О преобразовании грамматик Ламбека с одним делением в контекстно-свободные грамматики

С. Л. Кузнецов*
Математический институт им. В. А. Стеклова РАН
sk@mi.ras.ru

13 марта 2017 г.

УДК 519.766.23

Аннотация

В работе описан способ построения по грамматике Ламбека с одним делением контекстно-свободной грамматики, задающей тот же язык, размер которой ограничен полиномом от размера исходной грамматики. Известные ранее конструкции Бушковского и Пентуса приводили к экспоненциальному росту размера грамматики.

1 Грамматики Ламбека и контекстно-свободные грамматики

Грамматики Ламбека были введены у И. Ламбека [4] для математического описания синтаксиса фрагментов естественных языков и являются

*Работа выполнена на средства гранта Российского научного фонда, проект № 14-50-00005.
Это препринт статьи, принятой для опубликования в Трудах Математического института им. В. А. Стеклова. © С. Л. Кузнецов, 2016.
Сайт Издателя: http://www.maik.ru/

одной из разновидностей *категориальных грамматик.* Грамматики задают *формальные языки* (слово "формальный" мы далее будем опускать), т. е. множества *слов,* составленных из элементов (*букв,* или *символов*) некоторого конечного множества Σ, называемого *алфавитом.* Множество всех слов над данным алфавитом Σ обозначается Σ^*; множество всех слов, кроме пустого — Σ^+. Пустое слово обозначается ε. Мы будем рассматривать только грамматики, задающие языки без пустого слова (т.е. подмножества Σ^+). При этом пустое слово может возникать внутри грамматики во вспомогательных целях.

Язык может быть бесконечным как множество, поэтому задача его описания конечной грамматикой нетривиальна (в частности, из соображений мощности не для всякого языка существует задающая его формальная грамматика).

Категориальная грамматика представляет собой бинарное соответствие между буквами алфавита и логическими выражениями, называемыми *синтаксическими типами,* при этом каждой букве сопоставляется конечное число синтаксических типов (но не обязательно ровно один). Слово[1] $w = a_1 \dots a_n$ принадлежит языку, задаваемому грамматикой, если найдутся такие синтаксические типы $A_1, \dots, A_n$, что тип A_i находится в заданном соответствии с буквой a_i (для i от 1 до n) и в некотором логическом исчислении выводится *секвенция* $A_1 \dots A_n \to H$, где H — некоторый фиксированный синтаксический тип.

В грамматиках Ламбека в качестве логического исчисления, контролирующего принадлежность слов языку, задаваемому грамматикой, используется *исчисление Ламбека,* обозначаемое L. Синтаксические типы для исчисления Ламбека строятся из множества *примитивных типов* (переменных) Pr с помощью трех бинарных связок — $\cdot$ (называемой *умножением*), $\backslash$ и $/$ (*левого* и *правого делений*). Множество всех синтаксических типов обозначается Tp. Мы будем обозначать типы заглавными латинскими буквами (A, B, C, ...); заглавные греческие буквы (Γ, Δ, Π, ...) используются для обозначения конечных (возможно, пустых) последовательностей типов. Следуя лингвистической традиции, мы не

[1]Здесь следует отметить некоторое расхождение в терминологии между исследованиями "формальных" и "настоящих" языков. В лингвистических приложениях *буквам* из Σ соответствуют не буквы, а *слова* (словоформы) естественного языка; при этом *словам* над Σ соответствуют *предложения.* Таким образом, грамматики предназначены для описания не лексики, а синтаксиса языков. Здесь мы придерживаемся терминологии "буквы – слова", а не "слова – предложения".

разделяем типы в последовательности запятой.

Объекты, выводимые в исчислении L, называются *секвенциями* и имеют вид $\Pi \to B$, где B — синтаксический тип, а Π — непустая последовательность синтаксических типов.

Исчисление L задаётся аксиомами вида $A \to A$ и следующими правилами вывода:

$$\frac{A\Pi \to B}{\Pi \to A \setminus B}\text{, где } \Pi \text{ непусто} \qquad \frac{\Pi \to A \quad \Gamma B \Delta \to C}{\Gamma \Pi (A \setminus B) \Delta \to C}$$

$$\frac{\Pi A \to B}{\Pi \to B \,/\, A}\text{, где } \Pi \text{ непусто} \qquad \frac{\Pi \to A \quad \Gamma B \Delta \to C}{\Gamma (B \,/\, A) \Pi \Delta \to C}$$

$$\frac{\Gamma \to A \quad \Delta \to B}{\Gamma \Delta \to A \cdot B} \qquad \frac{\Gamma A B \Delta \to C}{\Gamma (A \cdot B) \Delta \to C}$$

Кроме того, в исчислении L допустимо [4] *правило сечения:*

$$\frac{\Pi \to A \quad \Gamma A \Delta \to C}{\Gamma \Pi \Delta \to C}$$

(Это означает, что добавление правила сечения не увеличивает множества выводимых секвенций.)

Если в исчислении L выводима секвенция $\Pi \to B$, пишем $\mathrm{L} \vdash \Pi \to B$.

Наконец, грамматикой Ламбека называется тройка $\mathcal{G} = \langle \Sigma, \rhd, H \rangle$, где Σ — алфавит, $H \in \mathrm{Tp}$ — выделенный тип, а $\rhd \subset \Sigma \times \mathrm{Tp}$ — конечное бинарное соответствие между буквами алфавита и синтаксическими типами. Как и говорилось ранее, слово $w = a_1 \dots a_n$ допускается грамматикой $\mathcal{G}$, если существуют такие типы $A_1, \dots, A_n$, что $a_i \rhd A_i$ (для всех i от 1 до n) и $\mathrm{L} \vdash A_1 \dots A_n \to H$.

Заметим, что определенные выше грамматики Ламбека явным образом не могут допускать пустое слово. Однако, если в правилах вывода отбросить условия непустоты левых частей секвенций, получится модифицированное исчисление L^*, и языки, порождаемые основанными на нем грамматиками, уже могут содержать пустое слово.

Другой способ описания синтаксиса языков, восходящим к Хомскому [10], — это *контекстно-свободные грамматики.* Контекстно-свободной грамматикой называется четвёрка $G = \langle N, \Sigma, P, S \rangle$, где Σ — алфавит, над которым грамматика задаёт язык, N — вспомогательный алфавит, элементы которого называются *нетерминальными символами* (требуется,

чтобы N и Σ не пересекались), $S \in N$ — *стартовый символ*, а P — множество *правил* вида $A \Rightarrow \alpha$, где $A \in N$, а α — слово в алфавите $(N \cup \Sigma)$. Далее мы будем считать, что в любом правиле α непусто. Символ A и слово α называются, соответственно, *левой* и *правой частями* правила $(A \Rightarrow \alpha)$. Пусть η и θ — произвольные (возможно, пустые) слова в алфавите $(N \cup \Sigma)$. Тогда, если $(A \Rightarrow \alpha) \in P$, слово $\eta\alpha\theta$ *непосредственно выводится* из слова $\eta A\theta$ в грамматике G; пишем: $\eta A\theta \Rightarrow_G \eta\alpha\theta$. Отношение $\Rightarrow_G^*$ ("выводится в грамматике G") определяется как рефлексивно-транзитивное замыкание отношения $\Rightarrow_G$. *Язык, задаваемый грамматикой G,* есть множество $\{w \in \Sigma^+ \mid S \Rightarrow_G^* w\}$. Такие языки называются *контекстно-свободными* (напомним, что мы рассматриваем только языки, не содержащие пустого слова).

Если G — грамматика (контекстно-свободная или грамматика Ламбека), то задаваемый ей язык обозначим через $\mathcal{L}(G)$.

Теорема 1. *Всякий язык, порождаемый грамматикой Ламбека, является контекстно-свободным. Всякий контекстно-свободный язык порождается некоторой грамматикой Ламбека.*

Первое утверждение этой теоремы доказано у Пентуса [5]; для случая, когда используется только одна операция деления, этот результат был доказан ранее у Бушковского [14]. Второе утверждение для более слабого формализма, называемого базовыми категориальными грамматиками, или грамматиками Айдукевича – Бар-Хиллела, доказал Гайфман [13]; Бушковский [14] отметил, что конструкция Гайфмана пригодна и для грамматик Ламбека. В этой конструкции используется только одна операция деления ($\backslash$ или, симметрично, $/$).

Теорема 1 утверждает эквивалентность грамматик Ламбека и контекстно-свободных грамматик *в слабом смысле.* Это означает, что классы языков как множеств слов, задаваемых этими двумя грамматическими формализмами, совпадают. На самом деле, грамматики (и контекстно-свободные, и грамматики Ламбека) способны не только проверять принадлежность слова языку, но также сообщать словам, принадлежащим языку, некоторую дополнительную структуру, кодирующую *семантику* ("смысл") данного слова. Если и эта дополнительная структура сохраняется при преобразовании грамматик из одного класса в другой, говорят об *эквивалентности в сильном смысле* [18][2]. Мы же рассматриваем здесь только эквивалентность в слабом смысле.

Помимо самого исчисления Ламбека L, рассматриваются также его расширения и фрагменты. Расширения исчисления Ламбека важны для лингвистических приложений и весьма разнообразны [17][19][16][20]. Поскольку связки исчисления Ламбека естественным образом интерпретируются как операции на формальных языках [6] (умножение и деления), естественно пытаться построить расширения исчисления Ламбека, охватывающие также другие подобные операции. Интересной открытой проблемой здесь является построение исчисления Ламбека, расширенного итерацией ("звёздочкой Клини"). Несмотря на кажущуюся простоту и естественность, для этого расширения не известно "хорошей" аксиоматизации; однако существуют полные системы с ω-правилом или бесконечными выводами [3]. Интересно, что аналогичные системы успешно "сворачиваются" в системы с конечными (циклическими) выводами в случае модальной логики Гёделя – Лёба GL и ее расширений [12][27]. Возможно, вариант такой стратегии может сработать и для исчисления Ламбека с итерацией.

Что же касается фрагментов L, то теорема об устранении сечения делает их аксиоматизацию очень легкой: достаточно оставить из правил те, которые относятся к выбранным связкам.

В этой статье нас будет интересовать фрагмент исчисления Ламбека только с одним делением, $\mathrm{L}(\backslash)$. В отличие от полного исчисления Ламбека (L) и его фрагментов с двумя операциями ($\mathrm{L}(\backslash, \cdot)$, $\mathrm{L}(/, \cdot)$, $\mathrm{L}(/, \backslash)$), проблемы выводимости в которых NP-полны [22][26][9], проблема выводимости в $\mathrm{L}(\backslash)$ разрешима за полиномиальное время [8]. С другой стороны, всякий контекстно-свободный язык (см. выше) может быть задан $\mathrm{L}(\backslash)$-грамматикой. Иначе говоря, для моделирования контекстно-свободного вывода достаточно только одной операции исчисления Ламбека, а именно одного из двух делений.

2 Критерий Саватеева выводимости секвенций в $\mathrm{L}(\backslash)$

В этом разделе формулируется комбинаторный критерий выводимости секвенций в исчислении $\mathrm{L}(\backslash)$, доказанный у Саватеева [25]. Этот критерий понадобится нам для дальнейших построений.

Пусть $\mathrm{Atn} = \{p^{(i)} \mid p \in \mathrm{Pr}, i \in \mathbb{N}\}$ — множество *атомов* (примитив-

ных типов с натуральными верхними индексами). Нас будут интересовать конечные непустые последовательности (цепочки) атомов; множество таких цепочек обозначается Atn^+.

Если $\mathbb{A} = p_1^{(i_1)} p_2^{(i_2)} \dots p_k^{(i_k)} \in \mathrm{Atn}^+$, положим $\mathbb{A}^{+2} = p_1^{(i_1+2)} p_2^{(i_2+2)} \dots p_k^{(i_k+2)}$.

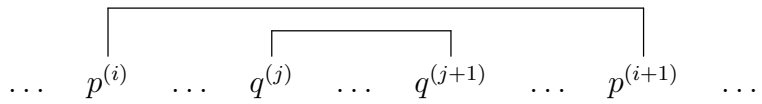
Определим два отображения $\gamma, \bar{\gamma} \colon \mathrm{Tp} \to \mathrm{Atn}^+$ типов исчисления $\mathrm{L}(\backslash)$ в цепочки атомов:

$$\begin{array}{ll} \gamma(p) = p^{(1)} & \bar{\gamma}(p) = p^{(2)} \\ \gamma(A \setminus B) = \bar{\gamma}(A)\gamma(B) & \bar{\gamma}(A \setminus B) = \bar{\gamma}(B)\big(\gamma(A)\big)^{+2} \end{array}$$

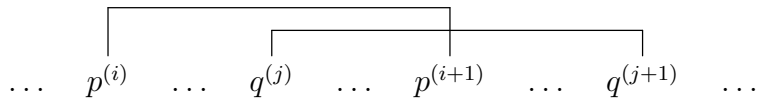
Пусть $\mathbb{A}$ — цепочка атомов. *Сетью доказательства* на $\mathbb{A}$ называется такое разбиение элементов $\mathbb{A}$ на пары, что

1. каждый элемент входит ровно в одну пару;
2. каждая пара состоит из $p^{(i)}$ и $p^{(i+1)}$ для некоторых $p \in \mathrm{Pr}$, $i \in \mathbb{N}$, причём $p^{(i)}$ стоит левее, чем $p^{(i+1)}$;
3. линии, соединяющие атомы в одной паре, можно нарисовать в верхней полуплоскости без пересечений; иначе говоря, две пары могут располагаться так:

или так:



но не так:



4. если левый атом в паре имеет чётный индекс, то между атомами этой пары найдётся атом с индексом меньшим, чем у атомов этой пары.

Заметим, что в другой работе Саватеева [8] используется немного иной критерий: требуется, чтобы между атомами в паре, где левый атом имеет чётный индекс 2ℓ, был атом с индексом *ровно* $2\ell - 1$. Эти два критерия эквивалентны.

Теорема 2. *Секвенция $A_1, \ldots, A_n \to B$ выводима в* L(\) *тогда и только тогда, когда на цепочке $\gamma(A_1) \ldots \gamma(A_n)\bar{\gamma}(B)$ существует сеть доказательства.* [25]

3 Рост размера грамматики Ламбека при преобразовании в контекстно-свободную грамматику

По теореме 1, между двумя грамматическими формализмами (грамматиками Ламбека и контекстно-свободными грамматиками) существуют преобразования в обе стороны, поддерживающие слабую эквивалентность формализмов. Однако в подобных случаях (см., например, [24]) интерес представляет не только существование такого преобразования, но также *изменение размера* объекта (в данном случае — грамматики) после применения к нему такого преобразования. Если объект при переходе в другой формализм сильно (экспоненциально) разрастается, то такое преобразование теряет практическую применимость.

К сожалению, преобразование грамматик Ламбека в контекстно-свободные, предложенное у Пентуса [5], приводит к экспоненциальному росту размера грамматики. По-видимому, в общем случае такая неэффективность неизбежна, поскольку проблема выводимости в исчислении L NP-полна (см. выше), в то время как задача проверки, допускается ли слово данной контекстно-свободной грамматикой, решается за полиномиальное время.

Для случая с одним делением еще один метод преобразования L(\)-грамматики в контекстно-свободную предлагался ранее у Бушковского [14], однако при этом преобразовании грамматика также экспоненциально разрастается. Мы же предъявим способ преобразования L(\)-грамматики в контекстно-свободную грамматику, размер которой ограничен полиномом от размера исходной грамматики.

Прежде всего уточним понятие *размера* для грамматик Ламбека и контекстно-свободных грамматик.

Пусть $A \in \mathrm{Tp}$. Определим размер типа A как количество вхождений примитивных типов в A; обозначение: $|A|$. Формально размер типа A определяется рекурсивным образом: $|p_i| = 1$; $|A \setminus B| = |B \,/\, A| = |A \cdot B| = |A| + |B|$. Размером грамматики Ламбека $G = \langle \Sigma, \rhd, H \rangle$ назовем величину $|H| + \sum\limits_{\langle a, A \rangle \in \rhd} |A|$.

Размером контекстно-свободной грамматики $G = \langle N, \Sigma, P, S \rangle$ назовем величину $\sum\limits_{(A \Rightarrow \alpha) \in P} |\alpha|$, где $|\alpha|$ — количество символов в слове α. Заметим, что $|\Sigma| \leqslant |G|$ (если считать, что в Σ нет лишних символов, никогда не появляющихся в $\mathcal{L}(G)$), $|N| \leqslant |G|$ и $|P| \leqslant |G|$.

Размер грамматики (как контекстно-свободной, так и грамматики Ламбека) G обозначим $|G|$.

Теорема 3. *Для любой* $\mathrm{L}(\setminus)$*-грамматики существует эквивалентная ей (в слабом смысле) контекстно-свободная грамматика, размер которой ограничен полиномом от размера исходной грамматики.*

4 Условия критерия Саватеева как контекстно-свободные правила

Основная идея предъявляемого нами преобразования грамматик — записать условия существования сети доказательства в виде контекстно-свободной грамматики. Однако фигурирующие в определении сети доказательства цепочки атомов являются словами в бесконечном алфавите Atn, поэтому для данного множества таких цепочек, формально говоря, невозможно составить контекстно-свободную грамматики. Чтобы обойти это затруднение, сузим множество атомов до конечного.

Пусть дана $\mathrm{L}(\setminus)$-грамматика G. Сначала удалим из множества Pr все примитивные типы, которые не встречаются в G. После этого множество Pr станет конечным; более того, его мощность не превосходит $|G|$.

Теперь определим индуктивно понятие *глубины* синтаксического типа из $\mathrm{Tp}(\setminus)$: $d(p_i) = 1$; $d(A \setminus B) = \max\{d(A) + 1, d(B)\}$. С другой стороны, обозначим через Atn_m ($m \in \mathbb{N}$) ограниченное множество атомов $\{p^{(i)} \mid p \in \mathrm{Pr}, i \leqslant m\}$.

Лемма 1. *Если* $A \in \mathrm{Tp}(\setminus)$*, то* $\gamma(A) \in \mathrm{Atn}^+_{d(A)}$ *и* $\bar{\gamma}(A) \in \mathrm{Atn}^+_{d(A)+1}$.

Доказательство. Индукция по построению A. Если $A = p_i$, то $d(A) = 1$, $\gamma(A) = p_i^{(1)} \in \mathrm{Atn}_1^+$ и $\bar{\gamma}(A) = p_i^{(2)} \in \mathrm{Atn}_2^+$.

Для случая $A = B \setminus C$ сначала отметим некоторые очевидные свойства множеств Atn_m^+. Во-первых, если $\mathbb{B} \in \mathrm{Atn}_{m_1}^+$ и $\mathbb{C} \in \mathrm{Atn}_{m_2}^+$, то $\mathbb{BC} \in \mathrm{Atn}_{\max\{m_1,m_2\}}^+$. Во-вторых, если $\mathbb{B} \in \mathrm{Atn}_m^+$, то $\mathbb{B}^{+2} \in \mathrm{Atn}_{m+2}^+$.

Далее,

$$\gamma(B \setminus C) = \bar{\gamma}(B)\gamma(C) \in \mathrm{Atn}_{\max\{d(B)+1,d(C)\}}^+ = \mathrm{Atn}_{d(B \setminus C)}^+;$$

$$\bar{\gamma}(B \setminus C) = \bar{\gamma}(C)\big(\gamma(B)\big)^{+2} \in \mathrm{Atn}_{\max\{d(C)+1,d(B)+2\}}^+ = \mathrm{Atn}_{d(B \setminus C)+1}^+.$$

□

Назовем *глубиной* L(\)-грамматики $G = \langle \Sigma, \rhd, H \rangle$ величину $d(G) = \max\{d(A) \mid a \rhd A$ для некоторой $a \in \Sigma$ или $A = H\}$. Заметим, что, поскольку $d(A) \leqslant |A|$ для любого типа A (это легко проверяется по индукции), для любой L(\)-грамматики G имеем $d(G) \leqslant |G|$. [2]

Положим $m = d(G) + 1$. Тогда каждой секвенции, используемой для проверки принадлежности некоторого слова языку, задаваемому грамматикой G, соответствует цепочка атомов из Atn_m^+. С другой стороны, $|\mathrm{Atn}_m| = |\mathrm{Pr}| \cdot m \leqslant |G|(d(G) + 1) \leqslant |G|(|G| + 1)$, т.е. рассматриваемые цепочки атомов суть слова в алфавите, мощность которого ограничена полиномом от размера исходной грамматики.

Теперь определим контекстно-свободную грамматику $\mathbf{S}_m$, формализующую существование сети доказательства для данного слова в алфавите Atn_m: $\mathbf{S}_m = \langle N, \mathrm{Atn}_m, P, S \rangle$, где $N = \{S, R_1, \ldots, R_{m-1}\}$, а P состоит из следующих правил:

$S \Rightarrow R_k$	для всех k от 1 до $m-1$;
$R_{k_1} \Rightarrow R_{k_2}$,	если $k_1 > k_2$;
$R_k \Rightarrow p^{(2\ell-1)} R_k p^{(2\ell)}$,	если $k < m$ и $2\ell < m$;
$R_k \Rightarrow p^{(2\ell)} R_k p^{(2\ell+1)}$,	если $2\ell + 1 < m$ и $k < 2\ell$;
$R_{2\ell-1} \Rightarrow p^{(2\ell-1)} S p^{(2\ell)}$,	если $2\ell < m$;
$R_{2\ell-1} \Rightarrow p^{(2\ell-1)} p^{(2\ell)}$,	если $2\ell < m$;
$R_k \Rightarrow R_k S$	для всех k от 1 до $m-1$;
$R_k \Rightarrow S R_k$	для всех k от 1 до $m-1$.

[2] В то же время, если в грамматике G содержится "очень глубокий" сложный синтаксический тип, величина $d(G)$ может быть по порядку близка к $|G|$.

Общее число правил не превосходит $m + m^2 + 2 \cdot |\mathrm{Pr}| \cdot m^2 + 2 \cdot |\mathrm{Pr}| \cdot m + 2m \leqslant 8 \cdot |\mathrm{Pr}| \cdot m^2 \leqslant 8|G|^3$. Правая часть каждого правила имеет длину не более 3. Значит, $|\mathbf{S}_m| \leqslant 24|G|^3$, т.е. размер грамматики $\mathbf{S}_m$ ограничен полиномом от размера исходной грамматики G.

Лемма 2. *Цепочка атомов $\mathbb{A} \in \mathrm{Atn}_m^+$ обладает сетью доказательства тогда и только тогда, когда она принадлежит языку, задаваемому грамматикой $\mathbf{S}_m$.*

Доказательство. Чтобы доказать эту лемму индукцией, дополним ее утверждение ($S \Rightarrow_{\mathbf{S}_m}^* \mathbb{A}$ тогда и только тогда, когда $\mathbb{A}$ обладает сетью доказательства) аналогичными утверждениями об остальных нетерминальных символах: $R_k \Rightarrow_{\mathbf{S}_m}^* \mathbb{B}$ тогда и только тогда, когда $\mathbb{B}$ обладает сетью доказательства и содержит атом с индексом, не превосходящим k.

Теперь эти утверждения легко доказываются совместной индукцией: в части "тогда" — по длине вывода в $\mathbf{S}_m$, в части "только тогда" — по длине рассматриваемой цепочки атомов. □

Теперь опишем способ, позволяющий получить из $\mathbf{S}_m$ контекстно-свободную грамматику, в слабом смысле эквивалентную грамматике G. Для этого нам потребуется понятие *гомоморфизма* формальных языков.

Пусть Σ_1 и Σ_2 — два алфавита. Гомоморфизмом называется функция $h\colon \Sigma_1^* \to \Sigma_2^*$, такая что $h(uv) = h(u)h(v)$ для любых двух слов $u, v \in \Sigma_1^*$. Ясно, что гомоморфизм можно задать произвольным образом на элементах Σ_1, а дальше он однозначно распространяется на более длинные слова из Σ_1^*; $h(\varepsilon)$ всегда равно ε (иначе нарушается равенство $h(a) = h(a\varepsilon) = h(a)h(\varepsilon)$).

Частным случаем гомоморфизма является *неудлиняющий гомоморфизм,* при котором для любой буквы $a \in \Sigma_1$ ее образ $h(a)$ есть либо пустое слово ε, либо буква алфавита Σ_2 (но не более длинное слово).

Всюду далее алфавит Atn_m будем обозначать Σ_1.

Напомним, что мы строим контекстно-свободную грамматику для языка, задаваемого L(\)-грамматикой $G = \langle \Sigma, \rhd, H \rangle$. Введем вспомогательный алфавит $\Sigma_2 = \{\langle a, A\rangle \mid a \rhd A\}$. Введем также новый символ \$, не принадлежащий введенным ранее алфавитам. Определим два гомоморфизма $g\colon \Sigma_2 \cup \{\$\} \to \Sigma_1$ и $h\colon \Sigma_2 \cup \{\$\} \to \Sigma$ следующим образом:

$$\begin{array}{ll} g(\langle a, A\rangle) = \gamma(A); & h(\langle a, A\rangle) = a; \\ g(\$) = \bar{\gamma}(H); & h(\$) = \varepsilon. \end{array}$$

Легко видеть, что если грамматика $\mathbf{S}_m$ задает язык M, то язык, задаваемый грамматикой G, есть $h\big(g^{-1}(M)\cap\{u\$ \mid u\in\Sigma_2^+\}\big)$. Для этого языка можно построить контекстно-свободную грамматику, поскольку операции взятия полного прообраза при гомоморфизме, пересечения с регулярным (автоматным) языком и применения гомоморфизма не выводят за пределы класса контекстно-свободных языков. Эти факты были независимо замечены несколькими авторами [15][21][28] и теперь входят в стандартные учебники теории формальных языков [1][11][7]. Однако нам придется аккуратно проанализировать их доказательства, чтобы получить оценку на размер полученной таким способом контекстно-свободной грамматики.

5 Оценка размера контекстно-свободной грамматики

Для применения "обратного гомоморфизма" (перехода от M к $g^{-1}(M)$) нам потребуется вспомогательное понятие *автомата с магазинной памятью* (коротко — *"МП-автомат"*). Мы будем пользоваться следующим вариантом определения МП-автомата: МП-автоматом называется шестерка $\mathfrak{M} = \langle Q, \Sigma, \Gamma, \Delta, q_0, Z_0\rangle$, где Q, Σ, Γ — конечные множества, $q_0 \in Q$, $Z_0 \in \Gamma$, а Δ — конечное подмножество декартова прозведения $Q \times (\Sigma \cup \{\varepsilon\}) \times \Gamma \times Q \times \Gamma^*$. Множество Q называется *множеством состояний,* Σ и Γ — *входной* и *магазинный алфавиты* соответственно, q_0 — *стартовое состояние,* Z_0 — *стартовый магазинный символ,* а Δ называется *множеством переходов.* Элемент $\langle p, a, Z, q, \beta\rangle \in \Delta$ (*переход* автомата) будем записывать так: $p \xrightarrow{a,\ Z:\beta} q$.

Конфигурацией МП-автомата называется тройка $\langle q, v, \gamma\rangle$, где $q \in Q$, $v \in \Sigma^*$, $\gamma \in \Gamma^*$. Неформально эта конфигурация означает следующее: автомат находится в состоянии q, на входе осталось непрочитанным слово v (окончание того слова, которое подавалось на вход изначально), а магазинная память содержит слово γ.

Изначальная конфигурация МП-автомата имеет вид $\langle q_0, w, Z_0\rangle$, где w — то слово, на котором мы хотим запустить автомат.

Переход $p \xrightarrow{a,\ Z:\beta} q$ изменяет конфигурацию автомата следующим образом: $\langle p, aw, Z\gamma\rangle \to_{\mathfrak{M}} \langle q, w, \beta\gamma\rangle$. Неформально говоря, при этом перехо-

де, если автомат находится в состоянии p и на вершине магазина[3] находится символ Z, из входного потока считывается один символ a (или ничего не считывается, если $a = \varepsilon$), символ Z снимается с вершины магазина, на вершину помещается слово β, и автомат переходит в состояние q. Заметим, что в этом определении при каждом переходе с вершины магазина необходимо снять ровно один символ (при этом, разумеется, его можно тотчас же положить назад, добавив к β). Как обычно, отношение $\to^*_{\mathfrak{M}}$ есть рефлексивно-транзитивное замыкание отношения $\to_{\mathfrak{M}}$.

Наконец, слово w *допускается* МП-автоматом $\mathfrak{M}$, или принадлежит *языку, задаваемому* МП-автоматом $\mathfrak{M}$, если $\langle q_0, w, Z_0\rangle \to^*_{\mathfrak{M}} \langle q, \varepsilon, \varepsilon\rangle$ для некоторого[4] $q \in Q$.

Заметим, что $\mathfrak{M}$, вообще говоря, работает *недетерминированно:* в некоторых конфигурациях возможно применить не один, а несколько различных переходов. Слово допускается автоматом, если хотя бы одна из цепочек переходов приводит к успеху (т.е. не обрывается и заканчивается в конфигурации с пустым магазином).

Дальнейший план построения контекстно-свободной грамматики для языка $\mathcal{L}(G)$ таков: сначала по грамматике $\mathbf{S}_m$ построим МП-автомат, задающий тот же язык $M = \mathcal{L}(\mathbf{S}_m)$. Далее, построим МП-автомат для языка $g^{-1}(M) \cap \{u\$ \mid u \in \Sigma_2^+\}$. После этого перейдем обратно к контекстно-свободным грамматикам, построив таковую для последнего. Наконец, применив к этой грамматике неудлиняющий гомоморфизм h, получим контекстно-свободную грамматику для языка $\mathcal{L}(G) = h\big(g^{-1}(M) \cap \{u\$ \mid u \in \Sigma_2^+\}\big)$.

Конструкции автоматов и грамматик, используемые при этих преобразованиях, взяты из [1] и [11]. В этих книгах изложены доказательства корректности этих конструкций (т.е. утверждений о том, что построенные автоматы и грамматики задают правильные языки). Мы же сосредоточимся на оценке размера полученной контекстно-свободной грамматики.

Сложность (размер) МП-автомата $\mathfrak{M} = \langle Q, \Sigma, \Gamma, \Delta, q_0, Z_0\rangle$ будем оценивать следующими параметрами: во-первых, числом переходов ($|\Delta| = \delta(\mathfrak{M})$); во-вторых, величиной $d(\mathfrak{M}) = \max\big\{|\beta| \mid \big(p \xrightarrow{a,\ Z:\beta} q\big) \in \Delta\big\}$, характеризующую наибольшую длину слова, помещаемого в магазин при

[3]Мы считаем, что магазин "растет" влево.

[4]Как видно, в этом определении критерий остановки МП-автомата — это опустошение магазина, а не переход в специальное "завершающее" состояние.

переходе автомата; в-третьих, числом состояний ($|Q| = q(\mathfrak{M})$).

Лемма 3. *Существует МП-автомат $\mathfrak{M}_1$, задающий язык $M = \mathcal{L}(\mathbf{S}_m)$* ([1], лемма 2.24). *При этом $\delta(\mathfrak{M}_1) \leqslant 2|\mathbf{S}_m|$, $q(\mathfrak{M}_1) = 1$, а $d(\mathfrak{M}_1) = 3$.*

Доказательство. Обозначим через N множество нетерминальных символов грамматики $\mathbf{S}_m$.

Автомат $\mathfrak{M}_1$ содержит всего одно состояние q_0 и имеет вид $\langle\{q_0\}, \Sigma_1, N \cup \Sigma_1, \Delta_1, q_0, S\rangle$. Здесь в качестве магазинного алфавита взято объединение основного и вспомогательного алфавитов грамматики $\mathbf{S}_m$, а стартовый магазинный символ — это стартовый символ грамматики.

Множество переходов Δ_1 строится следующим образом:

1. для каждого правила $A \Rightarrow \alpha$ грамматики $\mathbf{S}_m$ добавим переход $q_0 \xrightarrow{\varepsilon,\ A:\alpha} q_0$;
2. для каждого $a \in \Sigma_1$ добавим переход $q_0 \xrightarrow{a,\ a:\varepsilon} q_0$.

Число переходов первого типа равно числу правил в грамматике $\mathbf{S}_m$; число переходов второго типа равно $|\Sigma_1|$. Обе эти величины не превосходят $|\mathbf{S}_m|$. Значит, $\delta(\mathfrak{M}_1) \leqslant 2|\mathbf{S}_m|$.

Равенство $d(\mathfrak{M}_1) = 3$ следует из того, что правые части всех правил в $\mathbf{S}_m$ содержат не более трех символов. □

Лемма 4. *Пусть $\mathfrak{M}_1$ — МП-автомат для языка M, построенный в предыдущей лемме. Пусть также $g\colon \Sigma_2 \cup \{\$\} \to \Sigma_1$ — гомоморфизм, причем для всех $a \in \Sigma_2$ выполнено неравенство $|g(a)| \leqslant n$. Тогда существует МП-автомат $\mathfrak{M}_2$, задающий язык $g^{-1}(M) \cap \{u\$ \mid \Sigma_2^+\}$* ([11], теорема 7.30 и [1], лемма 2.22 и теорема 2.26). *При этом $q(\mathfrak{M}_2) \leqslant 2(n+1) \cdot |\Sigma_2| + 2$, $\delta(\mathfrak{M}_2) \leqslant (2n+6) \cdot |\mathbf{S}_m| \cdot |\Sigma_2| + 2|\mathbf{S}_m| + 2|\Sigma_2| + 2$ и $d(\mathfrak{M}_2) = 3$.*

Доказательство. Построим МП-автомат $\mathfrak{M}_2 = \langle Q_2, \Sigma_2 \cup \{\$\}, \Gamma_2, \Delta_2, q_0, S\rangle$ следующим образом. Положим $\Gamma_2 = N \cup \Sigma_1 \cup \{\#\}$. Новый символ $\#$ будет играть роль *маркера конца магазина,* запрещая автомату завершать работу раньше времени, даже если магазин пуст и все слово уже прочитано.

Множество Q_2 составим из пар вида $\langle i, x\rangle$, где $i \in \{0, 1\}$, а $x \in \Sigma_1^*$ — окончание слова $g(a)$ для некоторой буквы $a \in \Sigma_2$ (в частности x может быть пустым; состояния $\langle 0, \varepsilon\rangle$ и $\langle 1, \varepsilon\rangle$ будут играть в $\mathfrak{M}_2$ особую роль), а также двух особых состояний q_0 (стартового) и q_F (заключительного).

Наконец, переходы автомата $\mathfrak{M}_2$ (элементы Δ_2) суть следующие:

1. $q_0 \xrightarrow{\varepsilon,\ S:S\#} \langle 0, x\rangle$;

2. $\langle 0, \varepsilon\rangle \xrightarrow{a,\ X:X} \langle 0, g(a)\rangle$ для всех $X \in \Gamma_2$ и $a \in \Sigma_2$;

3. $\langle 0, \varepsilon\rangle \xrightarrow{a,\ X:X} \langle 1, g(a)\rangle$ для всех $X \in \Gamma_2$ и $a \in \Sigma_2$;

4. $\langle 1, \varepsilon\rangle \xrightarrow{\$,\ X:X} \langle 1, g(\$)\rangle$ для всех $X \in \Gamma_2$;

5. $\langle i, bv\rangle \xrightarrow{\varepsilon,\ X:\alpha} \langle i, v\rangle$, если в автомате $\mathfrak{M}_1$ есть переход $q_0 \xrightarrow{b,\ X:\alpha} q_0$;

6. $\langle 1, \varepsilon\rangle \xrightarrow{\varepsilon,\ \#:\varepsilon} q_F$.

Неформальный смысл $\mathfrak{M}_2$ следующий. Первым переходом он помещает специальный символ # на дно магазина. Поскольку извлечь его оттуда (и не положить тотчас же назад) можно только шестым переходом, автомат может закончить работу только в состоянии q_F. Далее с помощью перехода 2 автомат считывает символ a и сохраняет во "внутренней памяти" (вторая компонента состояния автомата) слово $g(a)$, на котором несколькими применениями перехода 5 имитируется работа $\mathfrak{M}_1$. После этого "внутренняя память" опять пуста, и $\mathfrak{M}_2$ готов считать следующую букву входного слова. Предпоследняя буква считывается уже переходом 3 с переменой первой компоненты состояния с 0 на 1. Наличие этого перехода гарантирует, что слово непусто и не состоит из одного лишь символа \$. После этого опять несколько раз применяется переход 5. Наконец, переход 4 и опять же несколько переходов 5 делают то же самое для символа \$. Тем самым гарантируется, что этот символ в слове ровно один и расположен в конце (т.е. условие пересечения с языком $\{u\$ \mid u \in \Sigma_2^+\}$).

Корректность этой конструкции легко следует из доказательств утверждений из [1] и [11], указанных в формулировке леммы.

Легко видеть, что $d(\mathfrak{M}_2) = d(\mathfrak{M}_1) = 3$. Оценим $\delta(\mathfrak{M}_2) = |\Delta_2|$. Переходов типа 2 и 3 имеется по $|\Gamma_2| \cdot |\Sigma_2|$. Переходов типа 4 — $|\Gamma_2|$. По условию, длина каждого слова $g(a)$ ($a \in \Sigma_2$) не превосходит n, поэтому у него не более $(n+1)$ окончаний (от пустого до всего слова). Всего таких окончаний не больше, чем $(n+1) \cdot |\Sigma_2|$. Каждый переход типа 5 задается таким окончанием и переходом автомата $\mathfrak{M}_1$, значит, их число не превосходит $(n+1) \cdot |\Sigma_2| \cdot \delta(\mathfrak{M}_1)$. Наконец, переходы типа 1 и 6 присутствуют в единственном экземпляре, всего 2.

Следовательно, общее число переходов в $\mathfrak{M}_2$ не превосходит $2|\Gamma_2| \cdot |\Sigma_2| + |\Gamma_2| + (n+1) \cdot |\Sigma_2| \cdot \delta(\mathfrak{M}_1) + 2$. Вспоминая, что $|\Gamma_2| = |N| + |\Sigma_1| + 1 \leqslant 2|\mathbf{S}_m| + 1$ и что $\delta(\mathfrak{M}_1) \leqslant 2|\mathbf{S}_m|$, получаем требуемую оценку $(2n+6) \cdot |\mathbf{S}_m| \cdot |\Sigma_2| + 2|\mathbf{S}_m| + 2|\Sigma_2| + 2$.

Наконец, оценка $q(\mathfrak{M}_2) = |Q_2| \leqslant 2(n+1) \cdot |\Sigma_2| + 2$ также следует из оценки числа окончаний слов вида $g(a)$. □

Лемма 5. *Пусть $\mathfrak{M}_2 = \langle Q_2, \Sigma_2 \cup \{\$\}, \Gamma_2, \Delta_2, q_0, S\rangle$ — МП-автомат, задающий язык над алфавитом $\Sigma_2 \cup \{\$\}$. Тогда существует контекстно-свободная грамматика G_2, задающая тот же язык* ([1], лемма 2.26). *При этом $|G_2| \leqslant (d(\mathfrak{M}_2) + 1) \cdot \delta(\mathfrak{M}_2) \cdot \big(q(\mathfrak{M}_2)\big)^{d(\mathfrak{M}_2)} + q(\mathfrak{M}_2)$.*

Доказательство. Искомая грамматика $G_2 = \langle N_2, \Sigma_2, P_2, S_2\rangle$ строится следующим образом.

Нетерминальные символы этой грамматики суть тройки $\langle q, Z, r\rangle$, где $q, r \in Q_2$ и $Z \in \Gamma_2$, а также особый символ S_2. Тройку $\langle q, Z, r\rangle$, следуя [1], будем обозначать $[qZr]$ (эта запись означает *один* символ из N_2).

Правила грамматики G_2 (элементы P_2) следующие:

1. $[qZr] \Rightarrow a[rX_1s_1][s_1X_2s_2]\ldots[s_{k-1}X_ks_k]$,
   если в $\mathfrak{M}_2$ есть переход $q \xrightarrow{a,\ Z:X_1\ldots X_k} r$; $a \in \Sigma_2 \cup \{\$, \varepsilon\}$, а s_1, …, s_k — произвольные состояния из Q_2;
   в частности, если переход имеет вид $q \xrightarrow{a,\ Z:\varepsilon} r$, добавляется правило $[qZr] \Rightarrow a$;

2. $S_2 \Rightarrow [q_0Sq]$ для всех $q \in Q_2$.

В каждом правиле типа 1 $k \leqslant d(\mathfrak{M}_2)$. Поскольку каждое такое правило задается переходом МП-автомата $\mathfrak{M}_2$ и набором $s_1, \ldots, s_k \in Q_2$, всего таких правил не больше, чем $\delta(\mathfrak{M}_2) \cdot \big(q(\mathfrak{M}_2)\big)^{d(\mathfrak{M}_2)}$, а суммарная длина их правых частей не превосходит $(d(\mathfrak{M}_2) + 1) \cdot \delta(\mathfrak{M}_2) \cdot \big(q(\mathfrak{M}_2)\big)^{d(\mathfrak{M}_2)}$.

Число правил типа 2 равно $|Q_2| = q(\mathfrak{M}_2)$, а правая часть каждого такого правила однобуквенна. Отсюда получаем требуемую оценку величины $|G_2|$. □

Упростим оценку $|G_2|$. По построению, $|\Sigma_2| \leqslant |G|$, а также для любого $a \in \Sigma_2$ имеет место неравенство $|g(a)| \leqslant |G|$, где G — исходная L(\)-грамматика. Кроме того, $|\mathbf{S}_m| \leqslant 24|G|^3$. Положим $n = |G|$. По лемме 4,

$d(\mathfrak{M}_2) = 3$, $q(\mathfrak{M}_2) \leqslant 2(n+1) \cdot n + 2 = 2n^2 + 2n + 2$ и $\delta(\mathfrak{M}_2) \leqslant (2n+6) \cdot 24n^3 \cdot n + 2 \cdot 24n^3 + 2n + 2 = 48n^5 + 144n^4 + 48n^3 + 2n + 2$.

Значит, $|G_2| \leqslant 4 \cdot (48n^5 + 144n^4 + 48n^3 + 2n + 2) \cdot (2n^2 + 2n + 2)^3 + 2n^2 + 2n + 2 = O(n^{11})$.

Наконец, применение неудлиняющего гомоморфизма h к правым частям грамматики G_2 не увеличивает ее размера и дает контекстно-свободную грамматику размера $O(|G|^{11})$ для языка $h(g^{-1}(M) \cap \{u\$ \mid u \in \Sigma_2^+\} = \mathcal{L}(G)$. Это завершает доказательство теоремы 3.

6 Заключение

Несмотря на то, что полное исчисление Ламбека L NP-полно, и поэтому теорема 3 вряд ли может быть обобщена на все грамматики Ламбека, для грамматик, все типы в которых имеют ограниченную константой глубину, существует полиномиальный ($O(n^4)$) алгоритм проверки принадлежности слова языку, порождаемому такой грамматикой [23]. Поэтому вопрос о преобразовании грамматики Ламбека с ограниченной глубиной типов в контекстно-свободную грамматику полиномиального размера представляет интерес для будущих исследований.

Полиномиальный ($O(n^3)$) алгоритм проверки выводимости секвенций в исчислении L($\backslash$) [8], по сути, состоит в применении стандартного метода динамического программирования (алгоритм Кока – Янгера – Касами, CYK) к грамматике $\mathbf{S}_m$. К сожалению, эта грамматика неоднозначна (некоторые цепочки в ней имеют несколько деревьев разбора), поэтому большинство более эффективных алгоритмов синтаксического анализа к ней неприменимы либо работают медленнее, чем CYK. Кроме того, размер грамматики $\mathbf{S}_m$ зависит от сложности исходной секвенции, поэтому даже универсальный алгоритм Валианта [29] работает на ней медленнее, чем CYK. Тем не менее, интерпретация критерия Саватеева в терминах контекстно-свободных правил может оказаться полезной при попытках построить более эффективный алгоритм проверки выводимости в L($\backslash$).

Благодарности

Автор благодарен профессорам Г. Морриллу (Барселона) и М.Р. Пентусу (Москва) за ценные обсуждения.

Список литературы

[1] *Ахо А., Ульман Дж.* Теория синтаксического анализа, перевода и компиляции. Т. 1. Синтаксический анализ: Пер. с англ. — М.: Мир, 1978.

[2] *Кузнецов С.Л.* О преобразовании контекстно-свободных грамматик в грамматики Ламбека // Тр. МИАН. 2015. Т. 290. С. 72–79.

[3] *Кузнецов С.Л., Рыжкова Н.С.* Фрагмент исчисления Ламбека с итерацией // Мальцевские чтения 2015. Тезисы докладов. Новосибирск: ИМ СО РАН, НГУ, 2015.

[4] *Ламбек И.* Математическое исследование структуры предложений. Пер. с англ. // Математическая лингвистика: сб. пер. / под ред. Ю.А. Шрейдера и др. — М.: Мир, 1964. — С. 47–68.

[5] *Пентус М.Р.* Исчисление Ламбека и формальные грамматики // Фунд. и прикл. математика. 1995. Т. 1, № 3. С. 729–751.

[6] *Пентус М.Р.* Полнота синтаксического исчисления Ламбека // Фунд. и прикл. математика. 1999. Т. 5, № 1. С. 193–219.

[7] *Пентус А.Е., Пентус М.Р.* Математическая теория формальных языков: Учебное пособие. — М.: Интернет-университет информационных технологий; БИНОМ. Лаборатория знаний, 2009.

[8] *Саватеев Ю.В.* Распознавание выводимости для исчисления Ламбека с одним делением // Вестн. Московск. ун-та. Сер. 1. Матем. Механ. 2009. № 2. С. 59–62.

[9] *Саватеев Ю.В.* Алгоритмическая сложность фрагментов исчисления Ламбека: дисс. ... канд. физ.-мат. наук. М.: МГУ, 2009.

[10] *Хомский Н.* Три модели описания языка: Пер. с англ. // Кибернетический сборник, вып. 2. — М.: ИЛ, 1961. — С. 237–266.

[11] *Хопкрофт Дж., Мотвани Р., Ульман Дж.* Введение в теорию автоматов, языков и вычислений, 2-е изд.: Пер. с англ. — М.: ИД “Вильямс”, 2002.

[12] *Шамканов Д.С.* Циклические выводы для логики Гёделя-Лёба // Матем. заметки. 2014. Т. 96. Вып. 4. С. 609–622.

[13] *Bar-Hillel Y., Gaifman C., Shamir E.* On the categorial and phrase-structure grammars. Bull. Res. Council Israel, Section F. 1960. Vol. 9F. P. 1–16.

[14] *Buszkowski W.* The equivalence of unidirectional Lambek categorial grammars and context-free grammars // Z. math. Logik Grundl. Math. 1985. Bd. 31. S. 369–384.

[15] *Evey J.* Application of pushdown store machines // Proc. Fall Joint Computer Conference. Montvale, NJ: AFIPS Press, 1963. P. 215–227.

[16] *Jäger G.* On the generative capacity of multi-modal categorial grammars // Research Lang. Comput. 2003. V. 1. No. 1–2. P. 105–125.

[17] *Kanazawa M.* The Lambek calculus enriched with additional connectives // J. Log. Lang. Inform. 1992. V. 1. P. 141–171.

[18] *Kanazawa M., Salvati S.* The string-meaning relations definable by Lambek grammars and context-free grammars // Formal Grammar: Proc. 17th and 18th Intl. Confs., FG 2012/2013 / Ed. by G. Morrill, M.-J. Nederhof. Berlin: Springer, 2013. P. 191–208. (Lect. Notes Comput. Sci.; V. 8036).

[19] *Moortgat M.* Multimodal linguistic inference // J. Log. Lang. Inform. 1996. V. 5. No. 3–4. P. 349–385.

[20] *Morrill G.* Categorial grammar. Logical syntax, semantics, and processing. Oxford Univ. Press, 2011.

[21] *Oettinger A.G.* Automatic syntactic analysis and pushdown store // Proc. Symp. Appl. Math. 1961. V. 12. AMS, Providence, RI.

[22] *Pentus M.* Lambek calculus is NP-complete // Theor. Comput. Sci. 2006. V. 357. P. 186–201.

[23] *Pentus M.* A polynomial-time algorithm for Lambek grammars of bounded order // Linguistic Analysis. 2010. V. 36. No. 1–4. P. 441–471.

[24] *Podolskii V.V.* Circuit complexity meets ontology-based data access // Proc. 10th Intl. Comput. Sci. Symp. in Russia, CSR 2015. Berlin: Springer, 2015. P. 7–26. (Lect. Notes Comput. Sci.; V. 9139).

[25] *Savateev Yu.* Lambek grammars with one division are decidable in polynomial time // Computer Science — Theory and Applications / Editors E.A. Hirsch et al. Berlin: Springer, 2008. P. 273–282. (Lect. Notes Comput. Sci., V. 5010).

[26] *Savateev Yu.* Product-free Lambek calculus is NP-complete // Ann. Pure Appl. Log. 2012. V. 163. Iss. 7. P. 775–788.

[27] *Shamkanov D.* Nested sequents for provability logic GLP // Log. J. IGPL. 2015. V. 3. No. 5. P. 789–815.

[28] *Schützenberger M.-P.* On context-free languages and pushdown automata // Information and Control. 1963. V. 6. No. 3. P. 246–264.

[29] *Valiant L.G.* General context-free recognition in less than cubic time // J. Comp. Syst. Sci. 1975. V. 10. P. 308–315.