10. Erosa A. M., Henron L. J. Taming Control Flow: A structured Approach to Elimination GOTO Statements // Proc. IEEE International Conf. on Computer Languages. May 1994. – P. 229–240.

11. Овчинников В. А., Иванова Г. С. Информационно-логическая модель алгоритма // Вестник МГТУ. Серия "Приборостроение". – 2005. – № 1. – С. 109–121.

Статья поступила в редакцию 27.10.2004

Иванова Галина Сергеевна родилась в 1954 г., окончила МВТУ им. Н.Э. Баумана в 1978 г. Канд. техн. наук, доцент кафедры "Компьютерные системы и сети". Автор 25 научных работ в области вычислительной техники и проектирования программных систем.

G.S. Ivanova (b. 1954) graduated from the Bauman Moscow Higher Technical School in 1978. Ph. D. (Eng.), assoc. professor of "Computer Systems and Networks" department of the Bauman Moscow State Technical University. Author of 25 publications in the field of computing technology and design of software systems.

УДК 004.89; 004.912; 004.5

В. А. Фомичев

КЛАСС ФОРМАЛЬНЫХ ЯЗЫКОВ И АЛГОРИТМ ДЛЯ ПОСТРОЕНИЯ СЕМАНТИЧЕСКИХ АННОТАЦИЙ ВЕБ-ДОКУМЕНТОВ

Предложена широко применимая и гибкая теория формального описания структурированных значений текстов на естественном языке (предложений и дискурсов) — теория стандартных К-языков (СК-языков). Анализ выразительной силы класса СК-языков дает возможность предположить, что СК-языки позволяют строить семантические аннотации произвольных Веб-документов и удобны для построения таких аннотаций. Теория СК-языков была использована при разработке широко применимой математической модели лингвистической базы данных (ЛБД) и сложного структурированного алгоритма семантико-синтаксического анализа текстов из представляющих практический интерес подъязыков естественного (русского) языка, базирующегося на построенной модели ЛБД. Алгоритм реализован в системе программирования Visual C++ и может быть широко использован для построения семантических аннотаций Веб-документов.

Благодаря бурному прогрессу компьютерной сети Всемирная Паутина (the World Wide Web, WWW, W3) пользователи сети во всем мире получили быстрый доступ к огромному количеству ЕЯ-текстов, относящихся к различным областям деятельности человека. С середины 1990-х годов специалисты в самых разных предметных областях работают не только с публикациями и базами данных (БД) своих орга-

низаций, но и стремятся использовать информационные ресурсы Паутины. Поэтому чрезвычайно актуальна задача организации взаимодействия на ограниченном естественном языке из различных предметных областей с огромным объемом накопленных информационных ресурсов Всемирной Паутины.

ЕЯ-интерфейсы для взаимодействия с информационными ресурсами Паутины необходимы не только специалистам для решения профессиональных задач, но и конечным пользователям, перед которыми стоят задачи получения медицинской или юридической информации, расширения культурного кругозора и т. д.

В феврале 2001 г. консорциум сети Всемирная Паутина, обозначаемый в большинстве документов сокращением W3C (the World Wide Web Consortium), официально объявил о широком развертывании исследований по преобразованию существующей сети в Семантическую Всемирную Паутину (Semantic Web) [1]. Один из наиболее важных аспектов реализации этого крупномасштабного проекта заключается в том, что компьютерные интеллектуальные агенты (КИА) смогут анализировать информацию, представленную на Веб-сайтах, взаимодействуя между собой. Часть КИА сможет выполнять смысловой анализ ЕЯ-компонентов электронных документов, представленных в Веб-сайтах. Это даст возможность конечным пользователям осуществлять поиск информации в Паутине не по ключевым словам, а по смыслу, с помощью КИА. Важные дополнительные возможности для пользователя предоставят речевые браузеры. Такие браузеры позволят использовать телефоны (в том числе мобильные) для взаимодействия с Семантической Паутиной на естественном языке [2].

Анализ ряда публикаций, посвященных семантическим аннотациям Веб-данных, позволяет придти к следующему заключению: идеальной конфигурацией Семантической Паутины будет являться совокупность взаимосвязанных ресурсов, у каждого из которых имеется как аннотация на естественном языке (ЕЯ), так и формальная аннотация, отражающая содержание или обобщенное содержание ресурса, т.е. семантическая аннотация. ЕЯ-аннотации будут очень удобны для конечных пользователей, а семантические аннотации будут использоваться вопросо-ответными и поисковыми системами.

Построение семантических аннотаций Веб-ресурсов по их ЕЯ-аннотациям должно осуществляться компьютерными системами обработки ЕЯ, или лингвистическими процессорами (ЛП). Поэтому сегодня широко признано существование значительного пересечения интересов теории систем искусственного интеллекта, включающей компьютерную лингвистику, и исследований по Семантической Паутине [3, 4].

Формальную структуру, отражающую содержание (или смысл, или значение, или семантическую структуру, или смысловую структуру) выражения на ЕЯ, в компьютерной и теоретической лингвистике называют семантическим представлением (СП) этого выражения.

Наиболее вероятно, первой идеей относительно формирования семантических аннотаций Веб-ресурсов будет идея использования формальных средств построения СП ЕЯ-текстов, предоставляемых математической и компьютерной лингвистикой. Однако анализ показывает, что выразительная сила наиболее популярных формальных подходов к построению СП ЕЯ-текстов, в частности, теории представления дискурсов [5], теории концептуальных графов [6], и эпизодической логики [7] недостаточна для эффективного представления содержания (смысла) произвольных Веб-документов, в том числе произвольных биологических, медицинских или относящихся к бизнесу документов.

Прежде всего, ограничения касаются описания семантической структуры: (а) неопре-деленных форм глаголов (инфинитивов) с зависимыми словами, используемых, в част-ности, для представления целей, рекомендаций, предложений, обязательств, назначений предметов и процессов; (б) конструкций, образованных из инфинитивов с зависимыми словами с помощью логических связок “и”, “или”, “не”; (в) составных обозначений множеств; (г) фрагментов, в которых логические связки “и”, “или”соединяют не обозначения высказываний, а обозначения объектов (“продукт А выпускается фирмами Ф1, Ф2 и Ф3”); (д) пояснений понятий, являющихся неизвестными прикладной интел-лектуальной системе; (е) фрагментов, содержащих ссылки на смысл фраз или более длинных фрагментов дискурсов (“этот метод” и т.д.); (ж) обозначений функций, аргументами и/или значениями которых могут быть множества объектов (“персонал фирмы А”, “поставщики фирмы Б” и т.д.).

Несмотря на описанную ситуацию, высказанная идея о том, где можно найти формальные средства для построения семантических аннотаций, является корректной. Основная цель данной работы заключается в том, чтобы указать на широкие возможности построения семантических аннотаций Веб-документов средствами новой теории формального описания структурированных значений ЕЯ-текстов (предложений и дискурсов) — теории стандартных К-языков (СК-языков) [8–10]. Таким образом, данная работа продолжает линию статьи [4].

Цель данной работы заключается также в кратком описании разработанного автором нового алгоритма семантико-синтаксического анализа текстов на русском языке, который может использоваться для построения семантических аннотаций Веб-документов, а также для пре-

образования вопросов пользователя прикладной интеллектуальной системы в СП, являющиеся выражениями СК-языков.

Центральные идеи широко применимой и гибкой формальной метаграмматики семантических аннотаций. Теория СК-языков является оригинальной теорией фор-мального описания структурированных значений (СЗ) предложений и сложных дискурсов на ЕЯ, представления знаний о мире и целей интеллектуальных систем и описания соответствия между ЕЯ-текстами и их СЗ. Поэтому определение класса СК-языков может интерпретироваться как широко применимая и гибкая (по-видимому, универсальная) формальная метаграмматика семантических аннотаций Веб-документов.

На основании проведенного системного исследования выразительных возможностей ЕЯ и искусственных языков представления знаний о мире была поставлена задача построения такой модели, чтобы ее формальные средства позволяли:

— (свойство 1) строить обозначения СЗ как фраз, выражающих высказывания, так и связных повествовательных текстов; такие обозначения обычно называют семантичес-кими представлениями (СП) ЕЯ-выражений;

— (свойство 2) строить и различать формальными средствами обозначения СЗ повествовательных текстов, СЗ целей (выраженных неопределенными формами глаголов с зависимыми словами, таких как “окончить с отличием МГУ, подготовить и защитить кандидатскую диссертацию по биохимии”) и СЗ вопросов;

— (свойство 3) строить и различать обозначения единиц, соответствующих а) объектам, ситуациям, процессам в реальном мире и б) понятиям, квалифицирующим (характеризующим) эти объекты, ситуации, процессы;

— (свойство 4) строить и различать обозначения: (3.1) объектов и множеств объектов; (3.2) понятий и множеств понятий; (3.3) СП текстов и множеств СП текстов;

— (свойство 5) различать формальным образом понятия, квалифицирующие объекты, и понятия, квалифицирующие множества объектов тех же видов;

— (свойство 6) строить составные обозначения понятий, т. е. строить формулы, отражающие поверхностно-семантическаю структуру ЕЯ-выражений, подобных выражению “человек, окончивший МГУ имени М.В. Ломоносова и являющийся биологом или химиком”;

— (свойство 7) строить объяснения более общих понятий с помощью менее общих; в частности, строить цепочки вида ($a = Des(b)$), где a обозначает некоторое понятие, которое необходимо объяснить, а

$Des(b)$ обозначает описание некоторой конкретизации известного понятия b;

—(свойство 8) строить обозначения упорядоченных n-местных наборов различных сущностей, где $n > 1$;

— (свойство 9) строить (9.1) формальные аналоги составных обозначений множеств (“эта группа, состоящая из 12 туристов, являющихся химиками или биологами” и т. п.), (9.2) обозначения множеств упорядоченных наборов сущностей, (9.3) обозначения множеств, состоящих из множеств, и т. д.;

— (свойство 10) описывать теоретико-множественные отношения и операции над множествами;

— (свойство 11) строить обозначения СЗ фраз, содержащих, в частности, (11.1) слова “произвольный”, “некоторый”, “все”, “каждый”, и т. д.; (11.2) выражения, полученные применением связок “и”, “или” к обозначениям (11.2а) предметов, событий; (11.2б) понятий; (11.2в) множеств; (11.3) выражения , где связка “не” стоит непосредственно перед обозначением предмета, события и т. д.; (11.4) косвенную речь; (11.5) причастные обороты и придаточные определительные предложения; (11.6) слово “понятие”;

— (свойство 12) строить обозначения СЗ дискурсов со ссылками на упомянутые объекты;

— (свойство 13) указывать явно в СП дискурсов причинно-следственные и временные отношения между описываемыми ситуациями (событиями);

— (свойство 14) Описывать СЗ дискурсов со ссылками на смысл фраз и более крупных фрагментов рассматриваемых текстов;

— (свойство 15) выражать суждения о тождественности двух сущностей;

— (свойство 16) Строить формальные аналоги формул логики предикатов первого порядка с кванторами существования и/или всеобщности;

— (свойство 17) рассматривать нетрадиционные функции (и другие нетрадиционные отношения) с аргументами и/или значениями, являющимися: (17.1) множествами предметов, ситуаций (событий); (17.2) множествами понятий; (17.3) множествами СП текстов;

— (свойство 18) строить концептуальные представления текстов как информационные объекты, отражающие не только смысл, но и значения внешних характеристик текста: авторов, дату, области применения результатов и т. д.

Решение этой задачи изложено в [9, 10], а начальные версии решения в работах [11, 12]. Первая часть построенной в работах [9, 10] модели, описывающей систему из 10 операций на концептуальных структу-

рах, определяет новый класс формальных объектов, называемых концептуальными базисами (к.б.). Каждый к.б. строится для формализации группы предметных областей и является сложным упорядоченным набором, задающим а) множество первичных информационных единиц и множество переменных, используемых для построения формул, интерпретируемых как СП ЕЯ-текстов, б) сведения, относящиеся к таким единицам и используемые для комбинирования этих единиц и нескольких специальных символов в составные единицы — СП ЕЯ-текстов.

Модель для каждого к.б. B задает множество формул $Ls(B)$, удобных для построения СП ЕЯ-текстов, называемое стандартным К-языком (концептуальным языком), или СК-языком, порождаемым базисом B. Выражения СК-языков будут называться К-цепочками. Множество $Ls(B)$ для произвольного к.б. B определяется совместной индукцией с помощью системы специальных правил $P[0], P[1], \ldots, P[10]$; они интер-претируются как правила построения семантических представлений (СП) ЕЯ-текстов из элементов первичного информационного универсума $X(B)$, переменных из $V(B)$ и нескольких специальных символов при условии, что B является концептуальным базисом для рассматриваемой области [9, 10].

Каждое из этих правил фактически задает некоторую операцию на множестве всевозможных наборов, компоненты которых являются СП простых или составных выражений естественного языка (ЕЯ). Имеются веские основания предположить, что всего 10 операций достаточно для построения формул, отображающих смысл (или структурированные значения) сколь угодно сложных ЕЯ-текстов. Для любого к.б. B правило $P[0]$ задает начальный запас формул.

Пример. Можно построить такой к.б. B, что выполняются соотношения *чел, П. Сомов, НПО_"Радуга", Друзья, Персонал, Поставщики* $\in$ *Ls(B).*

Правило $P[1]$ предназначено для присоединения информационных единиц, соответствующих словам "некоторый", "каждый" , "какой-нибудь", "все", "несколько", "большинство" (такие информационные единицы в данной работе называются интенсиональными кванторами) к простым или составным обозначениям понятий. Поэтому правило $P[1]$позволяет строить формальные аналоги выражений: "некоторый человек", "все люди", "большинство людей", "некоторый человек ростом 175 см", "все тридцатилетние люди", "все города Европы". Примерами l-формул (К-цепочек) для $P[1]$, как последнего примененного правила, являются цепочки *нек чел, все чел* $*$*(Возраст, <30, год>), все город*$*$*(Регион, Европа)* . Правило $P[2]$ предназначено для построения цепочек вида $f(a_1, \ldots, a_n)$, где f — обозначение функции, $n \geq 1$, $a_1, \ldots, a_n$ — l-формулы, построенные с применением каких-то правил

из списка $P[0], P[1], \ldots, P[10]$. Например, после применения правила на последнем шаге вывода можно получить цепочки *Города*(*Европа*), *Колич-элем*(*Города*(*Европа*)).

Правило $P[3]$ позволяет строить цепочки вида ($a_1 \equiv \quad a_2$), где a_1, a_2 — l-формулы, полученные при помощи любых правил из $P[0], \ldots, \ldots P[10]$, и a_1, a_2 обозначают сущности, являющиеся однородными в некотором смысле. Примеры К-цепочек для $P[3]$ как последнего примененного правила: ($y_1 \equiv$ *нек город**(*Название, 'Саратов'*), (*Директор*(*АО_"Салют"*)≡*П. Сомов*).

Правило $P[4]$позволяет строить К-цепочки вида $r(a_1, \ldots, a_n)$, где r — n-арное отношение, $n \geq 1$, $a_1, \ldots, a_n$ — К-цепочки, полученные при помощи некоторых правил из $P[0], \ldots, P[10]$. Примеры К-цепочек для $P[4]$: *Принадлеж(Намюр, Города*(*Бельгия*)), *Подмнож*(*Города* (*Бельгия*), *Города*(*Европа*)).

Правило $P[5]$ предназначено для построения К-цепочек вида $d : v$, где d — К-цепочка, не включающая v, v — переменная, и выполнены некоторые условия. При помощи правила $P[5]$ можно помечать переменными в СП текстов на естественном языке: а) описания различных сущностей, встречающихся в тексте (физических объектов, событий, понятий и др.), б) семантические представления предложений или более крупных фрагментов текста, на которые имеется ссылка в любой части текста. Примерами К-цепочек для правила $P[5]$, примененного на последнем шаге вывода, являются выражения *все чел : Z1*, *Меньше*(*Возраст*(*П. Сомов*), *<30, год>*) *: P1*. Это правило дает возможность создавать СП текстов таким образом, что они отражают референтеую структуру текста на ЕЯ.

Правило $P[6]$ позволяет строить К-цепочки вида $\neg d$, где d — К-цепочка, удовлетворяющая ряду условий. Примеры К-цепочек для $P[6]$: ¬*биолог,* ¬*Принадлеж*(*Бонн, Города*(*Бельгия*)). Здесь ¬ обозначает связку "не".

При помощи правила $P[7]$ можно строить К-цепочки вида ($a_1 \wedge \ldots$ $\ldots \wedge a_n$) или ($a_1 \vee \ldots \vee a_n$), где $n > 1$, $a_1, \ldots, a_n$ — К-цепочки, обозначающие однородные в некотором смысле сущности. В частности, $a_1, \ldots, a_n$ могут быть семантическими представлениями высказываний, описаниями физических объектов, описаниями множеств, состоящих из объектов одной природы, описаниями понятий. Следующие цепочки являются примерами К-цепочек (или l-формул) для $P[7]$: (*Финляндия* ∨ *Норвегия* ∨ *Швеция*), (*Принадлеж*((*Намюр* ∧ *Гент*), *Города*(*Бельгия*)) ∧ ¬*Принадлеж*(*Бонн, Города*((*Финляндия* ∨ *Норвегия* ∨ *Швеция*)))).

Назначение правила $P[8]$ состоит в том, что оно позволяет строить, в частности, К-цепочки вида $c * (r_1, b_1), \ldots, (r_n, b_n)$, где c — информационная единица из первичного универсума X, обозначающая по-

нятие, для $i = 1, \ldots, n$, r_i — функция одного аргумента или бинарное отношение, b_i обозначает возможное значение r_i для объектов, характеризующихся понятием c. Например, если выбрать соответствующим образом первичные информационные единицы, то после применения на последнем шаге вывода правила $P[8]$ можно получить К-цепочки *чел*∗(*Имя, 'Петр'*)(*Фамилия, 'Сомов'*), *поворот*∗(*Направление, левое*).

Правило $P[9]$ дает возможность строить, в частности, К-цепочки вида $\forall v(\textit{des})D$ и $\exists v(\textit{des})D$, где $\forall$ — квантор всеобщности, $\exists$ — квантор существования, *des* обозначает понятие "город", "целое число" и др.) или составные понятия ("целое число, большее 200" и др.). D можно интерпретировать как СП высказывания с переменной v о любой сущности, характеризуемой понятием *des*. Примеры К-цепочек для $P[9]$ как правила, примененного на заключительном шаге построения формулы: $\forall$*x1*(*нат.ч.*) $\exists$*x2*(*нат.ч.*) *Меньше*(*x1, x2*), $\exists$*y*(*страна*∗(*Регион, Европа*))*Больше*(*Колич*(*Города*(*y*)), *15*).

Правило $P[10]$ позволяет строить, в частности, К-цепочки вида $<a_1, \ldots, a_n>$, где $n > 1$, $a_1, \ldots, a_n$ — К-цепочки. Цепочки, получаемые с использованием правила $P[10]$ на последнем шаге вывода, интерпретируются как обозначения n-местных наборов. Компонентами таких наборов могут быть не только обозначения чисел, объектов, но и семантические представления выражений, множеств, понятий и др.

Некоторые возможности построения семантических аннотаций средствами СК-языков. Изложенная выше схема дает только очень упрощенное представление о большой серии определений, завершающейся определением класса СК-языков (см. [9, 10]). В связи с этим рассмотрим на примерах некоторые новые возможности, предоставляемые аппаратом СК-языков для построения семантических аннотаций Веб-документов.

Пусть T — это выражение на ЕЯ, $Semr$ — цепочка СК-языка в некотором концеп-туальном базисе, которую можно интерпретировать как семантическое представление выражения T. Тогда будем говорить, что $Semr$ — К-представление (КП) выражения T.

Пример 1. СК-языки позволяют строить составные обозначения различных сущностей. Так, антибиотик "Зиннат" может быть обозначен К-цепочкой *нек антибиотик*∗(*Назв1, "Зиннат")*:v, где элемент $нек$ интерпретируется как информационная единица, соответствующая словам "определенный", "некоторый", а v — переменная, помечающая этот конкретный антибиотик.

Пример 2. Рассмотрим определение Def1 = "A flock (английский язык) — это большое количество птиц или млекопитающих (например, овец или коз), собирающихся вместе с определенной целью, такой, как питание, миграция или оборона". Тогда определение Def1 может

иметь следующее К-представление (КП) *Expr1* (т. е. СП, являющееся выражением некоторого СК-языка):

Определение 1 (*flock, англ-яз, динамич-группа*∗(*Кач-состав,* (*птица* ∨ *млекопитающее*∗ (*Примеры,* (*овца* ∨ *коза*)))), *S1,* (*Оценка*(*Колич-элем*(*S1*), *большое*) ∧ *Цель-формирования* (*S1, нек намерение*∗(*Примеры,* (*питание* ∨ *миграция* ∨ *оборона*))))).

Анализ этой формулы позволяет сделать вывод о том, что при построении СП ЕЯ-текстов удобно использовать: 1) обозначение 5-арного отношения *Определение 1,* 2) составные обозначения понятий (в данном примере использованы выражения *млекопитающее*∗ (*Примеры,* (*овца* ∨ *коза*)) и *динамич-группа*∗ (*Кач-состав, (птица* ∨ *млекопитающее*∗ (*Примеры,* (*овца* ∨ *коза*))))), 3) имена функций, аргументами и/или значениями которых могут быть множества (в примере использовано имя одноместной функции *Колич-элем,* значением которой является количество элементов множества), 4) составные обозначения намерений, целей (в примере — выражение *нек намерение* ∗ (*Примеры,* (*питание* ∨ *миграция* ∨ *оборона*))).

Пример 3. Определение Def1 взято из издания "Longman Dictionary of Scientifc Usage", опубликованного в Москве в 1989 г. СК-языки позволяют представлять опреде-ления и другие модули знаний в объектно-ориентированной форме, т. е как как выражения, отражающие не только содержание модуля знаний, но и метаданные, т. е. значения таких внешних характеристик, как авторы, источник, дата опубликования и т.д. Например, информация об определении Def1 может быть представлена как К-цепочка *Нек информ-объектt*∗(*Вид, определение*)(*Инф-содержание, Expr1*)(*Источник 1, нек словарь 1*∗(*Назв, 'Longman Dictionary of Scientifc Usage'*)(*Издательство,* (*Longman-Group-Limited/Harlow* ∧ *Russky-Yazyk-Publishers/Moscow*))(*Город, Москва*)(*Год, 1989*)), где *Expr1* — это построенное выше К-представление определения Def1.

Пример 4. Пусть D1 — относящийся к биологии и медицине дискурс "Все гранулоциты являются полиморфонуклеарными. Это означает, что их ядра многодольны". Тогда дискурсу D1 можно поставить в соответствие следующее К-представление *Expr2*:

(*Свойство* (*произвольн гранулоцит* : *x1 , полиморфонуклеарный*)*: P1*) ∧ *Пояснение* (*P1, Следует-из* (*Ситуация (e1, обладание1*∗(*Агент1, x1*) (*Объект1, нек ядро1* : *x2*)), *Свойство* (*x2, многодольный*)))).

Ключевую роль в построении КП *Expr2* сыграло правило, позволившее ввести метку *x1* для обозначения произвольного гранулоцита, метку *x2* для обозначения ядра клетки, и метку $P1$ для обозначения СП

первого предложения из дискурса D1. Метка $P1$ позволяет в структуре СП текста D1 эксплицировать ссылку на смысл первого предложения текста, даваемую сочетанием “Это означает”.

Преимуществами теории СК-языков по сравнению с теорией представления дискурсов [5] и эпизодической логикой [7] являются, в частности, возможности: 1) различать формальным образом обозначения объектов, ситуаций и понятий, характеризующих эти объекты, ситуации, 2) строить составные обозначения понятий, 3) различать формальным образом объекты и множества объектов, понятия и множества понятий, 4) строить формальные аналоги составных обозначений множеств, а также множеств, состоящих из множеств, 5) описывать теоретико-множественные отношения, 6) эффективно описывать структурированные значения (СЗ) дискурсов со ссылками на смысл фраз и более крупных фрагментов дискурсов, 7) описывать СЗ предложений со словами “понятие”, “концепт”, 8) описывать СЗ выражений, полученных применением связок “и”, “или” нек обозначениям высказываний, а к обозначениям предметов, событий, понятий ; 9) строить составные обозначения объектов и множеств, 10) рассматривать нетрадиционные функции (и другие нетрадиционные отношения) с аргументами и/или значениями, являющимися множествами предметов, ситуаций, понятий, СП текстов, 11) строить формальные аналоги значений инфинитивов с зависимыми словами, т. е. обозначения целей, рекомендаций, предложений, обязательств, назначений предметов и процессов.

Пункты (3)–(8), (10), (11) указывают принципиальные преимущества теории СК-языков по сравнению с теорией концептуальных графов (ТКГ) [6]. Кроме того, выразительные возможности СК-языков значительно шире возможностей ТКГ в отношении пунктов (1), (2), (9).

Модель лингвистической базы данных и новый алгоритм семантико-синтаксического анализа ЕЯ-текстов. На протяжении 1990-х и первой половины 2000-х годов автором был выполнен цикл исследований, направленных на создание эффективных формальных средств и методов проектирования семантико-синтаксических анализаторов (ССА) текстов на русском, английском и многих других языках. Полученные результаты дают новую систему формальных средств и методов проектирования ССА вопросо-ответных и информационно-поисковых Интернет-систем нового поколения. Основные полученные результаты очень кратко можно охарактеризовать следующим образом.

1. Построена формальная модель лингвистической базы данных (ЛБД), содержащей такие сведения о лексических единицах и их взаимосвязях с информационными единицами, которые достаточны для семантико-синтаксического анализа интересных для приложений подъ-

языков русского языка. С этой целью определено понятие лингвистического базиса [11].

2. Предложен новый метод преобразования ЕЯ-текстов в их СП. Метод предусматривает использование предложенного автором матричного семантико-синтаксического представления (МССП) входного текста как промежуточного представления при переходе от ЕЯ-текста к СП текста, являющемуся выражением некоторого СК-языка (т. е. К-представлением текста). При этом не используется традиционное синтаксическое представление текста [11, 12].

3. Разработан структурированный алгоритм семантико-синтаксического анализа текстов из представляющих практический интерес подъязыков естественного (русского) языка (алгоритм SemSyn1), базирующийся на построенной формальной модели лингвистической базы данных (ЛБД) и на введенном понятии МССП ЕЯ-текста. Алгоритм устанавливает смысловые отношения между элементарными значащими единицами входного текста, отражая эти отношения посредством МССП, а затем строит семантическое представление (СП) текста, являющееся К-представлением текста. Входные ЕЯ-тексты могут выражать высказывания (сообщения), команды, специальные вопросы (т. е. вопросы с вопросительными словами), общие вопросы (т. е. вопросы с ответом “Да”/ “Нет”)и могут, в частности, включать причастные обороты и придаточные определительные предложения.

Алгоритм SemSyn позволяет устанавливать возможные смысловые отношения, в частности, в сочетаниях “Глагол + Предлог + Существительное”, “Глагол + Существительное”, “Существительное1 + + Предлог + Существительное2”, “Число + Существительное”, “Прилагательное + Существительное”, “Существительное1 + Существительное2”, “Причастие + Существительное”, “Причастие + Предлог + Существительное”, “Вопросительно-относительное местоимение или местоименное наречие, играющее роль вопросительного слова + Глагол”, “Предлог + Вопросительно-относительное местоимение + Глагол”. Алгоритм существенно использует ряд новых выразительных возможностей, предоставляемых определением класса СК-языков [11].

Пример 1. Пусть Т1 = “Антибиотик “Зиннат”, выпускаемый фирмой “GlaxoWelcome”, излечивает болезни, вызванные кокковой флорой”. Тогда алгоритм SemSyn построит по тексту Т1 его К-представление (*Ситуация(e1, выпуск1 *(Агент2, нек фирма1“* (*Назв, “GlaxoWelcome”*) : *x1*)(*Время, #сейчас#*)(*Продукция1, нек антибиотик ** (*Назв, “Зиннат”*) : *x2*) ∧ *Ситуация* (*e2, лечение1*∗(*Агент1, x2*)(*Процесс1, все болезнь1*∗(*Причина, произв флора*∗(*Вид1, кокк*)))). Таким образом, алгоритм SemSyn может использоваться для построения семантических аннотаций Веб-документов.

Пример 2. Пусть В1 = “Сколько английских университетов используют для дистанци-онного обучения Интернет-платформу Blackboard?”. Тогда для некоторой лингвистической базы данных алгоритм SemSyn построит по вопросу В1 его КП в виде цепочки Semrepr1 = *Вопрос* (*x1*, ((*x1* ≡ *Колич*(*S1*)) ∧ *Качеств-состав* (*S1*, *университет*∗(*Регион*, *Англия*)) ∧ *Описание*(*произв университет* ∗(*Элем*, *S1*) : *y1*, *Ситуация*(*e1*, *использование*∗(*Время*, *#сейчас#*) (*Агент1*, *y1*)(*Процесс*, *обучение*∗(*Вид*, *дистанцион*))(*Объект1*, *нек платформа3*∗(*Название*, *‘Blackboard’*))))))).

Фрагментами цепочки Semrepr1 являются: а) составное обозначение понятия *университет* ∗(*Регион*, *Англия*), б) семантическая характеристика произвольного элемента множества *произв университет* ∗(*Элем*, *S1*) : *y1*, в) составное обозначение объекта *нек платформа3*∗ (*Название*, *‘Blackboard’*). Одно из правил построения выражений СК-языков позволило связать метку (переменную) $y1$ с характеристикой произвольного элемента искомого множества $S1$, а затем использовать только эту метку для последующих ссылок на эту характеристику.

Пример 3. Пусть В2 = “Проходила ли в Азии международная научная конференция “COLING”?”. Тогда в рамках некоторой лингвистической базы данных алгоритм SemSyn построит КП вопроса В2 в виде цепочки:

Semrepr2 = *Вопрос*(*x1*, (*x1*≡ *Ист-знач* (*Ситуация* (*e1*, *прохождение2*∗(*Время*, *нек мом* ∗ (*Раньше* ,*#сейчас#*) : *t1*)(*Событие*, *нек конф*∗(*Вид1*, *междун*) (*Вид2*, *научная*) (*Название*, *‘COLING’*) : *x2*) (*Место*, *нек континент*∗ (*Название*, *‘Азия’*) :*x3*))))).

В выражении Semrepr2 цепочка $Ист-знач$ интерпретируется как обозначение функции, аргументом которой является СП высказывания, а значением — логическая величина Истина или Ложь.

Важная особенность нового метода и алгоритма SemSyn1 заключается в том, что они не предусматривают использования синтаксического уровня представления (как резуль-тата выполнения синтаксического анализа) текста. Существенным преимуществом разработанного алгоритма является явный учет многозначности слов, что чрезвычайно важно для приложений. Алгоритм SemSyn реализован в программной среде Visual C++.

Выводы. Разработана широко применимая система формальных инструментов для построения семантических аннотаций Веб-документов. Первым инструментом является теория стандартных К-языков (СК-языков). Анализ выразительной силы класса СК-языков позволяет высказать предположение о том, что СК-языки позволяют строить семантические аннотации произвольных Веб-документов и удобны

для построения таких аннотаций. Поэтому определение класса СК-языков можно интерпретировать как широко применимую и гибкую (по-видимому, универсальную) формальную метаграмматику семантических аннотаций Веб-документов.

Теория СК-языков была успешно использована при разработке широко применимой математической модели лингвистической базы данных (ЛБД) и сложного структурированного алгоритма семантико-синтаксического анализа текстов из представляющих практический интерес подъязыков естественного (русского) языка, базирующегося на построенной модели ЛБД. Алгоритм реализован в системе программирования Visual C++ и может быть широко использован для построения семантических аннотаций. Веб-документов.

СПИСОК ЛИТЕРАТУРЫ

1. S e m a n t i c Web Activity Statement. W3C, 2001, URL http://www.w3.org/2001/sw/activity.
2. “V o i c e Browser” Activity — Voice enabling the Web. W3C paper, 2001, http://www.w3.org/Voice/.
3. K a t z B., L i n J. Annotating the Semantic Web using natural language // Proc. of the 2^{nd} Workshop on NLP and XML (NLPXML 2002) in conjunction with COLING 2002, Taipei, Taiwan, 2002.
4. F o m i c h o v V. A. An Ontological Mathematical Framework for Electronic Commerce and Semantically-structured Web // Zhang, Y., Fomichov, V.A., Zeleznikar, A.P. (eds.): Special Issue on Database, Web, and Cooperative Systems. Informatica, Slovenia. – 2000. – Vol. 24. – № 1. – P. 39–49.
5. K a m p H., R e y l e U. A Calculus for First Order Discourse Representation Structures // Journal for Logic, Language and Information. – 1996. – Vol. 5. - -P. 297–348.
6. S o w a J.F. Conceptual Graphs: Draft Proposed American National Standard // Tepfenhart, W., Cyre, W. (eds.): Conceptual Structures: Standards and Practices. Lecture Notes in AI, Vol. 1640. Springer-Verlag, Berlin Heidelberg New York. – 1999. – P. 1–65.
7. S c h u b e r t L. K., H w a n g, C. H. Episodic Logic Meets Little Red Riding Hood: A Comprehensive, Natural Representation for Language Understanding // Iwanska, L., Shapiro, S.C (eds.): Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language. MIT/AAAI Press, Menlo Park, CA, and Cambridge, MA. – 2000. – P. 111–174.
8. F o m i c h o v V. A. Mathematical Model for Describing Structured Items of Conceptual Level // Informatica. An International Journal of Computing and Informatics (Slovenia). – 1996. – Vol. 20. – №. 1. – P. 5–32.
9. Ф о м и ч е в В. А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть I. Модель системы первичных единиц концептуального уровня // Информационные технологии. – 2002. – № 10. – С. 16–25.
10. Ф о м и ч е в В. А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть II. Система правил для построения семантических представлений фраз и сложных связных текстов // Информационные технологии. – 2002. – № 11. – С. 34–45.

11. Ф о м и ч е в В. А. Формализация проектирования лингвистических процессоров. – М.: МАКС Пресс, 2005. – 367 с.

12. F o m i c h o v V. A. The Method of Constructing the Linguistic Processor of the Animation System AVIAROBOT // Pohl, J. (ed.): Proceedings of the Focus Symposium on Collaborative Decision-Support Systems; InterSymp-2002, the 14^{th} International Conference on Systems Research, Informatics and Cybernetics, July 29 – August 3, 2002, Germany. CAD Research Center, Cal Poly, San Luis Obispo, CA, USA, 2002. – P. 91–102.

Статья поступила в редакцию 30.03.2005

Владимир Александрович Фомичев родился в 1950 г., окончил в 1973 г. МИЭМ. Канд. техн. наук, профессор кафедры “Информационные технологии” Российского государственного технологического университета им. К.Э. Циолковского — “МАТИ”, доцент кафедры “Математическое и программное обеспечение систем обработки информации и управления” Московского государственного института электроники и математики (технического университета). Почетный профессор Международного института передовых исследований по системному анализу и кибернетике (г. Виндзор, Онтарио, Канада). Автор более 125 научных работ в области дискретной математики, математической теории интеллектуальных систем, проектирования лингвистических процессоров, теории многоагентных систем, теории электронной коммерции, теории семантической “паутины”, когнитивной науки и теории образования.

V.A. Fomichov (b. 1950) graduated from Moscow Institute of Electronic Machine-Building in 1973. Ph. D. (Eng.), professor of “Information Technologies” department of the Russian State Technological University n. a. K.E. Tsiolkovsky — “MATI”, assoc. professor of “Mathematical Methods and Software for Information Processing and Control Systems” department of Moscow State Institute for Electronics and Mathematics (Technical University). Honoured Professor of the International Institute for Advanced Studies in Systems Research and Cybernetics (University of Windsor, Ontario, Canada). Author of more than 125 publications in the field of discrete mathematics, artificial intelligence, multi-agent systems, semantic web, e-commerce, cognitive science and theory of education.