УДК 519.766.23

# $(m, n)$-ЖЕСТКИЕ КАТЕГОРИАЛЬНЫЕ ГРАММАТИКИ

**Карлов Б.Н.**
Тверской государственный университет, г. Тверь

*Поступила в редакцию 05.11.2017, после переработки 02.12.2017.*

В статье определяются $(m, n)$-жесткие категориальные грамматики. Построен алгоритм, который по грамматике $G$ и числам $m$, $n$ определяет, является ли $G$ $(m, n)$-жесткой. Доказано, что в классе $(m, n)$-жестких языков существует бесконечная иерархия, а также что класс $(m, n)$-жестких языков не сравним с классом регулярных языков. Исследуется сложность проблемы принадлежности для $(m, n)$-жестких грамматик.

**Ключевые слова:** формальные грамматики, категориальные грамматики, жесткие грамматики, алгоритм проверки жесткости, иерархия жестких языков, проблема принадлежности.

*Вестник ТвГУ. Серия: Прикладная математика. 2017. № 4. С. 7–23.*
https://doi.org/10.26456/vtpmk185

## Введение

Одной из основных задач компьютерной лингвистики является построение грамматических формализмов, допускающих эффективный анализ. Алгоритмы Кока-Янгера-Касами и Эрли (см. [1]) для контекстно-свободных грамматик имеют временную сложность $O(n^3)$, где $n$ — длина входного слова. Несмотря на то, что эти алгоритмы имеют полиномиальную сложность, они работают слишком медленно для грамматик большого размера и длинных слов. Алгоритмы для анализа различных расширений КС-грамматик имеют еще большую сложность. Так, алгоритм для анализа ТАГ-грамматик и эквивалентных им комбинаторных категориальных грамматик имеет сложность $O(n^6)$ (см. [8,11]), а общая проблема принадлежности для категориальных грамматик зависимостей является NP-полной [6,7]. Поэтому значительный интерес представляет выделение специальных подклассов грамматик, для которых существуют более эффективные методы анализа. Примерами таких классов для КС-грамматик являются LR- и LL-грамматики, а также различные варианты грамматик предшествования (см. [1]). Еще один тип эффективно анализируемых грамматик — жесткие грамматики (см. [9]). Это классические категориальные грамматики, в которых каждому символу соответствует единственная категория. Для естественных языков условие жесткости не выполняется, так как одно и то же слово может употребляться многими способами. Однако в некоторых случаях категория слова может определяться его контекстом, т.е. стоящими рядом с ним словами. Например, если *прилагательное* стоит между *предлогом* и *существительным*, то оно имеет категорию [опред] («в глубоком снегу»: глубоком $\xleftarrow{\text{опред}}$ снегу). В других контекстах прилагательные могут иметь и другие категории, например, в словосочетании «в очень глубоком снегу» слово «глубоком» имеет категорию [огранич\опред]: очень $\xleftarrow{\text{огранич}}$ глубоком $\xleftarrow{\text{опред}}$ снегу. В этих примерах «опред» обозначает определительное синтаксическое отношение, а «огранич» — ограничительное отношение (см. [3]). В настоящей статье предлагается обобщение понятия жесткой грамматики.

В разделе 1 приводятся определения классических категориальных грамматик, порождаемых этими грамматиками языков и $(m, n)$-жестких грамматик. Это классические категориальные грамматики, в которых категория, приписываемая каждому символу, определяется этим символом и его $m$ левыми и $n$ правыми соседями. Также определяются понятия $(m, *)$-жестких, $(*, n)$-жестких и $(*, *)$-жестких грамматик. Символ $*$ означает, что при определении категории учитываются все символы, стоящие по заданную сторону от текущего символа. В разделе 2 описан алгоритм, который по произвольной классической категориальной грамматике $G$ и числам $m$, $n$ определяет, является ли она $(m, n)$-жесткой. В разделе 3 исследуются некоторые свойства $(m, n)$-жестких грамматик. Доказывается, что классы $(m, n)$-жестких языков образуют бесконечную иерархию, что класс $(1, 0)$-жестких языков не сравним с классом $(*, 0)$-жестких языков, а класс $(0, 1)$-жестких языков не сравним с классом $(0, *)$-жестких языков. Также доказывается, что всякий регулярный язык порождается некоторой $(*, 1)$-жесткой грамматикой, но существуют регулярные языки, не порождаемые никакой $(m, n)$-жесткой грамматикой с категориями первого порядка. В разделе 4 исследуется проблема принадлежности для $(m, n)$-жестких грамматик. Доказывается, что даже при условии жесткости возможен случай, когда слова имеют экспоненциальное число выводов. Однако если искать только один вывод, а не все, то проблема принадлежности для $(m, n)$-жестких категориальных грамматик зависимостей может быть решена за кубическое время.

## 1. Основные определения

Пусть $\Sigma$ — произвольный алфавит. Элементы $\Sigma$ называются символами или буквами. Слово в алфавите $\Sigma$ — это конечная последовательность символов. Число символов в слове $w$ называется длиной слова и обозначается $|w|$. Пустое слово обозначается $\varepsilon$. Через $\Sigma^*$ обозначается множество всех слов в алфавите $\Sigma$, а через $\Sigma^+$ — множество всех непустых слов в алфавите $\Sigma$.

Классические категориальные грамматики были определены в [4]. Основным объектом этих грамматик являются категории.

**Определение 1.** *Пусть* $\mathbf{C}$ *— произвольное конечное множество (множество элементарных категорий),* [, ], /, \ *— символы, не принадлежащие* $\mathbf{C}$. *Множество категорий над* $\mathbf{C}$ *(обозначается* $\mathrm{Cat}(\mathbf{C})$*) — это наименьшее множество слов в алфавите* $\mathbf{C} \cup \{ [, ], /, \backslash \}$ *такое, что:*
*1)* $\mathbf{C} \subseteq \mathrm{Cat}(\mathbf{C})$*;*
*2) если* $\Phi, \Psi \in \mathrm{Cat}(\mathbf{C})$*, то* $[\Phi/\Psi] \in \mathrm{Cat}(\mathbf{C})$ *и* $[\Psi\backslash\Phi] \in \mathrm{Cat}(\mathbf{C})$.

*Элементы множества* $\mathrm{Cat}(\mathbf{C})$ *называются категориями.*

Для каждой категории $\Phi$ определяется множество $\mathrm{Subcat}(\Phi)$ всех ее подкатегорий.

**Определение 2.** *1) Если* $\Phi \in \mathbf{C}$, *то* $\mathrm{Subcat}(\Phi) = \{\,\Phi\,\}$.
*2) Если* $\Phi = [\Psi/\Theta]$ *или* $\Phi = [\Theta\backslash\Psi]$, *то* $\mathrm{Subcat}(\Phi) = \mathrm{Subcat}(\Psi) \cup \mathrm{Subcat}(\Theta) \cup \{\,\Phi\,\}$.

Обозначим через $\mathrm{Cat}(\mathbf{C})^*$ множество всех конечных последовательностей (или строк) категорий. На множестве $\mathrm{Cat}(\mathbf{C})^*$ определяется отношение сокращения $\vdash$.

**Определение 3.** *Для любых двух категорий* $\Phi$, $\Psi$ *и любых двух строк категорий* $\Gamma_1$, $\Gamma_2$ *имеет место* $\Gamma_1[\Phi/\Psi]\Psi\Gamma_2 \vdash \Gamma_1\Phi\Gamma_2$ *и* $\Gamma_1\Psi[\Psi\backslash\Phi]\Gamma_2 \vdash \Gamma_1\Phi\Gamma_2$. *Через* $\vdash^*$ *обозначается рефлексивное транзитивное замыкание отношения* $\vdash$.

Теперь определим классические категориальные грамматики.

**Определение 4.** *Классическая категориальная грамматика — это четверка* $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$, *где:*
$\Sigma$ *— конечный алфавит,*
$\mathbf{C}$ *— конечное множество элементарных категорий,*
$S \in \mathrm{Cat}(\mathbf{C})$ *— главная категория,*
$\delta\colon \Sigma \to \mathcal{P}(\mathrm{Cat}(\mathbf{C}))$ *— отображение алфавита в множество всех подмножеств множества категорий такое, что для любого* $a \in \Sigma$ *множество* $\delta(a)$ *конечно (словарь).*

Если $\delta(a) = \{\,\Phi_1, \ldots, \Phi_n\,\}$, то мы будем записывать это в виде $a \mapsto \Phi_1, \ldots, \Phi_n$. Для слова $w = a_1 \ldots a_n \in \Sigma^*$ через $\delta(w)$ обозначим множество всех последовательностей категорий, которые могут быть приписаны слову $w$:

$$\delta(w) = \{\,\Phi_1 \ldots \Phi_n \mid \Phi_i \in \delta(a_i) \text{ для } i = 1, \ldots, n\,\}.$$

**Определение 5.** *Язык, порождаемый грамматикой* $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$, *состоит из всех слов* $a_1 \ldots a_n \in \Sigma^+$, *для которых существует последовательность категорий* $\Phi_1 \ldots \Phi_n$ *такая, что* $\Phi_i \in \delta(a_i)$ *для всех* $i$ *и* $\Phi_1 \ldots \Phi_n \vdash^* S$. *Язык, порождаемый грамматикой* $G$, *обозначается* $L(G)$.

Специальный случай категорий — категории первого порядка, в которых выполняется только «деление» на элементарные категории.

**Определение 6.** *1) Если* $\Phi$ *— элементарная категория, то* $\Phi$ *— категория первого порядка.*
*2) Если* $\Phi$ *— категория первого порядка, а* $\Psi$ *— элементарная категория, то* $[\Phi/\Psi]$ *и* $[\Psi\backslash\Phi]$ *— категории первого порядка.*

Известно, что классические категориальные грамматики порождают в точности контекстно-свободные языки, не содержащие пустого слова [5]. Кроме того, по любой категориальной грамматике можно построить эквивалентную ей категориальную грамматику, в которой используются только категории первого порядка (см. [2]). Поэтому рассмотрение только грамматик с категориями первого порядка не приводит к уменьшению класса порождаемых языков.

Обозначим через $F(\mathbf{C})$ свободную группу, порожденную множеством $\mathbf{C}$. В [10] было определено понятие интерпретации категорий и последовательностей категорий в свободной группе (обозначается $[\![\Phi]\!]$).

**Определение 7.** $[\![\Phi]\!] = \Phi$ *для всех* $\Phi \in \mathbf{C}$
$[\![[\Phi/\Psi]]\!] = [\![\Phi]\!][\![\Psi]\!]^{-1}$ *для всех* $\Phi, \Psi \in \mathrm{Cat}(\mathbf{C})$
$[\![[\Psi\backslash\Phi]]\!] = [\![\Psi]\!]^{-1}[\![\Phi]\!]$ *для всех* $\Phi, \Psi \in \mathrm{Cat}(\mathbf{C})$
$[\![\Phi_1 \dots \Phi_n]\!] = [\![\Phi_1]\!] \dots [\![\Phi_n]\!]$ *для всех* $\Phi_1, \dots, \Phi_n \in \mathrm{Cat}(\mathbf{C})$

Следующее свойство интерпретаций было доказано в [10] для исчисления Ламбека. Оно непосредственно переносится на классические категориальные грамматики.

**Лемма 1.** *Если* $\Phi_1 \dots \Phi_n \vdash^* \Psi$*, то* $[\![\Phi_1 \dots \Phi_n]\!] = [\![\Psi]\!]$.

Теперь определим жесткие категориальные грамматики (см. [9]).

**Определение 8.** *Классическая категориальная грамматика* $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$ *называется жесткой, если* $|\delta(a)| = 1$ *для всех* $a \in \Sigma$.

Определим понятие контекста символа в слове.

**Определение 9.** *Пусть* $m$ *и* $n$ *— натуральные числа,* $w$ *— слово,* $a$ *— некоторый символ слова* $w$*. Пусть зафиксировано некоторое вхождение* $a$ *в* $w$*.*
*1)* $(m, n)$*-контекст вхождения символа* $a$ *в* $w$ *— это пара слов* $(u, v)$ *такая, что* $w = w_1 u a v w_2$ *для некоторых слов* $w_1$*,* $w_2$*, где* $a$ *— данное вхождение* $a$ *в* $w$*,* $|u| \le m$*,* $|v| \le n$ *и при этом, если* $|u| < m$*, то* $w_1 = \varepsilon$*, а если* $|v| < n$*, то* $w_2 = \varepsilon$*.*
*2)* $(m, *)$*-контекст вхождения символа* $a$ *в* $w$ *— это пара слов* $(u, v)$ *такая, что* $w = w_1 u a v$ *для некоторого слова* $w_1$*, где* $a$ *— данное вхождение* $a$ *в* $w$*,* $|u| \le m$ *и при этом, если* $|u| < m$*, то* $w_1 = \varepsilon$*.*
*3)* $(*, n)$*-контекст вхождения символа* $a$ *в* $w$ *— это пара слов* $(u, v)$ *такая, что* $w = u a v w_2$ *для некоторого слова* $w_2$*, где* $a$ *— данное вхождение* $a$ *в* $w$*,* $|v| \le n$ *и при этом, если* $|v| < n$*, то* $w_2 = \varepsilon$*.*
*4)* $(*, *)$*-контекст вхождения символа* $a$ *в* $w$ *— это пара слов* $(u, v)$ *такая, что* $w = u a v$*, где* $a$ *— данное вхождение* $a$ *в* $w$*.*

Неформально, $(m, n)$-контекст — это $m$ символов, стоящих непосредственно слева от $a$, и $n$ символов, стоящих непосредственно справа от $a$. Если с одной из сторон нет достаточного числа символов, то в контекст включаются все имеющиеся символы. Звездочка означает, что в контекст включаются все символы, стоящие по данную сторону от вхождения $a$.

Мы обобщаем понятие жесткой грамматики следующим образом.

**Определение 10.** *Категориальная грамматика* $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$ *называется* $(m, n)$*-жесткой, если для любых двух слов* $u = w_1 a w_2$*,* $v = w_1' a w_2'$*, в которых символ* $a$ *стоит в одном и том же* $(m, n)$*-контексте, для любых последовательностей категорий* $\Gamma_1 \in \delta(w_1)$*,* $\Gamma_1' \in \delta(w_1')$*,* $\Gamma_2 \in \delta(w_2)$*,* $\Gamma_2' \in \delta(w_2')$ *и для любых двух категорий* $\Phi, \Psi \in \delta(a)$ *из* $\Gamma_1 \Phi \Gamma_2 \vdash^* S$ *и* $\Gamma_1' \Psi \Gamma_2' \vdash^* S$ *следует* $\Phi = \Psi$*.*

Говоря неформально, категориальная грамматика является $(m, n)$-жесткой, если категория любого символа в слове однозначно определяется самим символом, а также его $m$ левыми и $n$ правыми соседями. В частности, $(0, 0)$-жесткая грамматика эквивалентна жесткой, поскольку категория символа зависит только от этого символа. В определении $(m, n)$-жесткой грамматики длина контекста не превосходит $m + n$. Можно ослабить это требование и учитывать все символы, стоящие слева или справа от текущего символа.

**Определение 11.** *Категориальная грамматика $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$ называется $(m,*)$-жесткой, если для любых двух слов $u = w_1 a w_2$, $v = w_1' a w_2$, в которых символ $a$ стоит в одном и том же $(m,*)$-контексте, для любых последовательностей категорий $\Gamma_1 \in \delta(w_1)$, $\Gamma_1' \in \delta(w_1')$, $\Gamma_2, \Gamma_2' \in \delta(w_2)$ и для любых двух категорий $\Phi, \Psi \in \delta(a)$ из $\Gamma_1 \Phi \Gamma_2 \vdash^* S$ и $\Gamma_1' \Psi \Gamma_2' \vdash^* S$ следует $\Phi = \Psi$.*

Таким образом, грамматика является $(m,*)$-жесткой, если категория любого символа однозначно определяется символом, а также его $m$ левыми и всеми правыми соседями. Аналогично определяются $(*,n)$-жесткие грамматики.

**Определение 12.** *Категориальная грамматика $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$ называется $(*,n)$-жесткой, если для любых двух слов $u = w_1 a w_2$, $v = w_1 a w_2'$, в которых символ $a$ стоит в одном и том же $(*,n)$-контексте, для любых последовательностей категорий $\Gamma_1, \Gamma_1' \in \delta(w_1)$, $\Gamma_2 \in \delta(w_2)$, $\Gamma_2' \in \delta(w_2')$ и для любых двух категорий $\Phi, \Psi \in \delta(a)$ из $\Gamma_1 \Phi \Gamma_2 \vdash^* S$ и $\Gamma_1' \Psi \Gamma_2' \vdash^* S$ следует $\Phi = \Psi$.*

Наконец, можно еще сильнее ослабить ограничения и учитывать все символы, стоящие как слева, так и справа от текущего символа.

**Определение 13.** *Категориальная грамматика $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$ называется $(*,*)$-жесткой, если для любого слова $u = w_1 a w_2$, для любых последовательностей категорий $\Gamma_1, \Gamma_1' \in \delta(w_1)$, $\Gamma_2, \Gamma_2' \in \delta(w_2)$ и для любых двух категорий $\Phi, \Psi \in \delta(a)$ из $\Gamma_1 \Phi \Gamma_2 \vdash^* S$ и $\Gamma_1' \Psi \Gamma_2' \vdash^* S$ следует $\Phi = \Psi$.*

Фактически условие $(*,*)$-жесткости является условием однозначности: каждому слову из $L(G)$ сопоставляется единственная последовательность категорий.

Мы будем называть язык $L$ $(m,n)$-жестким, если он порождается некоторой $(m,n)$-жесткой грамматикой. Аналогично определяются $(m,*)$-жесткие, $(*,n)$-жесткие и $(*,*)$-жесткие языки. Приведенные в этом разделе определения различных типов жестких грамматик и языков непосредственно переносятся и на другие варианты категориальных грамматик.

## 2. Алгоритм для проверки $(m,n)$-жесткости грамматики

Мы опишем алгоритм, который по категориальной грамматике $G$ и натуральным числам $m$ и $n$ проверяет, является ли $G$ $(m,n)$-жесткой. Сначала дадим вспомогательные определения. Пусть зафиксировано произвольное конечное множество категорий $C \subseteq \mathrm{Cat}(\mathbf{C})$. Обозначим через $\mathrm{Subcat}(C)$ множество всех подкатегорий категорий из множества $C$. Так как множество $C$ конечно, то множество $\mathrm{Subcat}(C)$ также конечно.

**Определение 14.** *Через $I(C)$ обозначим множество всех категорий, которые могут получиться при сокращении некоторой последовательности категорий из множества $C^*$:*

$$I(C) = \{\, \Phi \in \mathrm{Cat}(\mathbf{C}) \mid \text{существует строка } \Gamma \in C^* \text{ такая, что } \Gamma \vdash^* \Phi \,\}.$$

**Определение 15.** *Через $L_n(C)$ обозначим множество пар вида $(\Phi_1 \dots \Phi_n, \Psi)$ таких, что $\Phi_1 \dots \Phi_n$ можно сократить до $\Psi$, приписав справа некоторые категории из множества $C$:*

$$L_n(C) = \{\, (\Phi_1 \dots \Phi_n, \Psi) \mid \Phi_i \in C, \Psi \in \mathrm{Cat}(\mathbf{C}), \text{ существует строка } \Gamma \in C^* \text{ такая, что } \Phi_1 \dots \Phi_n \Gamma \vdash^* \Psi \,\}.$$

**Определение 16.** *Через $R_n(C)$ обозначим множество пар вида $(\Phi_1 \dots \Phi_n, \Psi)$ таких, что $\Phi_1 \dots \Phi_n$ можно сократить до $\Psi$, приписав слева некоторые категории из множества $C$:*

$$R_n(C) = \{\, (\Phi_1 \dots \Phi_n, \Psi) \mid \Phi_i \in C, \Psi \in \mathrm{Cat}(\mathbf{C}),\ \text{существует строка } \Gamma \in C^* \\ \text{такая, что } \Gamma\Phi_1 \dots \Phi_n \vdash^* \Psi \,\}.$$

**Определение 17.** *Через $M_n(C)$ обозначим множество пар вида $(\Phi_1 \dots \Phi_n, \Psi)$ таких, что $\Phi_1 \dots \Phi_n$ можно сократить до $\Psi$, приписав слева и справа некоторые категории из множества $C$:*

$$M_n(C) = \{\, (\Phi_1 \dots \Phi_n, \Psi) \mid \Phi_i \in C, \Psi \in \mathrm{Cat}(\mathbf{C}),\ \text{существуют строки } \Gamma_1, \Gamma_2 \in C^* \\ \text{такие, что } \Gamma_1\Phi_1 \dots \Phi_n\Gamma_2 \vdash^* \Psi \,\}.$$

Поскольку при сокращении последовательностей категорий из $C$ получаются подкатегории категорий из $C$, а множество $C$ конечно, то множества $I(C)$, $L_n(C)$, $R_n(C)$ и $M_n(C)$ тоже конечны. Из определений сразу следует, что $L_n(C) \subseteq M_n(c)$, $R_n(C) \subseteq M_n(C)$. При $n = 0$ последовательность $\Phi_1 \dots \Phi_n$ всегда пуста, поэтому можно отождествить $L_0(C)$, $R_0(C)$ и $M_0(C)$ с $I(C)$.

Для вычисления множества $I(C)$ построим последовательность множеств $I_0, I_1, \dots$ следующим образом:
$I_0 = C$;
$I_{i+1} = I_i \cup \{\, \Phi \mid [\Phi/\Psi], \Psi \in I_i$ или $[\Psi\backslash\Phi], \Psi \in I_i$ для некоторых $\Phi, \Psi \,\}$.

В следующей лемме сформулировано свойство множеств $I_i$.

**Лемма 2.** *$\Phi \in I_i$ для некоторого $i$ тогда и только тогда, когда $\Phi \in I(C)$.*

*Доказательство.* [$\Rightarrow$] Предположим, что $\Phi \in I_i$ для некоторого $i$. Выберем наименьшее из таких $i$. Нужно доказать, что существует строка категорий $\Gamma \in C^*$ такая, что $\Gamma \vdash^* \Phi$. Проведем доказательство индукцией по $i$.
*Базис индукции.* Если $i = 0$, то можно положить $\Gamma = \Phi$.
*Индукционный шаг.* Пусть $\Phi \in I_{i+1}$, $\Phi \notin I_i$. Это значит, что в множестве $I_i$ есть либо категории $[\Phi/\Psi], \Psi$, либо категории $[\Psi\backslash\Phi], \Psi$ для некоторого $\Psi$. В первом случае по индукционному предположению существуют строки категорий $\Gamma_1$ и $\Gamma_2$ такие, что $\Gamma_1 \vdash^* [\Phi/\Psi]$, $\Gamma_2 \vdash^* \Psi$. Во втором случае существуют строки категорий $\Gamma_1$ и $\Gamma_2$ такие, что $\Gamma_1 \vdash^* \Psi$, $\Gamma_2 \vdash^* [\Psi\backslash\Phi]$. В обоих случаях $\Gamma_1\Gamma_2 \vdash^* \Phi$.
[$\Leftarrow$] Пусть $\Phi \in I(C)$, т.е. существует строка категорий $\Gamma \in C^*$ такая, что $\Gamma \vdash^* \Phi$. Индукцией по $i$ докажем, что если $|\Gamma| \le i$, то $\Phi \in I_{i-1}$.
*Базис индукции.* Если $|\Gamma| = 1$ и $\Gamma \vdash^* \Phi$, то $\Gamma = \Phi$. Так как $\Gamma \in C^*$, то $\Phi \in I_0$.
*Индукционный шаг.* Пусть $\Gamma \vdash^* \Phi$ и $|\Gamma| = i + 1$. Рассмотрим последний шаг сокращения. Возможны два случая.
1) Вывод имеет вид $\Gamma \vdash^* [\Phi/\Psi]\Psi \vdash \Phi$. Значит, $\Gamma$ имеет вид $\Gamma_1\Gamma_2$, где $\Gamma_1 \vdash^* [\Phi/\Psi]$, $\Gamma_2 \vdash^* \Psi$. По индукционному предположению $[\Phi/\Psi] \in I_{i-1}$, $\Psi \in I_{i-1}$. Следовательно, $\Phi \in I_i$ по определению 14.
2) Вывод имеет вид $\Gamma \vdash^* \Psi[\Psi\backslash\Phi] \vdash \Phi$. Значит, $\Gamma$ имеет вид $\Gamma_1\Gamma_2$, где $\Gamma_1 \vdash^* \Psi$, $\Gamma_2 \vdash^* [\Psi\backslash\Phi]$. По индукционному предположению $\Psi \in I_{i-1}$, $[\Psi\backslash\Phi] \in I_{i-1}$. Снова $\Phi \in I_i$ по определению 14. $\square$

Поскольку $I_i \subseteq I_{i+1}$ и $I_i \subseteq I(C)$ для всех $i$, то существует такое число $k$, что $I_k = I_{k+1}$. По лемме 2 $I(C) = I_k$. Алгоритм для вычисления множества $I(C)$ состоит в следующем: вычислять последовательность $I_0, I_1, \ldots$ до первого $I_k$ такого, что $I_k = I_{k+1}$, после этого вернуть $I_k$.

Теперь опишем метод нахождения множеств $L_n(C)$. Множества $L_n(C)$ вычисляются индукцией по $n$. Согласно замечанию после определения 17 $L_0(C) = I(C)$. Для $n > 0$ построим последовательность множеств $L_n^{(0)}, L_n^{(1)}, \ldots$ следующим образом:

$$\begin{aligned}
L_n^{(0)} = {} & \{ (\Phi_1 \ldots \Phi_n, \Psi) \mid \Phi_i \in C, \Psi \in \mathrm{Cat}(\mathbf{C}), \Phi_1 \ldots \Phi_n \vdash^* \Psi \} \cup \\
& \cup \bigcup_{j=1}^{n} \{ (\Phi_1 \ldots \Phi_n, \Psi) \mid \Phi_1 \ldots \Phi_j \vdash^* [\Psi/\Theta], (\Phi_{j+1} \ldots \Phi_n, \Theta) \in L_{n-j}(C) \\
& \quad \text{или } \Phi_1 \ldots \Phi_j \vdash^* \Theta, (\Phi_{j+1} \ldots \Phi_n, [\Theta\backslash\Psi]) \in L_{n-j}(C) \text{ для некоторого } \Theta \}; \\
L_n^{(i+1)} = {} & L_n^{(i)} \cup \{ (\Phi_1 \ldots \Phi_n, \Psi) \mid (\Phi_1 \ldots \Phi_n, [\Psi/\Theta]) \in L_n^{(i)}, \Theta \in I(C) \\
& \quad \text{или } (\Phi_1 \ldots \Phi_n, \Theta) \in L_n^{(i)}, [\Theta\backslash\Psi] \in I(C) \text{ для некоторого } \Theta \}
\end{aligned}$$

В следующей лемме сформулировано свойство множеств $L_n^{(i)}$.

**Лемма 3.** $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n^{(i)}$ *для некоторого* $i$ *тогда и только тогда, когда* $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n(C)$.

*Доказательство.* $[\Rightarrow]$ Пусть $(\Phi_1 \ldots \Phi_n, \Psi) \in L^{(i)}$. Выберем наименьшее из таких $i$. Нужно доказать, что существует строка категорий $\Gamma$ такая, что $\Phi_1 \ldots \Phi_n\Gamma \vdash^* \Psi$. Проведем индукцию по $i$.
*Базис индукции.* Пусть $i = 0$. Если $\Phi_1 \ldots \Phi_n \vdash^* \Psi$, то можно положить $\Gamma = \varepsilon$. Пусть существует такое $j$, что либо $\Phi_1 \ldots \Phi_j \vdash^* [\Psi/\Theta]$, $(\Phi_{j+1} \ldots \Phi_n, \Theta) \in L_{n-j}(C)$, либо $\Phi_1 \ldots \Phi_j \vdash^* \Theta$, $(\Phi_{j+1} \ldots \Phi_n, [\Theta\backslash\Psi]) \in L_{n-j}(C)$. Значит, существует строка категорий $\Gamma$ такая, что $\Phi_{j+1} \ldots \Phi_n\Gamma \vdash^* \Theta$ в первом случае и $\Phi_{j+1} \ldots \Phi_n\Gamma \vdash^* [\Theta\backslash\Psi]$ во втором случае. В обоих случаях $\Phi_1 \ldots \Phi_n\Gamma \vdash^* \Psi$.
*Индукционный шаг.* Пусть $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n^{(i+1)}$, $(\Phi_1 \ldots \Phi_n, \Psi) \notin L_n^{(i)}$. Это значит, что либо $(\Phi_1 \ldots \Phi_n, [\Psi/\Theta]) \in L_n^{(i)}$, $\Theta \in I(C)$, либо $(\Phi_1 \ldots \Phi_n, \Theta) \in L_n^{(i)}$, $[\Theta\backslash\Psi] \in I(C)$ для некоторого $\Theta$. В первом случае по индукционному предположению существует строка категорий $\Gamma_1$ такая, что $\Phi_1 \ldots \Phi_n\Gamma_1 \vdash^* [\Psi/\Theta]$, а по определению $I(C)$ существует строка категорий $\Gamma_2$ такая, что $\Gamma_2 \vdash^* \Theta$. Поэтому $\Phi_1 \ldots \Phi_n\Gamma_1\Gamma_2 \vdash^* \Psi$. Во втором случае по индукционному предположению существует строка категорий $\Gamma_1$ такая, что $\Phi_1 \ldots \Phi_n\Gamma_1 \vdash^* \Theta$, а по определению $I(C)$ существует строка категорий $\Gamma_2$ такая, что $\Gamma_2 \vdash^* [\Theta\backslash\Psi]$. Поэтому снова $\Phi_1 \ldots \Phi_n\Gamma_1\Gamma_2 \vdash^* \Psi$.
$[\Leftarrow]$ Пусть $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n(C)$, т.е. существует строка категорий $\Gamma$ такая, что $\Phi_1 \ldots \Phi_n\Gamma \vdash^* \Psi$. Рассмотрим последний шаг сокращения. Предположим сначала, что существует такое $j$, что $1 \le j \le n$ и либо $\Phi_1 \ldots \Phi_j \vdash^* [\Psi/\Theta]$, $\Phi_{j+1} \ldots \Phi_n\Gamma \vdash^* \Theta$, либо $\Phi_1 \ldots \Phi_j \vdash^* \Theta$, $\Phi_{j+1} \ldots \Phi_n\Gamma \vdash^* [\Theta\backslash\Psi]$. В первом случае $(\Phi_{j+1} \ldots \Phi_n, \Theta) \in L_{n-j}(C)$, а во втором случае $(\Phi_{j+1} \ldots \Phi_n, [\Theta\backslash\Psi]) \in L_{n-j}(C)$. В обоих случаях $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n^{(0)}$.

Теперь предположим, что $\Gamma$ имеет вид $\Gamma_1\Gamma_2$ и при этом либо $\Phi_1 \ldots \Phi_n\Gamma_1 \vdash^* [\Psi/\Theta]$, $\Gamma_2 \vdash^* \Theta$, либо $\Phi_1 \ldots \Phi_n\Gamma_1 \vdash^* \Theta$, $\Gamma_2 \vdash^* [\Theta\backslash\Psi]$ для некоторого $\Theta$. Индукцией по $i$ докажем, что если $|\Gamma| \le i$, то $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n^{(i)}$.
*Базис индукции.* Если строка $\Gamma$ пуста, то $\Phi_1 \ldots \Phi_n \vdash^* \Psi$ и $(\Phi_1 \ldots \Phi_n, \Psi) \in L_n^{(0)}$.
*Индукционный шаг.* Пусть $|\Gamma| = i+1$. Если $\Phi_1 \ldots \Phi_n\Gamma_1 \vdash^* [\Psi/\Theta]$, $\Gamma_2 \vdash^* \Theta$, то по ин-

дукционному предположению $(\Phi_1,\dots\Phi_n,[\Psi/\Theta])\in L_n^{(i)}$, а по определению множества $I(C)$ имеет место $\Theta\in I(C)$. Поэтому по определению $(\Phi_1\dots\Phi_n,\Psi)\in L_n^{(i+1)}$. Если же $\Phi_1\dots\Phi_n\Gamma_1\vdash^*\Theta$, $\Gamma_2\vdash^*[\Theta\backslash\Psi]$, то по индукционному предположению $(\Phi_1\dots\Phi_n,\Theta)\in L_n^{(i)}$, а по определению множества $I(C)$ имеет место $[\Theta\backslash\Psi]\in I(C)$. Поэтому снова по определению $(\Phi_1\dots\Phi_n,\Psi)\in L_n^{(i+1)}$. □

Поскольку $L_n^{(i)}\subseteq L_n^{(i+1)}$ и $L_n^{(i)}\subseteq\{\,(\Phi_1\dots\Phi_n,\Psi)\mid\Phi_i\in C,\Psi\in\mathrm{Subcat}(C)\,\}$ для всех $i$, то существует такое $k$, что $L_n^{(k)}=L_n^{(k+1)}$. По лемме 3 $L_n(C)=L_n^{(k)}$. Алгоритм для вычисления множества $L_n(C)$ состоит в следующем: вычислять последовательность $L_n^{(0)},L_n^{(1)},\dots$ до первого $L_n^{(k)}$ такого, что $L_n^{(k)}=L_n^{(k+1)}$, после этого вернуть $L_n^{(k)}$.

Алгоритмы для вычисления множеств $R_n(C)$ и $M_n(C)$ аналогичны алгоритму для вычисления $L_n(C)$. Мы определим последовательности множеств и сформулируем их свойства.

Согласно замечанию после определения 17 $R_0(C)=I(C)$ и $M_0(C)=I(C)$. Для вычисления множества $R_n(C)$ при $n>0$ построим последовательность множеств $R_n^{(0)},R_n^{(1)},\dots$ следующим образом:

$$R_n^{(0)}=\{\,(\Phi_1\dots\Phi_n,\Psi)\mid\Phi_i\in C,\Psi\in\mathrm{Cat}(\mathbf{C}),\Phi_1\dots\Phi_n\vdash^*\Psi\,\}\cup$$
$$\cup\bigcup_{j=1}^{n}\{\,(\Phi_1\dots\Phi_n,\Psi)\mid(\Phi_1\dots\Phi_j,[\Psi/\Theta])\in R_{n-j}(C),\Phi_{j+1}\dots\Phi_n\vdash^*\Theta$$
$$\text{или }(\Phi_1\dots\Phi_j,\Theta)\in R_{n-j}(C),\Phi_{j+1}\dots\Phi_n\vdash^*[\Theta\backslash\Psi]\text{ для некоторого }\Theta\,\};$$
$$R_n^{(i+1)}=R_n^{(i)}\cup\{\,(\Phi_1\dots\Phi_n,\Psi)\mid(\Phi_1\dots\Phi_n,[\Theta\backslash\Psi])\in R_n^{(i)},\Theta\in I(C)$$
$$\text{или }(\Phi_1\dots\Phi_n,\Theta)\in R_n^{(i)},[\Psi/\Theta]\in I(C)\text{ для некоторого }\Theta\,\}.$$

Для вычисления множества $M_n(C)$ при $n>0$ построим последовательность множеств $M_n^{(0)},M_n^{(1)},\dots$ следующим образом:

$$M_n^{(0)}=L_n(C)\cup R_n(C)\cup$$
$$\cup\bigcup_{j=0}^{n}\{\,(\Phi_1\dots\Phi_n,\Psi)\mid(\Phi_1\dots\Phi_j,[\Psi/\Theta])\in R_j(C),(\Phi_{j+1}\dots\Phi_n,\Theta)\in L_{n-j}(C)$$
$$\text{или }(\Phi_1\dots\Phi_j,\Theta)\in R_j(C),(\Phi_{j+1}\dots\Phi_n,[\Theta\backslash\Psi])\in L_{n-j}(C)\text{ для некоторого }\Theta\,\};$$
$$M_n^{(i+1)}=M_n^{(i)}\cup\{\,(\Phi_1\dots\Phi_n,\Psi)\mid(\Phi_1\dots\Phi_n,[\Psi/\Theta])\in M_n^{(i)},\Theta\in I(C),$$
$$\text{или }(\Phi_1\dots\Phi_n,[\Theta\backslash\Psi])\in M_n^{(i)},\Theta\in I(C),$$
$$\text{или }(\Phi_1\dots\Phi_n,\Theta)\in M_n^{(i)},[\Theta\backslash\Psi]\in I(C),$$
$$\text{или }(\Phi_1\dots\Phi_n,\Theta)\in M_n^{(i)},[\Psi/\Theta]\in I(C)\text{ для некоторого }\Theta\,\}.$$

В следующих двух леммах сформулированы утверждения о множествах $R_n^{(i)}$ и $M_n^{(i)}$, аналогичные лемме 3.

**Лемма 4.** $(\Phi_1\dots\Phi_n,\Psi)\in R_n^{(i)}$ *для некоторого* $i$ *тогда и только тогда, когда* $(\Phi_1\dots\Phi_n,\Psi)\in R_n(C)$.

**Лемма 5.** $(\Phi_1\dots\Phi_n,\Psi)\in M_n^{(i)}$ *для некоторого* $i$ *тогда и только тогда, когда* $(\Phi_1\dots\Phi_n,\Psi)\in M_n(C)$.

Как и для множеств $L_n^{(i)}$, существуют числа $k$ и $l$ такие, что $R_n^{(k)}=R_n^{(k+1)}$ и $M_n^{(l)}=M_n^{(l+1)}$. Тогда $R_n(C)=R_n^{(k)}$ и $M_n(C)=M_n^{(l)}$.

Теперь мы опишем алгоритм для проверки жесткости грамматики.

**Алгоритм ЖЕСТКАЯ_ГРАММАТИКА.**
*Вход*: классическая категориальная грамматика $G = \langle \Sigma, \mathbf{C}, S, \delta \rangle$, натуральные числа $m$ и $n$.
*Выход*: «да», если $G$ является $(m,n)$-жесткой; «нет» в противном случае.

1. Пусть $C$ — множество всех категорий, приписываемых грамматикой $G$ некоторому символу: $C = \{ \delta(a) \mid a \in \Sigma \}$.

2. Вычислить множества $L_0(C), L_1(C), \dots, L_{n+1}(C)$, $R_0(C), R_1(C), \dots, R_{m+1}(C)$, $M_{m+n+1}(C)$.

3. Для каждого слова $w = b_1 \dots b_m a c_1 \dots c_n$ проверить, что категория, приписываемая символу $a$, определяется однозначно. Если существуют две строки категорий $\Phi_1 \dots \Phi_{m+n+1} \in \delta(w)$, $\Phi'_1 \dots \Phi'_{m+n+1} \in \delta(w)$ такие, что $\Phi_{m+1} \neq \Phi'_{m+1}$, $(\Phi_1 \dots \Phi_{m+n+1}, S) \in M_{m+n+1}(C)$, $(\Phi'_1 \dots \Phi'_{m+n+1}, S) \in M_{m+n+1}(C)$, то вернуть «нет».

4. Для каждого слова $w = b_1 \dots b_k a c_1 \dots c_n$, где $k < m$, проверить, что категория символа $a$ определяется однозначно при условии, что перед словом $w$ нет других символов. Если существуют две строки категорий $\Phi_1 \dots \Phi_{k+n+1} \in \delta(w)$, $\Phi'_1 \dots \Phi'_{k+n+1} \in \delta(w)$ такие, что $\Phi_{k+1} \neq \Phi'_{k+1}$, $(\Phi_1 \dots \Phi_{k+n+1}, S) \in L_{k+n+1}(C)$, $(\Phi'_1 \dots \Phi'_{k+n+1}, S) \in L_{k+n+1}(C)$, то вернуть «нет».

5. Для каждого слова $w = b_1 \dots b_m a c_1 \dots c_k$, где $k < n$, проверить, что категория символа $a$ определяется однозначно при условии, что после слова $w$ нет других символов. Если существуют две строки категорий $\Phi_1 \dots \Phi_{m+k+1} \in \delta(w)$, $\Phi'_1 \dots \Phi'_{m+k+1} \in \delta(w)$ такие, что $\Phi_{m+1} \neq \Phi'_{m+1}$, $(\Phi_1 \dots \Phi_{m+k+1}, S) \in R_{m+k+1}(C)$, $(\Phi'_1 \dots \Phi'_{m+k+1}, S) \in R_{m+k+1}(C)$, то вернуть «нет».

6. Для всех контекстов, содержащих менее $m$ левых символов и менее $n$ правых символов проверить однозначность выбора категорий прямым перебором. Если для некоторого слова существуют две различные последовательности категорий, сокращающиеся до $S$, то вернуть «нет».

7. Вернуть «да».

Докажем корректность описанного алгоритма.

**Теорема 1.** *Алгоритм ЖЕСТКАЯ_ГРАММАТИКА возвращает «да» тогда и только тогда, когда грамматика $G$ является $(m,n)$-жесткой.*

*Доказательство.* [$\Rightarrow$] Пусть грамматика $G$ не является $(m,n)$-жесткой. Тогда существуют два слова $u$ и $v$ из языка $L(G)$, в которых некоторый символ $a$ стоит в одном и том же $(m,n)$-контексте, но получает разные категории. Возможны четыре случая:
1) и слева, и справа от символа $a$ имеется достаточное количество других символов (не менее $m$ слева и не менее $n$ справа, пункт 3 алгоритма);
2) число символов слева меньше $m$, а число символов справа больше или равно $n$ (пункт 4 алгоритма);

3) число символов слева больше или равно $m$, а число символов справа меньше $n$ (пункт 5 алгоритма);
4) число символов слева меньше $m$, а число символов справа меньше $n$ (пункт 6 алгоритма).

Рассмотрим первый случай (для остальных случаев доказательство аналогично). Пусть $u = w_1 b_1 \ldots b_m a c_1 \ldots c_n w_2$, $v = w'_1 b_1 \ldots b_m a c_1 \ldots c_n w'_2$. Пусть подслову $b_1 \ldots b_m a c_1 \ldots c_n$ в слове $u$ сопоставлена строка категорий $\Phi_1 \ldots \Phi_{m+n+1}$, а в слове $v$ — строка категорий $\Phi'_1 \ldots \Phi'_{m+n+1}$. По предположению $\Phi_{m+1} \neq \Phi'_{m+1}$. Поскольку оба слова принадлежат языку $L(G)$, то существуют строки категорий $\Gamma_1 \in \delta(w_1)$, $\Gamma_2 \in \delta(w_2)$, $\Gamma'_1 \in \delta(w'_1)$, $\Gamma'_2 \in \delta(w'_2)$ такие, что $\Gamma_1 \Phi_1 \ldots \Phi_{m+n+1} \Gamma_2 \vdash^* S$, $\Gamma'_1 \Phi'_1 \ldots \Phi'_{m+n+1} \Gamma'_2 \vdash^* S$. Но тогда по определению множества $M_{m+n+1}(C)$ имеет место $(\Phi_1 \ldots \Phi_{m+n+1}, S) \in M_{m+n+1}(C)$, $(\Phi'_1 \ldots \Phi'_{m+n+1}, S) \in M_{m+n+1}(C)$. Алгоритм обнаружит это на шаге 3 и вернет «нет».
[$\Leftarrow$] Пусть грамматика является $(m, n)$-жесткой. Докажем, что на шагах 3–6 алгоритм не вернет «нет». Рассмотрим пункт 4 алгоритма (для остальных случаев доказательство аналогично). Если алгоритм возвращает «нет», то существует слово $w = b_1 \ldots b_k a c_1 \ldots c_n$, где $k < m$, и две строки категорий $\Phi_1 \ldots \Phi_{k+n+1} \in \delta(w)$, $\Phi'_1 \ldots \Phi'_{k+n+1} \in \delta(w)$, такие что $\Phi_{k+1} \neq \Phi'_{k+1}$, $(\Phi_1 \ldots \Phi_{k+n+1}, S) \in L_{k+n+1}(C)$, $(\Phi'_1 \ldots \Phi'_{k+n+1}, S) \in L_{k+n+1}(C)$. Из определения $L_{k+n+1}(C)$ следует, что существуют две строки категорий $\Gamma$ и $\Gamma'$ такие, что $\Phi_1 \ldots \Phi_{k+n+1} \Gamma \vdash^* S$, $\Phi'_1 \ldots \Phi'_{k+n+1} \Gamma' \vdash^* S$. Поскольку $C$ содержит только те категории, которые приписаны грамматикой некоторым символам, то строки $\Gamma$ и $\Gamma'$ соответствуют некоторым словам $v$ и $v'$. Значит, слова $wv$ и $wv'$ принадлежат языку $L(G)$, но символ $a$ получает разные категории в одном и том же $(m, n)$-контексте. Это противоречит тому, что грамматика $(m, n)$-жесткая.

Итак, на шагах 3–6 алгоритм не возвращает «нет». Следовательно, он дойдет до шага 7 и вернет «да». □

Заметим, что в описанном алгоритме существенно используется тот факт, что в процессе сокращения категорий в классических категориальных грамматиках может получаться лишь конечное множество категорий. Поэтому описанный алгоритм нельзя непосредственно перенести на комбинаторные категориальные грамматики или на категориальные грамматики зависимостей.

## 3. Свойства $(m, n)$-жестких грамматик

В этом разделе мы исследуем некоторые свойства $(m, n)$-жестких грамматик. Сначала мы докажем, что для $(m, n)$-жестких языков, порождаемых грамматиками с категориями первого порядка, существует бесконечная иерархия.

**Теорема 2.** *1) Для любых натуральных чисел $m$ и $n$ существует конечный язык, порождаемый $(m, n+1)$-жесткой грамматикой, но не порождаемый никакой $(m, n)$-жесткой грамматикой с категориями первого порядка.*
*2) Для любых натуральных чисел $m$ и $n$ существует конечный язык, порождаемый $(m+1, n)$-жесткой грамматикой, но не порождаемый никакой $(m, n)$-жесткой грамматикой с категориями первого порядка.*

*Доказательство.* Пусть $L = \{ a^{m+n+1}, a^{m+n+2} \}$. Припишем символу $a$ следующие категории: $[S/A_1], [A_1/A_2], \ldots, [A_{m+n-1}/A_{m+n}], [A_{m+n}]$,
$[S/B_1], [B_1/B_2], \ldots, [B_{m+n}/B_{m+n+1}], [B_{m+n+1}]$.

Эта грамматика является как $(m, n+1)$-жесткой, так и $(m+1, n)$-жесткой, поскольку каждый символ $a$ имеет уникальный $(m, n+1)$-контекст и уникальный $(m+1, n)$-контекст. Докажем, что язык $L$ не порождается никакой $(m,n)$-жесткой грамматикой. Предположим противное: пусть $(m,n)$-жесткая грамматика $G = \langle \{ a \}, \mathbf{C}, S, \delta \rangle$ порождает язык $L$. Слово $a^{m+n+1}$ принадлежит языку $L$, следовательно, существует строка категорий $\Gamma \in \delta(a^{m+n+1})$ такая, что $\Gamma \vdash^* S$. Обозначим через $\Phi$ категорию, сопоставленную $(m+1)$-му символу $a$, так что $\Gamma = \Gamma_1 \Phi \Gamma_2$.

Рассмотрим слово $a^{m+n+2}$. Его $(m+1)$-й и $(m+2)$-й символы стоят в том же $(m,n)$-контексте, что и $(m+1)$-й символ слова $a^{m+n+1}$, а значит, им сопоставлена категория $\Phi$ в силу того, что грамматика $(m,n)$-жесткая. Поэтому слову $a^{m+n+2}$ сопоставляется строка категорий $\Gamma_1 \Phi \Phi \Gamma_2$. Поскольку обе строки $\Gamma_1 \Phi \Gamma_2$ и $\Gamma_1 \Phi \Phi \Gamma_2$ сокращаются до $S$, то по лемме 1 $[\![\Gamma_1 \Phi \Gamma_2]\!] = [\![\Gamma_1 \Phi \Phi \Gamma_2]\!] = [\![S]\!]$. Отсюда следует, что $[\![\Phi]\!] = [\![\varepsilon]\!]$. Поскольку грамматика содержит только категории первого порядка, то либо $\Phi = [A/A]$, либо $\Phi = [A\backslash A]$ для некоторого $A \in \mathbf{C}$. Рассмотрим случай $\Phi = [A/A]$. Так как $\Gamma_1 [A/A] \Gamma_2 \vdash^* S$, а все категории являются категориями первого порядка, то $\Gamma_2$ можно представить в виде $\Gamma_2' \Gamma_2''$, так что $\Gamma_2' \vdash^* A$, а сокращение можно выполнять следующим образом: $\Gamma_1 [A/A] \Gamma_2' \Gamma_2'' \vdash^* \Gamma_1 [A/A] A \Gamma_2'' \vdash^* S$. Следовательно,

$$\Gamma_1 \Phi \Phi \Phi \Gamma_2 = \Gamma_1 [A/A][A/A][A/A] \Gamma_2' \Gamma_2'' \vdash^* \Gamma_1 [A/A][A/A][A/A] A \Gamma_2'' \vdash \Gamma_1 [A/A] A \Gamma_2'' \vdash^* S.$$

Но $\Gamma_1 \Phi \Phi \Phi \Gamma_2$ — строка категорий, сопоставляемая слову $a^{m+n+3}$. Поэтому $a^{m+n+3} \in L(G)$, что противоречит предположению $L(G) = L$.

Случай $\Phi = [A\backslash A]$ рассматривается аналогично. Следовательно, язык $L$ не порождается никакой $(m,n)$-жесткой грамматикой. $\square$

Теперь докажем, что левый (соответственно правый) контекст могут позволить определить категорию символа, когда ее невозможно определить с помощью правого (соответственно левого) контекста.

**Теорема 3.** *1) Существует конечный $(1,0)$-жесткий язык, не являющийся $(0,*)$-жестким.*
*2) Существует конечный $(0,1)$-жесткий язык, не являющийся $(*,0)$-жестким.*

*Доказательство.* Докажем первое утверждение теоремы. Пусть $L = \{ ab, b \}$. Этот язык порождается $(1,0)$-жесткой грамматикой с категориями $a \mapsto [S/A]$, $b \mapsto [A], [S]$. Предположим, что $(0,*)$-жесткая грамматика $G = \langle \{ a, b \}, \mathbf{C}, S, \delta \rangle$ порождает язык $L$. Поскольку $b \in L(G)$, то $S \in \delta(b)$. В слове $ab$ символ $b$ имеет тот же $(0,*)$-контекст, что и в слове $b$, поэтому ему сопоставляется та же категория $S$. Поскольку $ab \in L(G)$, то символу $a$ должна сопоставляться категория $[S/S]$. Но тогда $aab \in L(G)$, так как $[S/S][S/S]S \vdash^* S$. Получилось противоречие.

Доказательство второго утверждения аналогично. Примером является язык $L = \{ ba, b \}$. $\square$

Наконец, докажем теорему о соотношении класса регулярных языков и класса языков, порождаемыми $(m,n)$-жесткими грамматиками.

**Теорема 4.** *1) Всякий регулярный язык порождается некоторой* $(*, 1)$*-жесткой грамматикой и некоторой* $(1, *)$*-жесткой грамматикой с категориями первого порядка.*
*2) Существует регулярный язык, не порождаемый никакой* $(m, n)$*-жесткой грамматикой с категориями первого порядка.*
*3) Существует нерегулярный* $(1, 1)$*-жесткий язык.*

*Доказательство.* 1) Пусть $L$ — регулярный язык. Он распознается некоторым детерминированным конечным автоматом $M = \langle \Sigma, Q, q_0, F, \delta \rangle$. Построим категориальную грамматику $G = \langle \Sigma, Q, q_0, \lambda \rangle$ следующим образом: $\lambda(a) = \{ [p/q] \mid \delta(p, a) = q \} \cup \{ p \mid \delta(p, a) = q \text{ и } q \in F \}$. Непосредственной индукцией по длине слова $w$ доказывается, что $(p, w) \vdash^* (q, \varepsilon)$ для некоторого $q \in F$ тогда и только тогда, когда существует строка категорий $\Gamma \in \lambda(w)$ такая, что $\Gamma \vdash^* p$. Построенная грамматика является $(*, 1)$-жесткой. Пусть $a$ — некоторый символ в слове $w$. Символы символы $a_1 \dots a_k$, стоящие левее $a$, позволяют определить состояние $p$, в которое переходит автомат, прочитав строку $a_1 \dots a_k$, а правый символ позволяет определить, является ли символ $a$ последним. Если $a$ — последний символ, то ему сопоставляется категория $p$, в противном случае — категория $[p/\delta(p, a)]$. Единственность строки категорий следует из того, что автомат детерминированный.

$(1, *)$-жесткая грамматика строится аналогично. Пусть $M = \langle \Sigma, Q, q_0, F, \delta \rangle$ — детерминированный конечный автомат, распознающий язык $L^{-1}$ — обращение языка $L$. Тогда словарь $\lambda$ грамматики определяется следующим образом: $\lambda(a) = \{ [q \backslash p] \mid \delta(p, a) = q \} \cup \{ p \mid \delta(p, a) = q \text{ и } q \in F \}$

2) Пусть $L = \{ a^{2k} \mid k > 0 \}$ — множество всех непустых слов четной длины. Предположим, что $(m, n)$-жесткая грамматика $G = \langle \{ a \}, \mathbf{C}, S, \delta \rangle$ порождает язык $L$. Пусть $k$ — четное число большее $m + n$. Рассмотрим три слова $a^k$, $a^{k+1}$ и $a^{k+2}$. $(m + 1)$-й символ слова $a^k$, $(m + 1)$-й и $(m + 2)$-й символы слова $a^{k+1}$ и $(m+1)$-й, $(m+2)$-й и $(m+3)$-й символы слова $a^{k+2}$ стоят в одном и том же $(m, n)$-контексте, поэтому им сопоставляется одна и та же категория $\Phi$. $m$-буквенным префиксам всех трех слов сопоставляется одна и та же строка категорий $\Gamma_1$, а $(k - m - 1)$-буквенным суффиксам — одна и та же строка категорий $\Gamma_2$. Итак, словам сопоставляются строки $\Gamma_1 \Phi \Gamma_2$, $\Gamma_1 \Phi \Phi \Gamma_2$ и $\Gamma_1 \Phi \Phi \Phi \Gamma_2$, причем $\Gamma_1 \Phi \Gamma_2 \vdash^* S$ и $\Gamma_1 \Phi \Phi \Phi \Gamma_2 \vdash^* S$ . Как и в теореме 2 доказывается, что $\Phi$ имеет вид либо $[A/A]$, либо $[A \backslash A]$, а значит, $\Gamma_1 \Phi \Phi \Gamma_2 \vdash^* S$. Но это противоречит тому, что $L$ содержит только слова четной длины.

3) Язык $L = \{ a^k b^k \mid k > 0 \}$ порождается грамматикой со следующими категориями: $a \mapsto [S/B], [A/B]$, $b \mapsto [A \backslash B], [B]$. Первому символу $a$ приписывается категория $[S/B]$, остальным символам $a$ — категория $[A/B]$, первому символу $b$ — категория $[B]$, а остальным символам $b$ — категория $[A \backslash B]$. Категория каждого символа определяется его двумя соседями. □

## 4. О сложности анализа

Несмотря на то, что $(m, n)$-жесткие грамматики приписывают каждому слову единственную строку категорий, число возможных сокращений этой строки может быть экспоненциально большим.

**Теорема 5.** *Существует жесткая грамматика $G$ такая, что строка категорий, сопоставленная слову длины $n$, имеет $O(2^n/\sqrt{n})$ сокращений до $S$ .*

*Доказательство.* Язык $L = \{\, ab^k cd^l \mid k,l \geq 0 \,\}$ порождается следующей жесткой грамматикой:

$$a \mapsto [S/A] \qquad b \mapsto [A/A] \qquad c \mapsto [A] \qquad d \mapsto [A\backslash A]$$

Слову $b^k cd^l$ сопоставляется единственная последовательность категорий $\underbrace{[A/A]\dots[A/A]}_{k}\, A\, \underbrace{[A\backslash A]\dots[A\backslash A]}_{l}$. Для сокращения этой последовательности до $A$ требуется $k+l$ шагов. Поскольку на каждом шаге можно выполнять сокращение слева или справа, то число возможных сокращений равно $C^k_{k+l}$. При $k = l$ получаем $C^k_{2k} = O(4^k/\sqrt{k})$ вариантов (по формуле Стирлинга). Поэтому для слова $w = ab^k cd^k$ также существует $O(4^k/\sqrt{k})$ вариантов сокращения строки категорий. Если обозначить через $n$ длину слова $w$, то $k = n/2-1$. Следовательно, число сокращений составляет $O(4^{n/2-1}/\sqrt{n/2-1}) = O(2^n/\sqrt{n})$. □

Если требуется найти только одно сокращение до $S$, то можно использовать следующий алгоритм. Сначала входному слову сопоставляется строка категорий. Это можно сделать за время $O(n)$, так как категория каждого символа определяется контекстом фиксированного размера. После этого алгоритм сокращает получившуюся строку категорий как в алгоритме Кока-Янгера-Касами. Поскольку для каждого слова рассматривается только одна категория, то такой метод анализа рассмотрит меньше вариантов сокращения, чем алгоритм Кока-Янгера-Касами, примененный к произвольной грамматике. Описанную процедуру можно применить не только к классическим категориальным грамматикам, но и к другим их вариантам.

В [6,7] рассматривается одно из обобщений классических категориальных грамматик — категориальные грамматики зависимостей (КГЗ). Категории в КГЗ — это категории классических категориальных грамматик, к которым добавлены списки дальних зависимостей, называемые потенциалами. Потенциал задает дальние зависимости, выходящие из слова и входящие в слово (в естественных языках число входящих зависимостей не превосходит 1). Потенциал называется сбалансированным, если для каждой выходящей зависимости имеется соответствующая ей входящая и наоборот (точные определения приведены в [6, 7]). В этих же статьях доказана теорема об анализе: слово $w = a_1 \dots a_k \in \Sigma^+$ принадлежит языку $L(G)$ тогда и только тогда, когда существует строка категорий $\gamma_1^{\theta_1} \dots \gamma_k^{\theta_k}$ такая, что $\gamma_i^{\theta_i} \in \delta(a_i)$, $\gamma_1 \dots \gamma_k \vdash^* S$ и потенциал $\theta_1 \dots \theta_k$ сбалансирован. Проблема принадлежности для КГЗ в общем случае является NP-полной.

Для КГЗ можно определить понятие $(m,n)$-жесткой грамматики так же, как и для классических категориальных грамматик: КГЗ называется $(m,n)$-жесткой, если категория любого символа однозначно определяется самим символом, а также его $m$ левыми и $n$ правыми соседями. Для $(m,n)$-жестких КГЗ проблема принадлежности разрешима за полиномиальное время. Если КГЗ является $(m,n)$-жесткой, то строка категорий, приписываемая слову, оказывается единственной. Сокращение строки $\gamma_1 \dots \gamma_k$ можно выполнить за время $O(k^3)$ с помощью алгоритма Кока-Янгера-Касами, а сбалансированность потенциала проверяется за линейное время (это задача проверки того, правильно ли расставлены скобки в слове). Поэтому для $(m,n)$-жестких КГЗ существует алгоритм анализа, имеющий временную сложность $O(k^3)$.

## Заключение

В статье мы ввели обобщение понятия жестких категориальных грамматик — $(m,n)$-жесткие грамматики. В разделе 2 мы доказали алгоритмическую разрешимость следующей проблемы: по классической категориальной грамматике $G$ и натуральным числам $m$, $n$ определить, является ли грамматика $G$ $(m,n)$-жесткой. В разделе 3 мы исследовали некоторые свойства языков, порождаемых $(m,n)$-жесткими грамматиками. В частности, мы доказали, что существует бесконечная иерархия $(m,n)$-жестких языков, а также, что классы всех $(m,n)$-жестких языков и регулярных языков не сравнимы. В разделе 4 мы исследовали проблему принадлежности для $(m,n)$-жестких грамматик. Мы доказали, что для $(m,n)$-жестких категориальных грамматик зависимостей существует полиномиальный алгоритм анализа (общая проблема принадлежности для них является NP-полной).

Ряд вопросов, относящихся к $(m,n)$-жестким грамматикам, остаются открытыми. Перечислим некоторые из них.

1. Разрешима ли проблема определения $(m,n)$-жесткости для различных обобщений классических категориальных грамматик?

2. Разрешима ли проблема определения $(m,*)$-, $(*,n)$- и $(*,*)$-жесткости для классических категориальных грамматик? Мы предполагаем, что проблема определения $(*,*)$-жесткости неразрешима, так как она похожа на проблему однозначности для КС-грамматик.

3. Справедливы ли теоремы 2 и 4 для грамматик с категориями не только первого порядка?

## Список литературы

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

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

[3] Синтаксически размеченный корпус русского языка: информация для пользователей [Электронный ресурс]. 2003–2017. URL: http://www.ruscorpora.ru/instruction-syntax.html (дата обращения 26.10.2017).

[4] Bar-Hillel Y. A quasi-arithmetical notation for syntactic description // Language. 1953. Vol. 29, № 1. Pp. 47–58.

[5] Bar-Hillel Y., Gaifman H., Shamir E. On categorial and phrase structure grammars // Bulletin of the Research Council of Israel. 1960. Vol. 9F. Pp. 1–16.

[6] Dekhtyar M., Dikovsky A. Generalized categorial dependency grammars // In: Pillars of Computer Science. Ed. by A. Avron, N. Dershowitz, A. Rabinovich. Series: Lecture Notes in Computer Science. Vol. 4800. Berlin, Heidelberg: Springer, 2008. Pp. 230–255. https://doi.org/10.1007/978-3-540-78127-1_13

[7] Dekhtyar M., Dikovsky A., Karlov B. Categorial dependency grammars // Theoretical Computer Science. 2015. Vol. 579. Pp. 33–63. http://dx.doi.org/10.1016/j.tcs.2015.01.043

[8] Kallmeyer L. Parsing Beyond Context-Free Grammars. Series: Cognitive Technologies. Berlin, Heidelberg: Springer-Verlag, 2010. https://doi.org/10.1007/978-3-642-14846-0

[9] Kanazawa M. Learnable Classes of Categorial Grammars. Series: Studies in Logic, Language, and Information. FoLLI & CSLI, 1998.

[10] Pentus M. Equivalent types in Lambek calculus and linear logic // Препринт №2 отдела математической логики математического института им. В.А. Стеклова РАН. Серия Математическая логика и теоретическая информатика. М., 1992. 21 с.

[11] Vijay-Shanker K., Weir D.J. The equivalence of four extensions of context-free grammars // Mathematical Study Theory. 1994. Vol. 27. Pp. 511–545.

## Образец цитирования

Карлов Б.Н. $(m, n)$-жесткие категориальные грамматики // Вестник ТвГУ. Серия: Прикладная математика. 2017. № 4. С. 7–23. https://doi.org/10.26456/vtpmk185

## Сведения об авторах

1. **Карлов Борис Николаевич**
доцент кафедры информатики Тверского государственного университета.

*Россия, 170100, г. Тверь, ул. Желябова, д. 33, ТвГУ.*
*E-mail: bnkarlov@gmail.com.*

# $(m, n)$-RIGID CATEGORIAL GRAMMARS

**Karlov Boris Nikolaevich**
Associate professor at Computer Science department,
Tver State University
*Russia, 170100, Tver, 33 Zhelyabova str., TverSU.*
*E-mail: bnkarlov@gmail.com*

*Received 05.11.2017, revised 02.12.2017.*

In the article $(m, n)$-rigid categorial grammars are defined. An algorithm is proposed that verifies for a given grammar $G$ and natural numbers $m$, $n$ whether $G$ is $(m, n)$-rigid. It is proved that an infinite hierarchy exists in the class of $(m, n)$-rigid languages, and that the class of $(m, n)$-rigid grammars is incomparable with the class of regular languages. The complexity of the membership problem for $(m, n)$-rigid grammars is studied.

**Keywords:** formal grammars, categorial grammars, rigid grammars, algorithm for rigidity checking, hierarchy of rigid languages, membership problem.

## Citation

Karlov B.N. $(m, n)$-rigid categorial grammars. *Vestnik TvGU. Seriya: Prikladnaya Matematika* [Herald of Tver State University. Series: Applied Mathematics], 2017, no. 4, pp. 7–23. (in Russian). https://doi.org/10.26456/vtpmk185

## References

[1] Aho A., Ulman D. *Teoriya Sintaksicheskogo Analiza, Perevoda i Kompilyatsii* [The Theory of Syntactic Analysis, Translation and Compilation]. Vol. 1. Mir Publ., Moscow, 1978. (in Russian)

[2] Gladkii A.V. *Formalnye Grammatiki i Yazyki* [Formal Grammars and Languages]. Nauka Publ., Moscow, 1973. (in Russian)

[3] *Syntactically marked corpus of the Russian language: information for users* [Electronic resource ]. 2003–2017. URL: http://www.ruscorpora.ru/instruction-syntax.html (accessed at 26.10.2017). (in Russian)

[4] Bar-Hillel Y. A quasi-arithmetical notation for syntactic description. *Language*, 1953, vol. 29(1), pp. 47–58.

[5] Bar-Hillel Y., Gaifman H., Shamir E. On categorial and phrase structure grammars. *Bulletin of the Research Council of Israel*, 1960, vol. 9F, pp. 1–16.

[6] Dekhtyar M., Dikovsky A. Generalized categorial dependency grammars. In: *Pillars of Computer Science*. Ed. by A. Avron, N. Dershowitz, A. Rabinovich. Series: Lecture Notes in Computer Science. Vol. 4800. Springer, Berlin, Heidelberg, 2008. Pp. 230–255. https://doi.org/10.1007/978-3-540-78127-1_13

[7] Dekhtyar M., Dikovsky A., Karlov B. Categorial dependency grammars. *Theoretical Computer Science*, 2015, vol. 579, pp. 33–63. http://dx.doi.org/10.1016/j.tcs.2015.01.043

[8] Kallmeyer L. *Parsing Beyond Context-Free Grammars*. Series: Cognitive Technologies. Springer-Verlag, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-14846-0

[9] Kanazawa M. *Learnable Classes of Categorial Grammars*. Series: Studies in Logic, Language, and Information. FoLLI & CSLI, 1998.

[10] Pentus M. *Equivalent Types in Lambek Calculus and Linear Logic*. Preprint No. 2 of the Department of Mathematical Logic of the Math. V.A. Steklov Institute of RAS. Series: Mathematical Logic and Theoretical Informatics. Moscow, 1992. 21 p. (in Russian)

[11] Vijay-Shanker K., Weir D.J. The equivalence of four extensions of context-free grammars. *Mathematical Study Theory*, 1994, vol. 27, pp. 511–545.