На правах рукописи

Карлов Борис Николаевич

О классах категориальных грамматик зависимостей

Специальность 01.01.06 — математическая логика, алгебра и теория чисел

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук.

Ярославль — 2012

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Тверской государственный университет».

Научный руководитель — доктор физико-математических наук, доцент Дехтярь Михаил Иосифович

Официальные оппоненты: Соколов Валерий Анатольевич, доктор физико-математических наук, профессор, заведующий кафедрой теоретической информатики Ярославского государственного университета им. П. Г. Демидова

Валиев Марс Котдусович, кандидат физико-математических наук, старший научный сотрудник Института прикладной математики им. М. В. Келдыша РАН

Ведущая организация — Московский государственный университет им. М. В. Ломоносова

Защита диссертации состоится 23 ноября 2012 г. в 14.00 на заседании диссертационного совета Д 212.002.03 при Ярославском государственном университете им. П. Г. Демидова по адресу: Российская Федерация, 150008, Ярославль, ул. Союзная, 144, ауд. 426.

С диссертацией можно ознакомиться в библиотеке Ярославского государственного университета им. П. Г. Демидова.

Автореферат разослан «___» октября 2012 г.

Учёный секретарь диссертационного совета — Яблокова Светлана Ивановна

Общая характеристика работы

Актуальность. Формальные способы описания синтаксической структуры предложения имеют первостепенную важность для большинства задач информатики, связанных с обработкой информации на естественном языке. После основополагающих работ Н. Хомского, определившего четыре базовых класса порождающих грамматик, был определен еще целый ряд типов грамматик, позволяющих вычислять синтаксическую структуру предложения в ходе его вывода или доказательства его правильности. В частности, *грамматики зависимостей* (специальный тип грамматик) присваивают структуры зависимостей (структуры подчинения) предложениям языка, который они определяют. Теории синтаксиса естественных языков, основанные на понятии *зависимости*, имеют давнюю традицию, восходящую к средним векам. Теньер впервые систематически описал структуру предложения в терминах именованных бинарных отношений между словами (*зависимостей*). Когда два слова w_1 и w_2 связаны в предложении посредством зависимости d (обозначение $w_1 \xrightarrow{d} w_2$), w_1 является *главным словом*, а w_2 — *зависимым словом*. Содержательно, зависимость d задаёт ограничения на грамматические и лексические свойства w_1 и w_2, на их порядок, контекст и т.п., которые вместе означают, что "w_1 управляет w_2". Например, в структуре зависимостей предложения *"Летом здесь играют дети"*, приведённой на рис. 1, отношение *играют* $\xrightarrow{пред}$ *дети* показывает предикатную зависимость между сказуемым *играют* и подлежащим *дети*, в котором главным словом является глагол.

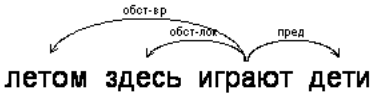

Рис. 1: Пример проективной структуры зависимостей

В этом предложении, как и в большинстве обычных предложений русского языка, структура зависимостей *проективная*, что, несколько упрощая, означает, что зависимости в структуре не пересекаются. Большинство грамматик, порождающих деревья зависимостей, имеют дело только с проективными структурами. С другой стороны, в языках достаточно часто встречаются предложения, имеющие *непроективные* структуры зависимостей.

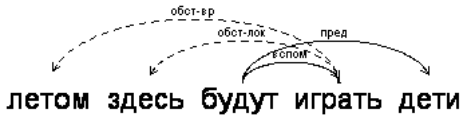

Рис. 2: Пример непроективной структуры зависимостей в русском

Например, использование будущего времени в предложении *"Летом здесь будут играть дети"* приводит к появлению двух разрывных зависимостей *играть* $\overset{обст\text{-}вр}{\longrightarrow}$ *летом* и *играть* $\overset{обст\text{-}лок}{\longrightarrow}$ *здесь*, показанных на рис. 2. Разрывные зависимости встречаются и в других языках. На рис. 3 изображена структура зависимостей французского предложения "Il n'en avait plus besoin" ("Он больше в этом не нуждался"), а на рис. 4 — английского предложения "The person to whom you must refer is Smith" ("Человек, к которому Вы должны обратиться, — Смит").

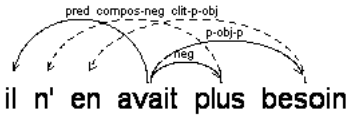

Рис. 3: Пример непроективной структуры зависимостей во французском языке

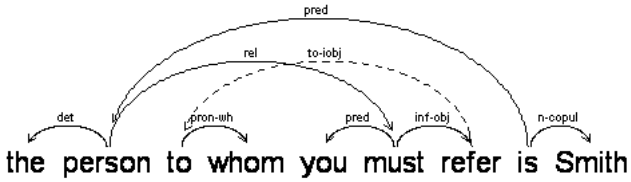

Рис. 4: Пример непроективной структуры зависимостей в английском языке

Современная лингвистическая теория синтаксических зависимостей была разработана Мельчуком. Первые точные определения грамматик зависимостей появились в работах Хейса и Гайфмана. Они имели много общего с классическими *категориальными грамматиками* Бар-Хиллела (которые восходят к работам Лесьневского и Айдукевича). Они полностью лексикализованы, используют синтаксические типы вместо правил вывода и естественно подходят для функциональных семантических структур. В 1960 году Бар-Хиллел, Гайфман и Шамир доказали, что формальный язык, не содержащий пустого слова, может быть задан классической категориальной грамматикой тогда и только тогда, когда

он является контекстно-свободным. В 1958 году Ламбек ввёл синтаксическое исчисление, расширяющее исчисление категориальных грамматик. В 1986 году Бушковский доказал, что грамматики Ламбека в неассоциативном варианте эквивалентны контекстно-свободным грамматикам (кс-грамматикам), а в 1993 году Пентус доказал эквивалентность кс-грамматик и исходных (ассоциативных) грамматик Ламбека. Однако сейчас признано, что кс-грамматики недостаточно выразительны для описания естественных языков. Например, кс-грамматики, как и классические категориальные грамматики, неспособны описывать в предложениях, подобных вышеприведённым, разрывные составляющие. В результате значительный интерес представляет разработка и изучение формальных грамматик, более выразительных, чем кс-грамматики. Например, ТАГ-грамматики Джоши, один из классов часто используемых для практических приложений, более выразительны, чем кс-грамматики, и позволяют выражать некоторые контекстные зависимости. Как показали Виай-Шенкер и Уэйр, ТАГ-грамматики эквивалентны некоторым грамматикам совершенно иной природы (линейным индексным грамматикам Ахо, комбинаторным категориальным грамматикам Стидмана и некоторым другим), разделяя с ними свойства, благодаря которым их называют слабо контекстными. Имеется несколько неэквивалентных понятий слабой контекстности класса грамматик. Одно из них состоит в том, что 1) грамматики этого класса порождают все кс-языки; 2) для них существует алгоритм синтаксического анализа за полиномиальное время; 3) эти грамматики позволяют выразить по меньшей мере некоторые пересекающиеся зависимости; 4) порождаемые ими языки обладают свойством линейного роста, т.е. если все предложения языка упорядочить по длине, то длины соседних предложений могут отличаться не более чем на заранее фиксированную константу. Первые грамматики зависимостей, выражающие неограниченные непроективные зависимости, были определены Диковским в 2001 году[1]. Эти грамматики являются порождающими, а непроективные зависимости между словами определяются в них через двойственные поляризованные валентности: одноимённые однонаправленные валентности с противоными знаками. Принцип спаривания двойственных валентностей (FA) аналогичен правильному спариванию скобок. В 2004 году Диковский

[1]Dikovsky A. Polarized Non-projective Dependency Grammars // Proc. of the Fourth Int. Conf. on Logical Aspects of Computational Linguistics (LACL), Springer, LNAI v. 2099, 2001. P. 139–157.

определил категориальные грамматики зависимостей (КГЗ)[2] . Подобно классическим категориальным грамматикам, КГЗ являются анализирующими, т.е. синтаксическая структура предложения — дерево зависимостей — извлекается в них из доказательства правильности приписывания типов словам. Непроективные зависимости, как и в порождающих грамматиках зависимостей, определяются через двойственные поляризованные валентности. В последующих совместных работах с Дехтярём определение КГЗ эволюционировало. Используемое ниже окончательное определение КГЗ содержится в работе Дехтяря и Диковского[3]. КГЗ хорошо зарекомендовали себя на практике. Например, в университете Нанта разработана и успешно используется для синтаксического анализа и для создания корпусов деревьев зависимостей весьма полная КГЗ французского языка. В 2007 году Диковский определил мультимодальные категориальные грамматики зависимостей (ммКГЗ), обобщающие КГЗ. Их отличие от КГЗ состоит в том, что каждому типу валентностей соответствует своё правило спаривания.

Настоящая диссертационная работа посвящена формальной теории КГЗ.

Цели диссертационной работы. Исследование свойств классов категориальных грамматик зависимостей и соответствующих классов порождаемых ими языков. Изучаемые вопросы включают существование нормальных форм КГЗ, существование класса автоматов, распознающих КГЗ-языки, замкнутость классов КГЗ-языков относительно различных операций над языками, полулинейность КГЗ-языков, а также проблему принадлежности для КГЗ.

Методы исследования. В диссертации использованы методы математической логики, теории формальных грамматик и языков, теории автоматов, теории алгоритмов и теории сложности вычислений.

Научная новизна и основные положения, выносимые на защиту. Все полученные в работе результаты являются новыми. На защиту выносятся следующие результаты.

[2]Dikovsky A. Dependencies as Categories // In “Recent Advances in Dependency Grammars". COLING'04 Workshop, 2004. P. 90–97.

[3]Dekhtyar M., Dikovsky A. Generalized Categorial Dependency Grammars // Trakhtenbrot/Festschrift, Springer Verlag, LNCS v. 4800, 2008. P. 230–255

1. Определена нормальная форма для классов КГЗ и ммКГЗ, аналогичная нормальной форме Грейбах для кс-грамматик, и установлена возможность эффективного приведения всякой КГЗ и ммКГЗ к этой нормальной форме.

2. Доказаны свойства замкнутости для классов КГЗ- и ммКГЗ-языков относительно операций объединения, пересечения с регулярными языками, неукорачивающего гомоморфизма, обращения гомоморфизма, конкатенации и усечённой итерации. Установлено, что класс ммКГЗ-языков образует абстрактное семейство языков.

3. Доказано, что любой КГЗ-язык можно представить как проекцию пересечения кс-языка и скобочного языка $L_{\text{ск}}$, состоящего из слов, проекции которых на каждый тип скобок являются правильными скобочными словами.

4. Построено расширение автоматов с магазинной памятью (МП-автоматов) — МП-автоматы с независимыми счётчиками. Эти автоматы дополняют обычные МП-автоматы конечным числом целочисленных счётчиков. Их отличие от машин Минского состоит в отсутствии проверки счётчиков на ноль. Доказано, что МП-автоматы с независимыми счётчиками распознают в точности КГЗ-языки.

5. Показано, что класс ммКГЗ-языков содержит неполулинейные языки, в отличие от класса кс-языков.

6. Доказано, что проблема принадлежности слова ммКГЗ-языку является NP-полной для конкретной ммКГЗ, в отличие от КГЗ, где для каждой грамматики существует полиномиальный алгоритм.

Теоретическая и практическая значимость. Работа носит теоретический характер. Материалы диссертации могут использоваться при чтении специальных курсов студентам и аспирантам, специализирующимся в области компьютерной лингвистики. Кроме того, результаты об автоматах, распознающих КГЗ-языки, и о сложности проблемы принадлежности могут использоваться при построении и анализе практических КГЗ для естественных языков.

Апробация работы Результаты диссертации докладывались автором на семинаре по теоретической информатике в Тверском государственном университете, на семинаре по математической логике в Петербургском отделении математического института РАН и на двенадцатой национальной конференции по искусственному интеллекту с международным участием "КИИ-2010". Некоторые результаты были представлены в совместном докладе на 15-й международной конференции по формальным грамматикам ("Formal Grammars 2010", Copenhagen, Denmark). Статья [1] была представлена в 2009 году на "Открытый конкурс на лучшую научную работу студентов по естественным, техническим и гуманитарным наукам в высших учебных заведениях Российской Федерации" и получила в этом конкурсе медаль "За лучшую научную студенческую работу". Работа поддерживалась грантами РФФИ 08-01-00241 и 10-01-00532а.

Публикации. Список публикаций по теме диссертации приведён в конце автореферата и включает 5 работ, 2 из которых опубликованы в издании, входящем в список рекомендованных ВАК ведущих рецензируемых изданий. Работы [4, 5] представляют доклады на международных конференциях и опубликованы в серии Lecture Notes in Computer Science.

Результаты совместной работы [4], вошедшие в диссертацию, принадлежат автору.

Структура и объём диссертации. Диссертация состоит из введения, трёх глав основной части, заключения и списка литературы. Общий объем работы — 103 стр. Список литературы содержит 40 наименований.

Краткое содержание работы

Содержание **введения** в основном совпадает с содержанием автореферата.

В **главе 1** приводятся основные определения и обозначения, связанные с категориальными грамматиками зависимостей, которые используются в остальной части работы.

Для формализации лингвистического понятия синтаксического типа используется понятие *категории*. Пусть $\mathbf{C}$ — непустое множество имен зависимостей (по аналогии с категориальными грамматиками, мы бу-

дем их называть элементарными категориями) (например, сказуемое, определение, дополнение и т. п.). Элементарные категории могут быть итерированы: для $C \in \mathbf{C}$ пусть C^* означает соответствующую итеративную категорию. Итерирование категории означает, что она может повторяться произвольное количество раз (возможно, ни одного). Например, в естественных языках у существительного может быть много определений, у глагола может быть много обстоятельств и т.п. Множество всех итеративных категорий будем обозначать $\mathbf{C}^*$.

Для описания непроективных зависимостей вводятся понятия полярности и поляризованной валентности. *Полярностью* v называется элемент множества $V = \{\searrow, \swarrow, \nwarrow, \nearrow\}$. *Поляризованная валентность* β — это выражение вида vC, где $v \in V$, $C \in \mathbf{C}$. Элементарная категория C называется именем валентности. Пары двойственных валентностей $(\nearrow C, \searrow C)$ и $(\swarrow C, \nwarrow C)$ назовём *правильными* (они являются соответствующими "скобками"). Валентности вида $\nearrow C$ и $\swarrow C$ называются левыми, а валентности вида $\searrow C$ и $\nwarrow C$ — правыми (они являются левыми и правыми скобками соответственно). Последовательность поляризованных валентностей называется *потенциалом*. Множество всех возможных потенциалов обозначается $Pot(C)$.

Потенциал называется *сбалансированным*, если его проекция на каждую пару двойственных поляризованных валентностей является правильным скобочным словом. Например, потенциал $\nearrow A \ \nearrow B \ \searrow A \ \searrow B \ \nearrow A \ \searrow A$ сбалансирован, а потенциал $\nearrow A \ \searrow B \ \searrow A$ не сбалансирован.

Определение 1.1. Множество локальных категорий $LCat(\mathbf{C})$ — это минимальное множество такое, что:
1) $\mathbf{C} \cup \{\varepsilon\} \subseteq LCat(\mathbf{C})$, где ε — символ пустой категории;
2) если $\alpha \in LCat(\mathbf{C}), A \in \mathbf{C} \cup \mathbf{C}^*$, то $[A\backslash\alpha], [\alpha/A] \in LCat(\mathbf{C})$.

Из локальных категорий и потенциалов строятся категории.

Определение 1.2. Категорией γ называется выражение вида α^θ, где $\alpha \in LCat(\mathbf{C}), \theta \in Pot(\mathbf{C})$. Множество всех категорий обозначим $Cat(\mathbf{C})$.

Полагаем, что конструкторы $\backslash$ и $/$ ассоциативны, например, $[B\backslash[A/C]] = [[B\backslash A]/C] = [B\backslash A/C]$. Поэтому любая категория γ может быть представлена в виде $[L_k\backslash\ldots\backslash L_1\backslash C/R_1/\ldots/R_m]^\theta$. Теперь приведём основное определение категориальной грамматики зависимостей.

Определение 1.3. Категориальной грамматикой зависимостей (КГЗ) называется система $G = \langle W, \mathbf{C}, S, \delta\rangle$, где:

W — конечное множество символов,
$\mathbf{C}$ — конечное множество элементарных категорий,
S — выделенная в $\mathbf{C}$ главная категория,
δ — словарь, функция на W такая, что $\delta(w) \subseteq Cat(\mathbf{C})$ и $\delta(w)$ конечно для каждого символа $w \in W$.

В КГЗ G после приписывания категорий символам слова происходит вывод главной категории S с помощью приведенных ниже правил сокращения. Эти правила имеют вид $\Gamma \vdash \Gamma'$, где Γ — сокращаемая строка категорий, а Γ' — строка категорий, получающаяся в результате сокращения. В приведённых ниже правилах Γ_1 и Γ_2 — это произвольные строки категорий из $Cat(\mathbf{C})^*$.

Определение 1.4. Правила сокращения

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

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

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

Правила L^l и L^r — это аналоги классических правила сокращения. При сокращении аргумента C каждое из них создаёт проективную зависимость C и выполняет конкатенацию потенциалов. Правила L^l_ε и L^r_ε — это правила сокращения пустой категории ε. Они не создают зависимостей. Правила I^l и I^r создают произвольное количество зависимостей C. Правила I^l_0 и I^r_0 служат для удаления итеративного подтипа C^*. Правило D сокращает две двойственные поляризованные валентности β и $\breve{\beta}$ и создаёт непроективную зависимость C. Для пары $\beta = {\nearrow}C$, $\breve{\beta} = {\searrow}C$ эта зависимость идёт слева направо от слова, содержащего в своём потенциале валентность ${\nearrow}C$, к слову, содержащему в потенциале ${\searrow}C$. Если

$\beta = \swarrow C$, $\breve{\beta} = \nwarrow C$, то зависимость C идёт в обратном направлении. Как обычно, через $\vdash^*$ обозначим рефлексивное и транзитивное замыкание отношения $\vdash$ на множестве $Cat(\mathbf{C})^*$.

Рассмотрим, например, предложение *"Летом здесь будут играть дети"*. Пусть словарь содержит следующие категории: $[\varepsilon]^{\swarrow\text{обст-вр}} \in \delta(\text{летом})$, $[\varepsilon]^{\swarrow\text{обст-лок}} \in \delta(\text{здесь})$, $[S/\text{пред}/\text{вспом}] \in \delta(\text{будут})$, $[\text{вспом}]^{\nwarrow\text{обст-вр}\nwarrow\text{обст-лок}} \in \delta(\text{играть})$, $[\text{пред}] \in \delta(\text{дети})$. Конкатенация этих категорий допускает сокращение до $[S]$ по приведённым правилам. Результирующая структура зависимостей показана на рис. 2 (локальные зависимости показаны сплошными линиями, а непроективные — пунктирными).

Определение 1.5. Слово $s = w_1 \dots w_n \in W^*$ принадлежит языку $L(G)$ тогда и только тогда, когда существует строка категорий Γ такая, что:
1) $\Gamma = \gamma_1 \dots \gamma_n$, где $\gamma_i \in \delta(w_i)$ для $i = 1, \dots, n$,
2) $\Gamma \vdash^* [S]$.

Класс всех языков, порождаемых категориальными грамматиками зависимостей, обозначим через $\mathcal{L}(CDG)$. Через $\mathcal{L}_k(CDG)$ обозначим класс всех языков, порождаемых категориальными грамматиками зависимостей, содержащими не более k типов валентностей.

В приведённом выше правиле сокращения D все валентности вида $\nearrow A$ и $\searrow A$ образуют правильные пары в потенциале $\nearrow A\theta \searrow A$, если θ не содержит валентностей типа $\nearrow A$ и $\searrow A$, т.е. $\searrow A$ является первой доступной валентностью, соответствующей $\nearrow A$. Это правило формирования пар называется "первый доступный" и обозначается FA (от англ. first available). Определение КГЗ можно расширить, разрешив дополнительные ограничения на возможность сокращения валентностей. Для этого на множестве всех элементарных категорий вводится *функция запретов* — отображение $\pi\colon \mathbf{C} \to 2^{\mathbf{C}}$. Эта функция описывает ограничения на расположение правильных пар валентностей. Если $Y \in \pi(X)$, то пара валентностей типа X не может создать зависимость, если между ними есть валентности типа Y. Например, если $\pi_1(X) = \{ Y \}$, $\pi_1(Y) = \{ X \}$, то в потенциале $\theta_1 = \nearrow X \nearrow Y \searrow X \searrow Y$ нет правильных пар. Действительно, $\nearrow Y$ не позволяет сформировать пару $\nearrow X \searrow X$, а $\searrow X$ не позволяет сформировать пару $\nearrow Y \searrow Y$. Если же $\pi_2(X) = \{ Y \}$, $\pi_2(Y) = \varnothing$, то пара $\nearrow Y \searrow Y$ является правильной в θ_1. Если удалить её из θ_1, то останется $\nearrow X \searrow X$ — тоже правильная пара.

С учётом функции запретов понятие сбалансированности потенциала изменяется следующим образом. Потенциал называется сбалансированным, если его проекция на каждую пару двойственных поляризованных валентностей является правильным скобочным словом и существует такой порядок правильных пар, что их можно удалять из потенциала, не нарушая ограничений функции запретов. Например, потенциал $\nearrow A \nearrow B \searrow A \searrow B \nearrow A \searrow A$ сбалансирован, если $\pi(A) = \{ B \}, \pi(B) = \varnothing$. Этот же потенциал не будет сбалансированным при $\pi(A) = \{ B \}, \pi(B) = \{ A \}$. Потенциал $\nearrow A \searrow B \searrow A$ не сбалансирован независимо от функции запретов. Таким образом, определение сбалансированности потенциала без функции запретов является частным случаем более общего определения и получается из него при $\pi(X) = \varnothing$ для всех X.

Функция запретов используется в следующем определении более широкого класса КГЗ являющемся важным специальным случаем общего определения, содержащегося в работе Диковского[4].

Определение 1.6. Мультимодальной категориальной грамматикой зависимостей (ммКГЗ) называется система $G = \langle W, \mathbf{C}, S, \delta, \pi \rangle$, где:
W — конечное множество символов,
$\mathbf{C}$ — конечное множество элементарных категорий,
S — выделенная в $\mathbf{C}$ главная категория,
δ — словарь, функция на W такая, что $\delta(w) \subseteq Cat(\mathbf{C})$ и $\delta(w)$ конечно для каждого $w \in W$ (мы будем задавать его в виде $w \mapsto \delta(w)$),
$\pi \colon \mathbf{C} \to 2^{\mathbf{C}}$ — функция запретов.

Как и в КГЗ, в ммКГЗ символам слова приписываются категории и происходит вывод главной категории с помощью правил сокращения. Правила сокращения локальных категорий для ммКГЗ точно такие же, что и для КГЗ. Изменяется лишь правило дальней зависимости.

Определение 1.7. Правило дальней зависимости:
$D : \Gamma_1 \alpha^{\theta_1 \beta \theta_2 \breve{\beta} \theta_3} \Gamma_2 \vdash \Gamma_1 \alpha^{\theta_1 \theta_2 \theta_3} \Gamma_2$,
где валентности $(\beta, \breve{\beta})$ образуют правильную пару и θ_2 не содержит β, $\breve{\beta}$, а также $\nearrow B$, $\nwarrow B$, $\swarrow B$, $\searrow B$ для всех $B \in \pi(C)$, где C — имя валентности β.

В простейшем случае ограничения означают, что зависимости некоторых типов не могут пересекаться. Например, в естественных языках

[4] Dikovsky A. Multimodal Categorial Dependency Grammars // Proc. of the 12th Conference on Formal Grammar. Dublin, Ireland, 2007. P. 1–12.

в предложениях с прямой речью зависимости не могут соединять слова, стоящие внутри кавычек, со словами, стоящими вне кавычек. Чтобы обеспечить это, можно соединить открывающие кавычки с закрывающими специальной зависимостью, которую не может пересечь ни одна другая зависимость.

Класс всех языков, порождаемых мультимодальными категориальными грамматиками зависимостей, обозначим $\mathcal{L}(mmCDG)$, а класс всех языков порождаемых такими грамматиками, содержащими не более k типов валентностей, обозначим через $\mathcal{L}_k(mmCDG)$.

Из определения следует, что КГЗ-языки содержат все языки, порождаемые классическими категориальными грамматиками, т.е. все кс-языки, не содержащие пустую строку ε. На самом деле этот класс существенно шире класса кс-языков. Мы рассматриваем примеры известных языков, не являющихся контекстно-свободными: $L_1 = \{\, a_1^k a_2^k \dots a_n^k \mid k > 0 \,\}$, $L_2 = \{\, w \in W^+ \mid |w|_{a_1} = \dots = |w|_{a_n} \,\}$ и $L = \{ww^{-1}w \mid w \in \{a, b\}^+\}$. Для языков L_1 и L_2 построены порождающие их КГЗ, а для языка L_3 — ммКГЗ.

Глава 2 посвящена свойствам категориальных грамматик зависимостей. В параграфе 2.1 рассматриваются нормальные формы для КГЗ.

Известно, что для кс-грамматик существуют различные специальные формы (нормальная форма Хомского, нормальная форма Грейбах), упрощающие анализ порождаемых ими языков. Мы определяем нормальную форму для КГЗ, аналогичную нормальной форме Грейбах.

Определение 2.2. Будем называть КГЗ $G = \langle W, \mathbf{C}, S, \delta \rangle$ грамматикой в нормальной форме, если все её категории имеют один из следующих видов:
1) $[X]^\theta$
2) $[X/Y]^\theta$
3) $[X/Y/Z]^\theta$,
где X, Y, Z — элементарные категории, θ — потенциал.

Основным результатом этого параграфа является следующая теорема.

Теорема 2.1. *Для любой КГЗ $G = \langle W, \mathbf{C}, S, \delta \rangle$ существует КГЗ $G' = \langle W, \mathbf{C}', S, \delta' \rangle$ в нормальной форме такая, что $L(G) = L(G')$. Существует алгоритм, который по произвольной КГЗ G строит КГЗ G', при этом размер G' является полиномиальным относительно раз-*

мера G.

Отметим, что эту теорему можно несколько усилить, построив для любой КГЗ G эквивалентную ей КГЗ G' в нормальной форме такую, что её списки зависимостей не содержат главной категории.

В параграфе 2.2 исследуются свойства замкнутости класса КГЗ-языков. Ранее Дехтярь, Диковский и Зайцев установили, что этот класс замкнут относительно операций объединения, конкатенации, неукорачивающего гомоморфизма и пересечения с регулярными языками. При этом грамматика для результата выполнения операции содержала существенно больше типов валентностей, чем грамматики для аргументов операции. В следующих теоремах мы устанавливаем более точные результаты, содержащие оценку числа типов валентностей в грамматиках для языков-результатов этих операций.

Теорема 2.2. *Если $L_1 \in \mathcal{L}_k(CDG)$ и $L_2 \in \mathcal{L}_k(CDG)$, то $L = L_1 \cup L_2 \in \mathcal{L}_k(CDG)$.*

Теорема 2.3. *Если $L_1 \in \mathcal{L}_k(CDG)$, а $L_1 \in \mathcal{L}_l(CDG)$, то $L = L_1 \cdot L_2 \in \mathcal{L}_{k+l}(CDG)$.*

Теорема 2.4. *Если $L \in \mathcal{L}_k(CDG)$ и R — регулярный язык, то $L \cap R \in \mathcal{L}_k(CDG)$.*

Теорема 2.5. *Пусть W и Σ — два словаря. Пусть определена функция (неукорачивающий гомоморфизм) $\tau\colon W^+ \to \Sigma^+$. Тогда, если язык $L \in \mathcal{L}_k(CDG)$, то $\tau(L) = \{\ \tau(w_1)\tau(w_2)...\tau(w_n) \mid w_1w_2...w_n \in L\ \} \in \mathcal{L}_k(CDG)$.*

В этом же параграфе получен новый результат о замкнутости класса КГЗ-языков относительно обращения гомоморфизма.

Теорема 2.6. *Если $L \in \mathcal{L}_k(CDG)$ и $h\colon W^* \to \Sigma^*$ — произвольный гомоморфизм, то $h^{-1}(L) \in \mathcal{L}_k(CDG)$.*

Таким образом, установлено, что для любого k класс $\mathcal{L}_k(CDG)$ замкнут относительно объединения, пересечения с регулярными языками, неукорачивающего гомоморфизма и обращения гомоморфизма.

В параграфе 2.3 мы получаем специальное представление КГЗ-языков. Для кс-языков известен следующий результат (теорема Хомского-Шютценберже): для любого кс-языка L существуют регулярный язык R, язык Дика D и гомоморфизм ϕ такие, что $L = \phi(R \cap D)$. КГЗ-языки обладают похожим свойством. Сначала определим понятие скобочного языка, который в нашем случае заменит язык Дика.

Определение 2.6. Пусть $\Sigma = \{\, a_1, \bar{a}_1, \ldots, a_n, \bar{a}_n \,\}$. Определим для ал-

фавита Σ скобочный язык $L_{\text{ск}}$, для которого его проекция на $\{ a_i, \bar{a}_i \}$ $(i = 1, \ldots, n)$ совпадает с языком правильной расстановки скобок $\{ a_i, \bar{a}_i \}$.

Аналогом теоремы Хомского-Шютценберже для КГЗ-языков является

Теорема 2.7. *Для любого КГЗ-языка L существуют КС-язык L_1, скобочный язык $L_{\text{ск}}$ и гомоморфизм ϕ такие, что $L = \phi(L_1 \cap L_{\text{ск}})$.*

В параграфе 2.4 решается задача определения класса автоматов, распознающих КГЗ-языки. Это класс автоматов, которые дополняют обычные автоматы с магазинной памятью конечным множеством целочисленных счётчиков. При этом выбор действия автомата на каждом шаге не зависит от состояния счётчиков.

Определение 2.7. Автомат с магазинной памятью (МП-автомат) с независимыми счётчиками — это семёрка $M = \langle \Sigma, Q, Z, q_0, z_0, P, n \rangle$, где:
Σ — конечный входной алфавит,
Q — конечное множество состояний,
Z — конечный алфавит магазина,
$q_0 \in Q$ — начальное состояние,
$z_0 \in Z$ — начальный символ магазина,
P — конечное множество правил,
n — число счётчиков.
Правила имеют вид $(q, a, z) \to (q', \alpha, \bar{v})$, где $q, q' \in Q$, $a \in \Sigma \cup \{ \varepsilon \}$, $z \in Z$, $\alpha \in Z^*$, $\bar{v} = (v_1, \ldots, v_n)$ — целочисленный вектор.

Эти автоматы используют магазинную память, чтобы проверять сокращение локальных категорий, а счётчики позволяют проверять сбалансированность потенциалов.

Определение 2.8. Конфигурация МП-автомата с независимыми счётчиками $M = \langle \Sigma, Q, Z, q_0, z_0, P, n \rangle$ — это четверка $\langle q, w, \gamma, \bar{u} \rangle$, где $q \in Q$, $w \in \Sigma^*$, $\gamma \in Z^*$, $\bar{u} = (u_1, \ldots, u_n)$ — вектор неотрицательных целых чисел. Отношение перехода за один шаг на множестве конфигураций M: $\langle q, w, \gamma, \bar{u} \rangle \vdash_M^1 \langle q', w', \gamma', \bar{u}' \rangle$ означает, что существует правило $(q, a, z) \to (q', \alpha, \bar{v}) \in P$ такое, что выполняются следующие три условия:
1) $w = aw'$,
2) $\gamma = z\gamma''$, $\gamma' = \alpha\gamma''$,
3) $\bar{u}' = \bar{u} + \bar{v}$.
Отношения $\vdash_M^n$ и $\vdash_M^*$ определяются стандартным образом.

Фактически числа в счётчиках — это избытки левых валентностей, т.е. количество левых валентностей, для которых пока не найдена со-

ответствующая правая валентность. Положительные числа в правилах соответствуют левым валентностям, а отрицательные — правым. Автомат работает подобно обычному МП-автомату. Дополнительно он изменяет значения счётчиков на каждом шаге, но сам шаг не зависит от этих значений.

Язык, распознаваемый автоматом M определяется опустошением магазина и обнулением счётчиков.

Определение 2.9. Слово w распознаётся МП-автоматом с независимыми счётчиками M тогда и только тогда, когда

$$\langle q_0, w, z_0, (0, \ldots, 0)\rangle \vdash_M^* \langle q, \varepsilon, \varepsilon, (0, \ldots, 0)\rangle.$$

Язык, распознаваемый автоматом M, — это множество всех слов, распознаваемых этим автоматом.

Мы будем рассматривать специальный подкласс автоматов без пустых циклов.

Определение 2.10. МП-автомат с независимыми счётчиками M называется автоматом без пустых циклов, если не существует состояний $q_1, \ldots q_k$ $(k > 1)$ таких, что $(q_i, \varepsilon, z_i) \to (q_{i+1}, \gamma_i, \bar{v}_i) \in P$ для $1 \leq i < k$, $(q_k, \varepsilon, z_k) \to (q_1, \gamma_k, \bar{v}_k) \in P$.

Заметим, что без этого ограничения возможна ситуация, когда автомат выполняет ε-команды в цикле и изменяет счётчики. Тогда он сможет неограниченно увеличить счётчики, не читая новых символов. Но все потенциалы в КГЗ имеют конечную длину. Определение автомата без пустых циклов позволяет избежать этой ситуации.

Основными результатами параграфа являются следующие две теоремы.

Теорема 2.8. *Любой КГЗ-язык распознаётся некоторым МП-автоматом с независимыми счётчиками и без пустых циклов. При этом число счётчиков автомата равно числу типов валентностей в КГЗ.*

Теорема 2.9. *Если язык L распознаётся МП-автоматом с независимыми счётчиками и без пустых циклов, то $L\backslash\{\varepsilon\}$ — КГЗ-язык. При этом число типов валентностей в КГЗ равно числу счётчиков автомата.*

Из построений, описанных в теоремах, следует, что для любого МП-автомата с независимыми счётчиками и без пустых циклов существует эквивалентный автомат без ε-команд.

В главе 3 изучаются мультимодальные категориальные грамматики зависимостей. В параграфе 3.1 показано, что результат о нормальной

форме для КГЗ непосредственно переносится на ммКГЗ.

В параграфе 3.2 исследуются свойства замкнутости класса ммКГЗ-языков. Устанавливается, что, как и класс КГЗ-языков, он замкнут относительно операций объединения, конкатенации, пересечения с регулярными множествами, неукорачивающего гомоморфизма и обращения гомоморфизма. Кроме того, мы доказываем, что класс ммКГЗ-языков замкнут относительно усечённой итерации.

Теорема 3.8. *Если* $L \in \mathcal{L}_k(mmCDG)$, *то* $L^+ \in \mathcal{L}_{k+1}(mmCDG)$.

Тогда справедлива следующая

Теорема 3.9. *Класс* $\mathcal{L}(mmCDG)$ *является абстрактным семейством языков.*

Далее в параграфе 3.2 доказывается, что класс ммКГЗ-языков замкнут относительно пересечения, но не замкнут относительно проекции (теоремы 3.10 и 3.11).

В параграфе 3.3 рассматривается вопрос о полулинейности ммКГЗ-языков. Известная теорема Парика утверждает, что все кс-языки являются полулинейными. Одним из известных свойств полулинейных языков является то, что длины входящих в них слов содержат бесконечные арифметические прогрессии. Мы доказываем, что язык $L = \{\, 101001\ldots10^{2^n} \mid n = 3, 5, 7, \ldots \}$ является ммКГЗ-языком. Поскольку множество длин слов этого языка не содержит бесконечных арифметических прогрессий, он неполулинейный. Отсюда выводим

Следствие 3.5. *Класс ммКГЗ-языков содержит неполулинейные языки.*

В параграфе 3.4 рассматривается сложность проблемы принадлежности для ммКГЗ-языков. Ранее в работе Дехтяря и Диковского (см. ссылку на стр. 6) было установлено, что проблема определения по КГЗ G и слову w, принадлежит ли w языку $L(G)$, является NP-полной. Для класса ммКГЗ-языков этот результат может быть усилен.

Теорема 3.14. *Существует ммКГЗ* G *с двумя типами валентностей такая, что проблема принадлежности для* $L(G)$ *является NP-полной.*

Таким образом, если в ммКГЗ есть хотя бы два типа валентностей, которые не могут пересекаться, то проблема принадлежности оказывается сложной. Отметим, что для грамматик без запретов на пересечение дальних связей существует полиномиальный относительно размеров входного слова алгоритм.

В **заключении** диссертации перечислены основные полученные в ней результаты и сформулирован ряд открытых проблем.

Публикации в рецензируемых научных журналах

[1] Карлов Б.Н. Нормальные формы и автоматы для категориальных грамматик зависимостей // Вестник Тверского государственного университета, серия "Прикладная математика", №35 (95), 2008. С. 23–43.

[2] Карлов Б.Н. О свойствах языков, задаваемых мультимодальными категориальными грамматиками зависимостей // Вестник Тверского государственного университета, серия "Прикладная математика", №34, 2011. С. 91–110.

Другие публикации

[3] Карлов Б.Н. О свойствах обобщённых категориальных грамматик зависимостей // Двенадцатая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. Т. 1. М.: Физматлит, 2010. С. 283–290.

[4] Dekhtyar M., Dikovsky A., Karlov B. Iterated dependencies and Kleene iteration // In Proc. of the 15th Conference on Formal Grammar (FG 2010), Copenhagen, Denmark, Lecture Notes in Computer Science №7395, Springer, 2012. P. 66–81.

[5] Karlov B. Abstract Automata and a Normal Form for Categorial Dependency Grammars // Proc. of the 7th International Conference on Logical Aspects of Computational Linguistics (LACL 2012), Nantes, France, Lecture Notes in Computer Science №7351, Springer, 2012. P. 86–102.