

# НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ  
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 5

Москва 2013

## ОБЩИЙ РАЗДЕЛ

УДК 001.102

Е. И. Полтавская

### Информация: к уточнению понятия

*Рассматриваются различные концепции информации и делается вывод: нет оснований для того, чтобы считать информацию атрибутом материи, а ее понятие – философской категорией.*

**Ключевые слова:** информация, когнитивная информация, идеальное, материальное, информационные науки

Все, что не относится к материи, с давних пор обозначается термином «идеальное». Решение проблемы взаимоотношенности материального и идеального – основной вопрос философии. Это та исходная точка, от которой отталкивается исследовательская мысль при решении других проблем. Проблема идеального рассматривалась в отечественной философии и науке в двух направлениях: при изучении сознания и процессов мышления, а также в историческом материализме – как изучение общественного сознания и культурных ценностей. Целостный подход к категории «идеальное» впервые был предпринят Д. И. Дубровским (1983) в книге «Проблема идеального. Субъективная реальность» [1], который вслед за классиками марксизма-ленинизма определил *идеальное в качестве субъективной реальности*. Причем понятие субъективной реальности относится не только к целостному внутреннему миру человека, но и к любому отдельному его прояв-

лению: мысли, представлению и т. п. [1, § 1.3]. При широком распространении термина «информация», который многими исследователями полагается как обозначение идеальной реальности, рассмотрим подробнее эти понятия.

Понятие «информация» неясно и многозначно: на запрос в Интернете поисковые системы выдают более миллиарда текстовых ответов и картинок. Подсчитано, что за жизнь человек накапливает объем информации, равный десяти терабайтам [2, с. 114]. Информация признана не только не убывающим, но и постоянно прирастающим ресурсом (хотя может искажаться и теряться при копировании). Поэтому естественно считать информацией приращение знания. Изучать информацию призваны различные информационные науки, среди которых – информатика.

Становление информатики в нашей стране как науки прошло ряд стадий. В начале своего развития информатика мыслилась по образу гуманитарной

науки, затем – подобно технической, естественной. Ныне информатика все чаще позиционируется как «*комплексное междисциплинарное научное направление*» – фундаментальная наука, имеющая первостепенное значение для естествознания и гуманитарных наук [3, с. 35]. Причиной эволюции информатики является развитие понятия «информация» (которую предложено понимать «как меру разнообразия в природе и обществе, т. е. как степень организационной сложности технических, природных и социальных систем») и осознание роли информационных процессов в развитии живой и неживой природы [3, с. 35].

Науки, ныне именуемые документо-коммуникационными, еще в конце 1970-х гг. были включены в сферу интересов информатики. Информатика претендовала тогда на статус обобщающей науки о документе. Некоторые специалисты документо-коммуникационных наук считают сверхзадачей своей отрасли знаний передачу информации, адекватной потребностям пользователей [4]. Поэтому предложение К. К. Колина о введении класса «Информационные науки» в Номенклатуру специальностей научных работников, куда планируется включить и документо-коммуникационные дисциплины, было воспринято с воодушевлением. Сейчас для предстоящего обсуждения в ВАКе подготовлен консолидированный вариант мнений разных специалистов, особая ценность которого в том, что в разделе 26.04.00 *Информационные системы и процессы в социальной сфере* объединены собственно документо-коммуникационные науки, ранее разбросанные по разным группам и отраслям.

Понятие «информация» в различных отраслях знания трактуется по-разному, и пока не удалось создать общую теорию информации. Отдельные исследователи, исходя из информационного единства окружающего мира, объясняют многоаспектность термина «информация» существованием иерархии множества структурных уровней Универсума, каждому из которых соответствуют свои информационные свойства [5, с. 76; 3, с. 72] (ранее эта точка зрения – из иерархического строения материи выводить и иерархичность информации – разделялось и мной). Распространены представления, что информация не является отдельной субстанцией, что одна и та же информация может быть на разных носителях; «на одном и том же носителе может одновременно находиться информация от нескольких никак не связанных друг с другом источников» и др. [5, с. 80; 81]. Согласно теории информации, информация существует, если появляются основания для избрания одного из вариантов решения вопроса: «информация – это некое идеализированное сообщение, уменьшающее или полностью исключающее неопределенность в выборе одной из нескольких возможных альтернатив» [6, с. 4]. При характеристике информации в профессиональной печати не обходится без ярких выражений: «информация – это беспокойство материи», «двигатель развития» [Цит. по: 5, с. 91]. Высказано также мнение, что термин «информация» – это «просто омоним, который используется для обозначения совершенно различных и не связанных между собой сущностей» [Цит. по: 3, с. 67].

Полагают, что информация может считываться (или не считываться) с объекта в зависимости от свойств окружающих его субъектов/объектов. При взаимодействии материальных систем на них остаются следы воздействий, которые считаются основанием для приобретения потенциальной информации об этих системах. Материальные системы, способные к активному поведению, ориентируются «на эффект воздействия как на сигнал включения такой активности .., потенциальная информация превращается в *актуальную информацию*» [7, с. 524]. Развивая эту мысль, К. К. Колин трактует след деформации на материальном объекте как возникший новый объект «идеальной реальности» [3, с. 70]. Однако претерпевший изменения в физическом процессе материальный объект неживой природы остается тем же самым. След деформации на нем имеет пространственную протяженность и представляет собой, например – если этот объект кристаллический – разрушение какой-то части кристаллической решетки. Ничего идеального – обладающего интенциональностью и не имеющего пространственных и физических свойств – при этом не возникает. Информацией это событие станет лишь для субъекта, который будет его анализировать. Для понимания таких явлений в неживой природе понятие «информация» не требуется. Поэтому при уточнении понятия «информация» *потенциальная информация* исключается из нашего рассмотрения как *неинформация*. Правда, надо осознать, что *потенциальная информация* – это удобная метафора, которая позволяет представить эволюцию материальных систем, их иерархию.

Поскольку активно реагировать на материальное воздействие могут лишь живые системы, а также созданные искусственно, то и информацией могут обладать только они, считают приверженцы *семантических* (культурологических, антропоцентристских, социоцентрических) концепций. Допускается, что информация присуща лишь людям, которые владеют языком, сознанием и самосознанием. Информации не бывает без смысла, убеждены сторонники этой группы теорий (А. И. Михайлов, А. И. Черный, Р. С. Гиляревский, Г. Г. Белоногов и др.). Она всегда имеет некое содержание, которое передается определенными единицами смысла (понятиями, предложениями). Возникая вместе с человеком, информация «представляет собой субъективный образ объективного мира, опосредованный генетической и социальной природой человека, но тем не менее более или менее адекватно отражающий объективную реальность» [8, с. 5].

Разделяя эту точку зрения, добавим, что по представлениям современных специалистов, занимающихся когнитивной наукой, не только люди, но и высшие животные способны к мыслительным актам (но оперируют они, в отличие от человека, невербальными репрезентациями) [9, с. 252]. Организмы, обладающие нервной системой, способны интерпретировать сигналы внешнего мира и сообразно на них реагировать, выбирая необходимую реакцию, – осуществлять информационный контроль среды, что позволяет им адаптироваться и выживать. Поэтому ин-

формацию, которую создают живые организмы, называют *когнитивной* (от лат. *cognitio* – познание, узнавание, познание) [6, с. 6]. Когнитивная информация формируется в согласии с эволюцией генома, который можно представить как «программное обеспечение», определяющее «формат внутренних ментальных репрезентаций» [6, с. 33].

Тем не менее многие исследователи, признавая существование специфического субъективного мира у животных, ограничивают сферу идеального субъективным миром человека – на том основании, что у животных она «работает» иначе (даже у приматов не выявлены явления рефлексии, абстрактного мышления, самопроектирования). Поскольку Д. И. Дубровский предложил обозначить сферы субъективного мира человека и животных разными терминами [1, § 4.1], предлагаю информацию, создаваемую в субъективном мире человека, называть *субъективной информацией* (в отличие от когнитивной, создаваемой всеми живыми организмами).

Основоположник отечественной теории семантической информации Ю. А. Шрейдер определял информацию как «знание, отчужденное от непосредственного владельца и обобществленное путем вербализации и закрепления на материальном носителе» [10, с. 23]. Его позиция согласуется с общенаучным представлением об информации: информация, по сути, толкуется как опредмеченное идеальное (опредмечивание и распредмечивание – два взаимосвязанных процесса) [11, с. 157].

Среди основных философских гипотез о природе информации выделяют также атрибутивные и функциональные.

*Функциональные* концепции связывают информацию с управлением, во многом пересекаясь с семантическими концепциями. Сторонники этой группы (П. В. Копнин, В. С. Тютин, Н. Н. Моисеев, Д. И. Дубровский и др.) признают информацию свойством самоорганизующихся систем (биологических, социальных и включенных в них био- и социотехнических) и опираются на «принцип инвариантности информации по отношению к физическим свойствам ее носителя» (химическим, субстратным, пространственным, временным) [12, с. 86].

Д. И. Дубровский подчеркивает, что, во-первых, информации не существует вне определенного физического объекта, процесса; во-вторых, одна и та же информация может быть передана разными по своим физическим свойствам носителями; в третьих, информация может быть причиной определенных изменений в самоорганизующейся системе, т. е. осуществлять информационное управление [13, с. 98]. Следствием сказанного является то, что информация (о чем-то) представляет собой явление субъективной реальности, «сложившуюся кодовую зависимость, которая носит функциональный характер» [12, с. 86]. Понимание информации означает процесс перекодирования материальных сигналов во внутренние коды индивида. Информация, по Дубровскому, дается человеку в виде явлений его субъективной реальности (образов, мыслей, целей, желаний и т. п.). При этом ни одно из явлений субъективной реальности не существует обособленно от материального носителя,

что исключает идеалистическое или дуалистическое толкование идеального. Идеальное философ связывает с тремя видами кодов, различая мозговой нейродинамический код (фундаментальный) и связанные с ним бихевиорально-экспрессивный (двигательные акты) и речевой [1, § 4.2; 13]. Отсюда напрашивается, например, вывод, что мышление осуществляется не обязательно вербально, а двигательные акты могут осуществляться без их осмысления.

Сторонники *атрибутивных* концепций (А. Д. Урсул, И. Б. Новик, Л. А. Петрушенко, К. К. Колин и др.) понимают информацию как аспект любого вида отражения, информация, как и материя, не мыслится вне движения, изменения, различия [12, с. 85, 91; 14, с. 40]. Поскольку «отражение – это воздействие одной материальной системы на другую, ведущее к установлению определенного (конкретного) тождества между системами, когда внутренние различия одной системы (отражающей) соответствуют внутренним различиям другой системы (отражаемой)» [15, с. 176], то информацию можно трактовать «как содержание отражения», которое – в разной мере – соответствует изменениям и отражаемой системе, и отражаемому. А. Д. Урсул делает вывод: «информация с позиций теории отражения может быть представлена как отраженное разнообразие» [15, с. 177]. При этом им подчеркивается, что такая информация не является полной, т. к. она ограничена определенным классом различий. А. Д. Урсул считает, что «если объективная (и потенциальная для субъекта) информация может считаться свойством материи, то идеальная, субъективная информация есть отражение объективной, материальной информации» [15, с. 229].

Сходная с А. Д. Урсулом позиция у К. К. Колина, который различает *первичную* (связанную) информацию, представляющую собой «организованное по определенным правилам пространственное размещение материи» (по сути, строение вещества), *вторичную* информацию, которая есть «отраженное разнообразие» и *третичную* информацию, «которая представляет некоторый образ наблюдаемого объекта» и которая формируется в сознании наблюдателя [3, с. 86].

По воззрениям некоторых ученых, придерживающихся атрибутивной концепции, вещество, энергия и информация являются понятиями одного уровня общности, характеризующими свойства материи [5, с. 84]. Причем, по Колину, именно информация, понимаемая как асимметрия, «стимулирует материальные и энергетические системы к движению в пространстве и определяет направленность этого движения», а также является «решающим фактором в развитии эволюционных процессов в материальных или же энергетических системах» [3, с. 92] (правда, возникает вопрос, что является причиной асимметрии).

Из-за того, что не удастся найти отличительные, соответствующие только ей, референты информации – объекты внеязыковой действительности, возник тезис *информация – везде* (т. е. информация и везде и нигде). Из него следует, что понятие «информация» должно считаться философской категорией, поскольку

ку «категории нашего познания являются всеобщими (философскими) тогда, когда они отражают атрибуты материи и их взаимосвязи, участвуют в решении основного вопроса философии» [16, с. 350]. Против такого взгляда выступает ряд авторов. Чрезмерно расширительное толкование информации, по Дубровскому, «полностью растворяет понятие информации либо в комплексе понятий классической физики, либо в комплексе философских категорий», затеняя сущность информационного подхода и его отличие от энергетического подхода [17, с. 245].

Поскольку сейчас появилась новая концепция о существовании темной материи, в которой нет разнообразия, а, значит, не должно быть и информации, то Урсул признал некорректным само название подхода, сторонником которого он сам является, – атрибутивный – и считает информацию общенаучной, а не философской категорией [12, с. 91]<sup>1</sup>.

Итак, согласно взглядам сторонников атрибутивной концепции, информация проявляется в разных формах. Под термином «информация» ими понимаются несколько «информаций», среди которых и «объективная материальная информация», и «содержание отражения», и «образ наблюдаемого объекта». Еще раз отметим, что «объективная (и потенциальная для субъекта) информация» нами понимается как строение материальных предметов, к которому некорректно применять термин «информация», поскольку субъект получит, т. е. *создаст* информацию о строении вещества лишь тогда, когда начнет осмысливать данные эксперимента с объектом в своей голове. Чем отличаются «содержание отражения» и «образ наблюдаемого объекта» не ясно, ведь воспринимает содержание информации тот же субъект, у которого возникает ее образ. Обе эти «информации», по-видимому, относятся авторами к идеальному, но выражение «объективная материальная информация» заставляет задуматься о различении материального и идеального, поскольку их совмещение в одном термине приводит к нивелированию содержания этих важных понятий и затеняет исходную позицию авторов, затрудняя решение проблемы.

Исходя из того, что многие исследователи считают информацию идеальным феноменом, обратимся к понятию «идеальное» и к противостоянию идеальному и материальному вообще. Представляется, что понятие идеального, разрабатываемое отечественными философами и учеными с различными позициями, несмотря на, казалось бы, несовместимость некоторых их взглядов, может дать «ключ» к пониманию информации, ограничению этого понятия от других.

Проблему идеального с историко-материалистических позиций исследовал Э. В. Ильенков. Согласно его концепции, «идеальность ... есть не что иное, как представленная в вещи форма общественно-человеческой деятельности. Или, наоборот, форма человеческой деятельности, представленная как вещь, как предмет» [18, с. 189]. Таким образом, Ильенков ограничивал сферу идеального только теми

духовными явлениями, которые признаны всеобщими и необходимыми. Так, все достижения социума и культуры, зафиксированные в устном творчестве, воплощенные материально – от предметов быта и технических достижений до величайших произведений литературы, искусства и т. д. (то, что называется социокультурным знанием), отнесены философом к *объективному идеальному*. Ибо, согласно Ильенкову, *идеальность* – «это своеобразная печать, наложенная на вещество природы общественно-человеческой деятельностью, это форма функционирования физической вещи в процессе общественно-человеческой жизнедеятельности. Поэтому-то все вещи, вовлеченные в социальный процесс, и обретают новую, в физической природе их никак не заключенную и совершенно отличную от последней “форму существования”, идеальную форму» [18, с. 189–190]. И далее: «идеальность есть характеристика вещей ... той определенности, которой они обязаны труду» [18, с. 210].

Термин «идеальное» понимается Дубровским диаметрально противоположно с Ильенковым – как субъективная реальность. Полностью разделяя критику позиции Ильенкова Дубровским и одобряя его аргумент, что идеальное не существует внеличностно [1, § 1.4], заметим, что понимание Ильенковым отражения можно интерпретировать как соответствие, *тождество различного* – идеи, идеального образа, понятия – реальному предмету [19, с. 22]. Причем образ этот мыслился не как статичный предмет, а как возникающий лишь при созидании вещи. «Представленная в вещи форма общественно-человеческой деятельности», «та определенность» вещи, которая создается в труде, деятельности – идеальное по Ильенкову, думается, и есть идея, а в современном понимании – абстрактный пропозициональный код, *схема* вещи, которую человек представляет, создавая вещь или продумывая предстоящие действия [9, с. 229, 254].

Получается, объективным идеальным (имеющим характер внеличностного и надличностного) Ильенков называл *всеобщую форму объекта* – значимые для социума интеллектуальные конструкты, логические структуры, которые человек строит в уме при осмыслении чего-либо, в процессе жизнедеятельности, чтобы затем передать знание другим людям. Этими наработанными схемами, отражающими инвариантные структуры окружающего мира (и отражающими нормы культуры), обладает – в понятиях – каждый человек, выросший в социуме, часто не задумываясь об этом. То, что Ильенков относит к объективному идеальному, действительно существует: это логические схемы социальной деятельности человека, закрепленные в понятиях. Объективно, т. е. реально существуя, они надличностны в том смысле, что не зависят от отдельного индивида. Однако в процессе деятельности из субъективных идеальных схем (мозга) они перекодированы на другой материальный носитель: в устное слово, текст, жест, вещь, и поэтому должны относиться к опредмеченному [11]. Такое накопленное веками знание сохраняется в культуре народа в виде материального. Форма существования мысли стала иной, «воплощенная в звуча-

<sup>1</sup> В своей ранней работе «Природа информации» (1968) А. Д. Урсул высказывал предположение, что понятие информации «лишь становится философской категорией». См. [15, с. 229].

ших и написанных словах мысль – это уже не мысль как таковая, это бывшая мысль», – убедительно пишет Дубровский [1, Приложение. Еще раз о проблеме идеального (в связи с книгой К. Любутина и Д. Пивоварова «Синтетическая теория идеального». Екатеринбург–Псков, 2000)].

Представление Ильенкова о существовании общего социального знания в виде «идей», логических структур опирается на философское наследие Спинозы, Канта, Шеллинга, Гегеля, Маркса. Представления же об идеальном как субъективной реальности Ильенкову, приверженцу исторического материализма, были чужды, он не считал это темой для философских обсуждений [18, с. 155].

Тем не менее, при всей своей значимости, структуры общественного сознания не самодостаточны, они диалектически взаимодействуют со структурами индивидуального сознания и взаимообогащаются, их нельзя отрывать друг от друга или преувеличивать роль одних в ущерб другим. Интеллектуальные конструкции как «всеобщую закономерность» воспринятого органами чувств и осмысленного предмета мыслящий человек (т. е. уже умеющий строить открытые идеальные схемы) выстраивает все-таки в *собственной* голове. Такие конструкции он может «найти» не только в предметах своей деятельности, но и в природе, в обществе, приобретая тем самым о них информацию и знание. Идеальное – это *представление* предмета с помощью собственной мысли. Дубровский отмечает: если материальное признается материалистами объективной реальностью, то категория идеального – как логически ей противопоставленная – имеет смысл лишь «в качестве субъективной реальности», когда отмечается уникальный тип «связи духовных феноменов с определенными объектами материальной действительности, качественно отличающийся от типов связей между самими материальными объектами» [1, § 1.2].

Процесс «нахождения» интеллектуальных конструкций в предметах культуры и природы можно трактовать как распремечивание. Следовательно, *реально* никакой когнитивной (в том числе, и субъективной) информации – информации-идеального – в природе и обществе нет. Человек в процессе своей социальной деятельности ее находит, делает актуальной и семантической, извлекая с помощью органов чувств или техники сигналы, которые интерпретирует и перерабатывает. В таком случае виды информации, различаемые сторонниками атрибутивного подхода («содержание отражения» и «образ наблюдаемого объекта»), есть одна и та же информация субъективной реальности – *субъективная* информация.

Есть основания полагать, что получение/неполучение информации и усвоение знания у человека связано с умением строить интеллектуальные конструкции (образно-схематические и абстрактные структуры). Тогда информацию и знание можно рассматривать как возникновение идеального *внутри* материальной системы, на материальном носителе, как результат построения неких внутренних репрезентаций, соответствующих изменениям внешней среды или внутреннего состояния. Схема предмета, которую

(предположительно) мысленно строит человек для понимания полученного сигнала и принятия решения о действии/недействии, и есть тот случай, когда «один объект отражает содержание другого объекта».

Можно допустить, что информация как нечто новое будет ощущаться тогда, когда возникшие идеальные схемы не совпадут с построенными ранее в подобной ситуации (когда прошлый опыт позволяет интерпретировать поступившую информацию лишь частично), а применение старых схем в новой ситуации приведет к иллюзиям. Вероятно, способность реагировать на сигналы зависит от состояния организма, физического и социального опыта человека. Опыт проявляется в разной скорости возникновения конструкций, их последовательности, наращивании и сочетании друг с другом, что создает определенные связи и возможность преобразования. Так понятие идеального является *свойством высокоорганизованной материи*. В таком случае идеального *реально* нет вне человека (и некоторых высших животных): ни в мире природы, ни в мире культуры, ни в обществе, ни (пока) в мире искусственных систем.

«Генетическая информация» также не относится к идеальному. Геном – это, по сути, свернутая программа развития живого организма, которая есть существование материального в материальном: молекулы ДНК и РНК имеют пространственные и физические характеристики. Главное, если информация присуща не всей материи, а только ее высокоорганизованной части, то *нет основания считать информацию философской категорией*. Таким образом, представление о термине «информация» как *ряде омонимов*, за которым скрывается несколько не связанных между собой сущностей, оправданно.

Субъект, реагируя на изменения внешнего мира, «находит» информацию повсюду, *приписывая* ее ему, поскольку осмысливает все *объектно*, что не мешает получать действительные знания. В этом смысле, информация потенциальная, или первичная, а также вторичная и третичная, согласно взглядам сторонников атрибутивной концепции, «находится» действительно везде – ее можно найти и в живой и неживой природе, и в обществе. Но это *не может быть основанием для объявления информации атрибутом материи*, поскольку создает информацию только самоорганизующаяся система, обладающая мозгом или его заменителем (на данном этапе развития науки). Информация объективно, реально возникает в мозгу мыслящего субъекта (возможно, как идеальная схема) и становится «субъективным образом объективного мира», а не «разлита» по миру. Думающий человек не записывает информацию из мира, никто ему не диктует ее из космоса. Мыслитель сам во взаимодействии с миром созидает ее, интерпретируя изменения реальности как сигналы об информации. Мышление – это активный процесс, который обеспечивает нашу адаптацию к внешней среде, его нельзя назвать пассивным копированием или прямым отражением внешних воздействий. Как считает И. П. Меркулов, «“об отражении” ... правомерно говорить (да и то лишь весьма условно, метафорически) только по отношению к полученным в ходе переработки когнитивной информации знаниям, но не о

самих мыслительных процессах оперирования внутренними ментальными репрезентациями» [9, сноска на с. 244]. Поэтому согласимся со сторонниками семантических концепций: информации не бывает без смысла, без содержания, ибо то, на что не обращается внимание (по поводу чего вольно или невольно не строятся идеальные схемы), проходит мимо, не замечается. Об информации начинают говорить в том случае, если на фоне обычного информационного шума выявляется чем-то необычный сигнал, который и анализируется самоорганизующейся системой.

## СПИСОК ЛИТЕРАТУРЫ

1. Дубровский Д. И. Проблема идеального. Субъективная реальность. – 2-е изд. – М.: Канон+, 2002. – 368 с. – URL: [http://www.globalistika.ru/dubrovsky/nauchnye\\_texty/probl\\_ideal.htm#P2](http://www.globalistika.ru/dubrovsky/nauchnye_texty/probl_ideal.htm#P2) (дата обращения 20.01.2013).
2. Журавлева Е. Ю. Эпистемический статус цифровых данных в современных научных исследованиях // Вопросы философии – 2012. – № 2. – С. 113-123.
3. Колин К. К. Философские проблемы информатики. – М.: БИНОМ. Лаборатория знаний, 2010. – 264 с.
4. Дворкина М. Я. Информационное обслуживание: социокультурный подход. – М.: Профиздат, 2003. – 112 с.
5. Романенко В. Н., Никитина Г. В. Многозначность понятия информации // Философия науки. – 2010. – № 4 (47). – 75-99.
6. Меркулов И. П. Мышление как информационный процесс // Эволюция. Мышление. Сознание. – М.: Канон+, 2004. – С. 228–260.
7. Швырев В. С. Сознание: гл. 3 // Философия. – М.: Норма, 2009. – С. 521–554.
8. Белоногов Г. Г., Гиляревский Р. С. Ещё раз о гносеологическом статусе понятия «информация» // НТИ. Сер. 2. – 2010. – № 2. – С. 1-6.
9. Меркулов И. П. Введение: формирование когнитивных представлений в эпистемологии // Эволюция. Мышление. Сознание. – М.: Канон+, 2004. – С. 3-34.
10. Шрейдер Ю. А. Социальные аспекты информатики // НТИ. Сер. 2. – 2010. – № 2. – С. 18-27.

11. Батищев Г. С. Предметивание и распределение // Новая философская энциклопедия: в 4 т. – М.: Мысль, 2010. – Т. 3. – С. 157.

12. Информационный подход в междисциплинарной перспективе: материалы круглого стола // Вопросы философии. – 2010. – № 2. – С. 84-112.

13. Дубровский Д. И. Проблема духа и тела: возможности решения (в связи со статьей Т. Нагеля «мыслимость невозможного и проблема духа и тела») // Вопросы философии. – 2002. – № 10. – С. 92-107.

14. Урсул А. Д., Урсул Т. А. Универсальный эволюционизм: концепции, подходы, принципы, перспективы. – М.: Изд-во РАГС, 2007. – 326 с.

15. Урсул А. Д. Природа информации: философский очерк. – 2-е изд. Челябинск, 2010. – 231 с.

16. Готт В. С. Эвристическое значение принципа единства симметрии и асимметрии // Неизбежность нелинейного мира. К 100-летию со дня рождения В. С. Готта. – М.: Гуманитарий, 2012. – С. 349-414.

17. Дубровский Д. И. Психические явления и мозг. Философский анализ проблемы в связи с некоторыми актуальными задачами нейрофизиологии, психологии и кибернетики. – М.: Наука, 1971. – 386 с.

18. Ильенков Э. В. Проблема идеального // Эвальд Васильевич Ильенков. – М.: РОССПЭН, 2008. – С. 153-214.

19. Ильенков Э. В. Вопрос о тождестве мышления и бытия в домарксистской философии // Диалектика – теория познания. Историко-философские очерки. – М.: Изд-во АН СССР, 1964. – С. 21-54.

*Материал поступил в редакцию 25.02.13.*

## Сведения об авторе

**ПОЛТАВСКАЯ Елена Игоревна** – кандидат педагогических наук, зав. Отделом хранения Научной музыкальной библиотеки им. С. И. Танеева Московской государственной консерватории им. П. И. Чайковского.  
E-mail: [poltavskaya.elen@gmail.com](mailto:poltavskaya.elen@gmail.com)

## Формализованная запись нормативных документов для компьютерной обработки

*Предложена модель формализованной записи правовых норм в различных актах. Описана структура базы данных для хранения нормативных документов в формализованной записи. Обозначены преимущества использования формализованной записи и компьютерных информационных технологий для консультирования и принятия управленческих решений, в судебной и нормотворческой деятельности. Приведен алгоритм, позволяющий описать должное поведение для заданной ситуации и выявить противоречия в нормативных документах.*

**Ключевые слова:** нормативные документы, формализованная запись, деонтическая логика, структура базы данных

### ВВЕДЕНИЕ

В деятельности различных структур часто возникает необходимость в контроле согласованности большого объема нормативных документов (в частности, системы локальных нормативных актов компании). Практика показывает наличие значительных трудностей при попытке увязать вручную документы, содержащие большое количество правовых норм. Необходимым же условием автоматизации согласования правовых норм является их логическое описание. Вопрос однозначного логического описания, или формализации, правовых норм в научной литературе поднимается достаточно давно [1-3].

Формализация правовых норм с учетом современных компьютерных технологий могла бы предоставить следующие возможности [4]:

- автоматизация разработки проектов нормативных документов;
- автоматизированный контроль (мониторинг) непротиворечивости нормативных правовых актов;
- автоматизированное выявление сфер, не урегулированных нормативными документами (так называемых пробелов нормативного регулирования);
- повышение степени однозначности положений нормативных актов за счет параллельного их формулирования средствами не только естественного языка, но и формализованной записи, что позволило бы уже на этапе разработки проектов документов если не устранить, то хотя бы минимизировать возможные неясности и противоречия;
- имитационное моделирование процессов правоприменения, полезное при экспериментальной проверке правовых норм в различных условиях;
- углубление представлений о правовой норме.

Таким образом, внедрение формализованной записи правовых норм позволило бы вывести на новую качественную ступень процесс разработки, толкования и применения нормативных правовых актов, уменьшить распространенность пробелов и противоречий, снизить риск ошибок при толковании правовых актов.

Вопрос заключается лишь в том, чтобы предложить соответствующую формализованную запись. Для этого в литературе предпринимаются попытки использовать для описания правовых норм аппарат математической логики, в том числе языковые средства логики предикатов или модальной логики [2, 4]. Трудность автоматизации при этом, очевидно, заключается в необходимости по каждому положению нормативного акта формировать свою, уникальную логическую конструкцию, а затем использовать специальный язык программирования (например, ПРОЛОГ) для работы с ней. Причем построение такой уникальной логической конструкции сопряжено с большим числом описательных трудностей, вызванных многогранностью предмета правового регулирования.

Следует отметить, что соответствующие затруднения оказались настолько серьезны, что, несмотря на их обсуждение в течение десятилетий, на сегодняшний день ни одно из формальных описаний правовых норм и, соответственно, ни один возможный на его основе алгоритм автоматизированного анализа не получили сколько-нибудь широкого практического применения. Более того, активных дискуссий по вопросам автоматизации процедур анализа правовых норм в литературе последних лет не наблюдается.

В качестве основного препятствия на пути математического описания правовых норм отмечается неоднозначность понятийного аппарата юридической науки, которая возрастает при некритичном исполь-

зовании математических средств для его анализа [5]. Проблема действительно существует, но, по нашему мнению, к ее решению можно приступить лишь тогда, когда разработан приемлемый способ математического описания правовых норм. Иными словами, возникает своего рода замкнутый круг, когда для разработки математического описания требуются однозначные правовые нормы, а их создание должно опираться на то понимание однозначности, которое диктуется математической моделью. Разрешить возникшее противоречие можно диалектически: вначале предложить такое описание, которое допускало бы существование неоднозначных, противоречивых положений законодательства и при этом позволяло бы их выявлять, а затем с его помощью совершенствовать тексты нормативных актов. Разумеется, при таком подходе описание правовых норм при помощи математической логики, на наш взгляд, существенно затруднено, и нужны более гибкие описательные средства.

Некоторые авторы [1] пришли к выводу, что математическая модель никогда не сможет заменить традиционное описание правовых норм, что математика и естественный язык дополняют, а не заменяют друг друга. Однако такая точка зрения в литературе уже не является бесспорной, достаточно отметить контраргументы, приведенные в [5].

Считается, что все многообразие правовых норм невозможно вместить в какую-либо одну формальную запись. Однако упомянутое разнообразие не может служить доказательством отсутствия единого начала и общих правил. В качестве примера можно привести музыку: всего из семи нот, владея соответствующими правилами, удастся построить огромное количество разнообразных мелодий.

Таким образом, на сегодняшний день существуют затруднения с созданием и практическим внедрением формализованной записи правовых норм. Однако принципиальная невозможность создания такой формализованной записи не является доказанной. Основной причиной трудностей при этом видится попытка построить описание правовых норм средствами математической логики, тогда как неисчерпанными остаются более простые (и одновременно более гибкие) описательные средства. По нашему мнению, инструментарий математической логики может использоваться в работе с правовыми нормами на более поздних этапах, после того как будут решены задачи однозначности формулировок правовых норм, создания строгой формальной системы права.

В настоящей статье предпринята попытка сформулировать модель формализованной записи правовых норм, опирающуюся не на аппарат математической логики, а на саму физическую структуру, если угодно, объектную модель правовой нормы. Представленный далее аппарат формализованной записи не претендует на универсальность и какое-либо совершенство. Его основная цель – послужить иллюстрацией самой возможности формализованной записи правовых норм без применения аппарата математической логики и отправной точкой для научной дискуссии по этому вопросу. Необходимо отметить, что в настоящий момент предлагаемый подход разработан теоретически и еще не прошел широкую практическую апробацию.

## ФОРМАЛИЗОВАННАЯ ЗАПИСЬ ПРАВОВЫХ НОРМ

Будем исходить из того, что право, с позиций его формализации, – это своего рода деонтическая модель реальности, описывающая, каким должно быть поведение участников правоотношений при тех или иных условиях, а также их воздействие друг на друга и объекты материального мира. Следовательно, можно сказать, что правовая модель описывает различные состояния каких-либо объектов регулирования при разных условиях, или что всякому нормативному акту должно соответствовать некоторое конечное множество объектов регулирования  $\{E_i\}$  – произвольной природы. Объектами регулирования могут, очевидно, выступать различные лица, предметы, действия – всё то, что в силу нормативного акта при выполнении тех или иных условий должно прийти в предписанное состояние, несоблюдение которого влечет применение соответствующих мер принуждения. Иными словами, употребляемое в настоящей работе понятие «объект регулирования» не следует смешивать с такими распространенными в праве понятиями как «объект правоотношений» и «предмет регулирования».

Далее, каждый из объектов регулирования должен обладать неким набором свойств  $\{p_j^{(i)}\}$ , которые тем или иным образом участвуют в правовых нормах и потому являются существенными в рамках соответствующего нормативного акта. Для каждого объекта регулирования имеется свой набор существенных свойств, который, безусловно, зависит от специфики объекта регулирования и содержания конкретного нормативного акта. Например, для объекта регулирования «физическое лицо» в Трудовом кодексе РФ существенным является свойство «пол», поскольку в нем существуют статьи, регулирующие особенности труда женщин, а в Гражданском кодексе РФ оно существенным не является, так как ни в одной статье не упоминается.

Конкретному свойству  $p_j^{(i)}$  ставится в соответствие некое множество значений  $\{x_{jk}^{(i)}\}$ , которое это свойство может или не должно принимать. Данное множество может состоять из любого числа  $K$  однородных элементов произвольной природы ( $K$ , в частности, может быть равно 1 или даже 0). Для удобства можно записывать это множество в виде  $K$ -мерного вектора  $\vec{x}_j^{(i)}$ .

Положим, что  $p_j^{(i)}$  соотносится с  $\vec{x}_j^{(i)}$  через некий деонтический оператор  $\nabla: p_j^{(i)} \nabla_j^{(i)} \vec{x}_j^{(i)}$ . Связку  $p_j^{(i)} \nabla_j^{(i)} \vec{x}_j^{(i)}$ , устанавливающую определенное соотношение между свойством объекта регулирования и неким множеством значений, которое данное свойство может или не должно принимать, будем именовать предписанием и обозначать  $R_j^{(i)}$ .

Сами свойства и их значения могут быть как своего рода константами, задаваемыми извне средствами обычного языка, так и терминами, определенными

ми и используемыми в смысле, который задается конкретным нормативным актом.

Чтобы определить возможные значения деонтического оператора, необходимо принять во внимание следующее. Регулирующее действие права, очевидно, проявляется в запрете на принятие объектами регулирования каких-либо значений из ограниченного списка, либо в разрешении каких-либо значений из такого же ограниченного списка. Причем в составе такого списка может находиться элемент типа «и другие, отвечающие таким-то признакам». В самом деле, если свойству того или иного объекта нормативным актом было бы разрешено принимать значения не только из ограниченного списка, но и любые другие значения, то и смысл регулирования оказался бы утраченным: все равно значения могут быть любыми. Аналогично, если нормативный акт запрещает объекту принятие не только значений из ограниченного списка, но и любых других значений, то тогда ни одно из мыслимых значений, в принципе, не будет правомерным. В итоге, если в основе права лежит принцип: «Разрешено все, что не запрещено законом», то его последовательное применение требует введения для всех участников правоотношений только ограничивающих предписаний.

Безусловно, нормативные акты порой содержат перечисления значений, сопровождаемые выражением «в частности, ...». Однако, по нашему мнению, нельзя сказать, что такие положения в силу изложенного выше регулируют состояние объектов регулирования, они скорее описывают или определяют что-либо.

Таким образом, у деонтического оператора, как нам представляется, могут быть только два значения: «разрешено исключительно» и «запрещено исключительно». Дефинитивные и декларативные предписания, по-видимому, также могут описываться при помощи данных операторов.

Каков логический смысл деонтического оператора? Зададим логическое выражение, которое строится на основе предписания  $R_j^{(i)}$ :

$$Q_j^{(i)} = (p_j^{(i)} = x_{j1}^{(i)}) \cap (p_j^{(i)} = x_{j2}^{(i)}) \cap \dots$$

Если  $\nabla_j^{(i)}$  = «разрешено исключительно», то в конкретном случае для выполнения предписания  $R_j^{(i)}$  выражение  $Q_j^{(i)}$  должно принимать значение «истина», а если  $\nabla_j^{(i)}$  = «запрещено исключительно», то – «ложь».

Теперь можно задать формализованную запись нормативных актов. Как известно, своего рода молекулами права являются правовые нормы, имеющие трехчастную структуру:

- гипотеза – субъекты правоотношений и их свойства, юридические факты и их свойства;
- диспозиция – правила поведения субъектов, состояние их прав и обязанностей, свойств объектов в случае, когда гипотеза соблюдена;
- санкция – описание мер принуждения при несоблюдении диспозиции, которые заключаются в изменении свойств субъектов и объектов правоотношений.

Правовая норма строится по принципу: «Если (гипотеза), то (диспозиция), иначе (санкция)». Нетрудно заметить, что каждая часть правовой нормы описывает те или иные значения свойств объектов регулирования: гипотеза – значения свойств, совокупность которых приводит в действие норму права; диспозиция – значения, которые предписываются нормативным актом свойствам объектов регулирования при выполнении гипотезы; санкция – значения, в которые при невыполнении диспозиции будут переведены свойства объектов регулирования силой государственного принуждения.

Но тогда можно говорить, что правовая норма описывает некую совокупность объектов регулирования, их свойств, значений и правил, причем правила в ней разделены на три блока: блок гипотезы (H), блок диспозиции (D) и блок санкции (S). Переход к блоку диспозиции, который может определять новые или изменять ранее заданные значения свойств, происходит при выполнении гипотезы, а переход к блоку санкции – при выполнении гипотезы и невыполнении диспозиции.

Поскольку, как свидетельствует практика, в нормативном акте могут содержаться те или иные определения, то целесообразно предусмотреть еще один блок – блок дефиниции (F).

Каждый из этих блоков может содержать отсылки к положениям того же самого или иного нормативного акта, в связи с чем в каждом из блоков необходимо предусмотреть указание на применение или неприменение в данном случае того или иного положения соответствующего нормативного акта.

Но тогда, с учетом изложенного, правовая норма в символической записи примет вид:

$$N_m = (r, \{E_i\}), E_i : \left\{ \begin{array}{l} F : \{R_j^{F(im)}\}, \{\delta(N_i^{F(im)})\}; \\ H : \{R_j^{H(im)}\}, \{\delta(N_i^{H(im)})\}; \\ D : \{R_j^{D(im)}\}, \{\delta(N_i^{D(im)})\}; \\ S : \{R_j^{S(im)}\}, \{\delta(N_i^{S(im)})\}. \end{array} \right.$$

где:

$N_m$  –  $m$ -я правовая норма;

$r$  – уровень юридической силы правовой нормы;

$E_i$  –  $i$ -й объект регулирования;

символы  $F, H, D, S$  означают соответственно пространства гипотезы, диспозиции и санкции;

$R_j^{U(im)} = p_j^{(im)} \nabla_j^{U(im)} x_j^{\rightarrow U(im)}$  – предписание, устанавливаемое  $U$ -м блоком ( $U \in \{F, H, D, S\}$ )  $m$ -й нормы для  $j$ -го свойства  $i$ -го объекта регулирования;

$p_j^{(im)}$  –  $j$ -е свойство  $i$ -го элемента  $m$ -й правовой нормы;

$x_j^{\rightarrow U(im)}$  – вектор (множество) значений свойства в  $U$ -м блоке, так что свойство может принимать какое-либо одно из этих значений (например, свойство «цель» элемента «собираться» может принимать значение «собрание, митинг, демонстрация, шествие, пикетирование»);

$\nabla_j^{U(im)}$  – значение деонтического оператора для  $j$ -го свойства  $i$ -го элемента  $m$ -й правовой нормы в  $U$ -м блоке;

$N_j^{U(im)}$  –  $l$ -я ссылка на норму данного или иного нормативного акта в  $U$ -м блоке  $m$ -й правовой нормы, применяемая к  $i$ -му элементу;

$\delta(\cdot)$  – отсылочный оператор, принимающий значения: «применять», «не применять».

Будем говорить, что в конкретном случае правовая норма выполнена, если все предписания  $R_j^{H(im)}$  гипотезы выполнены и при этом выполнены все предписания диспозиции  $R_j^{D(im)}$ . Правовая норма, имеющая отсылки, является выполненной, только если соблюдено указанное выше правило и выполнены все нормы при отсылочном операторе «применять». Если существует какая-либо норма той же или более высокой юридической силы, в которой содержится отсылка к данной норме при отсылочном операторе «не применять», то при выполнении всех предписаний ее гипотезы данная норма считается выполненной независимо от того, выполняется ли она на самом деле по условиям конкретной ситуации.

В известных на сегодняшний день нормативных актах их положения излагаются последовательно, т. е. всякий нормативный акт содержит цепочку аддитивно связанных друг с другом норм:

$$A = N_1 \cup N_2 \cup \dots$$

Аддитивная связь здесь означает, что для выполнения нормативного акта в конкретном случае необходимо и достаточно, чтобы были выполнены одновременно все содержащиеся в нем правовые нормы.

Из аддитивности правовых норм вытекает, что для получения формализованного представления нормативного акта совершенно необязательно вычленять в отдельности каждую правовую норму, которая иногда распределена по различным частям нормативного акта или даже по разным актам. Достаточно выполнить формализованную запись каждого положения нормативного акта, а все заключенные в нем правовые нормы сложатся сами собой. При необходимости может быть предложен алгоритм свертки нормативного акта в его формализованной записи, который позволит вычленить все правовые нормы, а заодно и выявить их незавершенность. Такой алгоритм может основываться на суммировании (конкатенации) для каждого объекта регулирования диспозиций и санкций, содержащихся в различных положениях нормативного акта, при наличии в них одинаковых гипотез.

Опираясь на изложенное, для формализованной записи нормативных актов можно предложить структуру данных, изображенную на рис. 1.

В качестве примера представим в формализованной записи положения статьи 31 Конституции РФ: «Граждане Российской Федерации имеют право собираться мирно, без оружия, проводить собрания, митинги и демонстрации, шествия и пикетирование».

Результат удобно представить в табличной форме. В приведенном разложении (таблица) использован не упомянутый в статье, но подразумеваемый объект регулирования «мирное собрание». Необходимость его применения устанавливается путем толкования положений данной статьи и возникает, очевидно, тогда, когда значение свойства само имеет свойства, которые устанавливаются или изменяются в результате правового регулирования.



Рис. 1. Структура базы данных для хранения формализованной записи нормативных актов

## Формализованная запись статьи 31 Конституции РФ

| Объект регулирования | Блок       | Свойство           | Деонтический оператор   | Множество значений                                                                                              | Отсылка |
|----------------------|------------|--------------------|-------------------------|-----------------------------------------------------------------------------------------------------------------|---------|
| Любое лицо           | Гипотеза   | Гражданство        | Разрешено исключительно | Гражданин РФ                                                                                                    | -       |
|                      | Диспозиция | Участие в собрании | Разрешено исключительно | Мирное собрание                                                                                                 | -       |
| Мирное собрание      | Дефиниция  | Форма              | Разрешено исключительно | Собрание. Митинг. Демонстрация. Шествие. Пикетирование. Другое публичное собрание. Всякое непубличное собрание. | -       |
|                      |            | Цель               | Разрешено исключительно | Мирно                                                                                                           | -       |
|                      |            | Предметы           | Запрещено исключительно | Оружие                                                                                                          | -       |

### ОБ АЛГОРИТМЕ ПРАВОПРИМЕНИТЕЛЬНОГО МОДЕЛИРОВАНИЯ С ПРИМЕНЕНИЕМ ФОРМАЛИЗОВАННОЙ ЗАПИСИ ПРАВОВЫХ НОРМ

Определим правоприменительное моделирование как получение модели правомерного поведения в конкретных условиях. Такая модель представляет собой определение при заданных условиях (значениях свойств):

- должного состояния всех объектов регулирования, т. е. возможных значений таких свойств каждого объекта, которые не определены в заданных условиях;
- фактов нарушения законодательства заданными условиями и последствий, т. е. случаев невыполнения диспозиций правовых норм при заданных условиях и описания соответствующих санкций;
- дополнительных вариантов поведения, т. е. ситуаций, когда дополнительное задание значений каких-либо свойств порождает необходимость применения ранее не описанных правовых норм;
- противоречий в подлежащих применению правовых нормах (если они имеются) и способа их разрешения по существу;
- имеющейся неопределенности правовых норм.

На основе предложенного формализованного описания правовых норм возможна автоматизация процесса правоприменительного моделирования. Его результаты, очевидно, могут быть полезны в деятельности судов, при разработке законов и подзаконных нормативных актов, в процессе юриди-

ческого консультирования, в научно-исследовательской работе.

Сформированный при помощи компьютера ответ на запрос может выглядеть, например, следующим образом (рис. 2).

С использованием алгоритмов, описанных в [7], табличное описание может быть заменено текстовым.

Автоматизация правоприменительного моделирования не отменяет функций юриста, связанных с выработкой вариантов решения применительно к заданным условиям, однако существенно изменяет содержание этих функций, позволяя юристу не тратить время на разбор положений законодательства вручную, давая ему возможность сосредоточиться на главном – на собственно выработке правовой позиции. Ближайшей аналогией такого изменения могло бы послужить, пожалуй, сравнение содержания профессиональной деятельности юристов до появления справочных правовых систем («Гарант», «Консультант Плюс», «Кодекс», «Референт» и др.) и после него. До внедрения справочных правовых систем от юриста требовалась хорошо развитая память, позволяющая запоминать положения нормативных актов, а после их внедрения главным качеством стали аналитические способности юриста – умение сопоставить между собой различные правовые нормы, системно их истолковать и правильно применить. Подобная точка зрения поддерживается и в работе [4], где отмечается, что за счет формализации содержания и психология юридической деятельности изменится коренным образом, но возникшая разница будет подобна разнице между содержанием работы и психологией ямщика и водителя современного автобуса.

Описаны следующие начальные условия:

| Объект | Свойство | Значение |
|--------|----------|----------|
|        |          |          |

Произведенный анализ показал следующее.

1. При заданных начальных условиях полностью соблюдены нормы, заданные в <Перечень ссылок на положения нормативных актов, которые полностью соблюдены>.

*{Если ни одна норма законодательства начальными условиями не соблюдена, то вместо приведенного текста делается запись «При заданных начальных условиях отсутствуют правовые нормы, которые были бы соблюдены полностью»}*

2. При заданных начальных условиях законодательство предусматривает необходимость соблюдения следующих дополнительных условий:

| Объекты | Свойства | Значения | Основание (ссылка на положения нормативного акта) |
|---------|----------|----------|---------------------------------------------------|
|         |          |          |                                                   |

*{При отсутствии дополнительных условий вместо приведенного текста делается запись «Каких-либо дополнительных условий действующее законодательство не предусматривает»}*

3. При заданных начальных условиях нарушены следующие положения законодательства:

| Объекты | Свойства | Значения                    |           | Санкция | Основание (ссылка на положения нормативного акта) |
|---------|----------|-----------------------------|-----------|---------|---------------------------------------------------|
|         |          | задано начальными условиями | требуется |         |                                                   |
|         |          |                             |           |         |                                                   |

*{При отсутствии нарушений законодательства вместо приведенного текста делается запись «Нарушений законодательства при заданных условиях не выявлено»}*

4. Если к заданным условиям добавить перечисленные ниже условия, то возникнет необходимость дополнительного применения следующих правовых норм:

| Объекты | Дополнительные условия |          | Дополнительные предписания |          | Санкция | Основание (ссылка на положения нормативного акта) |
|---------|------------------------|----------|----------------------------|----------|---------|---------------------------------------------------|
|         | свойство               | значение | свойство                   | значение |         |                                                   |
|         |                        |          |                            |          |         |                                                   |

*{При отсутствии дополнительно налагаемых условий вместо приведенного текста делается запись «Дополнительных условий, с введением которых законодательство требовало бы выполнения дополнительных предписаний, не выявлено»}*

5. Внутри нормативных актов обнаружены следующие противоречия:

| Исходное положение (ссылка) | Противоречащее положение (ссылка) | Объекты, имеющие противоречивые значения свойств | Свойства с противоречивыми значениями | Значения               |                              | Способ разрешения противоречия |
|-----------------------------|-----------------------------------|--------------------------------------------------|---------------------------------------|------------------------|------------------------------|--------------------------------|
|                             |                                   |                                                  |                                       | по исходному положению | по противоречащему положению |                                |
|                             |                                   |                                                  |                                       |                        |                              |                                |

*{При отсутствии противоречий вместо приведенного текста делается запись «Противоречий между положениями нормативных актов не выявлено»}*

Рис. 2. Примерная форма ответа на запрос

Опираясь на описанный аппарат, можно предложить несложную принципиальную схему алгоритма правоприменительного моделирования:

1. Задать начальные условия: перечислить все объекты регулирования, их свойства и значения этих свойств, характерные для рассматриваемой ситуации.

2. Задать рабочую таблицу со следующими полями:

- норма;
- выполнение гипотезы (истина/ложь);
- перечень предписаний гипотезы, не охваченный начальными условиями;

– выполнение диспозиции (истина/ложь)

– перечень предписаний диспозиции, не охваченный начальными условиями;

– перечень невыполненных предписаний диспозиции.

3. В базе данных отобрать и отсортировать по убыванию юридической силы все нормативные акты, в которых встречаются заданные объекты регулирования.

4. Для каждого нормативного акта:

4.1. Для каждой нормы, содержащейся в нормативном акте:

4.1.1. Если норма содержит хотя бы один из объектов, заданных в начальных условиях, то:

4.1.1.1. Добавить строку в рабочую таблицу. Записать ссылку на норму.

4.1.1.2. Проверить выполнение гипотезы для заданных начальных условий: гипотеза не считается выполненной, если хотя бы одно предписание при заданных условиях не выполнено (значение поля «выполнение гипотезы» = ложь), иначе гипотеза считается выполненной (значение поля «выполнение гипотезы» = истина). При проверке выполнения учитываются отсылки. Если существует предписание, для проверки которого в начальных условиях нет исходных данных, оно считается выполненным, но в поле «перечень предписаний гипотезы, не охваченный начальными условиями» добавляется ссылка на это предписание.

4.1.1.3. Если «выполнение гипотезы» = истина, то:

4.1.1.3.1. Проверить выполнение диспозиции для заданных начальных условий: диспозиция не считается выполненной, если хотя бы одно предписание при заданных условиях не выполнено (значение поля «выполнение диспозиции» = ложь), невыполненные предписания записываются в поле «перечень невыполненных предписаний диспозиции», иначе диспозиция считается выполненной (значение поля «выполнение диспозиции» = истина). Если существует предписание, для проверки которого в начальных условиях нет исходных данных, оно считается выполненным, но в поле «перечень предписаний диспозиции, не охваченный начальными условиями» добавляется ссылка на это предписание.

5. Для каждой строки рабочей таблицы:

5.1. Если значение поля «выполнение гипотезы» = истина, значение поля «выполнение диспозиции» = истина, значение поля «перечень предписаний гипотезы, не охваченный начальными условиями» пустое и значение поля «перечень предписаний диспозиции, не охваченный начальными условиями» пустое, то:

5.1.1. Добавить ссылки на элементы (положения) нормативных актов, содержащиеся в норме, в раздел I ответа.

5.1.2. Перейти к следующей строке рабочей таблицы.

5.2. Если значение поля «выполнение гипотезы» = истина, значение поля «выполнение диспозиции» = истина, значение поля «перечень предписаний гипотезы, не охваченный начальными условиями» пустое и значение поля «перечень предписаний диспозиции, не охваченный начальными условиями» непустое, то:

5.2.1. Для каждого предписания, не охваченного начальными условиями, добавить

строку в таблицу раздела 2 ответа, заполнить значения строки.

5.2.2. Перейти к следующей строке рабочей таблицы.

5.3. Если значение поля «выполнение гипотезы» = истина, значение поля «выполнение диспозиции» = ложь, значение поля «перечень предписаний гипотезы, не охваченный начальными условиями» пустое, то:

5.3.1. Для каждого невыполненного предписания диспозиции добавить строку в таблицу раздела 3 ответа, заполнить значения строки.

5.3.2. Перейти к следующей строке рабочей таблицы.

5.4. Если значение поля «выполнение гипотезы» = истина, значение поля «перечень предписаний гипотезы, не охваченный начальными условиями» непустое и значение поля «перечень предписаний диспозиции, не охваченный начальными условиями» непустое, то:

5.4.1. Для каждого невыполненного предписания гипотезы добавить строку в таблицу раздела 4 ответа, заполнить значения строки.

5.4.2. Перейти к следующей строке рабочей таблицы.

5.5. Если «выполнение гипотезы» = истина и значение поля «перечень предписаний гипотезы, не охваченный начальными условиями» пустое, то:

5.5.1. Для каждой из нижележащих строк таблицы:

5.5.1.1. Если гипотеза нижележащей строки совпадает с гипотезой строки рабочей таблицы и между предписаниями в поле «перечень предписаний диспозиции, не охваченный начальными условиями» строки рабочей таблицы и данной нижележащей строки есть противоречие (одна строка разрешает значение, а другая запрещает), то добавить строку в таблицу раздела 5 ответа, заполнить значения строки. Способ разрешения противоречия предлагается по принятым правилам, исходя из юридической силы сравниваемых норм, срока принятия и т.д.

6. Вывести отчет.

Следует отметить, что возможны два крайних варианта задания начальных условий<sup>1</sup>: заполнить таблицу запроса с полями «объект», «свойство», «значение» или задать полнотекстовый запрос и осуществить его перевод в указанную таблицу посредством лингвистического процессора [7].

## ЗАКЛЮЧЕНИЕ

Внедрение формализованной записи правовых норм обещает достаточно большие перспективы не только в сфере законотворчества, формирования нормативной базы новых отраслей (например, инноваци-

<sup>1</sup> Между ними существует определенный набор промежуточных вариантов.

онной отрасли, охватывающей различные технопарковые структуры [8]), но и в плане функционирования электронных библиотек, призванных обеспечивать пользователей полной, достоверной и непротиворечивой нормативно-правовой информацией [9, 10].

Однако для того, чтобы возможности применения формализованной записи раскрылись наиболее полно, необходимо, чтобы нормативный акт изначально разрабатывался в электронной среде как структурно упорядоченный формализованный документ (см. фрагмент базы данных, представленный на рис. 1), и уже затем для обсуждения переводился в текстовую форму (например, с помощью лингвистического процессора [5]).

## СПИСОК ЛИТЕРАТУРЫ

1. Гаврилов О.А. Изучение права методами математической логики // Вопросы кибернетики и право. – М.: Наука, 1967. – С. 36-60.
2. Ивин А.А. Логика норм. – М.: Изд-во МГУ, 1973. – 122 с.
3. Кнапп В., Герлох А. Логика в правовом сознании / пер. А.Б. Венгерова. – М.: Прогресс, 1987. – 312 с.
4. Костылёв В.М. Право: опыт системного анализа - Уфа: ООО «Виртуал», 2002. — URL: [http://www.kostylew.ru/pr\\_opyt1.htm](http://www.kostylew.ru/pr_opyt1.htm) (дата обращения: 25.08.2012).
5. Чубукова С.Г., Элькин В.Д. Основы правовой информатики (юридические и математические вопросы информатики). – М.: «Юридическая фирма «Контракт», 2004. – 252 с.
6. Павлов В.П. Проблемы теории собственности в российском гражданском праве. – М., 2000. – 428 с.

7. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992. – 256 с.

8. Шокин Ю.И., Гришняков Б.Ю., Бобров Л.К. Технопарк «Новосибирск» как звено инновационной инфраструктуры региона // Вестник НГУЭУ. – 2012. – № 2. – С. 10-20.

9. Антипов А.Г., Гузнер И.А., Донских О.А., Ермоленко С.М. и др. Круглый стол «Библиотека без книг - неизбежность будущего?» // Идеи и идеалы. – 2011. – Т. 1, № 2. – С. 2-22.

10. Берязев В.А., Бобров Л.К., Булышева О., Исаков С.П. и др. Круглый стол «Белые пятна в предоставлении доступа к информации» // Идеи и идеалы. – 2012. – Т. 2, № 1. – С. 154-159.

*Материал поступил в редакцию 11.01.13.*

## Сведения об авторах

**РЫЖКОВ Олег Юрьевич** - кандидат экономических наук, ведущий аудитор Департамента налогового и правового консалтинга ЗАО «Объединенные консультанты ФДП», Москва  
E-mail: [ory@ngs.ru](mailto:ory@ngs.ru)

**БОБРОВ Леонид Куприянович** – доктор технических наук, профессор кафедры экономической информатики, первый проректор Новосибирского государственного университета экономики и управления «НИНХ»  
E-mail: [bobrov@nsuem.ru](mailto:bobrov@nsuem.ru)

## База данных «ДСМ-эксперименты по фармакологии»\*

*Описывается концептуальная схема базы данных для представления сведений о ДСМ-экспериментах и их результатах. Разработана макетная версия интерфейса, включающая возможности просмотра, сортировки, фильтрации и редактирования данных, создания нового эксперимента и его запуск.*

*Разработана программа автоматического кодирования химических соединений, программа встраивания кодировщика в базу данных и автоматическая подготовка данных к проведению ДСМ-эксперимента.*

**Ключевые слова:** биологическая активность, фармакофор, антифармакофор, база данных, база знаний, язык кодирования химических соединений, ДСМ-метод

### ВВЕДЕНИЕ

В последние годы при поиске новых биологически активных веществ интенсивно применяются компьютерные методы анализа связи «структура-активность». Для решения такого рода задач используется интеллектуальная информационная ДСМ-система по фармакологии, созданная в ВИНТИ РАН. Разработанная система порождает гипотетические фармакофоры (фрагменты химической структуры, ответственные за проявление биологической активности) и антифармакофоры (фрагменты структуры, наличие которых в соединениях приводит к отсутствию биологического действия вещества), а также с помощью выделенных гипотез о фармакофорах и антифармакофорах производит прогноз свойств соединений, активность которых неизвестна [1].

В 90-х гг. прошлого века экспертным сообществом были выделены функциональные группы и субструктуры, наличие которых в молекулах соединений определяет высокую вероятность проявления канцерогенной и мутагенной активности. Такие функциональные группы и субструктуры называются структурными «тревожными сигналами». В 2006 г. для прогноза канцерогенности соединений была разработана система ToxTree, которая использовала информацию о «тревожных сигналах» при определении генотоксичности и канцерогенности соединений. ДСМ-система позволяет с помощью выделенных фармакофоров и антифармакофоров делать вывод о биологической активности веществ и использовать их для синтеза новых лекарственных средств с заданными свойствами. Со времени разработки ДСМ-системы не существовало единой базы данных, в которой хранилась бы информация о результатах проводимых ДСМ-экспериментов. Эти результаты были представлены отчётами об экспериментах, либо в публикуемых статьях. В настоящее время назрела острая необходимость создания базы данных, в которой фиксировались бы все проводимые эксперименты и их результаты.

### 1. МАКЕТ БАЗЫ ДАННЫХ

#### 1.1. Разработка концептуальной схемы базы данных

Основной сущностью разрабатываемой схемы является Эксперимент, который содержит информацию об исходном массиве соединений, целевом свойстве, примененной стратегии ДСМ-метода и полученных гипотезах (фрагментах химической структуры) в ходе эксперимента (см. рис. 1).

В процессе работы над базой данных была выявлена следующая сущность – Химическая структура, которая включает в себя структурные формулы исходных веществ и гипотез, а также запись их в кодах ФКСП (Фрагментарный код суперпозиций подструктур). На рис. 2 приведен фрагмент схемы базы данных и присутствуют все таблицы, содержащие информацию об основных сущностях.

Информация обо всех имеющихся в базе данных химических структурах хранится в таблице strCodeFormula, а также присутствуют поля свойств, общие для сущности «Химическая структура» – кода ФКСП (текстовое поле) и структурной формулы (поле объекта OLE со ссылкой на файл формата MOL), при этом у каждой структуры из таблицы strCodeFormula имеется универсальный идентификатор strID.

На рис. 2, в окне Initials хранится информация о соединениях, составляющих исходный массив того или иного эксперимента. При этом одно и то же соединение может, разумеется, принимать участие более чем в одном эксперименте. Информация, содержащаяся в таблицах окна Initials, является своеобразным посредником, с помощью которого мы разрываем прямую связь между экспериментом и химической структурой. Ключом к данной связи является пара «идентификатор эксперимента – идентификатор соединения», т.е. одно и то же соединение может иметь более одного вхождения в таблицу.

Информация, содержащаяся в таблицах strCodeFormula и Experiment соответствующих окон strCodeFormula и Experiment связаны с информационными таблицами окна Initials отношением «один

\* Работа выполнена при поддержке РФФИ (проект № 11-07-00618а) и Президиума РАН (П115, проект № 209)

ко многим». Благодаря этому данные, при участии одного и того же соединения в разных экспериментах, не дублируются. Каждая запись в таблице окна Initials хранит информацию о соединении, уникальную для данного эксперимента: значение целевого свойства, номер в эксперименте, прогноз активности соединения на основе порожденных в эксперименте гипотез, в то время как структурная формула и код ФКСП, не меняющиеся от эксперимента к эксперименту, хранятся отдельно и не повторяются.

Окно Hypotheses содержит таблицы с информацией о порожденных в эксперименте гипотезах 1-го рода – гипотетических фармакофорах и антифармакофорах исследуемой активности. При разработке базы данных важной задачей является хранение информации о родителях (химические структуры, породившие гипотезы) каждой гипотезы. Фрагмент схемы базы данных, демонстрирующий представление иерархии гипотез, показан на рис. 3.

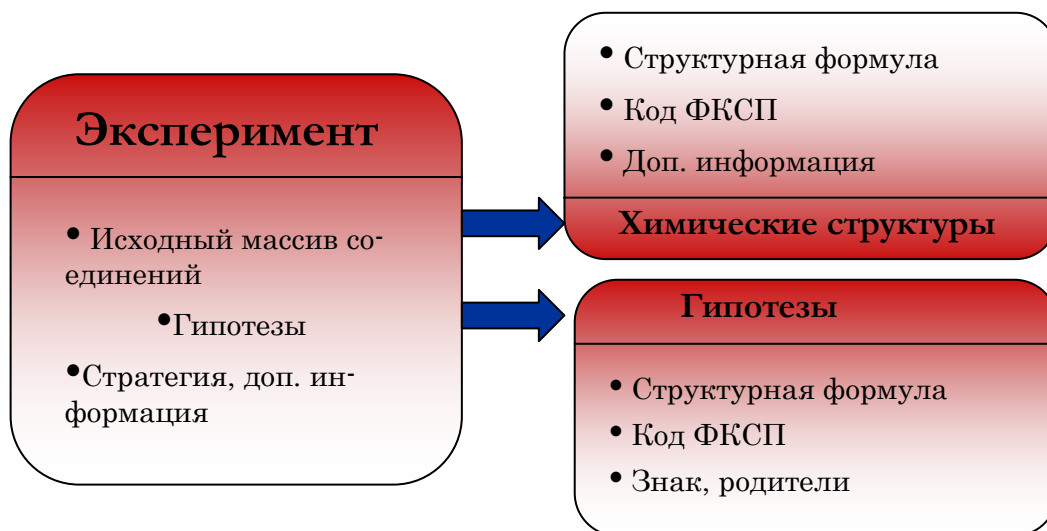


Рис. 1. Модель предметной области

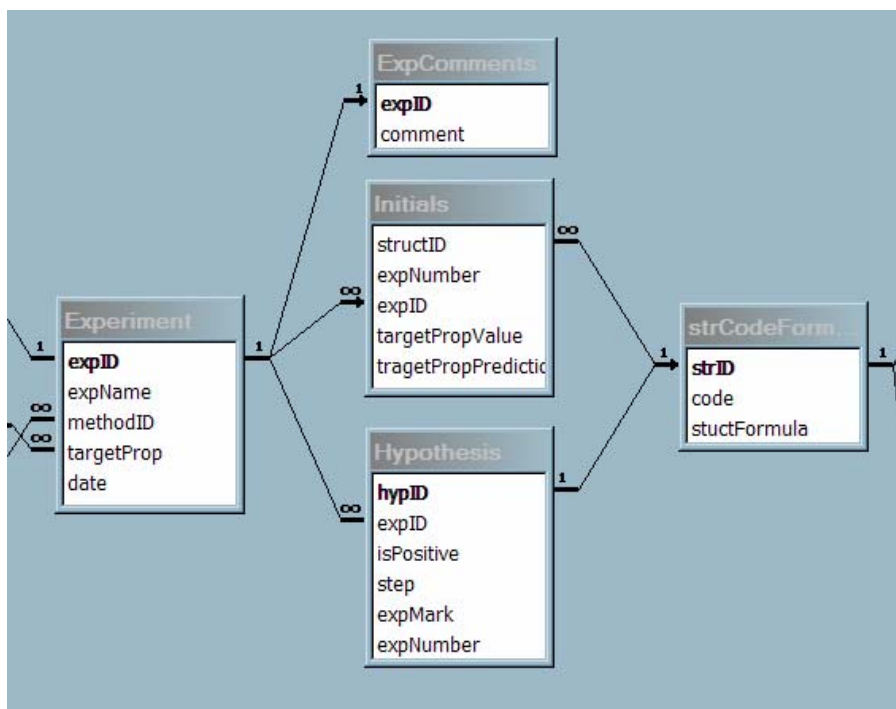


Рис. 2. Основные сущности на схеме базы данных

Таблица окна HypHierarchy имеет два поля: hypChildID и hypParentID – это идентификаторы гипотезы и ее родителя, соответственно. Благодаря наличию этой таблицы мы можем при разработке интерфейса предоставить пользователю возможность просмотреть список родителей гипотезы с кодами и структурными формулами.

Фрагмент схемы базы данных, с некоторыми служебными таблицами, а также таблицами, содержащими дополнительную информацию об эксперименте, представлен на рис. 4.

Таблицы окон Methods и ChemActivityPromNames являются словарными таблицами, содержащими названия различных стратегий ДСМ-метода и химических свойств соединений. При необходимости эти таблицы могут пополняться администратором базы данных.

Таблица окна Activity содержит дополнительную информацию о химической активности представленных в базе данных соединений. Информация, если таковая имеется, не принимает непосредственного участия в ДСМ-эксперименте, однако позволяет спе-

циалисту-химику получить определенное представление об интересующем его соединении.

И, наконец, таблица окна ExpComments содержит дополнительную информацию об эксперименте, не обозначенную в таблице окна Experiment.

## 2. Интерфейс базы данных

При разработке интерфейса были реализованы три основных режима работы с данными: просмотр, редактирование и добавление (см. рис. 5). В описываемой базе данных режимы существуют исключительно для удобства конечного пользователя – пошагового создания эксперимента, а также исключения возможности случайного изменения данных при просмотре. При переходе от локальной базы данных к базе данных, поддерживающей удаленный доступ, было бы целесообразным создать группы пользователей с различными правами доступа к базе данных.

Детальное описание работы с базой данных в каждом режиме приведено подробно далее.

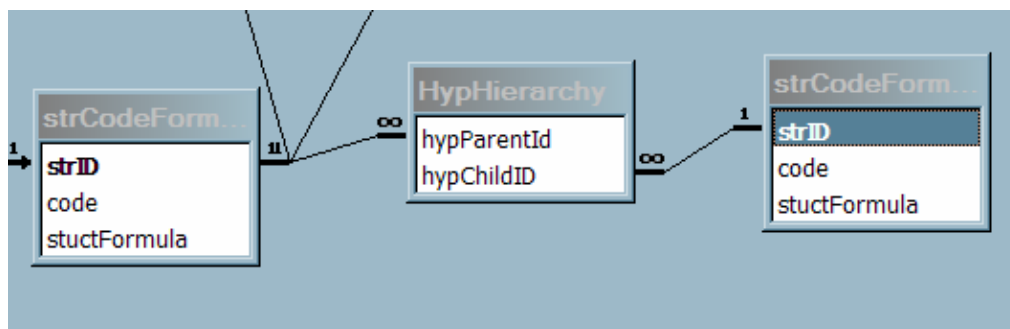


Рис. 3. Представление иерархии гипотез

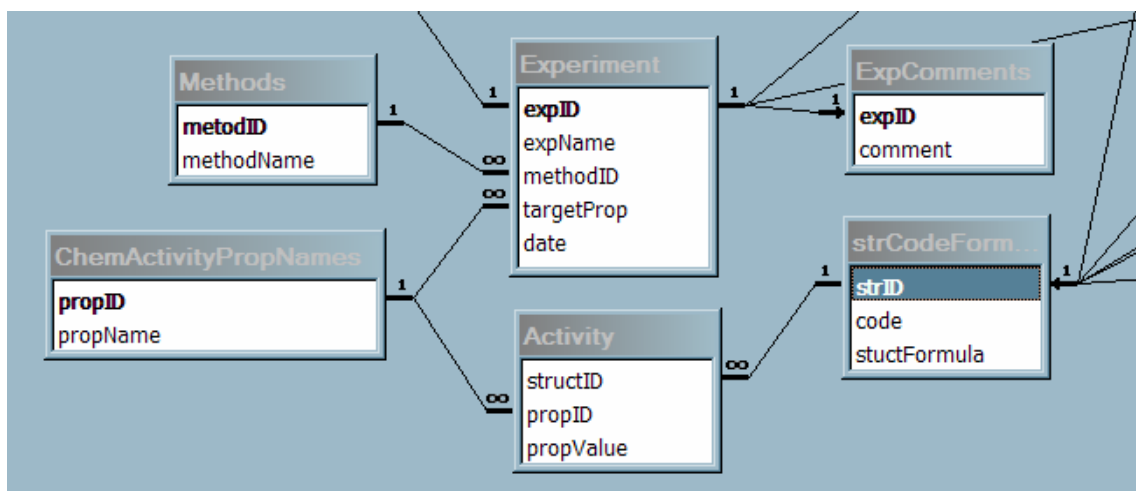


Рис. 4. Таблицы дополнительной информации

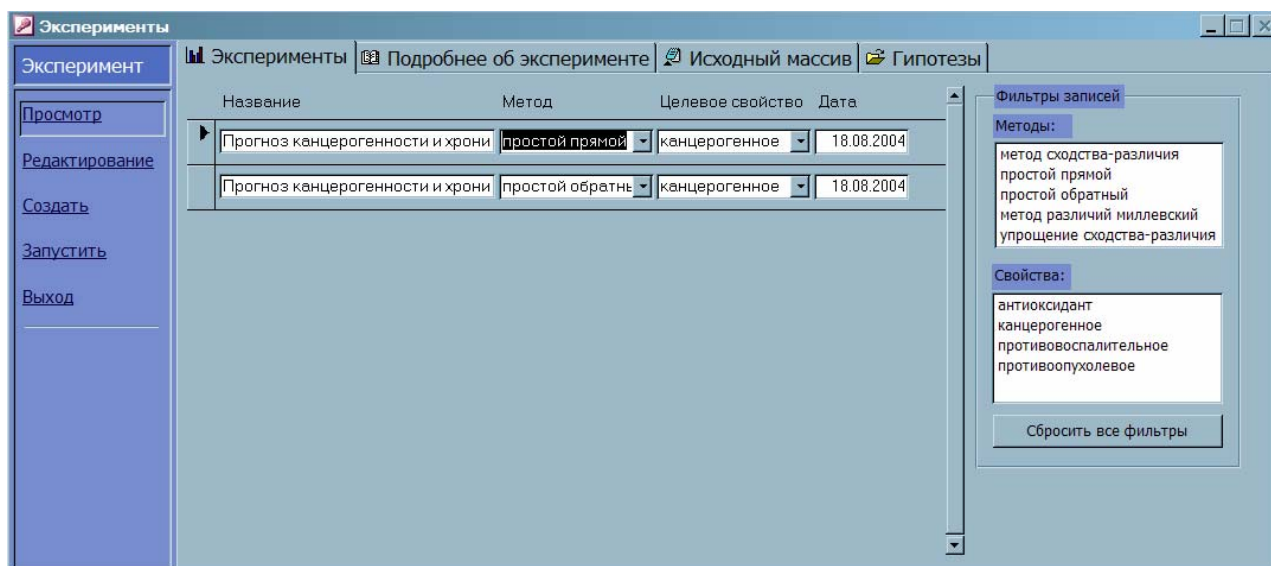


Рис. 5. Основная многостраничная форма.

## 2.1. Режим просмотра данных

Вся имеющаяся в базе данных информация об эксперименте, включая дополнительную информацию, содержащуюся в таблице ExpComment, размещена на вкладке «Подробнее об эксперименте» (рис. 6).

На вкладке «Исходный массив» (рис. 7) показаны все соединения, участвующие в эксперименте. Структурная формула хранится в виде связанного объекта OLE. База данных содержит не сам файл, а только ссылку на него. На этой же вкладке размещена кнопка вызова ФКСП-кодировщика для всех соединений.

Форма свойств соединений показана на рис. 8.

На вкладке «Гипотезы» (рис. 9) показаны все гипотезы первого рода, порожденные в данном эксперименте. Так как в результате работы ДСМ-решателя часто порождается большое количество гипотез, не все из которых являются значимыми, эксперт-химик выделяет наиболее значимые с его точки зрения гипотезы для данной активности. Это отражено наличием у каждой выбранной им гипотезы свойства expMark. На уровне интерфейса созданы фильтры, позволяющие работать либо со всеми гипотезами, либо с теми из них, которые, отобраны экспертом.

На рис.10 показана форма «Родители». Кнопка «Родители» открывает окно со списком родителей данной гипотезы. Каждый родитель описывается порядковым номером в эксперименте, кодом ФКСП и файлом структурной формулы.

## 2.2. Режим редактирования данных

Переход в режим редактирования данных предоставляет пользователю возможность вносить изменения в существующие записи. Достигается это изменением всех форм их свойств, отвечающих за модификацию и добавление данных.

В режиме редактирования есть возможность добавлять новые записи на вкладки «Исходный мас-

сив» и «Гипотезы», при регистрации в базе данных уже проведенного эксперимента, новые записи на вкладке «Эксперименты» добавлять нельзя.

СУБД MS Access снимает с разработчика труд по записи внесенных пользователем изменений в таблицы БД. Изменения оказываются в соответствующих таблицах сразу же после того, как редактируемая запись перестает быть активной.

## 2.3. Режим создания эксперимента

Переключение в режим «Создание эксперимента» (рис. 11) предоставляет пользователю возможность пошагового внесения в базу данных информации о новом эксперименте. При переключении в этот режим фокус перемещается на вкладку «Эксперименты», остальные вкладки скрываются от пользователя. На панели переключения режимов появляется кнопка «Продолжить», при нажатии которой пользователь переходит к следующему шагу создания нового эксперимента. Каждое нажатие кнопки делает видимой вкладку, соответствующую следующему шагу, и перемещает на нее фокус ввода.

Добавление записи о новом эксперименте осуществляется по следующему образцу:

1. Внесение основной информации об эксперименте.
2. Внесение дополнительной информации об эксперименте.
3. Добавление элементов исходного массива и внесение информации о них.
4. Добавление порожденных гипотез.

Обязательной является основная информация об эксперименте: название стратегии и целевое свойство. Без указания этой информации запись об эксперименте не будет зафиксирована в базе данных, остальную информацию допускается внести позднее, воспользовавшись режимом редактирования.

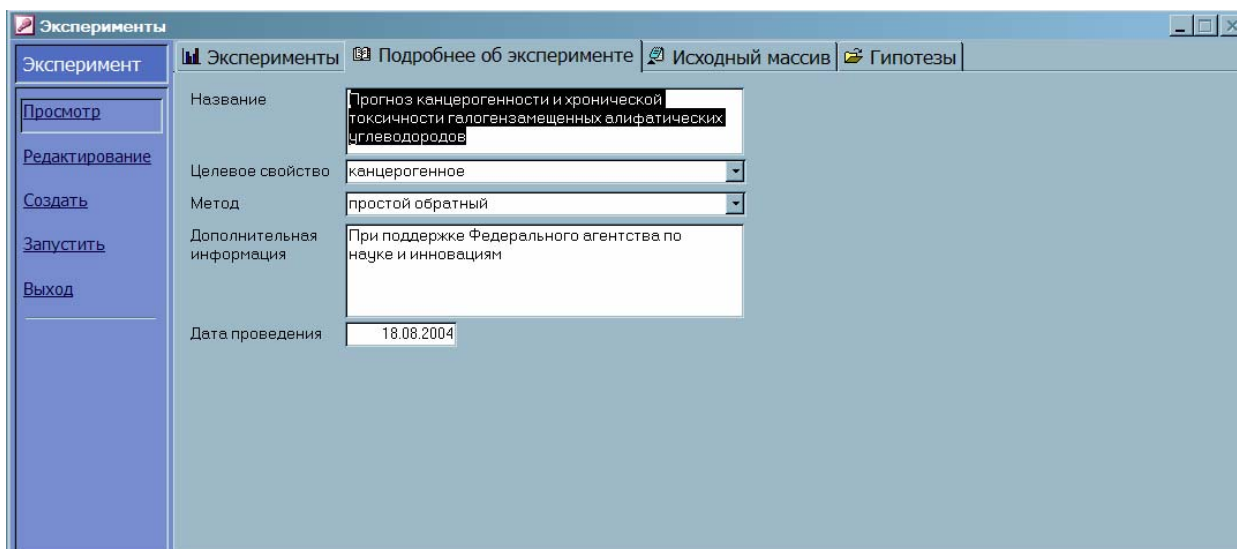


Рис. 6. Вкладка «Подробнее об эксперименте»

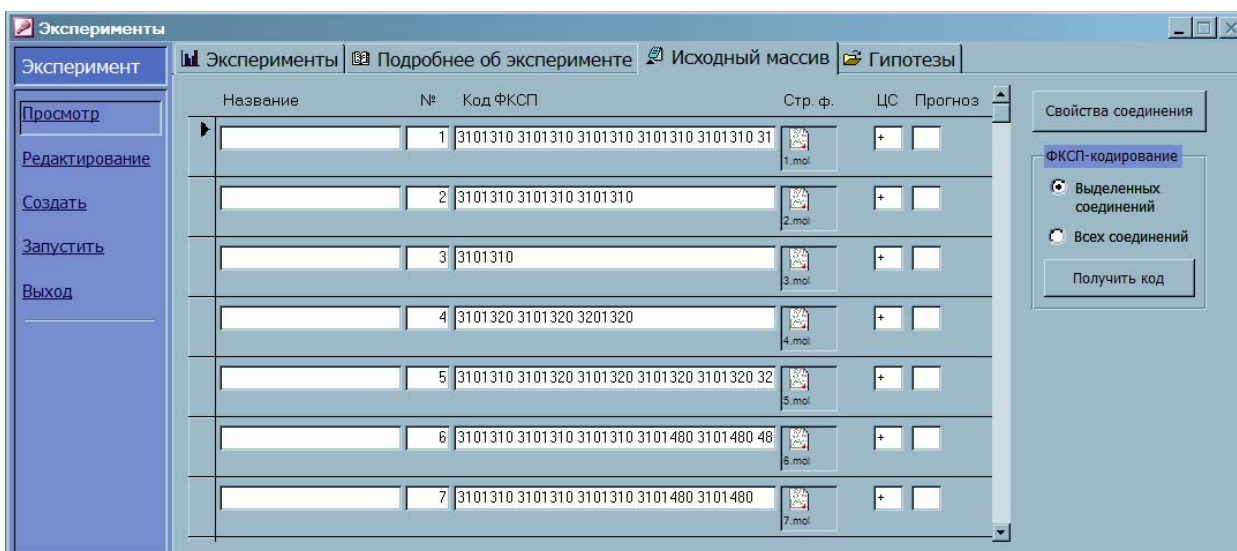


Рис. 7. Вкладка «Исходный массив»

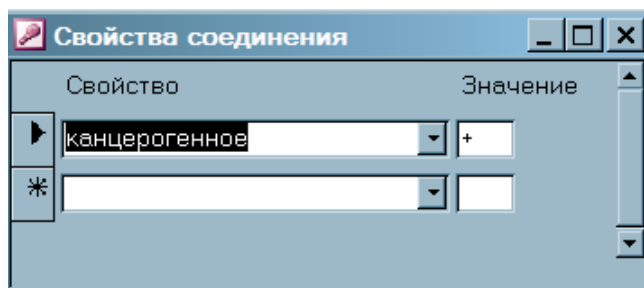


Рис. 8. Форма «Свойства соединения»

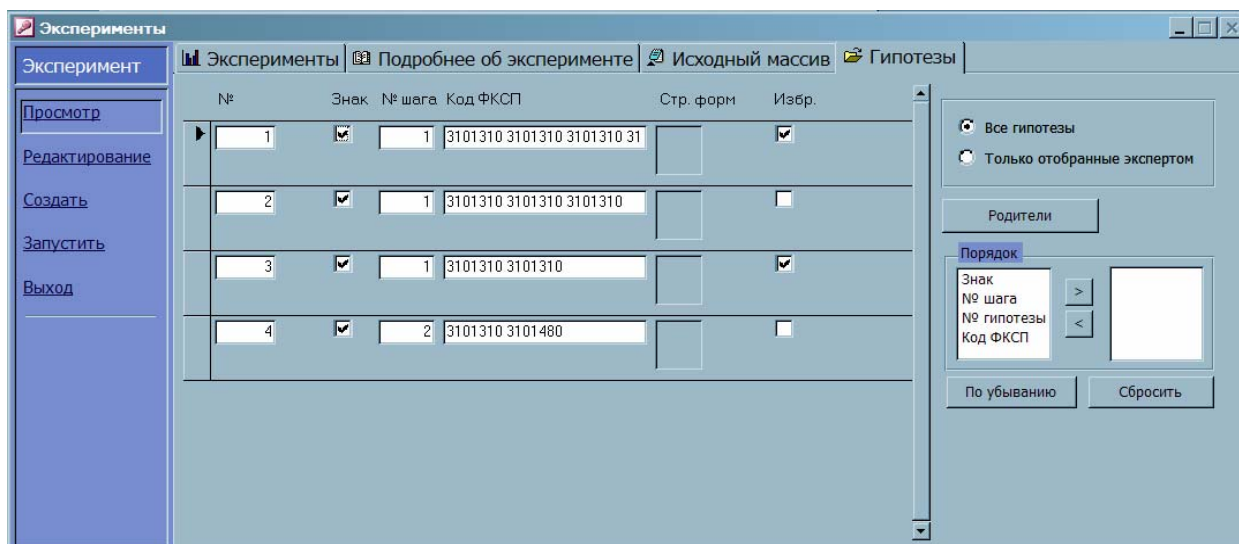


Рис. 9. Вкладка «Гипотезы»

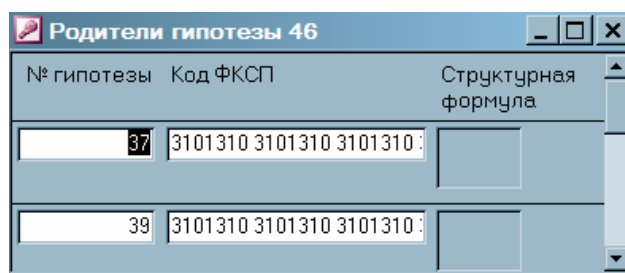


Рис. 10. Форма «Родители»

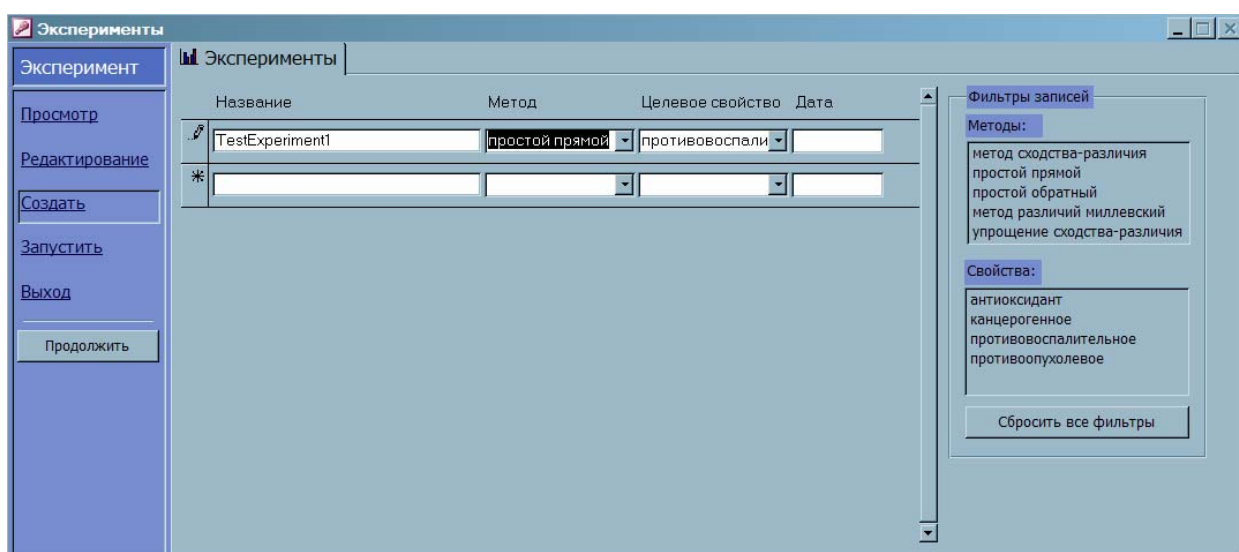


Рис. 11. Режим создания эксперимента.

### 3. Минимальные системные требования

В работе приведены системные требования, соответствующие базовым требованиям используемой СУБД Microsoft Access 2003 [2]:

- Процессор с тактовой частотой не ниже 233 МГц.
- Оперативная память не менее 128 МВ.
- HDD: 400 МВ свободного места.
- Устройство для чтения CD-дисков или совместимое устройство для чтения DVD-дисков.
- Монитор с разрешением 800х600 или более высоким, отображающий 256 и более цветов.
- Операционная система:  
Microsoft Windows Server 2003 или более поздняя версия,  
Microsoft Windows XP или более поздняя версия,  
Microsoft Windows 2000 с пакетом обновления 3 (SP3) или более поздним.

## II. АВТОМАТИЧЕСКОЕ КОДИРОВАНИЕ ХИМИЧЕСКИХ СТРУКТУР КОДОМ ФКСП

### 1. Общие сведения о языке ФКСП

Фрагментарный код суперпозиций подструктур является одним из проблемно-ориентированных языков описания структурной формулы и предназначен для дискретного описания химического соединения в виде набора всех имеющихся в нем подструктур. В настоящий момент ФКСП широко применяется для исследования корреляции «структура-свойство» и прогнозирования свойств соединений. Язык оказался алгоритмически удобен для выявления гипотетических фармакофоров и антифармакофоров с помощью ДСМ-метода автоматического порождения гипотез [3-5].

Структурная формула вещества кодируется кодом ФКСП на основе выделения в структуре дескрипторных центров – атомов или групп атомов, которые могут являться центрами слабого взаимодействия или реакционными центрами. К таковым относятся атомы и группы, содержащие  $\pi$ - и d-электроны или целый электростатический заряд, т.е. все гетероатомы (N, O, S, P, галоид, металлы и т.д.), пары атомов углерода, соединенные кратными связями, ароматические и циклические системы в целом.

Дескрипторные центры (ДЦ) ФКСП характеризуются типом атома (ДЦ первого типа) или же типом атома с его ближайшим окружением (ДЦ второго типа). За слово-дескриптор языка приняты цепочки атомов углерода, начинающиеся и кончающиеся ДЦ.

Так как основное внимание в данной работе уделяется кодированию ациклических соединений, упомянуты будут только принципы кодирования линейных дескрипторов.

Линейный дескриптор кодируется последовательностью из семи цифр следующим образом:

|            |            |            |             |
|------------|------------|------------|-------------|
| Номер ДЦ 1 | Расстояние | Номер ДЦ 2 | Сопряжение. |
|------------|------------|------------|-------------|

Под расстоянием здесь подразумевается количество атомов углерода между двумя ДЦ, составляющими дескриптор.

Наличие или отсутствие сопряжения обозначается цифрами 1 и 0 - соответственно. Формальным признаком сопряжения является наличие у каждого атома углерода в цепи свободного  $\pi$ -электрона.

Стоит отметить, что первым всегда записывается ДЦ с меньшим номером. Дескрипторы, описывающие соединение, располагаются в лексикографическом порядке по возрастанию.

## 2. Программная реализация

### 2.1. Минимальные системные требования

Системные требования соответствуют требованиям, используемым при разработке и необходимым при запуске программы платформы Microsoft .NET Framework 3.5 ServicePack 1 [6]:

- Операционные системы: Windows Server 2003, Windows Server 2008, Windows Vista, Windows XP.
- Процессор: Pentium с тактовой частотой 400 МГц или аналогичный процессор (рекомендуется Pentium с тактовой частотой 1 ГГц или аналогичный процессор).
- ОЗУ: не менее 96 МБ (рекомендуется 256 МБ).
- Жесткий диск: может потребоваться до 500 МБ доступного дискового пространства.
- Экран: разрешение 800 x 600, 256 цветов (рекомендуется разрешение 1024 x 768, 32-разрядный).

Стоит отметить, что в процессе разработки каких-либо ошибок при работе программы отмечено не было.

Необходимая для запуска и корректной работы программы платформа .NET Framework 3.5 доступна для бесплатного скачивания в сети Интернет.

### 2.2. Используемые пространства имен

При разработке помимо стандартных были использованы объекты, методы и свойства следующих пространств имен:

- System::IO – для работы с файловой системой;
- System::Collections::Generic – для работы со встроенными типами данных (списками и очередями);
- System::Windows::Forms – для вывода окон уведомлений;
- System::Data – для обеспечения доступа к данным из базы данных MS Access.

### 2.3. Формат входных данных

Для ввода данных используется формат файлов MOL. Этот формат является международным стандартом, что позволяет использовать для подготовки данных любой совместимый сторонний химический редактор.

Источником информации о структуре MOL-файла, необходимой для его парсирования и конверсии во внутренний формат разрабатываемого программного инструмента, была официальная документация на сайте разработчика формата [7].

### 2.4. Алгоритм работы программы

Кодировщик может осуществлять как кодирование структур в пакетном режиме, так и кодирование отдельного загруженного MOL-файла. При пакетной обработке в один момент времени кодируется один MOL-файл. В процессе работы кодировщик выполняет следующие действия:

1. Парсирует MOL-файл и преобразует структуры во внутренний формат, параллельно выделяя ДЦ 1-го типа и создавая массива индексов атомов данной молекулы.

2. Выделяет ДЦ 2-го типа, добавляет номера ДЦ и идентификаторов атома в индекс.

3. Выделяет ДЦ 45 и 46 – атомы углерода, соединенные кратными связями. Добавляет номера в индекс.

4. Сортирует индекс по возрастанию номеров ДЦ.

5. Осуществляет поиск путей от ДЦ к ДЦ, начиная с центров, имеющих меньший номер. Атомы, для которых все пути найдены, помечаются соответствующим маркером в целях сокращения количества операций.

6. При нахождении пути проверяет допустимость образования дескриптора, и если все условия выполнены, формирует линейный дескриптор.

## 2.5. Программная реализация алгоритма

Основным классом в разработанном приложении является класс *Atom*, содержащий информацию об атоме (имя, валентность, заряд), а также методы, конверсии объекта в другие форматы. От него наследуются следующие три основных класса:

- *BondedAtom* – элемент списка инцидентности атомов молекулы. Кроме основной информации содержит также информацию о связях.

- *DescCenterOne* – класс для представления ДЦ 1-го типа.

- *DescCenterTwo* – класс для представления ДЦ 2-го типа.

Молекула в программе представлена как список объектов класса *BondedAtom*.

Для поиска ДЦ в структуре используются методы реализованные в *System::Collections::Generic::List*. Чтобы обеспечить возможность использования встроенных методов, были реализованы следующие классы сравнения *DCOneComparer* и *DCTwoComparer*, наследуемые от стандартного *Comparer*.

Шаги 1 и 2 алгоритма реализованы в функции *molRead (fksp.cpp)*, принимающей на вход в числе прочих параметров MOL-файл, открытый для чтения, и возвращающей список объектов класса *BondedAtom*, представляющий молекулу.

Поиск ДЦ первого и второго типов ведется в функциях *findDCOne* и *findDCTwo(locateDC.cpp)* соответственно, вызываемых из *molRead*. Перед осуществлением поиска описания ДЦ загружаются из файлов, находящихся в папке установки программы. Это связано с тем, что одним из требований к кодировщику было сохранение имеющейся в программе Д.А. Добрынина возможности модифицировать и дополнять список ДЦ в соответствии с нуждами конкретной задачи. Файлы *decs1.dat* и *desc2.dat*, содержащие информацию о ДЦ, находятся в папке установки программы, и доступны для изменения. Формат хранения ДЦ описан в заголовке каждого из файлов.

Шаги 3, 4, 5 алгоритма реализованы в функции *createCode (fksp.cpp)*, которая для каждого атома, запись о котором есть в массиве индексов, запускает функцию *createLineDescriptor (fksp.cpp)*. Функция осуществляет поиск путей к соседним ДЦ, формирование и кодирование дескриптора.

## 2.6. Интерфейс программы

На рис. 12 представлено основное окно программы. При нажатии кнопки «Обзор» открывается стандартное окно выбора файла, в котором необходимо выбрать MOL-файлы для кодирования. Допускается выбор одного или нескольких файлов. Для выбора нескольких файлов необходимо зажать клавишу *Ctrl* или *Shift*. После того как все файлы будут выбраны, необходимо нажать кнопку «Открыть».

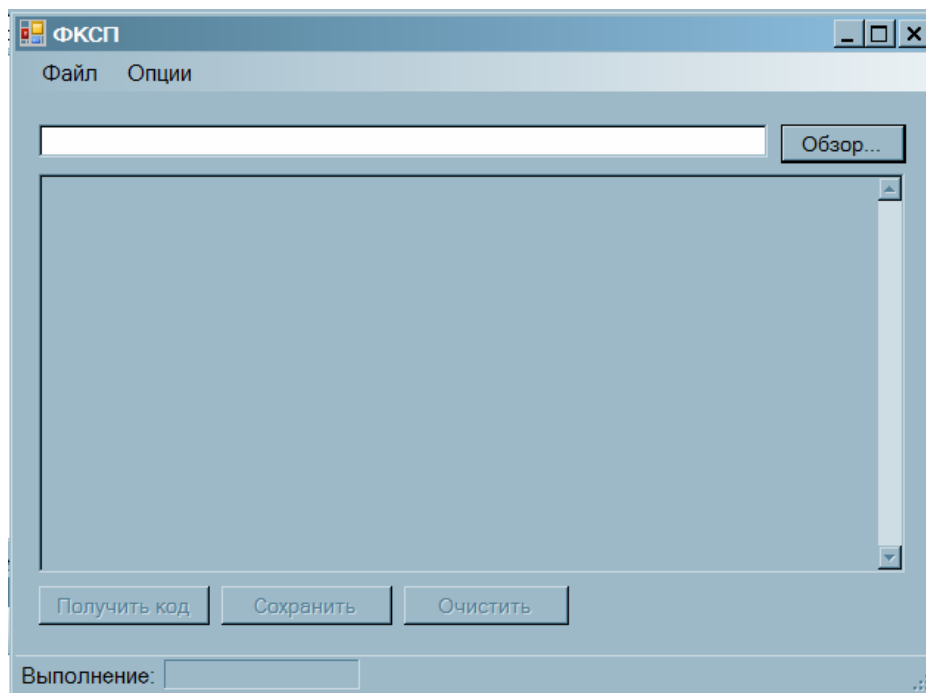


Рис. 12. Главное окно программы ФКСП-кодирования.

После выбора файлов необходимо использовать активную кнопку «Получить код», соединения кодируются. По завершению кодирования в окне вывода результатов печатается полный путь к файлу и коды для данного файла, что позволяет пользователю сохранить результаты кодирования в текстовом файле.

### **III. ИНТЕГРАЦИЯ ИМЕЮЩИХСЯ МОДУЛЕЙ В ЕДИНУЮ СИСТЕМУ**

#### **1. Взаимодействие БД и ФКСП-кодировщика**

Как упоминалось во введении, одним из основных требований к разрабатываемой системе была возможность автоматического заполнения БД из ФКСП-кодами.

Для этого к базе данных был добавлен модуль на VBA, осуществляющий экспорт данных в программу ФКСП-кодирования, а к программе кодирования данных, соответственно, модуль, отвечающий за доступ к данным, хранящимся в базе данных. Интерфейс базы данных был дополнен кнопкой вызова ФКСП-кодировщика (подробнее, см. параграф 2.1 главы II).

При нажатии данной кнопки производятся два основных действия:

1. Запись в файл launch.ini, находящийся в папке установки программы ФКСП-кодирования, маркера запуска DATABASE, а также пути к базе данных, информации о таблицы, в которую следует записать сгенерированные коды и кодируемых структурах.

Если записи каких-то из соединений, отображенных для кодирования, не содержат ссылки на MOL-файл с информацией о структуре, выдается соответствующее предупреждение; данная запись пропускается.

2. Запуск исполняемого файла fksppInterface.exe с помощью команды Shell языка VBA.

ФКСП-кодировщик при запуске считывает информацию из файла launch.ini. Если в файл записан маркер USER, открывается основное окно программы, описанное в параграфе 2.6. Однако, если запуск помечен маркером DATABASE, программа не предлагает пользователю возможности выбрать файлы для кодирования. Кодирование запускается сразу для файлов, информация о которых прописана в файле launch.ini.

По окончании работы кодировщика запускается программа импорта данных результатов в базу данных, которая создана на базе модели доступа к данным ADO.NET.

По завершении импорта окно программы ФКСП-кодирования автоматически закрывается.

#### **2. Взаимодействие базы данных и ДСМ-решателя**

Кроме описанных в разделе 2 главы II режимов работы с базой данных был реализован также режим запуска эксперимента, позволяющий передать ДСМ-решателю информацию из базы данных в соответствующем виде.

В данной работе был использован решатель, реализованный А.Ю. Волковой [8]. База данных и ДСМ-решатель взаимодействовали «через данные». Этот способ имеет ряд преимуществ, основным из которых является отсутствие необходимости вносить какие бы то, ни было изменения в исходный код существующего решателя. Для организации эффективного взаимодействия достаточно обеспечить экспорт и импорт данных.

Решатель принимает на вход файл формата .xls, содержащий представление данных, разработанное непосредственно А.Ю. Волковой. Результаты работы решателя записываются в тот же документ Excel на соответствующие листы. Задача автора данной работы, таким образом, сводится к созданию инструментов экспорта данных из БД в файл соответствующего формата.

Запуск выделенного эксперимента из интерфейса БД фактически запускает следующие действия:

1. Создается документ Excel нужного формата. Создание документа идет по имеющемуся шаблону, который хранится в папке служебных файлов БД и может, при необходимости, быть изменен или дополнен.

2. Запускается модуль экспорта данных в Excel из БД. Работа модуля включает следующие этапы:

- 2.1. Автоматическое кодирование структур, код для которых не был внесен в соответствующее поле БД заранее.

- 2.2. Формирование списка дескрипторов, которыми описываются все соединения исходного массива эксперимента.

- 2.3. Формирования описания исходных примеров: каждый пример описывается идентификатором, типом значения и информацией о повторяемости в соединении каждого дескриптора в соединении.

- 2.4. Экспорт информации на листы Descriptors и Examples созданного документа Excel. Экспорт осуществляется с использованием объектной модели MS Excel.

- 2.5. Запуск ДСМ-решателя с помощью команды Shell языка VBA.

После того, как решатель заканчивает работу, производится импорт результатов работы решателя с листов Hypotheses и T\_Result в базу данных. Для импорта используется объектная модель MS Excel и модель доступа к данным ADO. Полная схема данных ДСМ-экспериментов по фармакологии представлена на рис. 13.

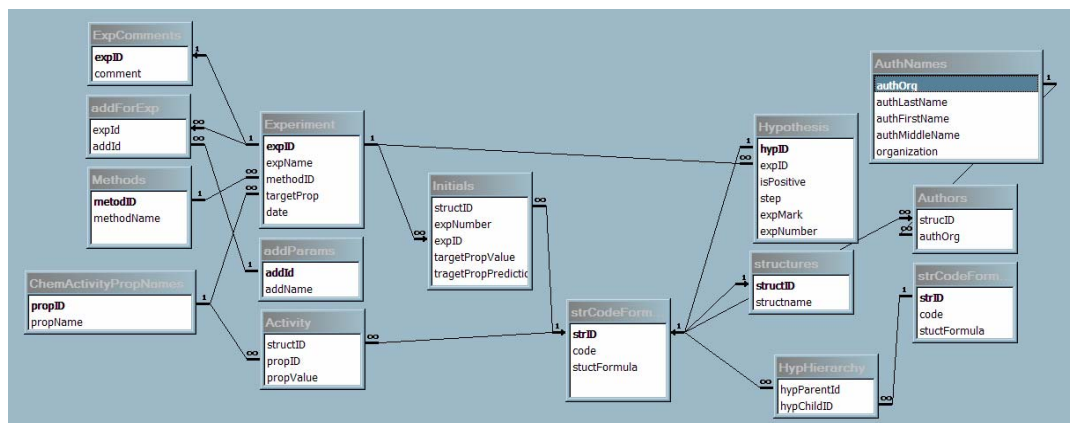


Рис. 13. Полная схема базы данных ДСМ-эксперимента по фармакологии

## ЗАКЛЮЧЕНИЕ

Итогом проделанной работы стала целостная система с дружественным пользователю интерфейсом, позволяющая осуществлять просмотр данных и работу с ними, подготавливать данные для проведения ДСМ-эксперимента и проводить эксперименты различными стратегиями. Эта система включает в себя не только базу фактов о наличии у тех или иных соединений какой-либо химической активности, но и базу знаний, содержащую информацию о фармакофорах и антифармакофорах, которые представляют собой большой практический интерес для специалистов-фармакологов.

## СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Автоматическое порождение гипотез в интеллектуальных системах / под ред. В.К. Финна. – М.: Книжный дом «Либроком», 2009. – С. 10-48.
2. Минимальные системные требования Microsoft Office 2003. – URL: <http://support.microsoft.com/kb/822129/ru>
3. Блинова В.Г., Добрынин Д.А. Языки представления химических структур в интеллектуальных системах для конструирования лекарств // НТИ. Сер.2. – 2000. – № 6. – С. 14-21.
4. Блинова В.Г., Решетникова В.В., Добрынин Д.А. и др. Интеллектуальный анализ цитотоксической активности химических соединений с помощью стратегий ДСМ-метода // НТИ. Сер. 2 – 2010. – № 6. – С.14-23.
5. Харчевникова Н.В., Блинова В.Г., Добрынин Д.А., Максин М.В., Жолдакова З.И. Применение ДСМ-метод и квантово-химических расчетов для прогноза хронической токсичности галогензамещенных алифатических углеводов // НТИ. Сер. 2. – 2004. – № 12. – С. 21-28.
6. Microsoft.NET Framework 3.5. – URL: <http://www.microsoft.com/downloadsru-rudetails.aspx?FamilyID=333325fd-ae52-4e35-b531-508d977d32a6>
7. CTFileformats. Документация на сайте разработчика. – URL: [www.symyx.com/downloads/public/ctfile/ctfile.pdf](http://www.symyx.com/downloads/public/ctfile/ctfile.pdf)
8. Волкова А. Ю. Анализ данных различных предметных областей с помощью процедур ДСМ-метода автоматического порождения гипотез // НТИ. Сер.2. – 2011 – № 6. – С. 9-18.

*Материал поступил в редакцию 13.09.12.*

## Сведения об авторах

**БЛИНОВА Валентина Георгиевна** – кандидат химических наук, старший научный сотрудник Отдела теоретических и прикладных проблем информатики ВИНТИ РАН, Москва  
e-mail: [blinova@viniti.ru](mailto:blinova@viniti.ru)

**РЕШЕТНИКОВА Вера Владимировна** – кандидат технических наук, старший научный сотрудник Российского онкологического научного центра им. Н.Н. Блохина Российской академии медицинских наук (ФГБУ РОНЦ им. Н.Н.Блохина РАМН), Москва  
e-mail: [veravr@gmail.com](mailto:veravr@gmail.com)

# АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.2'373.46 : [620.3:61]

И.О. Кузнецов

## Автоматическое извлечение двусловных терминов по тематике "Нанотехнологии в медицине" на основе корпусных данных\*

*Представлен обзор статистических методов извлечения двусловных терминов из текстов на естественном языке, а также предложена новая двухступенчатая модель, в которой связность и специфичность терминов оцениваются двумя независимыми параметрами. Производится экспертное сравнение качества извлечения и детальный анализ результатов, полученных при использовании каждого из рассматриваемых подходов.*

**Ключевые слова:** извлечение терминов, специфичность словосочетания, устойчивое словосочетание, корпусная лингвистика, онтологическое моделирование

### ВВЕДЕНИЕ

Автоматическая обработка естественного языка в последние десятилетия приобрела большое значение в связи с ростом объема информации, доступной в электронном виде. Лингвистические процессоры используются для извлечения информации из неструктурированных текстов, повышения качества поиска, автоматического перевода. Моделирование экстралингвистических фактов о внешнем мире – одна из важнейших задач, решаемых в рамках этого направления. От прикладных систем требуется не столько анализ текстов на уровне общеязыковых фактов, сколько извлечение информации, специфичной для конкретной предметной области или **домена** (например, "Экономика", "Биология", "Атомная энергетика", "Нанотехнологии"). Одним из способов представления такой информации в системах автоматической обработки текстов является **онтологическое моделирование** предметной области, при котором объекты внимания исчисляются и связываются семантическими отношениями. Онтологии для задач автоматической обработки узкоспециальных доменов зачастую создаются вручную с привлечением специалистов по онтологическому инжинирингу и экспертов предметной области. Создание онтологий вручную – крайне ресурсоёмкая процедура, в связи с чем возникает потребность в автоматизации процесса.

Качественное извлечение объектов предметной области и определение связей между ними – сложная задача, которая до сих пор не решена. Однако даже частичная автоматизация на уровне порождения списка кандидатов в предметную область способна значительно сократить время на разработку соответствующих экспертных ресурсов. Эксперт получает список объектов-кандидатов, после чего вручную удаляет из этого списка лишние объекты и определяет связи между оставшимися объектами. Подобный подход также может быть использован для расширения существующих онтологий и для адаптации систем к текстам новых типов [1].

В результате исследований в области автоматического извлечения терминов был предложен ряд методов, позволяющих на основе набора текстов из предметной области формировать списки терминов-кандидатов. Подобные методы можно условно разделить на три класса:

- лексико-грамматические шаблоны,
- статистические методы,
- машинное обучение.

В рамках **метода лексико-грамматических шаблонов** термины извлекаются из текстов с помощью правил, закодированных в виде конечного автомата [2]. Такие подходы опираются на лингвистические свойства прототипических терминов, позволяя по аналогии извлекать новые термины.

**Статистические методы** позволяют выделить терминологические сочетания из текстовой коллекции, основываясь на статистических свойствах естественного языка [3]. При статистическом подходе появление слова в тексте рассматривается как случайное событие. Используя такую трактовку, мы мо-

\* Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации в рамках государственного контракта № 07.524.11.4005 от «20» октября 2011 г., заключенного между Министерством образования и науки Российской Федерации и ЗАО «Эвентос»

жем оперировать языковыми явлениями в контексте аппарата теории вероятностей и математической статистики, в частности, оценивать уровень случайности совместных событий "появление слова А" и "появление слова В" в контекстах определенного размера. Если соседство двух слов в исходном корпусе является неслучайным со статистической точки зрения, эти слова могут рассматриваться как устойчивое словосочетание и, следовательно, как кандидат на попадание в предметную область. В то же время, в качестве случайных событий могут быть выбраны и события вида "появление конструкции Х в текстах предметной области" и "появление конструкции Х в языке в целом". При такой трактовке мы можем выбрать из конструкций те, которые являются специфичными для конкретной предметной области. Подобные конструкции также могут рассматриваться как потенциальные объекты онтологии.

Наконец, извлечение терминов может быть реализовано с помощью алгоритмов **машинного обучения**, опирающихся на контекстуальные и внутренние свойства слов (например, [4]). Подобные подходы, как правило, показывают сравнительно высокую точность, однако требуют для реализации наборов обучающих данных, которые могут отсутствовать на этапе прототипирования соответствующего ресурса.

Как правило, наиболее успешными для решения представленной задачи оказываются гибридные подходы, сочетающие в себе методы из разных классов.

В данной работе мы сравниваем результаты работы основных статистических методов выделения терминологических словосочетаний, а также предлагаем метод, опирающийся на трёхэтапную технологию обработки данных и внешний общезыковой корпус. В рамках исследования было решено ограничить область внимания двусловными терминологическими словосочетаниями. Исследования показывают, что двусловные термины составляют значительную часть терминологического аппарата независимо от предметной области [5], следовательно, оказывается возможным значительно уменьшить сложность задачи без ощутимых потерь в полноте. Кроме того, предложенные нами методы могут быть как сужены для анализа однословных терминов, так и расширены для анализа терминов произвольной длины.

## ТЕОРЕТИЧЕСКАЯ ОСНОВА

### Объект исследования

Понимание термина "устойчивое словосочетание" значительно варьируется у разных исследователей. По всей видимости, можно считать бесспорной континуальную природу этой "устойчивости" [6]. На одном конце этого континуума находятся свободные словосочетания, смысл которых определяется объединением смыслов входящих слов. На другом - идиомы, значение которых не может быть выведено подобным образом (*царь горы, бить баклуши, утечка мозгов*). Один из промежуточных вариантов - словосочетания, обозначающие единые фрагменты действительности (и в силу этого "устойчивые"), однако семантически разложимые с лингвистической точки зрения. В эту группу попадают как составные наиме-

нования объектов, неспецифичных для какой-либо предметной области (*телефонная книга, шариковая ручка*), так и терминологические словосочетания, которые образуют систему в рамках того или иного домена (*углеродная нанотрубка, файловое хранилище*). Возможна ситуация, когда некоторое словосочетание, будучи достаточно свободным в общезыковом контексте, оказывается связанным в рамках предметной области [7]. Так, например, словосочетание *иностранный язык* в общем употреблении будет менее устойчивым, чем в контексте деятельности вузов, где оно обозначает название учебной дисциплины и оказывается в семантическом поле других учебных дисциплин.

Принимая во внимание эти факты, мы можем формально определить тип словосочетаний, на извлечение которых будут направлены наши усилия. **Терминологическим словосочетанием** (или **термином**) мы будем называть **неразрывное устойчивое словосочетание, специфичное с точки зрения предметной области**. Так, например, при выборе предметной области "Нанотехнологии в медицине" словосочетания *иммобилизация протеиназы, гибридный эмбрион* и *левый желудочек* могут быть отнесены к корректным, а словосочетания *высокая температура* и *мушка дрозофила* корректными не будут (хотя в рамках других предметных областей они, несомненно, являются терминами).

### Связность слов

Устойчивость словосочетания может быть определена с помощью критериев лингвистического и статистического характера.

С одной стороны, неразрывное устойчивое словосочетание должно быть **грамматически связано**, т.е. образовывать синтаксическую группу. Как показано в работе [5], на материале [8], терминологические словосочетания как правило представляют собой группу существительного с зависимыми существительными или прилагательными, состоящую из двух-трех слов. В соответствии с упомянутым исследованием, доля терминов другой структуры составляет менее одного процента. Следовательно, установив лексико-грамматический фильтр нужного типа, можно значительно повысить качество извлекаемых словосочетаний и сократить набор кандидатов, который будет представлен эксперту.

С другой стороны, устойчивость также влияет на **совместную встречаемость** слов. Представив появление слов в тексте как случайное событие, мы можем измерить, насколько появление одного слова коррелирует с появлением другого, и на основании этого сделать вывод о неслучайности их совместного появления и, следовательно, о "статистической" устойчивости словосочетания. В качестве примера подобных методов можно упомянуть сравнение абсолютной частоты конструкций, оценку по совместной информации (*mutual information*), критерий *log likelihood* и *критерий хи-квадрат* (см. обзор [9] и сравнение в применении к нашей задаче в [3]).

Наилучшие результаты даёт комбинирование этих методов, при котором кандидаты сначала от-

бираются по лексико-грамматическим шаблонам, а затем ранжируются в соответствии со значениями выбранной оценочной функции.

### Специфичность словосочетания

Определение **специфичности** слов и словосочетаний - отдельная проблема, которая активно изучается в рамках исследований по информационному поиску. В результате этих исследований было разработано множество мер, позволяющих оценить, насколько хорошо то или иное слово или словосочетание характеризует документ, в котором оно содержится. На основе этих данных можно судить о соответствии документов запросам пользователя и ранжировать документы по выбранной метрике. Наряду с методами ссылочного ранжирования существует также обширный класс методов, основанных на сопоставлении частоты слова в документе и в текстовой коллекции. Наиболее известными метриками этого класса являются *tf\*idf* [10] и *Okapi BM-25* [11]. Будучи разработанными для повышения эффективности информационного поиска по документам, меры специфичности, основанные на частоте, могут быть использованы и для ранжирования устойчивых словосочетаний в той или иной предметной области. Представив корпус текстов предметной области как документ, а корпус языка - как общую коллекцию, можно оценить специфичность того или иного термина, пользуясь теми же метриками, что и в исходной задаче поиска по документам.

### МЕТОДОЛОГИЯ

Извлечение терминологических словосочетаний из текста может производиться как с помощью единственного метода, так и путём последовательного применения нескольких методов извлечения. Если говорить о статистических методах, извлечение производится, как правило, в два этапа. На первом этапе из всех последовательностей слов заданной длины формируется **список кандидатов**. На втором этапе производится **ранжирование кандидатов** в соответствии с выбранной метрикой. На выходе системы производится ранжированный список кандидатов, в котором, если извлечение было произведено успешно, присутствует высокая доля корректно отобранных терминов.

Применение единственной метрики при ранжировании зачастую позволяет достичь приемлемого результата. Однако мы предлагаем также рассмотреть возможность комбинированного применения методов ранжирования, которые по отдельности обеспечивали бы выполнение требований устойчивости и специфичности. Модели с использованием единственной метрики в дальнейшем будем называть **одноуровневыми моделями**, так как в данном случае специфичность и устойчивость измеряются одним общим показателем. Модели, включающие последовательное применение нескольких методов ранжирования, будем называть **многоуровневыми моделями**.

### Одноуровневые модели

В рамках исследования мы рассмотрим три популярных метрики, используемых при отборе терминологических словосочетаний на основании коллекции текстов из того или иного домена.

Естественный метод отбора подобных словосочетаний - ранжирование кандидатов по абсолютной частоте. Пусть  $c(x, y)$  - количество случаев, когда последовательность слов  $x$  и  $y$  встретилась в текстах предметной области. Тогда мера **Freq** определяется достаточно прямолинейно:

$$Freq(x, y) = c(x, y).$$

Несмотря на простоту, с помощью данного метода зачастую можно получить неплохие результаты в верхней части списка [7].

Другой широко распространенный метод ранжирования опирается на понятие взаимной информации о совместной встречаемости двух слов. Для каждой пары слов  $x$  и  $y$  вычисляется величина *pointwise mutual information (PMI)*:

$$PMI(x, y) = \frac{P(x, y)}{P(x) * P(y)} = \frac{N * c(x, y)}{c(x) * c(y)},$$

где  $N$  - число слов в корпусе,  $c(x)$ ,  $c(y)$ ,  $c(x, y)$  - число вхождений каждого из слов и словосочетания соответственно.

Третий из методов, которые мы рассмотрим в рамках анализа одноуровневых моделей, измеряет специфичность словосочетаний-кандидатов на основании сравнения относительной частоты употребления кандидата в корпусе текстов предметной области и в общеязыковом корпусе. Обозначим число слов в корпусе предметной области как  $N_D$ , а число слов в общеязыковом корпусе - как  $N_L$ . Далее, обозначим частоту рассматриваемого словосочетания в каждом из корпусов как  $c_D(x, y)$  и  $c_L(x, y)$  - соответственно. Тогда мера под названием *Странность (Weirdness)*, см. [12]) определяется как

$$Weirdness(x, y) = \frac{c_D(x, y) / N_D}{c_L(x, y) / N_L}.$$

Поскольку некоторые сочетания из корпуса предметной области могут отсутствовать в общем корпусе, для оценки частоты в общеязыковом корпусе предлагается использовать сглаживание Лапласа. Обозначим размер словаря как  $V$ . В нашем случае роль словаря выполняют словосочетания, так что в качестве величины  $V$  следует использовать количество уникальных пар слов. Формула меры *Weirdness* с учетом сглаживания приобретает следующий вид:

$$Weirdness_{LS}(x, y) = \frac{\frac{c_D(x, y)}{N_D}}{\frac{c_L(x, y) + 1}{N_L + V}}.$$

## Многоуровневые модели

Обратимся теперь к моделям, в которых последовательно применяется две метрики, обеспечивающие устойчивость и специфичность отобранных кандидатов. Вопрос о том, как комбинировать значения этих метрик, представляется достаточно сложным. В рамках исследования мы использовали отсечение по некоторому заданному порогу значения первой метрики с последующим ранжированием по второй метрике. Мы будем рассматривать две архитектуры, отличающиеся порядком применения метрик.

В первом случае сначала кандидаты ранжируются по устойчивости с использованием метрики *PMI*, а затем вершина списка дополнительно ранжируется по метрике *Weirdness*.

Во втором случае сначала по метрике *Weirdness* отбираются наиболее специфичные кандидаты, а затем эти кандидаты сортируются по устойчивости с использованием метрики *PMI*.

## Оценка результатов

Оценка результатов производилась экспертом на основе первых 100 словосочетаний в выдаче соответствующего метода. Словосочетания, полученные с помощью всех методов, были собраны в единый список, который был представлен эксперту. Эксперт отмечал, является ли данное словосочетание термином предметной области. Для оценки мы использовали бинарный параметр (корректно/некорректно). Существуют и более тонкие классификации, позволяющие детально оценить преимущества и недостатки методов (например, см. [13]).

Затем для каждого метода строился график зависимости кумулятивной точности в зависимости от ранга отсечения. Следует отметить, что в большинстве исследований по извлечению терминов приоритет отдаётся точности как более значимому показателю с практической точки зрения [14]. Мы также придерживаемся этого подхода. Пусть функция  $rank(x, y)$  возвращает значение ранга для словосочетания  $x, y$ . Пусть функция  $correct(x, y)$  истинна, если словосочетание  $x, y$  было признано релевантным. Тогда для произвольного ранга  $r$  значение точности вычисляется как

$$Precision(r) = \frac{|rank(x, y) < r \ \& \ correct(x, y)|}{|rank(x, y) < r|}.$$

В результате оценки формируются графики зависимости точности извлечения от ранга, которые позволяют наглядно представить информацию об изменении качества, а также сравнить предложенные методы.

## ЭКСПЕРИМЕНТ

### Корпуса

Для извлечения терминов был выбран корпус новостей и научно-популярных статей по тематике "Нанотехнологии в медицине" объёмом 1473147 словоупотреблений (далее **доменный корпус**). В качестве данных контрастного корпуса были использованы частотные списки двусловных и однословных

конструкций из **Основного корпуса Национального корпуса русского языка (НКРЯ)**, находящиеся в открытом доступе. Объём Основного корпуса составляет 209198275 слов.

### Предобработка

В целях повышения скорости прототипирования при реализации предварительной обработки доменного корпуса были использованы ресурсы платформы *NLTK*<sup>1</sup> и морфологический анализатор русского языка *Pymorphy*<sup>2</sup>. Тексты доменного корпуса были разбиты на слова и на предложения, затем каждое слово было подано на вход морфологическому анализатору, который приводил слово к нормальной форме, определял часть речи, а в случае неоднозначности и отсутствия слова в словаре генерировал возможные гипотезы (табл. 1).

Затем доменный корпус был помещён в базу данных MySQL для упрощения процедуры доступа к данным.

Поскольку слова и словосочетания в списках из НКРЯ представлены в ненормализованной форме, мы также выполнили предобработку этих списков, приведя слова к нормальной форме и автоматически определив их части речи, а также агрегировав данные о частотности для каждой нормализованной формы. Сделав это, мы получили дополнительное преимущество в силу унификации систем грамматической нотации и нормальных форм для списков из доменного корпуса и из НКРЯ.

### Отбор кандидатов

Кандидаты отбирались с помощью лексикограмматических шаблонов, полный список которых представлен в табл. 2:

В результате отбора из доменного корпуса были извлечены словосочетания, соответствующие хотя бы одному из представленных шаблонов. В случае неоднозначности словосочетание также извлекалось. Аналогичная процедура отбора была применена к словосочетаниям из Корпуса русского языка.

### Ранжирование

Затем для каждого словосочетания-кандидата были вычислены меры *Freq*, *PMI* и *Weirdness*. Пары были отсортированы по убыванию каждой из трех мер, и первые 100 пар были выбраны для оценки, в результате чего было сформировано три стословных списка кандидатов. Также была произведена многоуровневая обработка с приоритетом устойчивости и с приоритетом специфичности. В случае с устойчивостью пары ранжировались по значению *PMI*, и из ранжированного списка извлекались первые 1000 кандидатов. Затем мы ранжировали выбранные словосочетания по значению метрики *Weirdness*, и первые 100 кандидатов из полученного списка отправлялись на оценку. Эксперимент с приоритетом специфичности производился аналогичным образом: первые 1000 результатов из ранжирования по *Weirdness* упорядочивались по *PMI*,

<sup>1</sup> <http://nltk.org/>

<sup>2</sup> <http://pythonhosted.org/pymorphy/>

и из 100 словосочетаний с наибольшим значением *PMI* формировался финальный список, участвовавший в процедуре оценки.

## Результаты

В результате анализа качества работы каждого из предложенных методов были получены оценки точности на 100 первых кандидатах (табл. 3), а также графики зависимости качества извлечения от ограничения на ранг в рамках этого множества, представленные на рисунке.

Как можно видеть из графиков, представленных на рисунке, наилучшие результаты достигаются при использовании одноуровневой модели на основе *Weirdness* и двухуровневой модели *Weirdness+PMI*, в которой результаты ранжирования по специфичности дополнительно уточняются в ходе проверки на устойчивость. Также из графиков можно заключить, что значения точности для выбранных метрик стабилизируются при возрастании оценочного множества до 100, что позволяет говорить о релевантности полученных в ходе эксперимента оценок.

Таблица 1

### Результат морфологического анализа

| token                    | norm                    | class |
|--------------------------|-------------------------|-------|
| основополагающему        | ОСНОВОПОЛАГАЮЩИЙ        | П     |
| части                    | ЧАСТИТЬ+ЧАСТЬ           | Г+С   |
| светоперераспределяющего | СВЕТОПЕРЕРАСПРЕДЕЛЯЮЩИЙ | П+С   |

Таблица 2

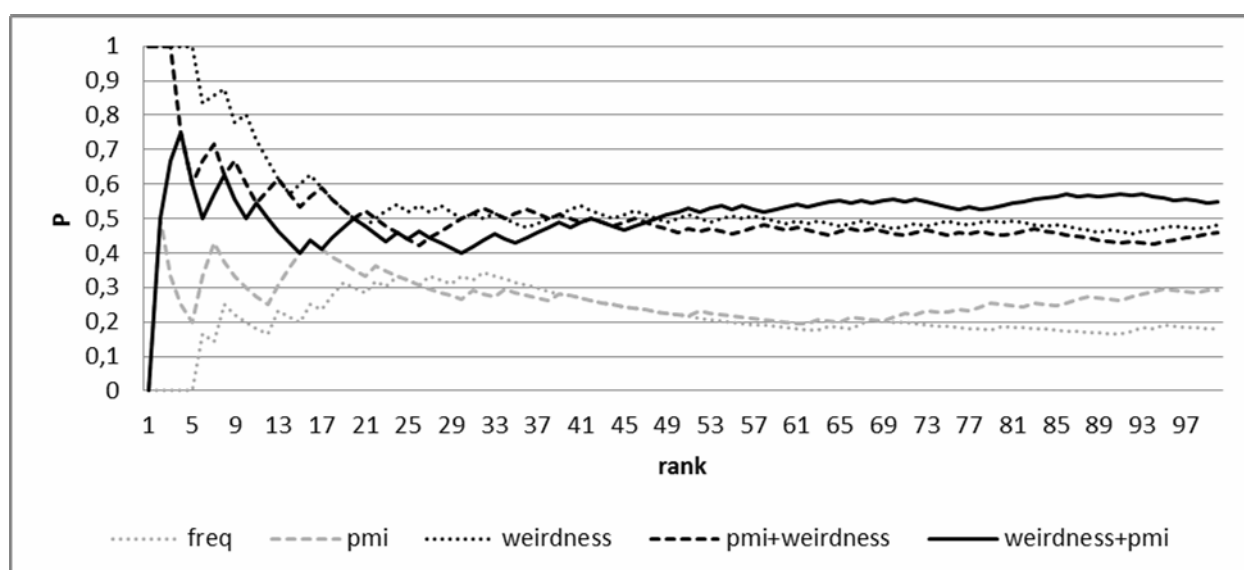
### Лексико-грамматические шаблоны

| шаблон       | пример                      |
|--------------|-----------------------------|
| Прил. + Сущ. | стволовая клетка            |
| Прич. + Сущ. | трансформированное вещество |
| Сущ. + Сущ.  | длина волны                 |

Таблица 3

### Оценка точности извлечения на 100 кандидатах

|        | Freq | PMI  | Weirdness | PMI+Weirdness | Weirdness+PMI |
|--------|------|------|-----------|---------------|---------------|
| P(100) | 0.18 | 0.29 | 0.48      | 0.46          | <u>0.55</u>   |



Зависимость суммарной точности извлечения от порога отсечения

## АНАЛИЗ РЕЗУЛЬТАТОВ

Для того чтобы интерпретировать полученные результаты, обратимся к ранжированным спискам, полученным с использованием рассматриваемых подходов.

В рамках нашего эксперимента худший итоговый результат показала метрика *Freq*, основанная на абсолютной частоте словосочетания. Первые 10 словосочетаний-кандидатов, выделенные с помощью этой метрики, представлены в табл. 4 (для наглядности в таблице также приведены абсолютные частоты употребления отдельных слов). Здесь и далее термины, которые эксперт определил как корректные, выделены полужирным шрифтом.

Как мы можем убедиться, ранжирование по частоте испытывает сильное влияние со стороны абсолютных частот слов, входящих в словосочетание. Кроме того, большинство отобранных словосочетаний является общими конструкциями русского языка и не может быть отнесено к специализированной терминологической лексике. В то же время следует отметить, что качество извлечения методом *Freq* на первых 40 позициях сопоставимо с качеством извлечения метода *PMI*, что хорошо согласуется с уже упомянутыми наблюдениями, сделанными в [7].

Следующий результат по итоговой точности показала метрика *PMI*, учитывающая статистическую устойчивость словосочетания. Первые 10 словосочетаний из выдачи этой метрики представлены в табл. 5.

Можно отметить, что на этот раз вершину списка действительно составляют устойчивые словосочетания, часть из которых также относится к терминологии области "Нанотехнологии в медицине". В то же время в результате повышения устойчивости возникает побочный эффект, связанный с попаданием в верхнюю часть списка редких словосочетаний, в том числе, комбинаций "Имя + фамилия". Также в полученном списке можно обнаружить общеязыковые устойчивые сочетания типа *губная помада*, релевантность которых для рассматриваемой предметной области вызывает сомнения.

Метрика *Weirdness* показывает почти идеальный результат на первых 10 терминах, однако обладает

недостатками, для демонстрации которых необходимо расширить окно внимания и обратиться к выгрузке первых 20 кандидатов (см. табл. 6).

Первые 10 отобранных словосочетаний в основном хорошо соответствуют поставленной задаче и действительно являются терминами предметной области. Однако качество извлечения на следующих 10 кандидатах страдает из-за отсутствия меры специфичности. Мы можем видеть, что попадание в верхнюю часть списка таких словосочетаний как *развитие нанотехнологий* или *индустрия наносистем* опирается исключительно на специфичность этих словосочетаний при том, что они являются достаточно свободными даже в рамках предметной области.

Комбинированная модель позволяет частично компенсировать недостаток специфичности и недостаток устойчивости, обнаруженные у последних двух методов.

Так, использование метода *PMI+Weirdness* (см. табл. 7) способствует увеличению доли специфичных словосочетаний в выдаче.

Словосочетания *автономный округ*, *губная помада*, *геотермальная ЭС* исчезли из списка 100 лучших кандидатов. В то же время использование этого метода не решило проблемы появления комбинаций "Имя + Фамилия", так как подобные словосочетания с формальной точки зрения действительно являются и специфичными, и устойчивыми. Фильтрация подобных словосочетаний может быть выполнена вручную или с использованием дополнительных словарных и орфографических фильтров.

Лучший суммарный результат на 100 кандидатах в ходе эксперимента продемонстрировал метод *Weirdness+PMI* (табл. 8).

Точность работы этого метода на первых 10 кандидатах сходна с точностью *PMI+Weirdness*, однако, при увеличении рангового порога этот метод показывает некоторый прирост качества. Полученные данные позволяют сделать предположение о сохранении этой тенденции при увеличении размера тестовых списков, однако, данная гипотеза нуждается в дополнительной проверке.

Таблица 4

Первые 10 результатов метода *Freq*

| bigram                            | Freq(x,y) | c(x) | c(y) |
|-----------------------------------|-----------|------|------|
| ОКРУЖАТЬ+ОКРУЖАЮЩИЙ#СРЕДА         | 1292      | 1452 | 2477 |
| РОССИЙСКИЙ#ФЕДЕРАЦИЯ              | 803       | 1965 | 848  |
| НАСТОЯЩЕЕ+НАСТОЯЩИЙ#ВРЕМЯ         | 776       | 1027 | 2284 |
| ТОТ+ТОМ+ТОМА#ЧИСЛО                | 752       | 1933 | 1430 |
| ЧРЕЗВЫЧАЙНЫЙ#СИТУАЦИЯ             | 682       | 708  | 1230 |
| <b>СТВОЛОВОЙ+СТВОЛОВЫЙ#КЛЕТКА</b> | 665       | 697  | 2254 |
| ИСТОЧНИК#ЭНЕРГИЯ                  | 548       | 1827 | 3491 |
| <b>КЛЕТОЧНЫЙ#ТЕХНОЛОГИЯ</b>       | 395       | 816  | 7068 |
| ВОЗОБНОВЛЯТЬ#ИСТОЧНИК             | 358       | 661  | 1827 |
| СИСТЕМА#УПРАВЛЕНИЕ                | 346       | 7249 | 1801 |

Первые 10 результатов метода PMI

| <b>bigram</b>                                         | <b>PMI</b>  |
|-------------------------------------------------------|-------------|
| АВТОНОМНЫЙ#ОКРУГ+ОКРУГА                               | 19.71286239 |
| <b>АКТИВИРОВАТЬ#ФОТОПРОТЕИН</b>                       | 19.49046997 |
| БЕРНХАРД#ВИНЗЕК                                       | 19.49046997 |
| БОГДАН+БОГДАНА#БУДЗУЛЯК                               | 19.49046997 |
| ГЕОТЕРМАЛЬНЫ#ЭС                                       | 19.49046997 |
| <b>ГЕТЕРИРОВАННЫЙ#НАНОПРИМЕСЬ</b>                     | 19.49046997 |
| <b>ГИПЕРВОЛЕМИЧЕСКИЙ+ГИПЕРВОЛЕМИЧЕСКА#ГЕМОДИЛЮЦИЯ</b> | 19.49046997 |
| ГРАНУЛОЦИТАРНЫЙ#КОЛОНИЕСТИМУЛИРУЮЩИЙ                  | 19.49046997 |
| ГУБНЫЙ+ГУБНОЙ#ПОМАДА                                  | 19.49046997 |
| ДЖОШ#ВОЛФ                                             | 19.49046997 |

Таблица 6

Первые 20 результатов метода Weirdness

| <b>bigram</b>                            | <b>Weirdness</b> |
|------------------------------------------|------------------|
| <b>КЛЕТОЧНЫЙ#ТЕХНОЛОГИЯ</b>              | 56300.95909      |
| <b>СИЛОВОЙ#ЭЛЕКТРОНИКА</b>               | 30787.35991      |
| <b>ЭНЕРГОСБЕРЕГАЮЩИЙ#ЛАМПА</b>           | 27936.67844      |
| <b>СТВОЛОВОЙ+СТВОЛОВЫЙ#КЛЕТКА</b>        | 18957.0318       |
| <b>БИОГАЗОВЫЙ#УСТАНОВКА</b>              | 14966.07773      |
| ИНДУСТРИЯ#НАНОСИСТЕМА                    | 13540.737        |
| <b>ПУПОВИННЫЙ#КРОВЬ</b>                  | 13113.13478      |
| <b>БИОМЕДИЦИНСКИЙ#ТЕХНОЛОГИЯ</b>         | 11830.32811      |
| ГЕОТЕРМАЛЬНЫЙ#ЭНЕРГИЯ                    | 11687.79404      |
| <b>ВОЗОБНОВЛЯТЬ#ЭНЕРГЕТИК+ЭНЕРГЕТИКА</b> | 10547.52145      |
| КОГНИТИВНАЯ+КОГНИТИВНЫЙ#ТЕХНОЛОГИЯ       | 9834.851082      |
| МИКРОСИСТЕМНЫЙ#ТЕХНИК+ТЕХНИКА            | 9407.248861      |
| ЛИКВИДАЦИЯ#ЧС                            | 8837.112566      |
| РАЗВИТИЕ#НАНОТЕХНОЛОГИЯ                  | 8266.976272      |
| <b>КОМПОНЕНТНЫЙ#БАЗ+БАЗА</b>             | 8160.075717      |
| <b>ШИРОКОПОЛОСНЫЙ#ДОСТУП+ДОСТУПА</b>     | 8124.442198      |
| ИСПОЛЬЗОВАНИЕ#ВИЭ                        | 7839.374051      |
| КРИТИЧЕСКИЙ#ВОЕННЫЙ                      | 7554.305904      |
| ПРОИЗВОДСТВО#БИОТОПЛИВО                  | 7554.305904      |
| СЕТЬ#ДОСТУП+ДОСТУПА                      | 7126.703683      |

Первые 10 результатов метода PMI+Weirdness

| <b>bigram</b>                            | <b>PMI</b> | <b>Weirdness</b> |
|------------------------------------------|------------|------------------|
| <b>СТВОЛОВОЙ+СТВОЛОВЫЙ#КЛЕТКА</b>        | 15.1444462 | 18957.0318       |
| <b>ВОЗОБНОВЛЯТЬ#ЭНЕРГЕТИК+ЭНЕРГЕТИКА</b> | 15.6137871 | 10547.5215       |
| <b>УПРЕЖДАТЬ+УПРЕЖДАЮЩИЙ#КЭШИРОВАНИЕ</b> | 15.0905394 | 3848.41999       |
| <b>МТЬ+М+М2М+МИЙ#ТЕЛЕМАТИКА</b>          | 15.5170118 | 3420.81777       |
| <b>АДРЕСНЫЙ#ДОСТАВКА+ДОСТАВОК</b>        | 16.2845657 | 3135.74962       |
| <b>ВЫЧИСЛИТЕЛЬНЫЙ#КЛАСТЕР</b>            | 15.3631511 | 2993.21555       |
| <b>АРТЕРИАЛЬНЫЙ#ГИПЕРТЕНЗИЯ</b>          | 15.2401716 | 2280.54518       |
| <b>ХАЙКО#ЮРГЕН</b>                       | 16.0754325 | 2138.0111        |
| <b>СУХОЙ#СЛАЙД</b>                       | 15.9879696 | 2138.0111        |
| <b>РЕГИСТРАЦИОННЫЙ#УДОСТОВЕРЕНИЕ</b>     | 16.1909097 | 1852.94296       |

Таблица 8

Первые 10 результатов метода Weirdness+PMI

| <b>bigram</b>                                   | <b>Weirdness</b> | <b>PMI</b>  |
|-------------------------------------------------|------------------|-------------|
| <b>ВСТРОИТЬ#ПУСКОРЕГУЛИРУЮЩИЙ</b>               | 1425.340737      | 18.22743556 |
| <b>УГЛЕРОДНЫЙ#НАНОТРУБ+НАНОТРУБА+НАНОТРУБЛЬ</b> | 997.7385156      | 18.21036205 |
| <b>РЕЗОНАНСНЫЙ#ТУННЕЛИРОВАНИЕ</b>               | 1140.272589      | 17.73558247 |
| <b>УГЛЕРОДНЫЙ#НАНОТРУБКА</b>                    | 1211.539626      | 17.65114266 |
| <b>ВОЗОБНОВЛЯТЬ#ИСТОЧНИК</b>                    | 1546.278738      | 17.13902483 |
| <b>ГАВРИЛОВ#Н</b>                               | 1425.340737      | 16.68311505 |
| <b>СЧИТЫВАТЬ#ГОЛОВКА</b>                        | 1567.87481       | 16.53204907 |
| <b>РОТАЦИОННЫЙ#ВЫТЯЖКА</b>                      | 1140.272589      | 16.36118695 |
| <b>АДРЕСНЫЙ#ДОСТАВКА+ДОСТАВОК</b>               | 3135.74962       | 16.28456567 |
| <b>ФЦ#ВНИИ</b>                                  | 1567.87481       | 16.24254245 |

## ЗАКЛЮЧЕНИЕ

В ходе исследования мы сравнили различные подходы к извлечению двусловных терминологических сочетаний из корпуса на примере текстовой коллекции для предметной области "Нанотехнологии в медицине". Результаты исследования позволяют сделать вывод о перспективности комбинированных подходов к извлечению терминов, при которых устойчивость и специфичность словосочетаний контролируются двумя независимыми показателями. В то же время пока не до конца ясно, следует ли, и если да, то каким образом следует комбинировать значения этих показателей для получения единой оценки. Одно из возможных решений заключается в использовании линейной комбинации значений метрик, при этом веса компонентов следовало бы определить на основании некоторой обучающей выборки. Подобная выборка может быть получена полуавтоматически с использованием методов, описанных в данной статье, и последующей экспертной корректировкой.

## СПИСОК ЛИТЕРАТУРЫ

1. Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008». – М., 2008.
2. Daille B. Combined approach for Terminology Extraction: Lexical statistics and linguistic filtering. – Universite Paris 7, 1994.
3. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двусловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». – М., 2006.
4. Лукашевич Н.В., Логачев Ю.М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. – 2010. – Т. 2.

5. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер. 2. – 1995. – № 3. – С. 21-24.
6. McKeown K.R., Radev D.R. Collocations / eds. R. Dale, H. Moisl, H. Somers // A Handbook of Natural Language Processing. – CRC Press, 1998.
7. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции" – 2003.
8. Тезаурус информационно-поисковый по законодательству. – М.: Министерство юстиции СССР, Научный центр правовой информации, 1980.
9. Manning C., Schütze H. Foundations of Statistical Natural Language Processing. – MIT Press, 1999.
10. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. – 1988. – Vol. 24. – Issue 5.
11. Spärck Jones K., Walker S., Robertson S.E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2) // Information Processing and Management. – 2000. – Vol. 36. – Issue 6.
12. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // Proceedings of TREC-8, 1999.
13. Браславский П., Соколов Е. Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2007». – М., 2007.
14. Drouin P. Term extraction using non-technical corpora as a point of leverage // Terminology. – 2003. – Vol. 9. – Issue 1.

*Материал поступил в редакцию 12.03.13.*

#### **Сведения об авторе**

**КУЗНЕЦОВ Илья Олегович** - аспирант факультета филологии Национального исследовательского университета – Высшая школа экономики (НИУ ВШЭ), младший научный сотрудник Центра семантических технологий НИУ ВШЭ, Москва  
e-mail: iokuznetsov@hse.ru

# Алгоритмы распознавания собственных имён

*Показывается влияние распознавания собственных имён на повышение эффективности функционирования систем автоматической обработки текста. Обсуждаются проблемы, возникающие в процессе разрешения анафоры. Предлагается позиционно-семантический подход к распознаванию собственных имён. Описываются алгоритмы распознавания антропонимов.*

**Ключевые слова:** собственные имена, антропонимы, алгоритмы распознавания, разрешение анафоры

Собственные имена включают лексические единицы, которые имеют конкретно-предметную референцию и выполняют функцию индивидуализации, позволяя уникально идентифицировать объект, выделив его из определённого класса [1, с. 98]. А.В. Суперанская [там же, с. 173-204] предлагает предметно-номинативную классификацию собственных имён, которая включает антропонимы, зоонимы, мифонимы, топонимы, космонимы и астронимы, фитонимы, хрематонимы, названия средств передвижения, сортовые и фирменные названия, названия предприятий и учреждений, названия органов периодической печати, хрононимы, названия праздников, названия мероприятий, войн, названия произведений литературы и искусства, названия стихийных бедствий, документонимы, фалеронимы. Исследование этих классов собственных имён имеет непосредственное значение как для лингвистики, так и для других предметных областей, например, истории и географии. Вместе с тем, не меньшее значение имеет изучение собственных имён и для разработки систем автоматического анализа текста.

Цель настоящей статьи – показать, как распознавание собственных имен влияет на эффективность функционирования систем автоматической обработки текста, а также предложить алгоритмы распознавания антропонимов.

Распознавание собственных имён лежит в основе разрешения анафоры [2] – направления автоматической обработки текста, в рамках которого разрабатываются алгоритмы замены анафорических местоимений предшествующими кореферентными именами объектов, а также замены кореферентных повторов стандартным термином. Алгоритмы разрешения анафоры имеют существенное значение для целого ряда направлений автоматической обработки текста. В системах автоматического реферирования и информационного поиска замена местоимений кореферентными именами позволяет произвести более адекватное взвешивание терминов и существенно повысить качество результата [3, 4]. Алгоритмы разрешения анафоры важны и для разработки вопросно-ответных систем [5], систем машинного перевода [6], систем интеллектуального анализа текста, а также и для систем искусственного интеллекта, поскольку

разрешение анафоры предполагает формулирование логического вывода в рассуждении [7].

Влияние разрешения анафоры на эффективность автоматической обработки текста можно показать, сравнив результаты автоматического реферирования оригинального газетного текста (см. **Текст 1** в *Приложении*) и реферирования того же текста, в котором вручную выполнены кореферентные замены (см. **Текст 2** в *Приложении*). Мы проводили реферирование с помощью системы *Copernic Summarizer*<sup>1</sup>; размер реферата определялся по умолчанию в 25% от оригинала.

Реферат оригинального текста.

*NEW YORK - Two-year-old Tangena Hussain vanished three months ago in the Detroit suburb of Hamtramck, Michigan, and police say the trail has gone cold.*

*Tangena's mother called police on the evening of October 2, 2008, to report her daughter missing after frantically searching the area where she was last seen.*

*The child's mother, Nilufa Begum, told police she had left the girl in the care of her boyfriend, Jamrul Hussain.*

*On the video, he makes some purchases and leaves within a few minutes.*

Реферат текста, в котором проведены кореферентные замены.

*NEW YORK - Two-year-old Tangena vanished three months ago in the Detroit suburb of Hamtramck, Michigan, and police say the trail has gone cold.*

*Tangena's mother, Begum, told police Begum had left Tangena in the care of her boyfriend, Hussain.*

*On the video, Hussain makes some purchases and leaves within a few minutes.*

*But there is no video of the area where Hussain parked Hussain's car, nor any video that could show whether Tangena had been there or how Tangena might have disappeared.*

В двух рефератах, состоящих из четырёх предложений, совпало два, при этом реферат текста, в котором было произведено разрешение анафоры, более связан и завершён. Последнее предложение называет

<sup>1</sup> <http://www.copernic.com/en/products/summarizer/index.html>

основное событие, описываемое в тексте (пропажа девочки), и подтверждает вывод, содержащийся в первом предложении текста (*the trail has gone cold*).

В табл. 1 показано, какие именно замены были проведены в процессе разрешения анафоры.

Таблица 1

### Кореферентные замены в тексте

| №/п | Термин           | Кореферентный термин |
|-----|------------------|----------------------|
| 1   | Tangena Hussain  | Tangena              |
| 2   | Tangena Hussain  | Tangena              |
| 3   | Tangena's mother | Begum                |
| 4   | her daughter     | Tangena              |
| 5   | She              | Tangena              |
| 6   | The child's      | Tangena's            |
| 7   | Nilufa Begum     | Begum                |
| 8   | She              | Begum                |
| 9   | the girl         | Tangena              |
| 10  | Jamrul Hussain   | Hussain              |
| 11  | He               | Hussain              |
| 12  | her              | Tangena's            |
| 13  | her daughter     | Tangena              |
| 14  | She              | Begum                |
| 15  | He               | Hussain              |
| 16  | They             | Tangena and Hussain  |
| 17  | He               | Hussain              |
| 18  | He               | Hussain              |
| 19  | He               | Hussain              |
| 20  | the little girl  | Tangena              |
| 21  | He               | Hussain              |
| 22  | She              | Begum                |
| 23  | Him              | Hussain              |
| 24  | She              | Tangena              |
| 25  | His              | Hussain's            |
| 26  | She              | Begum                |
| 27  | He               | Hussain              |
| 28  | He               | Hussain              |
| 29  | Her              | Begum                |
| 30  | her daughter     | Tangena              |
| 31  | He               | Hussain              |
| 32  | They             | Begum and Hussain    |
| 33  | her daughter     | Tangena              |
| 34  | He               | Hussain              |
| 35  | His              | Hussain's            |
| 36  | the little girl  | Tangena              |
| 37  | She              | Tangena              |
| 38  | the toddler's    | Tangena's            |

Анализ данных табл. 1 показывает, что разрешение анафоры – сложный процесс, включающий как местоименную, так и именную субституцию и предусматривающий решение ряда проблем:

1) разграничение мужских и женских имён для адекватной замены местоимений. Эта проблема особенно актуальна по отношению к английскому языку в силу неразвитости категории рода. Если в словарях русского языка для каждого существительного указывается его родовая категория, то в англоязычных словарях такая информация отсутст-

вует. Возникает необходимость составления специальных словарей, в которых существительным приписываются теги, указывающие на род, например *girl(f)*, *boy(m)*, *brother(m)*, *sister(f)* и т.д. Ещё одна проблема, связанная с категорией рода, – распознавание кореферентных существительных, которые могут относиться как к женскому, так и к мужскому роду, ср. *child*, *toddler*;

2) выбор стандартного имени для замены кореферентных имён. В **Тексте 1 Приложения** для обозначения одного лица используются имена: *Tangena Hussain*, *Tangena*, *girl*, *toddler*, *daughter*, *child*. В таком случае выбирается наиболее частотное имя. В данном тексте – *Tangena*, которое повторяется шесть раз. Стандартное имя может не предшествовать, а использоваться после первого упоминания объекта. Часто при первом упоминании одушевлённое лицо называется именем и фамилией, а затем только именем или только фамилией. Ср. в рассматриваемом тексте *Tangena Hussain* → *Tangena*; *Jamrul Hussain* → *Hussain*. Соответственно, речь должна идти не только об анафоре, но и о катафоре, что ещё более усложняет задачу кореферентной субституции;

3) неэквивалентность синтаксических конструкций. Стандартному имени, состоящему из одного слова, могут соответствовать словосочетания, состоящие из двух и более слов, см. пп. 1-4, 6, 9, 13, 20, 30, 33, 36, 38 в табл. 1. Поэтому возникает проблема определения границ синтаксической структуры заменяемых кореферентных имён. Возможна и противоположная ситуация: одно слово заменяется словосочетанием. Такая ситуация возникает, когда производится замена местоимений множественного числа, см. пп. 16, 32 в табл. 1. Данным местоимениям может соответствовать словосочетание из двух, трёх и более слов. По правилам английского языка их следует соединять сочинительным союзом. Возможны и случаи, когда одно местоимение требует замены не словосочетанием, а клаузой, предложением или группой предложений. В высказывании *It's difficult* местоимение отсылает к предыдущему событию, которое может обозначаться всеми указанными видами конструкций;

4) распознавание собственных имён. Виды имён, выделяемые в классификации А.В.Суперанской, неравномерно распределены по разным жанрово-стилевым типам текстов. В художественных текстах особое значение имеет распознавание антропонимов и топонимов, поскольку именно этими типами собственных имён обозначаются персонажи и места действия.

В разработанной нами программе ETS (Event Tracking Summarizer)<sup>2</sup>, специально оптимизированной для анализа художественного текста, реферирование выполняется на основе распознавания личных имён и географических названий. В проанализированном газетном тексте (см. *Приложение*) местоименная субституция выполнялась только для личных имён, хотя в газетных статьях другой тематики,

<sup>2</sup> ETS распространяется бесплатно на сайте <http://vetsky.narod2.ru/catalog/ets/>

например, посвящённых проблемам медицинского обслуживания, существенную роль могут играть и названия учреждений. В научных текстах описательного типа собственные имена могут встречаться редко, либо вообще не использоваться.

Алгоритмы, применяющиеся в существующем программном обеспечении разрешения анафоры, включают два основных этапа.

Вначале распознаются все именные группы (noun phrases – NP), которые используются в контексте, предшествующем местоимению. В качестве контекста могут выступать словосочетания в данном предложении/клаузе, предшествующем предложению/клаузе. Размер контекста определяется индивидуально в каждом проекте. В [8, с. 537] такой контекст составляют четыре предшествующих предложения. Распознавание NP проводится на основе предварительного аннотирования тегами частей речи и грамматик фразовой структуры, которые хорошо разработаны для английского языка [9].

Далее из этих NP выбирается наиболее вероятный антецедент текущего местоимения. Для этого по определённым правилам каждой NP начисляется весовой коэффициент, и в качестве антецедента выбирается NP с наиболее высоким коэффициентом [8, с. 558]. Такой статистический алгоритм является достаточно универсальным и не предусматривает специальной процедуры распознавания собственных имён, которые рассматриваются в качестве одного из возможных компонентов NP, обозначающегося специальным тегом (например, PNP).

Предлагаемый нами алгоритм не требует аннотирования тегами частей речи и включает следующие этапы.

1. Для входного текста составляется список всех токенов, которые пишутся с заглавной буквы. Из списка удаляются стоп-слова. Создаётся Список 1.

2. Список 1 сопоставляется со Словарём 1, который включает файлы со списками английских имён и фамилий. Эти списки были составлены нами на основе словарей, размещённых в Интернете. Список женских имён включает 1058, список мужских имён – 734 и список фамилий – 8420 единиц. Если токен находится в одном из списков, он помечается специальными тегами: отдельный тег – для имени, отдельный тег – для фамилии. Возможно и аннотирование тегами мужских и женских имён. Для списка фамилий учитывается окончание *s*, *es*. Если в тексте используется *Smiths*, то оно отождествляется с *Smith*, который есть в списке. В результате создаётся Список 2, включающий имена и фамилии, помеченные соответствующими тегами. Токены, которые не вошли в Список 2, отфильтровываются в отдельный список, создаётся Список 3.

Разграничение имён и фамилий имеет существенное значение для анализа художественного текста, персонажами которого могут выступать однофамильцы, члены одной семьи, например *John Forsyte*, *Soams Forsyte*, *June Forsyte*. Распознать эти сочетания как имена разных объектов можно именно с помощью разграничения имён и фамилий.

Заметим, что ни один список имён и фамилий не будет достаточно полным, тем более что могут использоваться имена с разной этимологией, что хорошо демонстрирует текст из *Приложения*. Поэтому следует применять дополнительные правила для распознавания антропонимов.

3. В Списке 3 ищутся токены, которые занимают позицию непосредственно после токена, помеченного тегом имени или непосредственно перед токеном, помеченным тегом фамилии. Если они находятся, то в первом случае помечаются тегом фамилии, во втором – тегом имени и включаются в Список 2, исключаясь из Списка 3. Например, *Alec Sidorov* – *Alec* есть в списке имён, и токен *Sidorov*, который отсутствует в списке имён и фамилий, распознаётся как фамилия, поскольку занимает позицию непосредственно после *Alec*.

В Списке 3 остаются токены, не распознанные как имена и фамилии.

4. Список 3 сопоставляется со Словарём 2 (табл. 2). Если перед токеном из Списка 3 используется слово из Словаря 2, то он распознаётся как фамилия и включается в Список 2. Регистр слов в словаре не учитывается. По этому правилу *Ivanov* в словосочетании *Mr Ivanov* или *mr Ivanov* будет распознано как фамилия, поскольку этот токен занимает позицию после слова из словаря. Если в этой позиции используются два слова из Списка 3, то первое слово распознаётся как имя, а второе – как фамилия. Если используется больше двух слов, то учитываются только первое (как имя) и последнее (как фамилия). Например, *Mr Ivan Ivanovich Ivanov* – *Ivan* будет распознано как имя, *Ivanov* – как фамилия, *Ivanovich* – проигнорировано. Токену *Ivanov* будет приписан тег фамилии, токену *Ivan* – тэг имени, и они будут удалены из Списка 3 и включены в Список 2. Отчества и средние имена игнорируются, поскольку они обычно используются только при первом упоминании лица или не используются вообще.

Словарь 2 включает слова и сокращения, обозначающие должность, звание, титул, и встречающиеся в позиции перед личным именем.

5. Список 3 сопоставляется со Словарём 3. Если после токена из Списка 3 используется слово из Словаря 3, то он распознаётся как фамилия и включается в Список 2. Регистр слов в словаре не учитывается. Например, во фразе *Smith, PhD* токен *Smith* будет распознан как фамилия, поскольку после него используется слово из Словаря 3. Если в этой позиции используются два слова, которые пишутся с большой буквы, то первое слово распознаётся как имя, а второе – как фамилия. Если используется больше двух слов, то учитываются только первое (как имя) и последнее (как фамилия). Например, во фразе *John Wilbur Smith, PhD* *John* будет распознано как имя, *Smith* – как фамилия, *Wilbur* – проигнорировано. Токену *Smith* будет приписан тег фамилии, токену *John* – тэг имени, и они будут удалены из Списка 3 и включены в Список 2.

Словарь 3 включает слова и сокращения, обозначающие должность, звание, титул, и встречающиеся в позиции после личного имени. Некоторые слова из Словаря 3 совпадают со словами из Словаря 2, поскольку могут использоваться как перед, так и после

фамилии. Ср. *O'Brien, Professor of Theology* и *Professor King*.

6. Список 3 сопоставляется со Словарём 4. Если после токена или перед токеном из Списка 3 используется слово из Словаря 4, то он распознаётся как имя и включается в Список 2. Например, "*Where the good-looking young man was,*" *said Sister. Sister* распознаётся как имя, потому что занимает позицию после слова *said*, которое есть в Словаре 4.

Словарь 4 включает глаголы речевого действия, ментальные и другие глаголы, которые обозначают действия и состояния, характерные для одушевлённых лиц.

Применение этого алгоритма можно проиллюстрировать на анализе следующего текста<sup>3</sup>.

*Mr. Blingham, and may he fry in his own cooking-oil, was assistant treasurer of the Flaver-Saver Company. He was driving from New York to Winnipeg, accompanied by Mrs. Blingham and their horrible daughter. When they came to their first prairie hamlet in Minnesota, six cottages, a garage, a store and a tall red grain elevator, Mrs. Blingham giggled, "Why, they've got an Empire State Building here!"*

*"And all the Svensons and Bensons and Hensons go up to the Rainbow Room every evening!" gurgled Sister.*

Вначале выписываются все слова, начинающиеся с большой буквы.

*Blingham, Flaver-Saver, Company, He, New, York, Winnipeg, Blingham, Minnesota, Blingham, Empire, State, Building, And, Svensons, Bensons, Hensons, Rainbow, Room, Sister.*

Из списка удаляются стоп-слова *And, He*.

Первое слово в списке *Blingham* сопоставляется со словарём имён и фамилий (Словарь 1). Там его нет. Смотрится, используется ли перед ним слово из списка 1. Такое слово находится – *Mr*. Слову *Blingham* приписывается тег фамилии. Второе слово в списке *Flaver-Saver* сопоставляется со списками имён и фамилий. Там его нет. Смотрится, используется ли перед ним слово из списка 1. Нет. Смотрится, используется ли после него слово из списка 2. Нет. Смотрится, используется ли перед ним слово из списка 3. Нет. Значит, это слово игнорируется.

Также игнорируются *Company, New, York, Winnipeg, Minnesota, Empire, State, Building, , Rainbow, Room*.

*Svensons, Bensons, Hensons* есть в списке фамилий, только без окончания *s*. Им приписывается тег фамилий. *Sister* получает тег имени, так как перед ним используется слово из списка 3 (*gurgled*).

Таким образом разграничиваются антропонимы и другие собственные имена (географические названия, названия компаний и т.п.).

Таблица 2

Словари для распознавания антропонимов (выдержка)

| Словарь 1     |               |          | Словарь 2                                                                                                                                                                                                                                                                                                                                                                                                                                              | Словарь 3                                                                                                                                                                                                             | Словарь 4                                                                                                                                                                                                                                                                     |
|---------------|---------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Женские имена | Мужские имена | Фамилии  |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Aaleigha      | Aaron         | Smith    | Mister, Mr, Mr., Miss, Ms.,<br>Mistress, Mrs, Mrs., Ms,<br>President, Prime-Minister,<br>Prime minister, P.M.,<br>Sgt., Sergeant, Lt.,<br>Lieutenant, Captain, Capt,<br>Capt., Major, Colonel, Col,<br>Col., General, Major-general,<br>Lieutenant-general,<br>C. J., Chief Justice,<br>C.-in-C., C-in-C,<br>Commander in Chief,<br>Comdr., Commander,<br>D. A., District Attorney,<br>Pte., Private, Detective,<br>Hon., Honorable, Rev.,<br>Reverend | Dr, Dr.<br>Doctor,<br>Professor,<br>Prof., Prof,<br>PhD, Ph.D.,<br>Senior, Sr.,<br>Junior, Jr<br>Abr.,<br>archbishop,<br>Asst.,<br>Assistant ,<br>PA, P.A., B.,<br>B,<br>B.A. , B.<br>M., B. Sc.,<br>Bt., Bt,<br>M.P. | Say, Says, Said,<br>Pronounce,<br>Pronounced,<br>Pronounces,<br>Exclaim, Exclaimed,<br>Exclaims,<br>Answer, Answered,<br>Answers,<br>Suppose, Supposed,<br>Supposes,<br>Conjecture,<br>Conjectures,<br>Conjectured,<br>Think, Thought,<br>Thinks<br>Snarl, Snarled,<br>Snarls |
| Aaryanna      | Abdul         | Jones    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Abbrielle     | Abel          | Taylor   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Abia          | Abner         | Williams |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Abida         | Abraham       | Brown    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Abigail       | Absalom       | Davies   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Abira         | Adam          | Evans    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Acacia        | Adair         | Wilson   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Ada           | Addison       | Thomas   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adah          | Adin          | Roberts  |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adelaide      | Adrian        | Johnson  |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adeline       | Aidan         | Lewis    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adin          | Akira         | Walker   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adina         | Alan          | Robinson |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Adriana       | Alex          | Wood     |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Agatha        | Alasdair      | Thompson |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Agnes         | Alastair      | White    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Ai            | Albert        | Watson   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Aile          | Alden         | Jackson  |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Ailsa         | Aled          | Wright   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |
| Aimee         | Alexander     | Green    |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                               |

<sup>3</sup> Отрывок из Kingsblood Royal, <http://gutenberg.net.au/ebooks02/0200171.txt>

Предложенный нами алгоритм распознавания антропонимов можно назвать позиционно-семантическим, поскольку он учитывает семантику слов, находящихся в определённой позиции по отношению к распознаваемому имени. В его основе лежит положение о том, что антропонимы можно распознать по сочетаемости с терминами, обозначающими состояния и действия, специфичные для одушевлённых объектов. Глаголы речи, ментальные глаголы, глаголы, обозначающие чувства, могут сочетаться только с антропонимами и не сочетаются с другими собственными именами.

Предлагаемый алгоритм не требует применения аннотирования тегами частей речи и грамматики для распознавания именных словосочетаний. Это позволяет использовать его в системах автоматической обработки текста, в которых обычно не предусматривается использование грамматик, в частности, в ИПС и системах реферирования. Для повышения быстродействия и выполнения алгоритма без применения морфологического анализа в словарях указываются все возможные словоформы лексических единиц. Распознавание антропонимов выполняется на этапе предварительной обработки текста [10] и позволяет повысить эффективность функционирования этих систем.

Особенность предложенного алгоритма состоит в последовательной ступенчатой фильтрации предварительной выборки токенов, начинающихся с большой буквы, на основе сопоставления с четырьмя типами словарей. Это позволяет достичь высокой точности распознавания антропонимов, которая, как показало предварительное тестирование, достигает 98 % в художественном тексте и 100% – в научном.

Очевидно, что позиционно-семантический подход может быть применён и для распознавания других типов имён собственных. В частности, географические названия могут распознаваться по такому контексту, как *is located/situated, place, area*; названия газет журналов – по контексту *print, monthly, publish*; названия организаций – по контексту *office, agency, organ* и т.д. Разработка алгоритмов распознавания этих видов собственных имён – задача последующих исследований.

## СПИСОК ЛИТЕРАТУРЫ

1. Суперанская А. В. Общая теория имени собственного. – М.: Наука, 1973. – 367 с.
2. Stuckardt R. An interdependency-sensitive approach to anaphor resolution // Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC). – Lancaster, 1996. – P. 400–413.
3. Baldwin B., Thomas S. M. Dynamic coreference-based summarization // Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP-3). – Granada, 1998. – P.1–6.
4. Kameyama M. Recognizing referential links: an information extraction perspective // Proceedings of the ACL'97/EACL'97 workshop on "Operational factors in practical, robust anaphora resolution". – Madrid, 1997. – P. 46–53.
5. Bobrow D. G. A question-answering system for high school algebra word problems // Proceedings of AFIPS conference. Part I. – New York, 1964. – P. 591–614.
6. Saggion H., Carvalho A. Anaphora resolution in a machine translation system // Proceedings of the International Conference "Machine translation, 10 years on". – Cranfield, 1994. – P. 1-14.
7. Cohen A., Kaminski M., Makowsky J. A. Notions of sameness by default and their application to anaphora, vagueness, and uncertain reasoning // Journal of logic, language and information. – 2008. – Vol. 17. – Issue 3. – P. 285-306.
8. Lappin S., Leass H. J. An algorithm for pronominal anaphora resolution // Computational linguistics. – 1994. – Vol. 20. – № 4. – P.535-561.
9. Börjars K., Burridge K. Introducing English Grammar. – London: Arnold, 2001. – 311 p.
10. Яцко В.А. Методы и алгоритмы автоматического анализа текста // НТИ. Сер. 2. – 2011. - № 9. – С. 12-19; Yatsko V. A. Methods and algorithms for automatic text analysis // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, № 5. – P. 224-231.

## ПРИЛОЖЕНИЕ

### Текст 1. Оригинальный газетный текст<sup>4</sup>

NEW YORK - Two-year-old Tangena Hussain vanished three months ago in the Detroit suburb of Hamtramck, Michigan, and police say the trail has gone cold.

Tangena Hussain, 2, has been missing since October 2 and police in Michigan are asking for the public's help.

Tangena's mother called police on the evening of October 2, 2008, to report her daughter missing after frantically searching the area where she was last seen.

The child's mother, Nilufa Begum, told police she had left the girl in the care of her boyfriend, Jamrul Hussain. Although Tangena and Hussain have the same last name, he is not her father.

Begum told investigators her daughter was with Hussain all day while she worked at the Northland Mall. Hussain said he stopped at a gas station with Tangena while on the way to the mall to pick up Begum.

They stopped at about 9 p.m. to buy gum and juice, he said. Hussain told police he left Tangena in the car for only a few moments and when he returned, the little girl was gone.

He did not call police and did not conduct a search, but instead drove to the mall and picked up Begum, investigators said.

<sup>4</sup> Взято с сайта <http://edition.cnn.com/2009/CRIME/01/13/grace.coldcase.hussain/index.html>; приводится в сокращении

When she saw Tangena was not with Hussain, Begum asked him where she was. His response was cryptic, she told police. He said he would take her to the place where her daughter was.

Begum became alarmed as Hussain drove to a gas station where, he said, Tangena disappeared. They searched for Tangena together, while Begum called the police to report her daughter missing.

A surveillance camera shows the boyfriend, Hussain, going into the gas station's store, police said. On the video, he makes some purchases and leaves within a few minutes. But there is no video of the area where Hussain parked his car, nor any video that could show whether the little girl had been there or how she might have disappeared.

Police have not named a person of interest in the toddler's disappearance.

**Текст 2.** Текст с заменёнными анафорическими терминами

NEW YORK - Two-year-old Tangena vanished three months ago in the Detroit suburb of Hamtramck, Michigan, and police say the trail has gone cold.

Tangena, 2, has been missing since October 2 and police in Michigan are asking for the public's help.

Begum called police on the evening of October 2, 2008, to report Tangena missing after frantically searching the area where Tangena was last seen.

Tangena's mother, Begum, told police Begum had left Tangena in the care of her boyfriend, Hussain. Although Tangena and Hussain have the same last name, Hussain is not Tangena's father. Begum told investigators Tangena was with Hussain all day while Begum worked at the Northland Mall. Hussain said Hussain stopped at a gas station with Tangena while on the way to the mall to pick up Begum.

Tangena and Hussain stopped at about 9 p.m. to buy gum and juice, Hussain said. Hussain told police Hussain left Tangena in the car for only a few moments and when Hussain returned, Tangena was gone.

Hussain did not call police and did not conduct a search, but instead drove to the mall and picked up Begum, investigators said.

When Begum saw Tangena was not with Hussain, Begum asked Hussain where Tangena was. Hussain's response was cryptic, Begum told police. Hussain said Hussain would take Begum to the place where Tangena was.

Begum became alarmed as Hussain drove to a gas station where, Hussain said, Tangena disappeared. Begum and Hussain searched for Tangena together, while Begum called the police to report Tangena missing.

A surveillance camera shows the boyfriend, Hussain, going into the gas station's store, police said. On the video, Hussain makes some purchases and leaves within a few minutes. But there is no video of the area where Hussain parked Hussain's car, nor any video that could show whether Tangena had been there or how Tangena might have disappeared.

Police have not named a person of interest in Tangena's disappearance.

*Материал поступил в редакцию 25.02.13.*

#### **Сведения об авторе**

**ЯЦКО Вячеслав Александрович** – доктор филологических наук, профессор Хакасского государственного университета им. Н.Ф.Катанова, г. Абакан  
e-mail: viatcheslav-yatsko@rambler.ru

## БАЗА ДАННЫХ ВИНИТИ РАН

ВИНИТИ предлагает к использованию через WWW-сервер (<http://www.viniti.ru>) крупнейшую Федеральную базу отечественных и зарубежных публикаций по естественным, точным и техническим наукам. БД ВИНТИ РАН генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. БД ВИНТИ представлена ретроспективными тематическими фрагментами и единой политематической БД (ретроспектива с 2001 г.), объединяющей все тематические фрагменты БД ВИНТИ.

### БД ВИНТИ РАН в сети INTERNET

**Сервер ВИНТИ** – <http://www.viniti.ru> – обеспечивает on-line доступ к Базе данных ВИНТИ РАН круглосуточно без выходных.

На основе БД ВИНТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации в **режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты ВИНТИ** выполняют тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

### БД ВИНТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНТИ или их разделов могут быть предоставлены на **CD-ROM в поисковой системе (ИПС) "Сокол"**, обеспечивающей все поисковые функции, доступные в режиме on-line:

- Поиск можно вести в годовом или ретроспективном массиве (за несколько лет сразу) в одном или нескольких тематических фрагментах .
- Поиск по словам и любым словосочетаниям из заглавия, реферата, ключевых слов.
- Использование года, языка, рубрик, шифров тематических разделов БД для уточнения поиска.
- Поиск по словарю, выполняющему функции многоаспектного указателя, в том числе авторского, предметного, источников, индексов МПК, номеров патентных документов и депонированных рукописей и т.д.
- Возможность запоминания запросов для последующего использования и/или редактирования их.
- Чтение документов не только как в РЖ (последовательный просмотр документов одного номера за другим), но и чтение документов нужных тематических фрагментов (разделов) по оглавлению за весь период заказанной ретроспективы.

ИПС "Сокол" является прикладной программой Microsoft Windows.

Любые наборы тематических фрагментов БД ВИНТИ или их разделов могут быть подготовлены в **коммуникативных форматах** ISO-2709, МЕКОФ, txt на любых видах электронных носителей.

Продукты предоставляются на договорной основе.

**Информационная служба БД ВИНТИ:** 125190, Москва, ул. Усиевича 20, ВИНТИ

**Телефон:** (499) 155-45-01, 155-45-02, **Факс:** (499) 152-62-31 **e-mail:** csbd@viniti.ru