. По индукционному предположению $Y \Rightarrow^* w_2^{\theta_2} \dots w_k^{\theta_k}$, $Z \Rightarrow^* w_{k+1}^{\theta_{k+1}} \dots w_n^{\theta_n}$. Следовательно, $X \Rightarrow w_1^{\theta_1} YZ \Rightarrow^* w_1^{\theta_1} \dots w_n^{\theta_n}$. □

Следствие 2. *Пусть G' — кс-грамматика в сильной нормальной форме Грейбах, $G = gCDG(G')$. Тогда $w_1 \dots w_n \in L(G)$ тогда и только тогда, когда $w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G')$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован.*

Доказательство. Пусть слово $u \in \Sigma^*$ таково, что $pot(u)$ сбалансирован. По лемме 2 $[S] \Rightarrow_{G'}^* u$ тогда и только тогда, когда существует строка категорий $\Gamma \in \delta(word(u))$ такая, что $\Gamma \vdash_G^* [S]^{pot(u)}$. Последнее по теореме об анализе означает, что слово $word(u)$ принадлежит языку, порождаемому грамматикой G. □

Теперь можно доказать теорему о возможности приведения всякой оКГЗ к нормальной форме.

Теорема 2. *Для любой оКГЗ $G =< W, \mathbf{C}, S, \delta >$ существует оКГЗ $G' =< W, \mathbf{C}', S, \delta' >$ в нормальной форме такая, что $L(G) = L(G')$.*

Доказательство. Построим кс-грамматику $G_1 = CF(G)$ по исходной грамматике G. Далее построим кс-грамматику G_2 в сильной нормальной форме Грейбах, эквивалентную G_1. В качестве G' возьмём $gCDG(G_2)$. Имеют место три утверждения.
1) $w_1 \dots w_n \in L(G) \Leftrightarrow w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_1)$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован (по следствию 1)
2) $w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_1) \Leftrightarrow w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_2)$
3) $w_1 \dots w_n \in L(G') \Leftrightarrow w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_2)$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован (по следствию 2)
Из этих утверждений следует, что $w_1 \dots w_n \in L(G) \Leftrightarrow w_1 \dots w_n \in L(G')$, т. е. $L(G) = L(G')$. По построению грамматика G' записана в нормальной форме. □

Теорему 2 можно усилить следующим образом.

Теорема 3. *Для любой оКГЗ $G =< W, \mathbf{C}, S, \delta >$ существует оКГЗ $G' =< W, \mathbf{C}', S, \delta' >$ в сильной нормальной форме такая, что $L(G) = L(G')$.*

Доказательство. Построим грамматику $\tilde{G} =< W, \mathbf{C_1}, S, \delta' >$, где $\mathbf{C_1} = \mathbf{C} \cup \{ A_1 \mid A \in \mathbf{C} \}$, а все категории получаются из категорий G заменой валентностей $\swarrow A$ и $\nwarrow A$ на $\nearrow A_1$ и $\searrow A_1$ соответственно. Легко видеть, что $L(G) = L(\tilde{G})$.

Как и при доказательстве теоремы 2 построим кс-грамматику $G_1 = CF(\tilde{G})$. Обозначим через p число элементарных категорий в грамматике $\tilde{G}$. Для G_1 можно построить эквивалентную грамматику $G_2 =< \Sigma, N', S, R' >$, все правила которой имеют один из следующих видов:
$A \to x$, $A \to yB$, $A \to yBC$, где $A, B, C \in N'$, $x, y \in \Sigma^+, |x| \leqslant 2p, |y| = 2p$ (см. [2]). Преобразуем G_2. Сначала удалим все правила вида $A \to x$, для которых $pot(x)$ не сбалансирован. Далее рассмотрим произвольное правило $r \to y\alpha$. Пусть $\theta = pot(y)$. Выполним в θ все возможные сокращения. Получившийся потенциал θ' перепишем в виде $(\searrow A_1)^{i_1} \dots (\searrow A_p)^{i_p} (\nearrow A_1)^{j_1} \dots (\nearrow A_p)^{j_p}$. Заменим правило r на правило $r' = A \to y_1^{\theta_1} \dots y_p^{\theta_p} y_{p+1}^{\theta'_1} \dots y_{2p}^{\theta'_p}$, где $\theta_k = (\searrow A_k)^{i_k}$, $\theta'_k = (\nearrow A_k)^{j_k}$. Неформально это означает, что мы распределили потенциал из стрелок $2p$ типов по $2p$ буквам, так что каждая буква получила стрелки одного типа. По получившейся грамматике G_3 построим эквивалентную грамматику G_4 в сильной нормальной форме Грейбах. По ней, как и в теореме 2, построим $G' = gCDG(G_4)$.

Пусть $w_1 \dots w_n$ — произвольное слово в алфавите W.
$w_1 \dots w_n \in L(G) \Leftrightarrow$
$w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_1)$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован $\Leftrightarrow$
$w_1^{\theta'_1} \dots w_n^{\theta'_n} \in L(G_2)$ и потенциал $\theta'_1 \dots \theta'_n$ сбалансирован $\Leftrightarrow$
$w_1^{\theta''_1} \dots w_n^{\theta''_n} \in L(G_3)$ и потенциал $\theta''_1 \dots \theta''_n$ сбалансирован $\Leftrightarrow$
$w_1^{\theta'''_1} \dots w_n^{\theta'''_n} \in L(G_4)$ и потенциал $\theta'''_1 \dots \theta'''_n$ сбалансирован $\Leftrightarrow w_1 \dots w_n \in L(G')$
Это и означает, что $L(G) = L(G')$. □

4. Замкнутость относительно обращения гомоморфизма

В работе [6] отмечено, что класс оКГЗ-языков замкнут относительно операций объединения, конкатенации, пересечения с регулярными множествами и неукорачивающего гомоморфизма. Там же дан набросок достаточно сложного доказательства замкнутости этого класса языков относительно обращения гомоморфизма. В этом разделе мы докажем это утверждение, используя существование нормальной формы оКГЗ.

Прежде всего заметим, что достаточно доказать замкнутость для гомоморфизмов
$h\colon W \to \Sigma^*$, $W \cap \Sigma = \varnothing$, отличающихся от биекции $h\colon W \leftrightarrow \Sigma$ не более чем одним присваиванием в форме либо $h(x) = ab, a, b \in \Sigma$, либо $h(x) = \varepsilon$. Действительно, пусть $h = \tau \circ \rho$, где $\tau\colon W \to \Delta^*$ и $\rho\colon \Delta \to \Sigma^*$ — произвольные гомоморфизмы и $L \subseteq \Sigma^*$ — любой язык в алфавите Σ. Тогда $\tau^{-1}(\rho^{-1}(L)) = \{ v \in W^* \mid \tau(v) \in \rho^{-1}(L) \} =$
$\{ v \in W^* \mid \rho(\tau(v)) \in L \} =$
$\{ v \in W^* \mid h(v) \in L \} = h^{-1}(L)$. Теперь справедливость нашего замечания следует из того, что любой гомоморфизм можно представить в виде композиции гомоморфизмов специального вида.

Теорема 4. *Если $L \subseteq \Sigma^*$ является оКГЗ-языком и $h\colon W \to \Sigma^*$ — гомоморфизм специального вида, то $h^{-1}(L)$ также является оКГЗ-языком.*

Доказательство. Пусть язык L задаётся оКГЗ $G =< \Sigma, \mathbf{C}, S, \delta >$. В соответствии с теоремой 2 можно считать, что G — оКГЗ в нормальной форме. Построим оКГЗ $G' =< W, \mathbf{C}', S, \delta' >$.

I. Пусть $h(x) = ab, a, b \in \Sigma, C_1 \in \delta(a), C_2 \in \delta(b)$. Тогда положим $\delta' = \delta \cup \delta_x$, где $\delta_x = \{ x \mapsto [\alpha/\beta]^{\theta_1\theta_2}$ для всех $C_1 = [\alpha/A]^{\theta_1}, C_2 = [A/\beta]^{\theta_2}$. Неформально мы «склеили» категории букв a и b в категорию буквы x. Имеет место следующая лемма.

Лемма 3. *Пусть $w \in \Sigma^*, u \in W^*, h(u) = w$, X — элементарная категория. Существует строка категорий $\Gamma \in \delta(w)$ такая, что $\Gamma \vdash_G^* [X]^\theta$ тогда и только тогда, когда существует строка категорий $\Gamma' \in \delta'(u)$ такая, что $\Gamma' \vdash_{G'}^* [X]^\theta$.*

Доказательство. [$\Leftarrow$] Индукция по длине u.
Базис индукции.
$|u| = 1$
Возможны два случая.
1) $u = u_1 \neq x$
Тогда $w = h(u) = u_1$ и $\Gamma = \Gamma'$.
2)$u = x$
Тогда $w = h(u) = h(x) = ab$. Γ' есть $[X]^\theta$, что возможно только если $a \mapsto [X/A]^{\theta'}, b \mapsto [A]^{\theta''}$, где $\theta = \theta'\theta''$. Возьмём $\Gamma = [X/A]^{\theta'}[A]^{\theta''}$.
Индукционный шаг.
Пусть для слов длины меньше n утверждение верно, $u = u_1 \dots u_n$ и существует строка категорий $\Gamma' \in \delta'(u)$ такая, что $\Gamma' \vdash_{G'}^* [X]^\theta$. Рассмотрим несколько возможных случаев.
1) $u_1 \neq x$
а) $u_1 \mapsto [X/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u') = u_1h(u') = w_1w'$, $\delta'(u') \vdash_{G'}^* [Y]^{\theta'}$. По индукционному предположению существует $\Gamma'' \in \delta(w')$ такая, что $\Gamma'' \vdash_G^* [Y]^{\theta'}$. В качестве Γ возьмём $[X/Y]^{\theta'}\Gamma''$.
б) $u_1 \mapsto [X/Z/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u'u'') = u_1h(u')h(u'') = w_1w'w''$, $\delta'(u') \vdash_{G'}^* [Y]^{\theta'}$, $\delta'(u'') \vdash_{G'}^* [Z]^{\theta''}$. По индукционному предположению существуют $\Gamma'' \in \delta(w')$, $\Gamma''' \in \delta(w'')$ такие, что $\Gamma'' \vdash_G^* [Y]^{\theta'}$, $\Gamma''' \vdash_G^* [Z]^{\theta''}$. В качестве Γ возьмём $[X/Z/Y]^{\theta_1}\Gamma''\Gamma'''$.
2) $u_1 = x$
а) $x \mapsto [X/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u') = abh(u') = abw', \delta'(u') \vdash_{G'}^* [Y]^{\hat{\theta}}$. По построению существует $A \in \mathbf{C}$ такое, что $a \mapsto [X/A]^{\theta'} \in \delta', b \mapsto [A/Y]^{\theta''} \in \delta', \theta_1 = \theta'\theta''$. По индукционному предположению существует $\Gamma'' \in \delta(w')$ такое, что $\Gamma'' \vdash_G^* [Y]^{\hat{\theta}}$. В качестве Γ возьмём $[X/A]^{\theta'}[A/Y]^{\theta''}\Gamma''$.
б) $x \mapsto [X/Z/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u'u'') = abh(u')h(u'') = abw'w''$, $\delta'(u') \vdash_{G'}^* [Y]^{\hat{\theta}}$, $\delta'(u'') \vdash_{G'}^* [Z]^{\tilde{\theta}}$. По построению существует $A \in \mathbf{C}$ такое, что либо $a \mapsto [X/Z/A]^{\theta'} \in \delta'$, $b \mapsto [A/Y]^{\theta''} \in \delta'$, либо $a \mapsto [X/A]^{\theta'} \in \delta'$, $b \mapsto [A/Z/Y]^{\theta''} \in \delta'$, где $\theta = \theta'\theta''$. По индукционному предположению существуют $\Gamma'' \in \delta(w')$ и $\Gamma''' \in \delta(w'')$ такие, что $\Gamma'' \vdash_G^* [Y]^{\hat{\theta}}$, $\Gamma''' \vdash_G^* [Z]^{\tilde{\theta}}$. В качестве Γ возьмём либо $[X/Z/A]^{\theta'}[A/Y]^{\theta''}\Gamma''\Gamma'''$, либо $[X/A]^{\theta'}[A/Z/Y]^{\theta''}\Gamma''\Gamma'''$.
в) $x \mapsto [X/T/Z/Y]^{\theta_1}$

Аналогично.
[⇒] Индукция по длине u.
Базис индукции.
$|u| = 1$
1) $u = u_1 \neq x$
Тогда $w = h(u) = u_1$ и $\Gamma' = \Gamma$.
2) $u = x$
Тогда $w = h(u) = h(x) = ab$. Строка Γ имеет вид $[X/A]^{\theta'}[A]^{\theta''}$. Поэтому можно взять $\Gamma' = [X]^{\theta'\theta''}$.
Индукционный шаг.
Пусть для слов длины меньше n утверждение верно, $u = u_1 \dots u_n$ и существует строка категорий $\Gamma \in \delta(w)$ такая, что $\Gamma \vdash^*_{G'} [X]^{\theta}$.
1) $u_1 \neq x$
а) $w_1 \mapsto [X/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u') = u_1h(u') = w_1w'$, $\delta(w') \vdash^*_G [Y]^{\theta'}$. По индукционному предположению существует $\Gamma'' \in \delta'(u')$ такая, что $\Gamma'' \vdash^*_{G'} [Y]^{\theta'}$. В качестве Γ' возьмём $[X/Y]^{\theta_1}\Gamma''$.
б) $w_1 \mapsto [X/Z/Y]^{\theta_1}$
В этом случае $h(u) = h(u_1u'u'') = u_1h(u')h(u'') = w_1w'w''$, $\delta(w') \vdash^*_G [Y]^{\theta'}$, $\delta(w'') \vdash^*_G [Z]^{\theta''}$. По индукционному предположению существуют $\Gamma'' \in \delta'(u')$ и $\Gamma''' \in \delta'(u'')$ такие, что $\Gamma'' \vdash^*_{G'} [Y]^{\theta'}$ и $\Gamma''' \vdash^*_{G'} [Z]^{\theta''}$. В качестве Γ' возьмём $[X/Z/Y]^{\theta_1}\Gamma''\Gamma'''$.
2) $u_1 = x$
В этом случае $h(u) = h(u_1u') = abh(u') = abw'$.
а) $a \mapsto [X/Y]^{\theta'}, b \mapsto [Y/Z]^{\theta''}$
Тогда существует строка $\Gamma_1 \in \delta(w')$ такая, что $\Gamma_1 \vdash^*_G [Z]^{\hat{\theta}}$. По индукционному предположению существует $\Gamma_2 \in \delta'(u')$ такая, что $\Gamma 2 \vdash^*_{G'} [Z]^{\hat{\theta}}$. По построению $x \mapsto [X/Z]^{\theta'\theta''} \in \delta'$. Поэтому $\Gamma' = [X/Z]^{\theta'\theta''}\Gamma_2$.
б) $a \mapsto [X/Y]^{\theta'}, b \mapsto [Y/Z/T]^{\theta''}$
Тогда $w' = w'_1w'_2$ и существуют $\Gamma_1 \in \delta(w'_1), \Gamma_2 \in \delta(w'_2)$ такие, что $\Gamma_1 \vdash^*_G [T]^{\hat{\theta}}, \Gamma_2 \vdash^*_G [Z]^{\tilde{\theta}}$. По индукционному предположению существуют $\Gamma_3 \in \delta'(u'_1)$, $\Gamma_4 \in \delta'(u'_2)$ такие, что $\Gamma_3 \vdash^*_{G'} [T]^{\hat{\theta}}$, $\Gamma_4 \vdash^*_{G'} [Z]^{\tilde{\theta}}$. По построению $x \mapsto [X/Z/T]^{\theta'\theta''} \in \delta'$. Поэтому можно взять $\Gamma' = [X/Z/T]^{\theta'\theta''}\Gamma_3\Gamma_4$.
Случаи $a \mapsto [X/Y/Z]^{\theta'}, b \mapsto [Z/T]^{\theta''}$ и $a \mapsto [X/Y/Z]^{\theta'}, b \mapsto [Z/T/R]^{\theta''}$ рассматриваются аналогично. □

Из леммы 3 следует, что существует $\Gamma \in \delta(h(u))$ такая, что $\Gamma \vdash^* [S]$, тогда и только тогда, когда существует $\Gamma' \in \delta'(u)$ такая, что $\Gamma' \vdash^* [S]$, т. е. $u \in L(G')$ тогда и только тогда, когда $h(u) \in L(G)$. Это и означает, что $L(G') = h^{-1}(L(G))$.

II. Пусть $h(x) = \varepsilon$. Тогда $\mathbf{C}' = \mathbf{C} \cup \{ d_x \}, \delta'(x) = \{ d_x \}$ и присваивания заменяются следующим образом:
$\delta\colon [X]^{\theta} \Rightarrow \delta'\colon [d^*_x \backslash X/d^*_x]$
$\delta\colon [X/Y]^{\theta} \Rightarrow \delta'\colon [d^*_x \backslash X/Y/d^*_x]$
$\delta\colon [X/Y/Z]^{\theta} \Rightarrow \delta'\colon [d^*_x \backslash X/Y/Z/d^*_x]$
Слова языка $h^{-1}(L)$ — это слова L с добавлением произвольным образом буквы x.
Пусть $w \in L, h(u) = w$. Существует $\Gamma \in \delta(w)$ такая, что $\Gamma \vdash^* [S]$. Припишем буквам слова u категории, соответствующие категориям букв из w. В получившейся

строке категорий Γ' сначала выполним сокращения для d_x. После этого уберём все итеративные категории. Оставшаяся строка категорий совпадает с Γ. Следовательно, $\Gamma' \vdash^* [S]$.
Обратно, пусть существует строка $\Gamma' \in \delta'(u)$ такая, что $\Gamma' \vdash^* [S]$. Категории $[d_x^* \backslash X/Y/Z/d_x^*]$, $[d_x^* \backslash X/Y/d_x^*]$ и $[d_x^* \backslash X/d_x^*]$ не могут сокращаться. Поэтому сначала были сокращены все категории d_x, а затем были удалены итеративные категории. Получившаяся строка Γ является строкой категорий для слова $w = h(u)$. При этом $\Gamma \vdash^* [S]$, так как $\Gamma' \vdash^* [S]$.
Отсюда следует, что $L(G') = h^{-1}(L(G))$. $\square$

Из теоремы 4 с учётом вышеприведённого замечания получаем окончательный результат.

Теорема 5. *Класс оКГЗ-языков замкнут относительно обращения гомоморфизма.*

5. Теорема о представлении

Для контекстно-свободных языков известен следующий результат (теорема Хомского-Шютценберже): для любого КС-языка L существуют регулярный язык R, язык Дика D и гомоморфизм ϕ такие, что $L = \phi(R \cap D)$ (см. [2]). Мы докажем, что оКГЗ-языки обладают похожим свойством. Сначала определим понятие скобочного языка, который в нашем случае заменит язык Дика.

Определение 12. *Пусть* $\Sigma = \{ a_1, \bar{a}_1, \ldots, a_n, \bar{a}_n \}$. *Будем называть язык* $L \subseteq \Sigma^*$ *скобочным, если для любого* $i = 1, \ldots, n$ *его проекция на* $\{ a_i, \bar{a}_i \}$ *представляет собой язык правильной расстановки скобок.*

Теперь сформулируем аналог теоремы Хомского-Шютценберже для оКГЗ-языков.

Теорема 6. *Для любого оКГЗ-языка* L *существуют КС-язык* L_1, *скобочный язык* $L_{\text{ск}}$ *и гомоморфизм* ϕ *такие, что* $L = \phi(L_1 \cap L_{\text{ск}})$.

Доказательство. Пусть $G =< W, \mathbf{C}, S, \delta >$, $L = L(G)$. Пусть $G_1 =< \Sigma, N, S, R >$ эквивалентна $CF(G)$ (определение 9) и находится в сильной нормальной форме Грейбах. Определим алфавит $\Delta = W \cup \{ \overline{w} \mid w \in W \} \cup V(\mathbf{C})$. Построим кс-грамматику $G_2 =< \Delta, N, S, R_1 >$, где R_1 определим по R.
$X \to a^\theta \in R \Leftrightarrow X \to a\bar{a}\theta \in R_1$
$X \to a^\theta Y \in R \Leftrightarrow X \to a\bar{a}\theta Y \in R_1$
$X \to a^\theta YZ \in R \Leftrightarrow X \to a\bar{a}\theta YZ \in R_1$
Справедлива следующая лемма.

Лемма 4. $X \Rightarrow_{G_1}^* a_1^{\theta_1} \ldots a_n^{\theta_n}$ *тогда и только тогда, когда* $X \Rightarrow_{G_2}^* a_1\bar{a}_1\theta_1 \ldots a_n\bar{a}_n\theta_n$.

Лемма доказывается непосредственной индукцией по длине вывода. В частности, из леммы следует, что $S \Rightarrow_{G_1}^* a_1^{\theta_1} \ldots a_n^{\theta_n}$ тогда и только тогда, когда $S \Rightarrow_{G_2}^* a_1\bar{a}_1\theta_1 \ldots a_n\bar{a}_n\theta_n$.

Теперь можно получить требуемое представление языка L. В качестве L_1 возьмём $L(G_2)$, а в качестве $L_{\text{ск}}$ — скобочный язык в алфавите Δ, в котором правильными парами скобок являются $(a, \bar{a})$ и $(v, \breve{v})$, где $a \in W$, $v \in \swarrow \mathbf{C} \cup \nearrow \mathbf{C}$. Гомоморфизм ϕ определим как проекцию Δ на W.
$a_1 \dots a_n \in L \Leftrightarrow$
$S \Rightarrow^*_{G_1} a_1^{\theta_1} \dots a_n^{\theta_n}$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован $\Leftrightarrow$
$S \Rightarrow^*_{G_2} a_1\bar{a}_1\theta_1 \dots a_n\bar{a}_n\theta_n$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован $\Leftrightarrow$
$a_1\bar{a}_1\theta_1 \dots a_n\bar{a}_n\theta_n \in L_1$ и $a_1\bar{a}_1\theta_1 \dots a_n\bar{a}_n\theta_n \in L_{\text{ск}} \Leftrightarrow$
$a_1\bar{a}_1\theta_1 \dots a_n\bar{a}_n\theta_n \in L_1 \cap L_{\text{ск}} \Leftrightarrow$
$a_1 \dots a_n \in \phi(L_1 \cap L_{\text{ск}})$
Следовательно, $L = \phi(L_1 \cap L_{\text{ск}})$. □

6. Счётчиковые МП-автоматы

6.1 Основные определения

В этом разделе мы покажем, что оКГЗ-языки можно задавать с помощью специальных расширений автоматов с магазинной памятью — счётчиковых МП-автоматов.Содержательно, счётчиковый МП-автомат — это МП-автомат, дополнительно оборудованный конечным числом счётчиков. Автомат работает так же как и обычный МП-автомат (за исключением того, что он не делает ε-тактов). Дополнительно на каждом шаге автомат изменяет значения счётчиков (однако сам шаг от этих значений не зависит).

Дадим формальное определение.

Определение 13. *k-счётчиковым автоматом с магазинной памятью (k-СМП-автоматом) называется семёрка $M =< W, Q, Z, q_0, z_0, P, k >$, где:*
W — множество входных слов;
Q — алфавит состояний;
Z — магазинный алфавит;
$q_0 \in Q$ — начальное состояние;
$z_0 \in Z$ — начальный символ магазина;
P — правила;
k — натуральное число (количество счётчиков).
Правила имеют вид $< q, w, z, < q', \alpha, v >>$, где $q, q' \in Q$, $w \in W$, $z \in Z$, $\alpha \in Z^*$, $v \in \mathbb{Z}^k$.

Теперь опишем работу СМП-автомата.

Определение 14. *Конфигурацией СМП-автомата $M =< W, Q, Z, q_0, z_0, P, k >$ называется четвёрка $< q, s, \gamma, u >$, где $q \in Q$, $s \in W^*$, $\gamma \in Z^*$, $v \in \mathbb{N}^k$.*
На множестве конфигураций определено отношение перехода за один шаг:
$< q, s, \gamma, u > \vdash^1_M < q', s', \gamma', u' >$, если существует правило $< q, s, z, < q', \alpha, v >> \in P$ такое, что выполнены условия:
1) $s = ws'$;
2) $\gamma = z\gamma''$, $\gamma' = \alpha\gamma''$;
3) $u' = u + v$.
Если $\gamma = \varepsilon$ или некоторая компонента вектора u' оказывается отрицательной,

то сделать очередной такт нельзя.
Как обычно определяются отношения $\vdash_M^n$ *и* $\vdash_M^*$.
Будем говорить , что предложение s *распознаётся СМП-автоматом* M*, если* $< q_0, s, z_0, (0, \ldots, 0) > \vdash_M^* < q, \varepsilon, \varepsilon, (0, \ldots, 0) >$. *Язык* $L(M)$*, распознаваемый автоматом, — это множество всех предложений, распознаваемых автоматом.*

6.2 Пример

В качестве примера построим СМП-автомат, распознающий язык

$$L = \{\, w_1^n w_2^n w_3^n \mid n = 0, 1, \ldots \}.$$

$M =< W, Q, Z, q_0, z_0, P, k > W = \{\, w_1, w_2, w_3 \,\}$, $Q = \{\, q_0, q_1, q_2 \,\}$, $Z = \{\, z_0, w_1, w_2, w_3 \,\}$, $k = 1$
Правила автомата будут следующими:

$< q_0, w_1, z_0, < q_0, w_1 z_0, 1 >>$ $\quad$ $< q_0, w_1, w_1, < q_0, w_1 w_1, 1 >>$
$< q_0, w_2, w_1, < q_1, \varepsilon, 0 >>$ $\quad$ $< q_1, w_2, w_1, < q_1, \varepsilon, 0 >>$
$< q_1, w_3, z_0, < q_2, z_0, -1 >>$ $\quad$ $< q_2, w_3, z_0, < q_2, z_0, -1 >>$
$< q_2, w_3, z_0, < q_2, \varepsilon, -1 >>$

При работе автомат проверяет в стеке, что блоки, состоящие из w_1 и w_2, имеют одинаковую длину. Затем с помощью счётчика он проверяет, что длина блока из w_3 такая же. **Утверждение.** $L(M) = L$

6.3 СМП-автоматы и оКГЗ

Интуитивно понятно, что стек СМП-автомата проверяет сокращение локальных категорий, а счётчики действуют так же, как и поляризованные валентности в оКГЗ. Эти аналогии можно формализовать, чтобы доказать, что СМП-автоматы распознают в точности оКГЗ-языки.

В доказательстве этого факта будут использоваться следующие обозначения. Пусть потенциал θ является префиксом правильного скобочного слова в алфавите $\{\nearrow A_1, \searrow A_1, \ldots, \nearrow A_n, \searrow A_n\}$. Обозначим через $c(\theta)$ вектор избытков левых скобок: $c(\theta) = (|\theta|_{\nearrow A_1} - |\theta|_{\searrow A_1}, \ldots, |\theta|_{\nearrow A_n} - |\theta|_{\searrow A_n})$ (в соответствии с теоремой 3 можно считать, что в грамматике нет стрелок из $\swarrow \mathbf{C} \cup \nwarrow \mathbf{C}$). Например, $c(\nearrow A_1 \nearrow A_1 \nearrow A_2 \searrow A_1 \nearrow A_1 \nearrow A_2 \searrow A_2) = (2, 1)$. Так как θ — префикс правильного скобочного слова, то все компоненты вектора $c(\theta)$ неотрицательны. Заметим также, что если потенциалы θ_1 и θ_2 — префиксы правильных скобочных слов, то $c(\theta_1\theta_2) = c(\theta_1) + c(\theta_2)$.
Пусть $u = (u_1, \ldots, u_n) \in \mathbb{Z}^n$. Тогда через $\theta(u)$ обозначим потенциал, «соответствующий» вектору u: $\theta(u) = \theta_1^{|u_1|} \ldots \theta_n^{|u_n|}$, где

$$\theta_i = \begin{cases} \nearrow A_i & \text{при } u_i > 0; \\ \searrow A_i & \text{при } u_i < 0; \\ \varepsilon & \text{при } u_i = 0. \end{cases}$$

Покажем вначале, что класс языков, допускаемых СМП-автоматами, содержит все оКГЗ-языки.

Теорема 7. *Каждый оКГЗ-язык задаётся некоторым СМП-автоматом.*

Доказательство. Пусть $L = L(G)$, где $G =< W, \mathbf{C}, S, \delta >$. Будем считать в соответствии с теоремой 3, что грамматика G находится в сильной нормальной форме. Построим по грамматике автомат $M =< W, \{ q \}, \mathbf{C}, q, S, P, k >$, где $k = |\mathbf{C}|$ — число типов стрелок в грамматике. Правила автомата определим следующим образом.
$w \mapsto [X]^{v^n} \in \delta \Leftrightarrow < q, w, X, < q, \varepsilon, u >> \in P$
$w \mapsto [X/Y]^{v^n} \in \delta \Leftrightarrow < q, w, X, < q, Y, u >> \in P$
$w \mapsto [X/Y/Z]^{v^n} \in \delta \Leftrightarrow < q, w, X, < q, ZY, u >> \in P$
Во всех случаях u определяется следующим образом:
1) если $v = \nearrow A_i$, то $u_i = n, u_j = 0$ при $i \neq j$;
2) если $v = \searrow A_i$, то $u_i = -n, u_j = 0$ при $i \neq j$.
Рассмотрим кс-грамматику $G_1 = CF(G)$, построенную согласно определению 9.

Лемма 5. *Пусть потенциал $\theta = \theta_1 \dots \theta_j$, где $\theta_i = v_i^{n_i}$, является префиксом правильного скобочного слова. Тогда $S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p$ тогда и только тогда, когда*
$< q, w_1 \dots w_j s, S, (0, \dots, 0) > \vdash_M^j < q, s, Z_1 \dots Z_p, c(\theta) >$.

Доказательство. [$\Rightarrow$] Индукция по j.
Базис индукции.
$j = 0$
По определению $< q, s, S, (0, \dots, 0) > \vdash_M^0 < q, s, S, (0, \dots, 0) >$.
Индукционный шаг.
Пусть $S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p$, тогда по индукционному предположению
$< q, w_1 \dots w_i w_{j+1} s, S, (0, \dots, 0) > \vdash_M^j < q, w_{j+1} s, Z_1 \dots Z_p, c(\theta) >$, где $\theta = \theta_1 \dots \theta_j$.
Далее было применено правило одного из трёх видов.
1) $Z_1 \to w_{j+1}^{\theta_{j+1}}$
Тогда $S \Rightarrow_{G_1}^{j+1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} Z_2 \dots Z_p$ и $< q, w_{j+1} s, Z_1 \dots Z_p, c(\theta) > \vdash_M^1$
$< q, s, Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, а значит, $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1}$
$< q, s, Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$.
2) $Z_1 \to w_{j+1}^{\theta_{j+1}} X$
Тогда $S \Rightarrow_{G_1}^{j+1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} X Z_2 \dots Z_p$ и $< q, w_{j+1} s, Z_1 \dots Z_p, c(\theta) > \vdash_M^1$
$< q, s, X Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, а значит, $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1}$
$< q, s, X Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$.
3) $Z_1 \to w_{j+1}^{\theta_{j+1}} XY$
Тогда $S \Rightarrow_{G_1}^{j+1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} XY Z_2 \dots Z_p$ и $< q, w_{j+1} s, Z_1 \dots Z_p, c(\theta) > \vdash_M^1$
$< q, s, XY Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, а значит, $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1}$
$< q, s, XY Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$.
[$\Leftarrow$] Индукция по j.
Базис индукции.
$j = 0$
По определению $S \Rightarrow_{G_1}^0 S$.
Индукционный шаг.
Пусть $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^j < q, w_{j+1} s, Z_1 \dots Z_p, c(\theta) >$, тогда по индукционному предположению $S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p$. Далее автомат использо-

вал правило одного из трёх видов.
1) $< q, w_{j+1}, Z_1, < q, \varepsilon, u >>$
Тогда $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1} < q, s, Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, где θ_{j+1} — потенциал, соответствующий u. Поэтому

$$S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p \Rightarrow_{G_1}^{1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} Z_2 \dots Z_p.$$

2) $< q, w_{j+1}, Z_1, < q, X, u >>$
Тогда $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1} < q, s, X Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, где θ_{j+1} — потенциал, соответствующий u. Поэтому

$$S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p \Rightarrow_{G_1}^{1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} X Z_2 \dots Z_p.$$

3) $< q, w_{j+1}, Z_1, < q, XY, u >>$
Тогда $< q, w_1 \dots w_j w_{j+1} s, S, (0, \dots, 0) > \vdash_M^{j+1} < q, s, XY Z_2 \dots Z_p, c(\theta\theta_{j+1}) >$, где θ_{j+1} — потенциал, соответствующий u. Поэтому

$$S \Rightarrow_{G_1}^{j} w_1^{\theta_1} \dots w_j^{\theta_j} Z_1 \dots Z_p \Rightarrow_{G_1}^{1} w_1^{\theta_1} \dots w_j^{\theta_j} w_{j+1}^{\theta_{j+1}} XY Z_2 \dots Z_p.$$

□
Из леммы следует, что $S \Rightarrow_{G_1}^{*} w_1^{\theta_1} \dots w_n^{\theta_n}$ тогда и только тогда, когда
$< q, w_1 \dots w_n, S, (0, \dots, 0) > \vdash_M^{*} < q, \varepsilon, \varepsilon, c(\theta) >$.
$w_1 \dots w_n \in L(G) \Leftrightarrow w_1^{\theta_1} \dots w_n^{\theta_n} \in L(G_1)$ и потенциал $\theta_1 \dots \theta_n$ сбалансирован $\Leftrightarrow$
$< q, w_1 \dots w_n, S, (0, \dots, 0) > \vdash_M^{*} < q, \varepsilon, \varepsilon, (0, \dots, 0) > \Leftrightarrow w_1 \dots w_n \in L(M)$
Это и означает, что $L(G) = L(M)$. □

Для СМП-автоматов имеют место свойства, аналогичные свойствам МП-автоматов.

Лемма 6. *Пусть* $< q, s, \alpha\beta, u > \vdash_M^{n} < q', \varepsilon, \varepsilon, u' >$. *Тогда* $s = s_1 s_2$, *так что* $< q, s_1, \alpha, u > \vdash_M^{n_1} < q_1, \varepsilon, \varepsilon, u_1 >$, $< q_1, s_2, \beta, u_1 > \vdash_M^{n_2} < q', \varepsilon, \varepsilon, u' >$, *причём* $n = n_1 + n_2$.

Доказательство. Индукция по n.
Базис индукции.
$n = 1$. Тогда $|s| = 1$, $s = w$. Если $\alpha\beta = \varepsilon$, то $< q, s, \alpha, u > \vdash_M^{1} < q', \varepsilon, \beta, u' >$, $< q', \varepsilon, \varepsilon, u' > \vdash_M^{0} < q', \varepsilon, \varepsilon, u' >$. Если же $|\alpha\beta| = 1$, то пусть $\alpha = z$, $\beta = \varepsilon$. Тогда в автомате есть правило $< q, w, z, < q', \varepsilon, v >>$ и поэтому $< q, s, \alpha, u > \vdash_M^{1}$ $< q', \varepsilon, \varepsilon, u' >$, $< q', \varepsilon, \beta, u' > \vdash_M^{0} < q', \varepsilon, \varepsilon, u' >$.
Индукционный шаг.
Пусть для n утверждение верно и пусть $< q, s, \alpha\beta, u > \vdash_M^{n+1} < q', \varepsilon, \varepsilon, u' >$. На первом шаге использовалось правило $< q, w, z, < q'', \gamma, v >>$. Тогда $< q, s, \alpha\beta, u > \vdash_M^{1}$ $< q'', s', \gamma\alpha'\beta, u + v >$, $< q'', s', \gamma\alpha'\beta, u + v > \vdash_M^{n} < q', \varepsilon, \varepsilon, u' >$, где $s = ws'$, $\alpha = z\alpha'$. По индукционному предположению $s' = s_1' s_2'$, так что $< q'', s_1', \gamma\alpha', u + v > \vdash_M^{n_1}$ $< q_1, \varepsilon, \varepsilon, u_1 >$, $< q_1, s_2', \beta, u_1 > \vdash_M^{n_2} > < q', \varepsilon, \varepsilon, u' >$. Поэтому $< q, w s_1', \alpha, u > \vdash_M^{n_1+1}$ $< q_1, \varepsilon, \varepsilon, u_1 >$ и можно взять $s_1 = w s_1'$, $s_2 = s_2'$. □

Обобщение этого свойства описано в следующей лемме.

Лемма 7. *Пусть* $< q, s, \alpha_1 \dots \alpha_k, u >\vdash_M^n < q', \varepsilon, \varepsilon, u' >$. *Тогда* $s = s_1 \dots s_k$, *так что*
$< q, s_1, \alpha_1, u >\vdash_M^{n_1} < q_1, \varepsilon, \varepsilon, u_1 >$, $\dots$, $< q_{k-1}, s_k, \alpha_k, u_{k-1} >\vdash_M^{n_k} < q', \varepsilon, \varepsilon, u' >$, *причём*
$n = n_1 + \dots + n_k$.

Доказательство. Индукция по k.
Базис индукции.
Если $k = 1$, то посылка совпадает с заключением.
Индукционный шаг.
Пусть для k утверждение верно и пусть $< q, s, \alpha_1 \dots \alpha_{k+1}, u >\vdash_M^{n+1} < q', \varepsilon, \varepsilon, u' >$. По лемме 6 $s = s_1 s'$, так что

$$< q, s_1, \alpha_1, u >\vdash_M^{n_1} < q_1, \varepsilon, \varepsilon, u_1 >, < q_1, s', \alpha_2 \dots \alpha_{k+1} >\vdash_M^{n-n_1} < q', \varepsilon, \varepsilon, u' > .$$

По индукционному предположению $s' = s_2 \dots s_{k+1}$, так что $< q_1, s_2, \alpha_2, u_1 >\vdash_M^{n_2} < q_2, \varepsilon, \varepsilon, u_2 >$, $\dots$, $< q_k, s_{k+1}, \alpha_{k+1}, u_k >\vdash_M^{n_{k+1}} < q_{k+1}, \varepsilon, \varepsilon, u' >$, причём $n = n_1 + \dots + n_{k+1}$, что и требовалось доказать. □

Лемма 8. *Пусть* $< q, s_1, \alpha_1, u >\vdash_M^n < q', \varepsilon, \varepsilon, u' >$. *Тогда* $< q, s_1 s, \alpha_1 \alpha, u >\vdash_M^n < q', s, \alpha, u' >$.

Доказательство. Индукция по n.
Базис индукции.
$n = 1$. Было применено правило $< q, w, z, < q', \gamma, v >>$, где $\alpha_1 = z$, $\gamma = \varepsilon$, $s_1 = a$, $u' = u + v$. Тогда $< q, ws, z\alpha, u >\vdash_M^1 < q', s, \alpha, u' >$.
Индукционный шаг.
Пусть для n утверждение верно и пусть $< q, s_1, \alpha_1, u >\vdash_M^{n+1} < q', \varepsilon, \varepsilon, u' >$. Первым было применено правило $< q, w, z, < q'', \gamma, v >>$, где $s = ws'$, $\alpha_1 = z\alpha'$. Тогда имеем $< q, ws', z\alpha', u >\vdash_M^1 < q'', s', \gamma\alpha', u + v >\vdash_M^n < q', \varepsilon, \varepsilon, u' >$. По индукционному предположению $< q'', s's, \gamma\alpha'\alpha, u + v > \vdash_M^n < q', s, \alpha, u' >$. Следовательно, $< q, s_1 s, \alpha_1 \alpha, u >\vdash_M^{n+1} < q', s, \alpha, u' >$. □

Следствием доказанной леммы является следующее утверждение.

Лемма 9. *Пусть*

$$< q_0, s_1, \alpha_1, u_0 >\vdash_M^{n_1} < q_1, \varepsilon, \varepsilon, u_1 >, \dots, < q_{m-1}, s_m, \alpha_m, u_{m-1} >\vdash_M^{n_m} < q_m, \varepsilon, \varepsilon, u_m > .$$

Тогда $< q_0, s_1 \dots s_m, \alpha_1 \dots \alpha_m, u_0 >\vdash_M^{n_1 + \dots + n_m} < q_m, \varepsilon, \varepsilon, u_m >$.

Доказательство. Индукция по m.
Базис индукции.
Если $m = 1$, то посылка совпадает с заключением.
Индукционный шаг.
Пусть для m утверждение верно и пусть $< q_0, s_1, \alpha_1, u_0 >\vdash_M^{n_1} < q_1, \varepsilon, \varepsilon, u_1 >$, $\dots$, $< q_m, s_{m+1}, \alpha_{m+1}, u_m >\vdash_M^{n_{m+1}} < q_{m+1}, \varepsilon, \varepsilon, u_{m+1} >$. По индукционному предположению
$< q_0, s_1 \dots s_m, \alpha_1 \dots \alpha_m, u_0 >\vdash_M^{n_1 + \dots + n_m} < q_m, \varepsilon, \varepsilon, u_m >$. По лемме 8
$< q_0, s_1 \dots s_m s_{m+1}, \alpha_1 \dots \alpha_m \alpha_{m+1}, u_0 >\vdash_M^{n_1 + \dots + n_m} < q_m, s_{m+1}, \alpha_{m+1}, u_m >\vdash_M^{n_{m+1}}$

$< q_{m+1}, \varepsilon, \varepsilon, u_{m+1} >$, что и требовалось доказать. □

Теперь докажем теорему, обратную теореме 7.

Теорема 8. *Пусть язык L задаётся k-СМП-автоматом $M =< W, Q, Z, q_0, z_0, P, k >$. Тогда L порождается некоторой оКГЗ.*

Доказательство. Определим по автомату M кс-грамматику $G_1 =< \Sigma, N, S, R >$:
$\Sigma = \{ w^{\theta(u)} |< q, w, z, < q', \alpha, u >>\in P \}$
$N = \{ [qzq'] \mid q, q' \in Q, z \in Z \} \cup \{ S \}$ множество правил R включает следующие правила:
$S \to [q_0 z_0 q]$ для всех $q \in Q$,
$[qzq'] \to w^{\theta(u)} \in R \Leftrightarrow < q, a, z, < q', \varepsilon, u >>\in P$,
$[qzq_p] \to w^{\theta(u)}[q'zq_1][q_1z_1q_2]\dots[q_{p-1}z_pq_p] \in R$ для всех $q_1, \dots, q_p \in Q \Leftrightarrow$
$< q, w, z, < q', z_1 \dots z_p, u >>\in P$.

Выводы в грамматике G_1 связаны с переходами автомата M следующим образом.

Лемма 10. *Пусть $\theta(u)\theta_1 \dots \theta_j$ — префикс правильного скобочного слова. Тогда* $[qzq'] \Rightarrow^j_{G_1} w_1^{\theta_1} \dots w_j^{\theta_j} \Leftrightarrow < q, w_1 \dots w_j, z, u >\vdash^j_M < q', \varepsilon, \varepsilon, u + c(\theta_1 \dots \theta_j) >$.

Доказательство. [⇒] Индукция по j.
Базис индукции.
Пусть $[qzq'] \Rightarrow^1_{G_1} w_1^{\theta_1}$. Тогда в автомате есть правило $< q, w_1, z, < q'\varepsilon, c(\theta_1) >>$. Поэтому
$< q, w_1, z, u >\vdash^1_M < q', \varepsilon, \varepsilon, u + c(\theta_1) >$.
Индукционный шаг.
Пусть $[qzq'] \Rightarrow^{j+1}_{G_1} w_1^{\theta_1} \dots w_{j+1}^{\theta_{j+1}}$. Выделив в этом выводе первый шаг, получим
$[qzq'] \Rightarrow^1_{G_1} w_1^{\theta_1}[q_1z_1q_2][q_2z_2q_3]\dots[q_pz_pq_{p+1}] \Rightarrow^j_{G_1} w_1^{\theta_1} \dots w_{j+1}^{\theta_{j+1}}$, где $q_{p+1} = q'$. Поэтому для всех $i = 1, \dots, p$ $[q_iz_iq_{i+1}] \Rightarrow^{j_i}_{G_1} v_i \in \Sigma^*$, где $v_1v_2 \dots v_p = w_2^{\theta_2} \dots w_{j+1}^{\theta_{j+1}}$, причём $j_1 + \dots + j_p = j$, $j_i < j$. По индукционному предположению

$$< q_1, word(v_i), z_i, u + c(\theta_1 pot(v_1) \dots pot(v_{i-1})) >\vdash^{j_i}_M$$
$$< q_{i+1}, \varepsilon, \varepsilon, u + c(\theta_1 pot(v_1) \dots pot(v_i)) > .$$

Здесь вместо u в первой конфигурации стоит $u = c(\theta_1 pot(v_1) \dots pot(v_{i-1}))$, так как этот вектор также соответствует префиксу правильного скобочного слова. Тогда для автомата по лемме 9 имеем $< q, w_1 word(v_1) \dots word(v_p), z, u >\vdash^1_M < q_1, word(v_1) \dots word(v_p), z_1 \dots z_p, u + c(\theta_1) >\vdash^j_M < q', \varepsilon, \varepsilon, u + c(\theta_1 \dots \theta_{j+1}) >$.
[⇐] Индукция по j.
Базис индукции.
Пусть $< q, w_1, z, u >\vdash^1_M < q', \varepsilon, \varepsilon, u + c(\theta_1) >$. Тогда в грамматике есть правило $[qzq'] \to w_1^{\theta_1}$. Поэтому $[qzq'] \Rightarrow^1_{G_1} w_1^{\theta_1}$.
Индукционный шаг.
Рассмотрим последовательность конфигураций при работе автомата:

$$< q, w_1 \dots w_{j+1}, z, u >,$$
$$< q_1, w_2 \dots w_{j+1}, z_1 \dots z_p, u + c(\theta_1) >,$$

$$\ldots,$$

$$< q', \varepsilon, \varepsilon, u + c(\theta_1 \ldots \theta_{j+1}) > .$$

По лемме 7 слово $w_2 \ldots w_{j+1}$ можно представить в виде $s_1 \ldots s_p$, так что

$$< q_1, s_1, z_1, u + c(\theta_1) > \vdash_M^{j_1}$$

$$< q_2, \varepsilon, \varepsilon, u + c(\theta_1\theta'_1) >, \cdots < q_p, s_p, z_p, u + c(\theta_1\theta'_1 \ldots \theta'_{p-1}) > \vdash_M^{j_s}$$

$$< q', \varepsilon, \varepsilon, u + c(\theta\theta'_1 \ldots \theta'_p) >,$$

где $\theta_1\theta'_1 \ldots \theta'_p = \theta_1 \ldots \theta_{j+1}$, $j_1 + \ldots j_p = j$, $j_i \leqslant j$. По индукционному предположению $[q_i z_i q_{i+1}] \Rightarrow_{G_1}^{j_i} v_i$, причём $word(v_i) = s_i$, $pot(v_i) = \theta'_i$. Поэтому существует вывод $[qzq'] \Rightarrow_{G_1}^{1} w_1^{\theta_1}[q_1 z_1 q_2][q_2 z_2 q_3] \ldots$
$[q_p z_p q'] \vdash_{G_1}^{j}$
$w_1^{\theta_1} v_1 \ldots v_p = w_1^{\theta_1} \ldots w_{j+1}^{\theta_{j+1}}$. □
Для завершения доказательства теоремы построим по G_1 эквивалентную ей грамматику G_2 в сильной нормальной форме Грейбах. Затем преобразуем ее в оКГЗ $G = gCDG(G_2)$ согласно определению 11. Тогда с учетом леммы 10 получаем, что $w_1 \ldots w_n \in L(M) \Leftrightarrow$
$< q_0, w_1 \ldots w_n, z_0, (0, \ldots, 0) > \vdash_M^* < q, \varepsilon, \varepsilon, (0, \ldots, 0) > \Leftrightarrow [q_0 z_0 q] \Rightarrow_{G_1}^* w_1^{\theta_1} \ldots w_n^{\theta_n}$ и потенциал $\theta_1 \ldots \theta_n$ сбалансирован $\Leftrightarrow S \Rightarrow_{G_1}^* w_1^{\theta_1} \ldots w_n^{\theta_n}$ и потенциал $\theta_1 \ldots \theta_n$ сбалансирован $\Leftrightarrow S \Rightarrow_{G_2}^* w_1^{\theta_1} \ldots w_n^{\theta_n}$ и потенциал $\theta_1 \ldots \theta_n$ сбалансирован $\Leftrightarrow w_1 \ldots w_n \in L(G)$. Это означает, что $L(M) = L(G)$. □

7. Заключение

В этой работе мы определили нормальные формы для оКГЗ и доказали, что любую оКГЗ можно представить в эквивалентной нормальной форме. Используя эту нормальную форму, мы передоказали теорему о замкнутости класса оКГЗ-языков относительно обращения гомоморфизма. Также доказано, что каждый оКГЗ-язык можно получить с помощью гомоморфизма из пересечения кс-языка и скобочного языка (это аналог известной теоремы Хомского-Шютценберже для кс-языков). Введено понятие счётчикового автомата с магазинной памятью (СМП-автомата) и доказано, что СМП-автоматы распознают в точности класс оКГЗ-языков.

Некоторые интересные вопросы об оКГЗ-языках остаются открытыми. В частности, неясно, замкнут ли класс этих языков относительно итерации. Неизвестно также, являются ли все оКГЗ-языки полулинейными.

Автор благодарен М.И. Дехтярю за постановки задач и внимание к работе и А.Я. Диковскому за полезное обсуждение результатов.

Список литературы

[1] А. Ахо, Дж. Ульман. Теория синтаксического анализа, перевода и компиляции. М.:Мир,1978. — 1 т.

[2] А. В. Гладкий. Формальные грамматики и языки. М.:Наука, 1973.

[3] А. В. Гладкий, И. А. Мельчук. Элементы математической лингвистики. М.:Наука, 1969.

[4] Y. Bar-Hillel and H. Gaifman and E. Shamir, On categorial and phrase structure grammars, phBull. Res. Council Israel, 9F,1960, pp. 1–16.

[5] M. Dekhtyar, A. Dikovsky. Categorial Dependency Grammars. B Proc. of Int. Conf. on Categorial Grammars, стр. 76-91, 2004.

[6] M. Dekhtyar, A. Dikovsky. Generalized Categorial Dependency Grammars. Pillars of Computer Science: Essays Dedicated to Boris (Boaz) Trakhtenbrot on the Occasion of His 85th Birthday LNCS, N 4800, 2008, 230-255.

[7] A. Dikovsky.Polarized Non-projective Dependency Grammars. Proc. of the Fourth Intern. Conf. on Logical Aspects of Computational Linguistics, Lecture Notes in Artificial Intelligence. vol. 2099, Springer, 2001,pp. 139–157.

[8] H. Gaifman, Dependency systems and phrase structure systems, Information and Control, 1965, v. 8, n 3, pp. 304-337.