S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Том 22, № 2, стр. 1408–1420 (2025)
https://doi.org/10.33048/semi.2025.22.086

УДК 510.63, 519.765
MSC 03B25, 68Q45

О ТЕОРИИ КОНКАТЕНАЦИИ КЛАССА БЕСПРЕФИКСНЫХ ЯЗЫКОВ

Б.Н. Карлов

Представлено С.В. Судоплатовым

Abstract: This paper studies the set $\mathcal{L}_p(\Sigma)$ of prefix-free languages over some alphabet Σ. It is proved that Levi's lemma holds for such languages, and also that the theory of the algebra $\mathcal{L}_p(\Sigma)$ with concatenation is undecidable. It is established that the algebra $\mathcal{L}_p(\Sigma)$ without the empty language is elementarily equivalent to the algebra of all words over an infinite alphabet.

Keywords: prefix-free language, concatenation, Levi's lemma, undecidable theory, elementary equivalence.

1 Введение

Известно, что теория конкатенации слов неразрешима, даже если алфавит содержит только два символа. В [1] была исследована слабая теория конкатенации TC, определяемая следующим конечным множеством аксиом (здесь α и β — константы, обозначающие символы алфавита):

TC1. $x(yz) = (xy)z$,
TC2. $xy = zw \to ((x = z \wedge y = w) \vee$
$\vee (\exists u)((xu = z \wedge y = uw) \vee (x = zu \wedge uy = w)))$,
TC3. $\neg(\alpha = xy)$,
TC4. $\neg(\beta = xy)$,
TC5. $\neg(\alpha = \beta)$.

Karlov, B.N., On concatenation theory of class of prefix-free languages.

Поступила 12 марта 2025 г., опубликована 3 декабря 2025 г.

В [1] было доказано, что теория TC неразрешима, а в [2] была доказана существенная неразрешимость этой теории, то есть неразрешимость любого её непротиворечивого расширения. Аксиомы TC3–TC5 утверждают, что α и β не представимы в виде конкатенации каких-либо слов и различны. Двумя основными аксиомами теории TC являются TC1 и TC2. Аксиома TC1 утверждает, что операция конкатенации ассоциативна. Аксиома TC2 выражает лемму Леви (см. [3]): если некоторое слово представлено в виде конкатенации двух слов двумя способами, то у этих представлений имеется некоторая общая часть. Из аксиом TC3 и TC4 следует, что рассматриваются только непустые слова. Но, как показано в [2], теории конкатенации слов с пустым словом и без него интерпретируются друг в друге, поэтому результаты о существенной неразрешимости переносятся и на случай универсума, содержащего пустое слово.

Операцию конкатенации можно определить не только на отдельных словах, но и на языках, и можно рассматривать теории конкатенации алгебр, в которых основным множеством является некоторое семейство языков в каком-нибудь фиксированном алфавите. Некоторые такие теории были исследованы в [4, 5]. В частности, было доказано, что теории класса всех языков или класса всех регулярных языков алгоритмически эквивалентны элементарной арифметике и, следовательно, неразрешимы. Эти результаты не являются следствием существенной неразрешимости теории TC, поскольку даже для множества всех конечных языков лемма Леви не выполняется.

В настоящей статье мы продолжаем исследование теории конкатенации для различных классов языков. Мы изучаем беспрефиксные языки, то есть такие, в которых ни одно слово не является собственным префиксом другого. В доказательстве неразрешимости теории конкатенации из [4, 5] существенным образом используются языки вида $\{\varepsilon, w^{mn}\}$ и $\{\varepsilon, w^n, w^{2n}, \ldots, w^{mn}\}$, не являющиеся беспрефиксными, поэтому описанный в этих статьях метод доказательства неприменим, если рассматривать только беспрефиксные языки. В [6] исследовался более узкий класс языков, являющихся одновременно беспрефиксными и бессуффиксными, и был сформулирован результат о неразрешимости теории конкатенации этого класса языков.

Наш интерес к беспрефиксным языкам обусловлен тем, что они обладают некоторыми свойствами, которые делают их полезными на практике. Например, в префиксных кодах (см. [7]) ни одно кодовое слово не является префиксом другого, что позволяет однозначно декодировать сообщения. Другое полезное свойство беспрефиксных языков связано с задачей сопоставления с образцом. В [8] было доказано, что если регулярное выражение r длины m задаёт беспрефиксный язык, то поиск всех подстрок строки $s[1 \ldots n]$, удовлетворяющих выражению r, может быть выполнен за время $O(mn)$, в то время как в случае произвольного r временная сложность увеличивается до $O(mn^2)$. Беспрефиксные языки обладают и интересными теоретическими свойствами. Так, любой язык,

распознаваемый детерминированным МП-автоматом через опустошение магазина, является беспрефиксным, а множество всех беспрефиксных детерминированных КС-языков совпадает с множеством строгих детерминированных КС-языков (см. [9]).

Настоящая статья содержит три основных раздела. В разделе 2 приводятся базовые определения. В разделе 3 мы доказываем наш главный результат: в классе беспрефиксных языков выполняется лемма Леви, а значит, к нему применимы результаты из [2]. В разделе 4 мы доказываем, что алгебра непустых беспрефиксных языков в алфавите, содержащем хотя бы два символа, элементарно эквивалентна алгебре всех слов в бесконечном алфавите.

2 Предварительные сведения

Алфавитом называется непустое множество символов. *Слово* в алфавите Σ — это конечная последовательность символов из Σ. Число символов в слове w называется *длиной* слова w и обозначается $|w|$. Слово длины 0 называется *пустым* и обозначается ε. Через Σ^* обозначается множество всех слов в алфавите Σ. Множество Σ^* всегда бесконечно, а если алфавит Σ тоже бесконечен, то Σ и Σ^* равномощны. *Конкатенацией* слов u и v называется слово, которое получается приписыванием слова v после u. Конкатенация слов u и v обозначается $u \cdot v$ или просто uv. i-й *степенью* слова v называется слово $v^i = \underbrace{v \cdot \ldots \cdot v}_{i \text{ раз}}$, в частности, $v^0 = \varepsilon$, $v^1 = v$. Слово u называется *префиксом* слова v, если $v = uw$ для некоторого слова w. Слово u называется *суффиксом* слова v, если $v = wu$ для некоторого слова w. В этом определении допустимы случаи $u = v$ и $u = \varepsilon$. Префикс слова v называется *собственным*, если он не равен v. Если u является префиксом слова v, то мы будем также говорить, что v начинается на u или что v является продолжением u.

Язык L в алфавите Σ — это произвольное множество слов в Σ. *Конкатенацией* языков L_1 и L_2 называется язык $L_1 \cdot L_2 = \{\, uv \colon u \in L_1, v \in L_2 \,\}$. Как и для слов, вместо $L_1 \cdot L_2$ можно писать просто $L_1 L_2$. i-я *степень* языка L определяется аналогично степени слова: $L^i = \underbrace{L \cdot \ldots \cdot L}_{i \text{ раз}}$, в частности, $L^0 = \{\, \varepsilon \,\}$, $L^1 = L$. Язык L называется *беспрефиксным*, если не существует двух различных слов $x, y \in L$, одно из которых является префиксом другого. Непосредственно из определения следует, что если L — беспрефиксный язык и $\varepsilon \in L$, то $L = \{\, \varepsilon \,\}$.

Теория T — это множество замкнутых формул первого порядка, замкнутое относительно логического следования. Теория алгебраической системы $\mathfrak{A}$ — это множество всех замкнутых формул истинных в $\mathfrak{A}$. Две алгебраические системы $\mathfrak{A}$ и $\mathfrak{B}$ *элементарно эквивалентны*, если $\mathfrak{A} \models \varphi$ тогда и только тогда, когда $\mathfrak{B} \models \varphi$ для любой замкнутой формулы φ, то есть их теории совпадают.

3 Лемма Леви для беспрефиксных языков

Поскольку операция конкатенации языков ассоциативна, то из выполнения леммы Леви для некоторого класса языков будет следовать неразрешимость теории этого класса в силу существенной неразрешимости теории TC (если только существуют хотя бы два языка, не представимых в виде конкатенации). Однако даже если алфавит содержит только один символ, то для множества всех языков лемма Леви не выполняется: из равенства $L_1L_2 = L_3L_4$ не следует существования языка L_5 такого, что $L_1L_5 = L_3$, $L_5L_4 = L_2$ или $L_3L_5 = L_1$, $L_5L_2 = L_4$. Пусть, например, $L_1 = L_4 = \{ a \}$, $L_2 = L_3 = \{ \varepsilon, a \}$. Тогда справедливо равенство $\{ a \} \cdot \{ \varepsilon, a \} = \{ \varepsilon, a \} \cdot \{ a \}$, однако язык L_5 не существует. Действительно, равенство $\{ a \} = \{ \varepsilon, a \} \cdot L_5$ невозможно, так как язык в правой части содержит не менее двух слов. Равенство $\{ \varepsilon, a \} = \{ a \} \cdot L_5$ также невозможно, так как язык в правой части не содержит пустого слова.

Однако если рассматривать не все языки, а только некоторое их подмножество, то ситуация может измениться. Например, очевидно, что лемма Леви выполняется для множества всех языков, содержащих ровно одно слово. В этом разделе мы докажем, что лемма Леви выполняется также для множества всех непустых беспрефиксных языков в некотором фиксированном алфавите Σ. Обозначим это множество через $\mathcal{L}_p(\Sigma)$.

Следующая лемма утверждает, что множество $\mathcal{L}_p(\Sigma)$ замкнуто относительно конкатенации и потому образует алгебру (см. также [10]).

Лемма 1. *Если L_1 и L_2 — беспрефиксные языки, то L_1L_2 — тоже беспрефиксный язык.*

Доказательство. Предположим, что языки L_1 и L_2 беспрефиксные, а язык L_1L_2 — нет. Это значит, что существуют слова $u, v \in L_1$, $x, y \in L_2$ такие, что ux — собственный префикс vy, то есть $vy = uxz$ для некоторого слова $z \neq \varepsilon$. Если $|u| \neq |v|$, то одно из слов u, v является собственным префиксом другого, что невозможно, так как L_1 беспрефиксный. Следовательно, $|u| = |v|$, $u = v$ и $y = xz$. Но из $z \neq \varepsilon$ следует, что x — собственный префикс y. Это противоречит, тому что язык L_2 беспрефиксный. Значит, предположение неверно, и язык L_1L_2 также беспрефиксный. □

Дальше покажем, что на непустые беспрефиксные языки можно сокращать слева.

Лемма 2. *Пусть $L_1 \neq \varnothing$ — беспрефиксный язык, L_2 и L_3 — произвольные языки. Тогда $L_1L_2 = L_1L_3$ тогда и только тогда, когда $L_2 = L_3$.*

Доказательство. Пусть $x \in L_2$. Возьмём произвольное слово $u \in L_1$, тогда $ux \in L_1L_2$, а значит, $ux \in L_1L_3$. Следовательно, существуют слова $v \in L_1$, $y \in L_3$ такие, что $ux = vy$. Но тогда $u = v$, так как в противном случае одно из слов u, v было бы собственным префиксом другого.

Поэтому $x = y$, $x \in L_3$ и $L_2 \subseteq L_3$. Обратное включение доказывается аналогично. □

Следующая лемма выражает «симметричное» свойство. Она утверждает, что допустимо сокращение справа на любой непустой язык, если левые языки беспрефиксные.

Лемма 3. *Пусть L_1 и L_2 — беспрефиксные языки, $L_3 \neq \varnothing$ — произвольный язык. Тогда $L_1L_3 = L_2L_3$ тогда и только тогда, когда $L_1 = L_2$.*

Доказательство. Предположим, что $L_1 \neq L_2$, и рассмотрим кратчайшее слово u, входящее ровно в один из языков L_1, L_2. Без потери общности можно считать, что $u \in L_1$, $u \notin L_2$. Пусть x — кратчайшее слово из языка L_3. Тогда $ux \in L_1L_3$, а значит, $ux \in L_2L_3$. Следовательно, существуют слова $v \in L_2$, $y \in L_3$ такие, что $ux = vy$. Если $|u| > |v|$, то из $v \in L_2$ следует $v \in L_1$, поскольку u — кратчайшее слово, которым различаются L_1 и L_2. Но тогда L_1 содержит слово u и его собственный префикс v, что невозможно, так как L_1 беспрефиксный. Если $|u| = |v|$, то $u = v$ и $u \in L_2$, что противоречит выбору u. Если $|u| < |v|$, то $|y| < |x|$, поэтому x не является кратчайшим словом языка L_3. Итак, во всех трёх случаях получилось противоречие, следовательно, $L_1 = L_2$. □

Пусть для некоторых языков выполняется равенство $L_1L_2 = L_3L_4$. Сначала рассмотрим случай, когда языки L_1 и L_3 имеют хотя бы одно общее слово.

Лемма 4. *Если $L_1L_2 = L_3L_4$, все L_i непусты, языки L_1 и L_3 беспрефиксные и $L_1 \cap L_3 \neq \varnothing$, то $L_1 = L_3$, $L_2 = L_4$.*

Доказательство. Пусть $u \in L_1 \cap L_3$ — произвольное общее слово двух языков. Докажем сначала, что $L_2 = L_4$. Пусть $v \in L_2$. Тогда $uv \in L_1L_2$, а значит, и $uv \in L_3L_4$. Поэтому $uv = xy$ для некоторых слов $x \in L_3$, $y \in L_4$. Если бы оказалось, что $|u| \neq |x|$, то одно из слов u, x было бы собственным префиксом другого. Но это невозможно, поскольку язык L_3 беспрефиксный. Следовательно, $|u| = |x|$, $u = x$, $v = y$ и $v \in L_4$. Поэтому $L_2 \subseteq L_4$. Обратное включение $L_4 \subseteq L_2$ доказывается симметрично. Следовательно, $L_2 = L_4$. Тогда по лемме 3 из равенства $L_1L_2 = L_3L_2$ следует, что $L_1 = L_3$. □

Остаётся рассмотреть случай, когда языки L_1 и L_3 не имеют общих слов. Обозначим через $m(L)$ длину кратчайшего слова из непустого языка L.

Лемма 5. *Если $L_1L_2 = L_3L_4$ и все L_i непусты, то $m(L_1) \leqslant m(L_3)$ тогда и только тогда, когда $m(L_2) \geqslant m(L_4)$.*

Доказательство. Пусть w_i — кратчайшее слово языка L_i для $1 \leqslant i \leqslant 4$. Тогда w_1w_2 и w_3w_4 — кратчайшие слова языков L_1L_2 и L_3L_4, поэтому их длины равны. Следовательно, из неравенства $|w_1| \leqslant |w_3|$ следует неравенство $|w_2| \geqslant |w_4|$ и наоборот. □

Следующая техническая лемма утверждает, что мы можем удалить из языка L_1 все кратчайшие слова, а из L_3 — их продолжения, и при этом равенство $L_1L_2 = L_3L_4$ по-прежнему будет выполняться.

Лемма 6. *Пусть $L_1L_2 = L_3L_4$, все L_i непусты, языки L_1 и L_3 беспрефиксные, $L_1 \cap L_3 = \varnothing$ и $m(L_1) \leqslant m(L_3)$. Пусть L_0 — множество всех кратчайших слов языка L_1, а L'_0 — множество всех слов из L_3, для которых какое-нибудь слово из L_0 является префиксом. Тогда выполняется равенство $(L_1 \setminus L_0) \cdot L_2 = (L_3 \setminus L'_0) \cdot L_4$.*

Доказательство. Если $w \in (L_1 \setminus L_0) \cdot L_2$, то $w = uv$ для некоторых слов $u \in L_1 \setminus L_0$, $v \in L_2$. Так как $uv \in L_1L_2$, то существуют такие слова $x \in L_3$, $y \in L_4$, что $uv = xy$. Предположим, что x имеет вид x_1x_2, где $x_1 \in L_0$, $x_2 \in \Sigma^*$, так что $uv = x_1x_2y$. Так как $x_1 \in L_0$, $u \in L_1 \setminus L_0$, то $|x_1| < |u|$, а значит, x_1 является собственным префиксом слова u. Это противоречит тому, что язык L_1 беспрефиксный. Поэтому x не начинается ни на какое слово из L_0, а значит, $x \in L_3 \setminus L'_0$ и $w \in (L_3 \setminus L'_0) \cdot L_4$.

Наоборот, пусть $w \in (L_3 \setminus L'_0) \cdot L_4$, то есть $w = xy$ для некоторых $x \in L_3 \setminus L'_0$, $y \in L_4$. Так как $xy \in L_3L_4$, то $xy = uv$ для некоторых $u \in L_1$, $v \in L_2$. Если $u \in L_0$, то $|u| \leqslant |x|$, так как $m(L_1) \leqslant m(L_3)$. Но тогда u является префиксом слова x, а значит, $x \in L'_0$ и $x \notin L_3 \setminus L'_0$. Снова получилось противоречие. □

Теперь мы можем доказать лемму Леви для случая, когда языки L_1 и L_3 не имеют общих слов.

Лемма 7. *Пусть $L_1L_2 = L_3L_4$, все L_i непусты, языки L_1 и L_3 беспрефиксные, $L_1 \cap L_3 = \varnothing$ и $m(L_1) \leqslant m(L_3)$. Тогда существует единственный язык L_5 такой, что $L_1L_5 = L_3$, $L_2 = L_5L_4$. При этом L_5 является беспрефиксным.*

Доказательство. По лемме 5 справедливо неравенство $m(L_2) \geqslant m(L_4)$. Сначала докажем, что любое слово $u \in L_1$ является собственным префиксом некоторого слова из L_3. Пусть L_0 и L'_0 определены так же, как в формулировке леммы 6. Проведём индукцию по $|u| - m(L_1)$.

Базис индукции. Пусть $|u| = m(L_1)$, то есть $u \in L_0$. Пусть v — произвольное слово из L_2, тогда $uv \in L_1L_2$, а значит, $uv \in L_3L_4$, то есть $uv = xy$ для некоторых $x \in L_3$, $y \in L_4$. Так как $m(L_1) \leqslant m(L_3)$ и $L_1 \cap L_3 = \varnothing$, то u является собственным префиксом слова x.

Индукционный шаг. Пусть $|u| > m(L_1)$. Пусть $L'_1 = L_1 \setminus L_0$, $L'_3 = L_3 \setminus L'_0$. Тогда по лемме 6 $L'_1L_2 = L'_3L_4$, а по лемме 5 $m(L'_1) \leqslant m(L'_3)$. Кроме того, $m(L'_1) > m(L_1)$ и $u \in L'_1$. Следовательно, по индукционному предположению u является собственным префиксом некоторого слова из L'_3, а значит, и из L_3.

Из доказанного непосредственно следует, что любое слово из L_3 является продолжением некоторого слова из L_1 и притом только одного. Действительно, если $x \in L_3$, $y \in L_4$, то $xy = uv$ для некоторых $u \in L_1$, $v \in L_2$. Если $|x| \leqslant |u|$, то согласно доказанному выше язык L_3 содержит

некоторое слово x', собственным префиксом которого является u. Но тогда x является собственным префиксом слова x', что невозможно, так как язык L_3 беспрефиксный. Если бы слово $x \in L_3$ было продолжением разных слов $u_1, u_2 \in L_1$, то одно из слов u_1, u_2 было бы собственным префиксом другого, что противоречит тому, что язык L_1 беспрефиксный.

Теперь докажем главное утверждение леммы. Через $A \sqcup B$ будем обозначать объединение множеств A и B при условии, что $A \cap B = \varnothing$. Рассмотрим произвольное слово $x \in L_3$. Для него существует единственное слово $u \in L_1$ такое, что $x = uz$ для некоторого $z \in \Sigma^*$. Поэтому язык L_3 можно представить в виде

$$L_3 = \bigsqcup_{u \in L_1} uL_{3,u} \tag{1}$$

для некоторых языков $L_{3,u}$. Тогда конкатенация L_3L_4 может быть записана в виде

$$L_3L_4 = \Big(\bigsqcup_{u \in L_1} uL_{3,u} \Big) \cdot L_4 = \bigsqcup_{u \in L_1} u(L_{3,u}L_4).$$

Эта запись корректна, поскольку языки $u(L_{3,u}L_4)$ не имеют общих слов. Действительно, если бы выполнялось равенство $u_1w_1 = u_2w_2$ для некоторых $u_1, u_2 \in L_1$, $w_1 \in L_{3,u_1}L_4$, $w_2 \in L_{3,u_2}L_4$, то одно из слов u_1, u_2 было бы собственным префиксом другого. Конкатенацию L_1L_2 можно записать в виде

$$L_1L_2 = \bigsqcup_{u \in L_1} uL_2.$$

Следовательно, равенство $L_1L_2 = L_3L_4$ принимает вид

$$\bigsqcup_{u \in L_1} uL_2 = \bigsqcup_{u \in L_1} u(L_{3,u}L_4),$$

откуда $uL_2 = uL_{3,u}L_4$ для всех $u \in L_1$. Сокращая слева на слово u, получаем $L_2 = L_{3,u}L_4$ для всех $u \in L_1$. Так как левые части всех равенств одинаковы, то $L_{3,u}L_4 = L_{3,v}L_4$ для всех u и v. Кроме того, все языки $L_{3,u}$ являются беспрефиксными. Действительно, если бы некоторый язык $L_{3,u}$ содержал слова w и wz, где $z \neq \varepsilon$, то L_3 содержал бы слова uw и uwz, а значит, не был бы беспрефиксным. Следовательно, по лемме 3 получается, что все языки $L_{3,u}$ равны. Теперь возьмём в качестве L_5 язык $L_{3,u}$. Непосредственно из равенства $L_2 = L_{3,u}L_4$ следует $L_2 = L_5L_4$. Из (1) получаем

$$L_3 = \bigsqcup_{u \in L_1} uL_5 = \Big(\bigsqcup_{u \in L_1} \{\, u \,\} \Big) \cdot L_5 = L_1L_5.$$

Теперь докажем единственность языка L_5. Если существует другой язык L_5' такой, что $L_1L_5' = L_3$, то $L_1L_5 = L_1L_5'$. Но тогда $L_5 = L_5'$ по лемме 2. □

Объединяя частные случаи, мы получаем главный результат.

Теорема 1 (Лемма Леви для беспрефиксных языков)**.** *Пусть $L_1L_2 = L_3L_4$, где все L_i непусты, L_1 и L_3 — беспрефиксные языки. Тогда существует единственный язык L_5 такой, что $L_1L_5 = L_3$, $L_5L_4 = L_2$ или $L_3L_5 = L_1$, $L_5L_2 = L_4$. При этом L_5 является беспрефиксным.*

Доказательство. Если $L_1 \cap L_3 \neq \varnothing$, то $L_1 = L_3$, $L_2 = L_4$ по лемме 4, поэтому можно положить $L_5 = \{\varepsilon\}$. Если $L_1 \cap L_3 = \varnothing$, то справедливость утверждения следует из леммы 7. □

Заметим, что языки L_2 и L_4 могут не быть беспрефиксными. Пусть, например, $L_1 = \{a, b\}$, $L_2 = \{w \in \{a, b\}^* : |w| \geqslant 2\}$, $L_3 = \{aa, ab, ba, bb\}$, $L_4 = \{w \in \{a, b\}^* : |w| \geqslant 1\}$. Тогда $L_1L_2 = L_3L_4$, $L_5 = \{a, b\}$, однако ни L_2, ни L_4 не является беспрефиксным. Отметим также, что беспрефиксности только одного из языков L_1, L_3 может быть недостаточно. Рассмотрим снова пример из начала этого раздела: $L_1 = L_4 = \{a\}$, $L_2 = L_3 = \{\varepsilon, a\}$. Язык L_1 беспрефиксный, однако лемма Леви не выполняется.

Для языка L_5 можно записать и явное выражение с использованием операции деления языков (см. [3]). Левое частное языков L_1 и L_2 определяется как

$$L_1^{-1}L_2 = \{w \in \Sigma^* : uw \in L_2 \text{ для некоторого } u \in L_1\}.$$

Мы не используем обозначение $L_1 \backslash L_2$, чтобы отличать частное от разности множеств.

Лемма 8. *В условиях леммы 7 $L_5 = L_1^{-1}L_3$.*

Доказательство. Пусть $w \in L_1^{-1}L_3$. Тогда существует слово $u \in L_1$ такое, что $uw \in L_3$. Так как $L_3 = L_1L_5$, то $uw \in L_1L_5$. Если $w \notin L_5$, то $uw = u_1w_1$ для некоторых $u_1 \in L_1$, $w_1 \in L_5$, при этом $w_1 \neq w$. Но тогда одно из слов u, u_1 является собственным префиксом другого, что невозможно. Следовательно, $w \in L_5$ и $L_1^{-1}L_3 \subseteq L_5$.

Обратно, пусть $w \in L_5$. Возьмём произвольное слово $u \in L_1$. Тогда $uw \in L_1L_5$, а значит, $uw \in L_3$. По определению получаем $w \in L_1^{-1}L_3$, поэтому $L_5 \subseteq L_1^{-1}L_3$. □

Аналогично левому частному можно определить и правое частное как

$$L_1L_2^{-1} = \{w \in \Sigma^* : wu \in L_1 \text{ для некоторого } u \in L_2\}.$$

Однако симметричная формула $L_5 = L_2L_4^{-1}$ может не иметь места. Пусть $L_1 = \{a\}$, $L_2 = \{a^ib : i > 0\}$, $L_3 = \{aa\}$, $L_4 = \{a^ib : i \geqslant 0\}$. Тогда $L_1L_2 = L_3L_4$, $L_1^{-1}L_3 = \{a\}$, но $L_2L_4^{-1} = \{a^i : i \geqslant 0\}$.

Лемма Леви справедлива и для бессуффиксных языков, то есть для таких, в которых ни одно слово не является собственным суффиксом другого слова.

Теорема 2 (Лемма Леви для бессуффиксных языков)**.** *Пусть $L_1L_2 = L_3L_4$, где все L_i непусты, L_2 и L_4 — бессуффиксные языки. Тогда существует единственный язык L_5 такой, что $L_1L_5 = L_3$, $L_5L_4 = L_2$ или $L_3L_5 = L_1$, $L_5L_2 = L_4$. При этом L_5 является бессуффиксным.*

Доказательство. Обозначим через w^R обращение слова w: если $w = a_1a_2\dots a_n$, то $w^R = a_n\dots a_2a_1$. Через L^R обозначим обращение языка L, то есть множество обращений всех слов из L: $L^R = \{\, w^R : w \in L \,\}$. Непосредственно из определений следует, что язык L является беспрефиксным тогда и только тогда, когда L^R является бессуффиксным, а также что $(L_1L_2)^R = L_2^RL_1^R$. Поэтому равенство $L_1L_2 = L_3L_4$ эквивалентно равенству $L_2^RL_1^R = L_4^RL_3^R$. По теореме 1 существует язык L_5' такой, что либо $L_2^RL_5' = L_4^R$, $L_5'L_3^R = L_1^R$, либо $L_4^RL_5' = L_2^R$, $L_5'L_1^R = L_3^R$. Полагая $L_5 = (L_5')^R$ и применяя обращение к этим равенствам, мы получим утверждение теоремы. □

По аналогии с леммой 8 можно доказать, что в этом случае $L_5 = L_4L_2^{-1}$, если $m(L_2) \leqslant m(L_4)$.

4 Неразрешимость теории конкатенации

С помощью леммы Леви легко доказать неразрешимость теории конкатенации для множества беспрефиксных языков.

Теорема 3. *Если $|\Sigma| \geqslant 2$, то теории алгебр $\mathfrak{A} = (\mathcal{L}_p(\Sigma), \cdot)$ и $\mathfrak{B} = (\mathcal{L}_p(\Sigma) \cup \{\, \varnothing \,\}, \cdot)$ неразрешимы.*

Доказательство. В алгебре $(\mathcal{L}_p(\Sigma) \setminus \{\, \{\, \varepsilon \,\} \,\}, \cdot)$ выполняются все аксиомы TC1–TC5. Следовательно, её теория неразрешима в силу существенной неразрешимости теории TC (см. [2]). Пустое множество определимо формулой $(\forall y)xy = x$, а множество $\{\, \varepsilon \,\}$ — формулой $(\forall y)xy = y$ (здесь переменные x и y обозначают не слова, а языки). Следовательно, теории алгебр $\mathfrak{A}$ и $\mathfrak{B}$ также неразрешимы. □

Мы уже отмечали, что лемма Леви неверна в классе всех языков, а значит, алгебры слов и языков с операцией конкатенации не являются элементарно эквивалентными. Если рассматривать все беспрефиксные языки, включая пустой, то более простым примером формулы, различающей алгебры слов и языков, является формула, выражающая существования нуля моноида: $(\exists x)(\forall y)xy = x$. Она ложна на множестве слов, но истинна на множестве языков. Далее мы докажем, что алгебры $(\Delta^*, \cdot)$ и $(\mathcal{L}_p(\Sigma), \cdot)$ элементарно эквивалентны, если $|\Sigma| \geqslant 2$, а алфавит Δ бесконечен.

Сначала мы определим специальные языки, которые будут играть роль символов в алгебре языков.

Определение 1. *Пусть $L \neq \varnothing$ — беспрефиксный язык. L называется* элементарным*, если не существует беспрефиксных языков L_1 и L_2, отличных от $\{\, \varepsilon \,\}$ и таких, что $L = L_1L_2$.*

Элементарные языки неразложимы подобно символам. В лемме 10 мы докажем, что любой язык из $\mathcal{L}_p(\Sigma)$ выражается через них подобно тому, как любое слово является конкатенацией своих символов. Заметим, что требование беспрефиксности языков L_1 и L_2 нельзя опустить. Например, $\{ ab, aabb \} \in \mathcal{L}_p(\Sigma)$ для $\Sigma = \{ a, b \}$, но этот язык представляется в виде конкатенации $\{ a, aab \} \cdot \{ b \}$, в которой первый язык уже не является беспрефиксным.

Лемма 9. *Если $|\Sigma| \geqslant 2$, то множества элементарных языков и всех языков в алфавите Σ равномощны.*

Доказательство. Для любого целого $i = 0, 1, 2, \ldots$ любой язык L, содержащий слова a и ba^ib, является элементарным. Действительно, если $L = L_1L_2$, то один из языков L_1, L_2 должен содержать слово a, а другой — пустое слово ε. Но тогда второй язык равен $\{ \varepsilon \}$, поэтому язык L элементарный. Остальные слова языка L можно выбрать произвольным образом, что и даст требуемую мощность. □

В частности, из этой леммы следует, что множество элементарных языков в не более чем счётном алфавите, содержащем хотя бы две буквы, имеет мощность континуум.

Лемма 10. *Любой беспрефиксный язык L представляется в виде конкатенации элементарных языков и при этом однозначно с точностью до множителей $\{ \varepsilon \}$.*

Доказательство. Существование представления докажем индукцией по длине k кратчайшего слова в L.
Базис индукции. Если $k = 0$, то $\varepsilon \in L$, а значит, $L = \{ \varepsilon \}$, так как в противном случае L не был бы беспрефиксным. L является искомым представлением.
Индукционный шаг. Пусть длина кратчайшего слова языка L равна $k+1$. Если L элементарный, то L является искомым представлением. Если L не является элементарным, то $L = L_1L_2$ для некоторых $L_1, L_2 \neq \{ \varepsilon \}$. Если $x \in L_1$, $y \in L_2$ — кратчайшие слова, то $|xy| = k + 1$. Поскольку $\varepsilon \notin L_1, L_2$, то $|x|, |y| \leqslant k$. Поэтому для L_1 и L_2 искомые представления существуют по индукционному предположению.

Теперь докажем единственность представления. Предположим, что для языка L существуют два различных представления

$$L_1 \cdot L_2 \cdot \ldots \cdot L_n = L'_1 \cdot L'_2 \cdot \ldots \cdot L'_m,$$

в которых все языки отличны от $\{ \varepsilon \}$. Предположим сначала, что $L_i \neq L'_i$ для некоторого i, и выберем наименьшее из таких i. Тогда

$$L_1 \cdot \ldots \cdot L_{i-1} \cdot L_i \cdot L' = L_1 \cdot \ldots \cdot L_{i-1} \cdot L'_i \cdot L'',$$

где через L' и L'' обозначены конкатенации языков, стоящих правее. Сокращая слева на $L_1, \ldots, L_{i-1}$ по лемме 2, получаем $L_iL' = L'_iL''$. По теореме 1 существует беспрефиксный язык L''' такой, что $L_iL''' = L'_i$ или

$L_i'L''' = L_i$. Но так как L_i и L_i' элементарны, то это возможно только при $L''' = \{\varepsilon\}$, откуда $L_i = L_i'$.

Теперь предположим, что $L_1 = L_1', \ldots, L_n = L_n'$, но $m > n$. Равенство двух разложений можно переписать в виде

$$L_1 \cdot L_2 \cdot \ldots \cdot L_n \cdot \{\varepsilon\} = L_1 \cdot L_2 \cdot \ldots \cdot L_n \cdot L_{n+1}' \cdot \ldots \cdot L_m'.$$

Сокращая слева по лемме 2 на беспрефиксные языки $L_1, \ldots, L_n$, получаем $L_{n+1}' \cdot \ldots \cdot L_m' = \{\varepsilon\}$, откуда $L_{n+1}' = \cdots = L_m' = \{\varepsilon\}$. Следовательно, в обоих случаях разложения совпадают. □

Леммы 9 и 10 позволяют описать строение алгебры беспрефиксных языков.

Теорема 4. *Для любого алфавита Σ моноид $(\mathcal{L}_p(\Sigma), \cdot)$ является свободным. Моноиды $(\mathcal{L}_p(\Sigma_1), \cdot)$ и $(\mathcal{L}_p(\Sigma_2), \cdot)$ изоморфны при $|\Sigma_1|, |\Sigma_2| \geqslant 2$ тогда и только тогда, когда либо алфавиты Σ_1 и Σ_2 не более чем счётны, либо они несчётны и имеют одинаковую мощность.*

Доказательство. Известно (см. [11]), что полугруппа S является свободной тогда и только тогда, когда каждый элемент из S может быть однозначно представлен в виде произведения элементов из некоторого множества X. Поэтому из леммы 10 следует, что моноид $\mathcal{L}_p(\Sigma)$ является свободным, причём множеством образующих являются все элементарные языки. Второе утверждение теоремы следует из леммы 9 и из того, что свободные моноиды изоморфны тогда и только тогда, когда их множества образующих равномощны. □

В частности из этой теоремы следует, что моноид беспрефиксных языков изоморфен моноиду всех слов в некотором подходящем алфавите Δ бесконечной мощности. Поскольку свободные моноиды с бесконечным числом образующих элементарно эквивалентны, то справедлив также следующий результат.

Следствие 1. *Пусть Δ — бесконечный алфавит, Σ — алфавит, содержащий не менее двух символов. Тогда моноиды $\mathfrak{A} = (\Delta^*, \cdot)$ и $\mathfrak{B} = (\mathcal{L}_p(\Sigma), \cdot)$ элементарно эквивалентны.*

Если использовать стандартное кодирование для случая счётного Δ, когда символ $a_i \in \Delta$ записывается в виде a^ib, то можно сформулировать другой вариант этого следствия, в котором все алфавиты конечны.

Следствие 2. *Пусть L_0 — множество всех слов в алфавите $\{a, b\}$, заканчивающихся на b, а также пустое слово, и пусть $|\Sigma| \geqslant 2$. Тогда алгебры $\mathfrak{A} = (L_0, \cdot)$ и $\mathfrak{B} = (\mathcal{L}_p(\Sigma), \cdot)$ элементарно эквивалентны.*

Доказательство. Любое слово из языка L_0 однозначно представляется в виде конкатенации слов вида a^ib, поэтому $\mathfrak{A}$ — свободный моноид со счётным множеством образующих $\{a^ib : i \geqslant 0\}$. Он элементарно эквивалентен свободному моноиду $\mathfrak{B}$ с бесконечным множеством образующих. □

В [12] была доказана разрешимость теории слов в произвольном алфавите со счётным множеством операций возведения в степень x^i для $i \geqslant 2$. Из следствия 1 немедленно получается аналогичный результат для $\mathcal{L}_p(\Sigma)$.

Теорема 5. *Теория алгебры $\mathcal{L}_p(\Sigma)$ с операциями x^i, $i \geqslant 2$, разрешима.*

Отметим также, что все доказанные в этом разделе результаты о разрешимости и элементарной эквивалентности непосредственно переносятся и на бессуффиксные языки.

В заключение сформулируем некоторые вопросы для дальнейшего исследования.

- Для каких более широких классов языков также выполняется лемма Леви?
- Разрешима ли теория всех языков в произвольном фиксированном алфавите Σ с операциями возведения в степень x^i, $i \geqslant 2$?

Автор выражает благодарность М. Е. Вишникину за обсуждение результатов и ценные замечания.

References

[1] A. Grzegorczyk, *Undecidability without arithmetization*, Stud. Log., **79**:2 (2005), 163–230. Zbl 1080.03004

[2] A. Grzegorczyk, K. Zdanowski, *Undecidability and concatenation*, in A. Ehrenfeucht (ed.) et al., *Andrzej Mostowski and foundational studies*, IOS Press, Amsterdam, 2008, 72–91. Zbl 1150.03014

[3] J. Shallit, *A second course in formal languages and automata theory*, Cambridge University Press, Cambridge, 2009. Zbl 1163.68025

[4] S.M. Dudakov, *On undecidability of concatenation theory for one-symbol languages*, Lobachevskii J. Math., **41**:2, (2020) 168–175. Zbl 1486.03073

[5] S. Dudakov, B. Karlov, *On decidability of theories of regular languages*, Theory Comput. Syst. **65**:3 (2021), 462–478. Zbl 1517.68185

[6] B.N. Karlov, *On concatenation theory of one class of languages*, in Proc. of International Conference, *Algebra and Mathematical Logic: Theory and Applications*, KFU, Kazan, 2024, 206–207.

[7] V.D. Kolesnik, G.Sh. Poltyrev, *Course of information theory*, Nauka, Moscow, 1982. (1980, Zbl 0511.94010)

[8] Y.-S. Han, Y. Wang, D. Wood, *Prefix-free regular languages and pattern matching*, Theor. Comput. Sci., **389**:1-2 (2007), 307–317. Zbl 1143.68037

[9] M.A. Harrison, I.M. Havel, *Strict deterministic grammars*, J. Comput. Syst. Sci., **7**:3 (1973), 237–277. Zbl 0261.68036

[10] M. Krausová, *Prefix-free regular languages: closure properties, difference, and left quotient*, in Z. Kotásek (ed.) et al., *Mathematical and engineering methods in computer science*, MEMICS 2011. Lect. Notes in Comput. Sci., **7119**, Springer, Berlin, 2012, 114–122.

[11] A.H. Clifford, G.B. Preston, *The algebraic theory of semigroups. Volume II,* American Mathematical Society, Providence, 1967. Zbl 0178.01203

[12] B. Karlov, *Algorithmic properties of some fragments of concatenation theory*, J. Phys., Conf. Ser., **1902** (2021), Paper No. 012117.

Boris Nikolaevich Karlov
Tver State University,
Zhelyabova str., 33,
170100, Tver, Russia
Email address: bnkarlov@gmail.com