РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 2 Москва 2014

ОБЩИЙ РАЗДЕЛ

УДК 81’42 – 027.22

С.В.Лихачев

Поле утилитарного дискурса

Утилитарный дискурс представляет собой поле общественных текстов на указателях, ценниках, табличках, не предназначенных для рекламы, украшения или самовыражения. Такие тексты образуют информационную среду, позволяющую человеку ориентироваться в мире вещей. Вопрос о границе между утилитарным дискурсом и рекламой, граффити, орнаментальными текстами решается с помощью исследования утилитарных функций: называния объекта, отсылки к объекту, пояснения объекта, а также лингвистических и экстралингвистических особенностей текста. В результате выявлены тексты ядра утилитарного дискурса, а также трех зон пересечения: с рекламным дискурсом, граффити и орнаментальными текстами.

Ключевые слова: *дискурс, текст, утилитарный, реклама, орнаментальный, информация*

Утилитарный дискурс представляет собой управляющее повседневными действиями человека для решения его собственных задач поле взаимосвязанных, позиционируемых на неоплачиваемом пространстве: указателях, ценниках, табличках и т.д. – текстов, в которых письменная часть функционально и семантически взаимодействует с расположенным в актуальной доступности предметом антропогенного происхождения (артефактом). Утилитарный дискурс – явление столь же необходимое, как и данные компьютерных сетей, систем глобального позиционирования и других средств информационного освоения пространства. Утилитарный дискурс информирует об артефактах среды, созданной человеком – антропогенном пространстве.

Утилитарный дискурс, в отличие от телепередачи или интернет-ресурса, может существовать при весьма ограниченных технических возможностях, поэтому появился раньше, и можно считать, что утилитарный дискурс был предшественником электронных и иных

современных баз данных с тем отличием, что информация в нем погружена в культурное пространство.

В современной лингвистике изучение повседневных утилитарных надписей на вывесках перестало быть необычным. К наиболее ранним публикациям относится материал Б.З. Букчиной и Г.А. Золотовой «Слово на вывеске» [1, с. 49-56], где внимание уделено соблюдению речевых норм. Масштабное изучение «письменности городской среды» более двадцати лет ведет Т.В. Шмелева [2]. На смежном с утилитарным дискурсом материале построены работы по эргонимам [3, 4]. В некоторых работах утилитарный дискурс изучается без использования специального термина, под названием «письменных жанровых форм, появившихся на улицах города в последнее время. Это городские плакаты нерекламного характера и репликинаклейки» [5, с. 257]. Однако не всегда понятно, где проходит граница между собственно утилитарными, рекламными и орнаментальными текстами надписей. Цель нашей статьи – провести эту границу.

Поле утилитарного дискурса состоит из текстов, проявляющих утилитарность в большей или меньшей степени, причем место утилитарного текста в организации поля определяется выполняемыми функциями.

Утилитарные тексты (общественно полезные надписи на предметах) могут выполнять различные функции:

- собственно утилитарные: сообщать о предмете, указывать направление, инструктировать о действиях с предметом;
- рекламные: привлечение непроизвольного внимания к объекту, убеждение в привлекательности объекта, мотивация использовать объект;
- функцию самовыражения автора, включая реализацию его творческих идей и возможностей, самохарактеристику и самооценку;
- орнаментальную, т. е. служить украшением, например архитектурных сооружений и предметов интерьера.

К ядру утилитарного дискурса относятся тексты, свободные от решения любых задач, кроме выполнения трех внутренних утилитарных функций: 1) сообщать о наличии и сути предмета (артефакта), 2) указывать на путь перехода между объектами и уровнями антропогенного пространства и 3) инструктировать о действиях с артефактом. Соответственно, в ядре утилитарного дискурса оказываются, во-первых, официальные таблички на зданиях с названиями организаций, в основном государственных, подобные же таблички на дверях с названиями подразделений организаций, надписи на папках с документацией и других артефактах официального обихода; во-вторых, указатели, в особенности, автодорожные, и иные транспортные тексты, т. е. свободные от эмоциональности и образности, например, в метрополитене; в-третьих, тексты, инструктирующие о мерах безопасности, а также о правилах поведения, например, «вытирайте ноги» на коврике у двери.

Тексты ядра утилитарного дискурса не выполняют рекламных задач (в смысле внушения адресату воли автора), не нарушают общественного порядка, а напротив, обеспечивают его, и не служат украшением. Таким образом, они не пересекаются с рекламным дискурсом, граффити или художественными текстами

Остальные три зоны утилитарного дискурса диффузны и многослойны, то есть находятся в состоянии пересечения с другими дискурсами. К ним относятся тексты, совмещающие: а) утилитарную и рекламную функции; б) утилитарную, рекламную функции и функцию самовыражения; в) все вышеперечисленные функции и художественную. Рассмотрим их в ходе последовательного анализа.

Тексты, совмещающие утилитарную и рекламную функции. В первой (околоядерной) зоне находятся утилитарные, т. е. связанные с артефактом, тексты, которые наряду с утилитарными функциями реализуют функцию привлечения внимания (рекламную). Среди них:

вывески торговых, обслуживающих и иных предприятий и учреждений,

тексты на упаковке товара, самом товаре, бирках, ярлыках, ценниках, бейджах, этикетках, прилавках и любые иные утилитарные тексты в торговле и ином бизнесе,

тексты на табличках экспонатов музеев, галерей, выставок,

этикетные фразы, которые становятся утилитарными текстами, когда позиционируются в соответствующем фрагменте антропогенного пространства, например, «Приятного аппетита» – в столовой,

временные отказы в чем-либо, указания на отсутствие чего-нибудь или кого-нибудь и административные извинения за неудобства,

инструкции по использованию объекта, например, «Нажмите. Говорите», тексты кнопок управления технических устройств.

Все эти тексты находятся в отношениях пересечения с рекламным дискурсом, но соотношение рекламного и утилитарного начала в них неравное. Например, официальные извинения косвенно рекламируют администрацию, но в большей степени в них проявлено именно утилитарное сообщение о временности неудобств и сути проводимых работ. В зоне пересечения утилитарного и рекламного дискурса становятся возможными эмоциональность, выразительность и нестандартность текстов.

В связи с полифункциональностью утилитарных текстов возникает вопрос о границах утилитарного дискурса, определение которого затруднено его взаимопроникновением (диффузией) с иными дискурсами, чаще всего с рекламным. Значит – перед нами стоит задача найти не одну границу, а как минимум две: между областью собственно утилитарного дискурса и областью взаимопроникновения, а также между областью взаимопроникновения и областью собственно рекламного дискурса, не имеющего утилитарности, причем, последний к объекту нашего исследования уже отношения не имеет. Кроме того, можно попытаться найти относительно обособленные области большей или меньшей утилитарности и их границы внутри диффузной зоны.

Для определения искомых границ необходимо найти отличительные признаки зон утилитарного дискурса и диффузных зон. Такие признаки могут

быть как лингвистическими, так и экстралингвистическими. К лингвистическим относятся наличие в тексте эргонимов, названий товарных знаков, а, кроме того, оценочных слов, т. е. таких лексем, дескриптивное содержание которых варьируется в зависимости от выбора оцениваемого объекта, например, *хороший автомобиль* и *хорошая мебель*. К внелингвистическим признакам принадлежат: способ оформления текста, удаленность текста от артефакта, наличие коммерческой составляющей в отношениях отправителя и изготовителя текста, а также наличие интенции управлять действиями адресата текста помимо его собственных планов и замыслов.

Нет причин считать фактом утилитарного или рекламного дискурса некоторое речевое произведение на основании одного только функционального критерия. С одной стороны, рекламная функция – пробудить интерес потребителя к продукту и пробудить в потребителе желание воспользоваться продуктом реализуется в многочисленных речевых произведениях, *a priori* не относящихся к рекламному дискурсу, например, в речи мамы, кормящей ребенка кашей «Кашка сладенькая, кашка мягонькая, скушай ложечку за маму». С другой стороны, реклама, стилизованная под дорожный указатель и присоединенная к нему, несмотря на наличие утилитарности, не относится к ядру утилитарного дискурса, например, как в выделенной курсивом части дорожной надписи в Петербурге: «Богатырский проспект / Коломяжский проспект / Б. Самсониевский проспект/ *Капитолий / Торгово-развлекательный центр* →». Основанием разграничения здесь будет строгое экономическое определение рекламы: «Реклама – это оплаченная, неперсонализированная коммуникация, осуществляемая идентифицированным спонсором и использующая средства массовой информации с целью склонить (к чему-то) или повлиять (как-то) на аудиторию» [6, с. 32], т. е. к ядру утилитарного дискурса не относится оплаченное с целью манипуляции адресатом распространение информации.

Действительно, утилитарный дискурс ведет пользователя к результату по кратчайшему пути. Вопреки тому выделенный особым цветом оплаченный рекламный указатель «*Капитолий / Торгово-развлекательный центр* →» уводит человека от намеченного пути, если ближе есть другой подходящий адресату подобный объект.

Итак, определен первый признак разграничения утилитарного и рекламного сообщения. Все части утилитарного дискурса взаимосвязаны и соразмерены с иерархическим устройством антропогенного пространства. Например, на первом этаже трехэтажного супермаркета не появится текст «На третьем этаже скользкий пол!», особенно, когда он скользкий и на втором. Все предупреждения будут даваться последовательно, по мере продвижения посетителя. Если же на первом этаже будет висеть неоплаченное объявление «Мужские костюмы – 3 этаж», при наличии торговли костюмами как на втором, так и на третьем этаже, – это будет зона пересечения рекламного и утилитарного дискурса с преобладанием рекламной (манипулятивной) функции. При наличии указания «Мужские костюмы – 2 и 3 этаж» наблюдается зона пересечения рекламного и утилитарного дискурса с преобладанием утилитарной (ориентирующей) функции. При наличии указания «Мужские костюмы «Сударь» 2 этаж / Мужские костюмы «Большевичка» 3 этаж» - опять же обнаруживается зона пересечения, где утилитарная и рекламная функции представлены раздельно, что потребовало затратить больше ресурсов на изготовление указателя.

В зоне пересечения рекламного и утилитарного дискурса оказываются вывески и надписи на упаковках. За их размещение обычно не надо платить собственнику места расположения текста – следовательно, они не относятся к рекламе в узком экономическом понимании [6, с. 32] и не являются безраздельно рекламным дискурсом. При этом, в отличие от указателей, критерии соответствия устройству антропогенного пространства здесь неприменимы – и надпись на упаковке, и вывеска расположены в непосредственной близости со своим артефактом (предметом), однако полностью утилитарными не являются, так как привлекают внимание к продукту и стимулируют его потребление, т. е. выполняют рекламную функцию. Чтобы разграничить степень утилитарности и рекламности придется учитывать наличие эргонимов, фирменных торговых названий товаров, оценочных слов, цвета размера и оформления шрифта.

Тексты на упаковке, очевидно, выполняют рекламную функцию, но в разной мере: «Молоко» и «Открывать здесь» рекламируют товар лишь в том смысле, что своим наличием указывают на заботу производителя об удобстве потребителя, т. е. являются скорее утилитарными. Тексты типа: «Состав: «Молоко коровье пастеризованное» и «ОАО “Молочный комбинат “Воронежский”, г. Воронеж» хотя и появились по требованию надзорных органов, а не по желанию маркетологов, выполняют в равной мере утилитарную и рекламную функции, так как удостоверяют качество и происхождение продукта, привлекая потребителя. В данном случае можно констатировать, что детализация информации усиливает рекламную составляющую интенционально утилитарного текста. Тексты «Вкуснотеево» и «Настоящее всегда вкуснее» содержат название торговой марки и оценочные слова *настоящее*, *вкуснее* – что говорит о преобладании в них рекламной (манипулятивной) функции, но некоторую утилитарную информацию о товаре содержат. То же касается и вывесок с эргонимами и оценочными словами, в отличие от типовых утилитарных вывесок «Магазин продукты» или «Аптека».

Избыточную для стопроцентного молока информацию «Не содержит ГМО» и «Для питания детей от 3-х лет» (от трех лет законодательных ограничений для продуктов питания нет) следует также считать в большей мере принадлежащей рекламному дискурсу и в меньшей мере – утилитарному. Утилитарность состоит в том, что информацию о ГМО получит среди прочих человек, не осведомленный об их заведомом отсутствии в натуральном молоке. Из второго текста можно эксплицировать тот факт, что нельзя использовать данное молоко для детей до трех лет.

Итак, для первой диффузной зоны утилитарного дискурса можно констатировать, что в нее не входят тексты рекламы в узком специальном (экономическом) понимании [6, с. 32]. Собственно первая диффузная зона утилитарного дискурса, включающая маркетинговые, этикетные и технические инструктирующие тексты, делится на три части:

- в первой – преобладает утилитарность, сюда относятся типовые вывески, безоценочные сообщения о содержимом упаковки, технические инструкции и неоплаченные как реклама торговые указатели, свободные от эргонимов и не отсылающие к дальнейшему объекту, минуя ближайший;
- во второй – утилитарность и рекламная функция уравновешены. Сюда относятся сообщения о происхождении, составе, дате выпуска товара, технические характеристики устройства, извинения, написанные в связи с артефактом, а также цена на ценнике, фамилия и имя на бейдже, информация на табличках о непродаваемых объектах – музейных, архитектурных и т.д.;
- в третьей – утилитарность представлена ограниченно, поскольку преобладает рекламная функция. Сюда относятся бесплатно размещаемые указатели, вывески и надписи на упаковке, содержащие эргонимы, названия торговых марок и оценочные слова, указатели, отсылающие к дальнейшему объекту, минуя ближайший, слоганы на упаковке, а также этикетные пожелания, связанные с предметным окружением.

Тексты, совмещающие утилитарную и рекламную функции и функцию самовыражения. Во второй, далее отстоящей от центра, диффузной зоне находятся утилитарные тексты, выполняющие, наряду с функцией привлечения внимания (рекламной) и одной из трех собственно утилитарных функций, еще одну функцию – средства самовыражения автора. Объектом этих текстов становится человек, позиционирующий сообщения в своем личном пространстве и от своего лица. Условно подобные тексты можно разделить на «чужие», т. е. рекламирующие не только носителя, например, названия фирмы на одежде, и «свои», отправляемые от лица носителя, что часто маркировано грамматически первым лицом. К последним относятся тексты прямо или косвенно характеризующие владельца на их личных вещах: на футболках, бейсболках, пляжных полотенцах, сумках, копилках, чашках и кружках, тапочках и т.д. Поскольку тексты данной зоны характеризуются эгоцентризмом и нередко – декларативным нонконформизмом, их можно сблизить с граффити.

Следует отметить, что утилитарные тексты второй зоны выполняют помимо утилитарной функции функцию привлечения внимания (рекламную). Следовательно, в рамках нашего анализа необходимо найти признаки, отличающие их от собственно рекламных текстов, с одной стороны, и от граффити – с другой. Если носитель текста на личных вещах никак не отождествлен с отправителем, то текст не утилитарный, а рекламный. В уникальных случаях в ограниченных масштабах рекламные тексты отправителя-рекламодателя, «говорящие» от первого лица (т. е. формально связанные с носителем и имитирующие утилитарность) пытались распространять, например, в Новосибирске владельцы книжного магазина, осуществляющего распродажу с 50%-ной скидкой, обязывали персонал носить футболки с текстом на спине «Отдаемся за полцены». Однако сотрудники обратились к юристам, и те констатировали нарушение прав трудящихся.

Итак, не будет утилитарным текст на личных вещах, если носитель никак не связывает его с собой, а данную вещь носит за деньги, например, в качестве корпоративной униформы. Такой текст будет собственно рекламным.

Чтобы побудить частное лицо демонстрировать на своей одежде и других вещах текст, направленный на привлечение внимания к организации или торговой марке, можно использовать не только деньги. Возможно стимулировать частных лиц носить одежду, наклеивать стикеры на автомобили или хотя бы пользоваться ручками с фирменным логотипом, чтобы создать мотив как-либо связывать себя с владельцем логотипа. В частности, радио «Максимум» в прямом эфире проводит викторины и победителям вручает футболки с текстом «Максимум», при этом те, вероятно, будут носить подарки, демонстрируя посвященным слушателям свой успех в мероприятии радиостанции.

Ярким примером подобной непрямой рекламы, эксплуатирующей утилитарную связь текста с носителем, является упаковочный пакет торговой сети «Билла», который содержит, помимо названия магазина, фразу «Сегодня я помог детям» (подразумевается благотворительная деятельность магазина). Эти пакеты покупатели подолгу носят, становясь, таким образом, отправителями рекламного текста, имитирующего утилитарный. Так как фраза характеризует отправителя с положительной стороны, можно считать принятие рекламной игры добровольным актом носителя-отправителя текста.

Описанная форма массовой коммуникации называется в маркетинге широким по значению термином промоушен (варваризм от англ. *promotion* – продвижение), в отличие от платной, по определению, рекламы, которая является лишь его частью. Тексты, распространяемые описанным способом, не являются безраздельной частью рекламного дискурса, но обладают минимальной утилитарностью, которая заключена в связи текста с носителем.

Утилитарность текстов логотипов и слоганов, т. е. «чужих» текстов на личных вещах, может проявляться и в указании на те или иные социальные признаки носителя. Например, использование школьных принадлежностей с эмблемой «Winx» (название мультипликационного сериала) маркирует носителя – девочку младшего школьного возраста для адресата, знающего этот сериал и его фабулу. Текст на футболке «Iron Maiden» маркирует человека, имеющего пристрастие к музыке названой рок-группы и т.д., т. е. фирменный логотип сообщает значимую социально информацию о носителе, а также характеризует и самого носителя этой надписи. Поскольку люди физически присутствуют в антропогенном пространстве, они являются его частью, а их разделение на

социальные группы (социоразметка) – это собственно утилитарная функция текстов на личных вещах.

Утилитарная функция логотипов может состоять еще и в том, что они указывают на производителя данной вещи, т. е. это свойство артефакта. В таком случае утилитарная и рекламная функция уравновешены. Если же логотип или слоган к производителю вещи отношения не имеют, как в случае с портфелями, карандашами, тетрадями и дневниками с текстом «Winx», то рекламная функция явно преобладает над утилитарной.

«Свои» тексты на личных вещах, которые не содержат названия фирмы производителя, торговой марки или слогана, а также татуированные тексты, в большей степени реализуют утилитарную функцию и в меньшей – рекламную, если не иметь в виду функцию саморекламы. Носитель текста является и его отправителем, и говорящим (от грамматической формы первого лица) – он сообщает нечто о себе как о члене общества, индивидууме – и эта информация утилитарна для адресата, поскольку позволяет идентифицировать некие важные качества коммуницирующей таким образом личности.

Почти непременным атрибутом таких текстов является игра, в частности языковая, например, многими способами читаемые татуированные аббревиатуры. При этом, сообщая о себе, носитель привлекает к себе внимание, в чем и заключается самореклама, примером чему может служить надпись на футболке девушки: «Готовлю вкусно / Говорю мало / Голова не болит», где исключены все недостатки фольклорного (из анекдотов) образа плохой жены. Разумеется, текст шутливый и не может восприниматься буквально – как пожелание выйти замуж, но он сообщает хотя бы то, что девушка одинока, так как ее муж вряд ли согласился бы с публикацией артефакта с подобным текстом.

Вопрос о границе между утилитарным и рекламным дискурсом в сфере текстов, совмещающих утилитарную и рекламную функции с функцией самовыражения, решен следующим образом: 1) если текст на личной вещи носитель размещает за плату, но не связывает со своими личными особенностями содержательно – он сугубо рекламный, но таких примеров фактически нет в бытовой сфере; 2) если же текст, пусть и рекламирующий стороннюю организацию или личность, носят на себе по причине симпатий и убеждений (бесплатно) – возникает минимальная утилитарность, так как носитель сообщает тем самым о себе: своих пристрастиях; однако рекламная функция здесь также преобладает; 3) рекламная и утилитарная функции уравновешиваются, когда рекламирующий текст избирается носителем вместе с фирменной вещью, которую произвел рекламируемый субъект. В этом случае текст сообщает сразу о двух предметах антропогенного пространства: о носителе и о вещи; 4) во всех случаях, когда в тексте на личной вещи нет явного рекламируемого объекта, кроме самого носителя, утилитарная функция преобладает, так как сам объект (носитель) актуален в пространстве и любая информация о нем утилитарна a priori.

Следующий вопрос, требующий решения, – разграничение утилитарных текстов, с одной стороны, и текстов, служащих только для самовыражения граффити, – с другой. Основным признаком такого разграничения является асоциальный характер граффити, которые располагаются на предметах, не принадлежащих отправителю, вопреки воле хозяина и уже тем самым нарушают общественный порядок, к тому же, часто содержат непристойности.

Итак, основным формальным критерием разграничения граффити и утилитарных текстов на личных вещах будем считать, соответственно, недоступность или доступность для наблюдения носителя-отправителя текста. При всем очевидном сходстве с граффити, утилитарные тексты на личных вещах имеют очень малую зону взаимопроникновения, которая реализуется в том случае, если носитель позволяет себе демонстрировать асоциальные тексты, что наказуемо и поэтому бывает нечасто. В остальных случаях тексты на личных вещах не являются граффити уже потому, что носитель характеризует сам себя как объект антропогенного пространства.

Что касается утилитарных текстов не на личных вещах, то их отличить от граффити по признаку наблюдаемости носителя-отправителя нельзя, кроме того, в них возможны прямые угрозы, например, в собственно утилитарных текстах о порядке парковки: «Машины не парковать / Штраф лопатой по стеклу», однако и здесь можно предложить ряд классификационных признаков. Утилитарными следует считать нерекламные тексты в антропогенном пространстве, буквы которых изготовлены промышленным, типографским способом или на принтере. В равной мере утилитарными являются тексты, отправитель или адресат которых предположительно известен, например, водитель или владелец в текстах автомобилях «Продаю 8-903-ххх-хх-хх» или «Куплю 8-916-ххх-хх-хх» . Но главными признаками утилитарности остаются цели: утилитарный текст указывает на свойства предмета, операции с ним («Машины не парковать»), т.е. основной целью утилитарного текста является распространение актуальной информации об антропогенном пространстве. Граффити может имитировать утилитарность указателей, например, в туалетной кабинке напротив двери «Посмотри направо», на правой стене «Посмотри налево», на левой – «Посмотри назад», на двери – «Посмотри вверх», на потолке «Что вертишься?». Однако в конечном стете граффити не приводит к актуальной информации о предмете, обманывая ожидания адресата и тем самым проявляя свой асоциальный характер.

В итоге можно сказать, что никакого отношения к граффити не имеют нейтральные полезные тексты, связанные с антропогенным пространством, даже когда они выполнены кустарно, а также неасоциальные тексты на личных вещах, которые дают информацию о носителе, являющемся частью антропогенного пространства. Кроме того, существуют три типа взаимопроникновения утилитарных текстов и граффити. Во-первых, отмечаются утилитарные тексты, содержащие угрозы и этим имитирующие граффити, но утилитарность в них преобладает, так как они направлены, прежде всего, на стимуляцию адре-

сата выполнить заданные действия в антропогенном пространстве. Во-вторых, редкие случаи полного взаимопроникновения граффити и утилитарных текстов отмечаются на личных вещах, если там содержатся угрозы, нецензурные слова. Однако это преследуется обществом и встречается редко. В-третьих, существуют граффити, имитирующие утилитарность внешне, например указывающие на что-то, но не передающие актуальной информации в антропогенном пространстве, поэтому истинная утилитарность у них нулевая, т. е. указание намеренно бесцельно.

Тексты, совмещающие функции самовыражения, утилитарную, рекламную и художественную. На самой дальней периферии утилитарного дискурса находится третья диффузная зона – утилитарные тексты, совмещающие и функцию привлечения внимания, и функцию самовыражения и еще одну функцию – художественную. Эта последняя заключается в том, что текст на этническом языке художественно связан с особым артефактом – книгой или журналом, произведением прикладного и пространственного искусства: монументального, включая и памятники в некрополях; прикладного искусства; живописи, включая плакаты и лубочные изображения; архитектуры, включая интерьер и фасады с памятными досками. Таким текстам свойственна образность и художественная выразительность. Они весьма далеки от утилитарных разъяснений к предмету, вместе с тем, без предмета и изображения, на котором они расположены, они не реализуют авторского замысла, что говорит о некоторой утилитарности. Тексты этой крайней периферии утилитарного дискурса находятся в зоне пересечения с рекламными текстами, граффити и художественными текстами. При этом выполнение утилитарной функции присуще некоторой части художественных надписей, что и следует обосновать.

Преобладание утилитарной функции над художественной

При всей существенности художественной и рекламной функций для текстов на обложках изданий гораздо важнее функция содержательная: сообщить о наличии под обложкой некоторого графического текста, назвать его, отослать к нему, кратко отразить его суть, – т. е. налицо утилитарная функция сообщения о наличии и сути предмета. И эта функция заглавий книг уже отмечена в лингвистической литературе: «Названия книг, журналов, картин и т. д., состоящие из слова или словосочетания с номинативным значением, выражают суждение существования и употребляются для того, чтобы обозначить, выделить какой-либо предмет или явление, указать на него, раскрыв его назначение наименованием» [7, с. 136- 137].

В частности, утилитарная функция будет преобладать в надписях на обложках учебных книг, так как книги эти используются не в эстетических целях, а для решения насущных задач образования. Редкий случай эстетической нагрузки можно найти в названиях учебников по гуманитарным предметам: «Русская словесность: От слова к словесности». В использованной как прием речевой выразительности тавтологии «от слова к словесности» реализована собственно языковая орнаментальность текста. В большинстве надписей на учебниках орнаментальность представлена лишь в области графической формы, например, «Задачи на развитие математического мышления с решениями и ответами». Тексты на обложках художественных книг будут в большей степени орнаментальными, но и они содержательно представляют книгу, т. е. утилитарны.

Итак, в третьей диффузной зоне утилитарного дискурса утилитарная функция преобладает над орнаментальной в текстах на обложках книг.

Преобладание художественной функции над утилитарной

Если малый письменный текст, наблюдаемый в контексте произведения искусства – архитектуры, живописи, кино и т. д., изменяет сам артефакт, т. е. придает ему новый социальный смысл, новый статус, новое назначение, то эта функция является утилитарной. По своему действию такие тексты функционально являются перформативами наречения. Текст на монете, указывающий ее достоинство, превращает металлический диск в платежное средство своей эпохи и страны при условии соблюдения соответствующих конвенций. Наличие на монете дополнительного текста, не имеющего отношения к денежной системе, превращает монету в памятный или юбилейный, иногда без указания денежного номинала, знак, который изменяет социальное назначение: функцию платежного средства выполняет крайне редко, становясь предметом коллекционирования, экспонатом выставок, средством государственной пропаганды – т. е. так называемой «юбилейной» монетой. Так, с помощью текста: «80 років з дня проголошення незалежності 1918 Українська Народна Республіка» монета становится пропагандистским знаком. Следует обратить внимание, что текст не только изменил социальное назначение монеты, но продолжает постоянно напоминать о нем, т. е. выражает идею, что монета именно юбилейная, сообщая о социальной сути артефакта, а эта функция собственно утилитарная. Таким образом, у орнаментального текста обнаруживается утилитарная и пропагандистская (рекламная) функция. Именно утилитарная функция обеспечивает возможность пропагандистской, что встречается не только на юбилейных монетах, но и на указателях названий улиц и других географических объектов. Разумеется, не все орнаментальные надписи в равной мере проявляют свою утилитарность, поэтому рассмотрим их по-отдельности, последовательно.

Орнаментальные тексты на памятниках – эпитафии – не только выражают скорбь близких об усопшем, но и сообщают нечто новое о нем: что у него есть близкие, часто кто они, что они скорбят и какие светлые качества усопшего помнят; кроме того, эпитафии часто указывают на религиозную принадлежность усопшего и его родственников, например: «Всякий, кто призовет имя Господне, спасется». Поскольку памятник поставлен усопшему при участии его родственников, то, сообщая о них, текст разъясняет предназначение памятника как артефакта, т. е.

выполняет утилитарную функцию даже в большей мере, чем текст на юбилейной монете, поскольку не просто указывает на мемориальный характер артефакта, но и сообщает некоторые из приведенных обстоятельств, касающихся мемориального объекта.

На предмете прикладного искусства, например на деревянной вилке, текст «Великий Новгород» показывает, что артефакт имеет сувенирное назначение, хотя отнюдь не лишает его функции столового прибора. Подтвердить переход вещи с орнаментальной надписью в разряд сувениров можно тем, что такие вещи чаще продаются в сувенирных магазинах, а не в хозяйственных. Причем в реальной практике предмет может быть использован, как правило, для украшения помещения. Таким образом, текст до некоторой степени изменяет социальную сущность артефакта (повседневный предмет → сувенир) и сообщает об этой измененной сущности, а также о происхождении сувенира – из Великого Новгорода, в этих двух сообщениях собственно и состоит утилитарность текста. Однако сувенирную вилку можно использовать и с практической целью и, в отличие от монеты, она при этом не перейдет к другому человеку, ценящему в нем орнаментальность, поэтому идея использования ее именно как сувенира не управляет действиями пользователя в полной мере.

Итак, артефакту в значительной мере свойственна первоначальная бытовая функция, а текст выполняет преимущественно художественную функцию, поэтому предмет и расположенный на нем текст весьма слабо связаны в коммуникативном плане. Таким образом, утилитарность текстов на предметах прикладного искусства менее существенна, чем текстов на монетах.

На архитектурных сооружениях орнаментальные надписи обладают еще меньшей утилитарностью, но и в этом случае они придают объекту новый статус. В частности, в интерьере станции метрополитена «Пушкинская» в Москве имеются медальоны с цитатами из произведений А.С. Пушкина, одна из которых такова: «Моя студенческая келья вдруг озарилась: Муза в ней открыла пир младых затей». Основная функция текста орнаментальная – он служит элементом декора и выполнен технически так, что прочитать его довольно трудно: буквы рельефные, одного цвета с фоном и значительно мельче тех, что используются на указателях. Но следует учесть и то обстоятельство, что без текстов с цитатами станция метрополитена оставалась бы исключительно транспортным объектом, а с добавлением цитат из произведений А.С. Пушкина приобретает дополнительный статус объекта мемориального, культурного. Тексты, разнообразящие монотонные белые стены станции, в числе прочих украшений, показывают пассажиру, что он проезжает через центр города, дополняя навигацию в метро, в чем реализована утилитарность.

Несколько отличаются от других орнаментальных текстов на предметах архитектуры тексты на памятных досках. Во-первых, они появляются не вместе с сооружениями архитектуры, а позднее. Во-вторых, они не превращают архитектурный объект в мемориальный, а лишь констатируют, что он стал таковым в процессе истории. В-третьих, они сообщают конкретные факты из истории данного сооружения, что может иметь туристическое значение, т. е. помогать бизнесу одних людей и отдыху других, направлять их деятельность, что и является их утилитарной функцией, которая проявлена таким образом сильнее, чем у всех других орнаментальных надписей.

Утилитарные тексты, вторично использованные с художественной целью

Утилитарные тексты, относящиеся к ядру утилитарного дискурса и ближним диффузным зонам, например, располагающиеся на указателях или вывесках, могут попадать в контекст произведения кинематографа или изобразительного искусства. Там они выполняют художественные задачи, что, в широком смысле, тоже относится к орнаментальной функции. Такие тексты образуют самую дальнюю периферию утилитарного дискурса.

В кинофильмах буквенные тексты появляются в двух видах: собственном и заимствованном. В собственном виде это титры или вставки на черном слайде, иногда разделяющие фильм на части, как в фильме П. Лунгина «Царь»: «Молитва царя», Война царя», «Гнев царя», «Веселье царя». Эти тексты раскрывают содержание произведения или его части. При этом они, как утилитарные тексты, сообщают об объекте - но не антропогенного пространства, а скорее, виртуального, представляя собой промежуточное звено между утилитарными текстами и активными компьютерными текстами, которые обладают способностью открывать новые виртуальные объекты. В собственном виде буквенные тексты в кинофильмах весьма схожи с утилитарными, но таковыми не являются из-за отсутствия связанного с ними артефакта.

Среди заимствованных буквенных текстов, или, точнее, вторично воспроизведенных в кинофильме есть достаточное количество утилитарных в виде показанных в кадре табличек, вывесок и т.д. Артефакт, с которым связан текст, можно наблюдать и в кадре. Эти тексты не случайно попадают в фильм, доказательством чему служит специальное создание таких текстов при съемке фильмов. В частности, в иностранных фильмах при появлении русской или советской тематики часто встречаются явно искусственные тексты кириллицей с орфографическими и иными ошибками, что указывает на преднамеренный характер появления текста в кинофильме. Например, в фильме Ксавьера Генса «Хитман» на спине сотрудника написано «ФЭБ» с зеркальной буквой «С». В кинофильме такие тексты выполняют художественную функцию создания местного колорита, но сохраняют и некоторую утилитарность, указывая зрителю на национальный источник предметов, изображенных на экране. Итак, тексты в кинофильмах могут сохранять остаточную утилитарность, если являются заимствованными.

Тексты на живописном полотне или выполняют только утилитарную функцию, например, автограф, или совмещают ее с орнаментальной.

Орнаментальная функция присуща тем буквенным текстам, которые размещены в композиции картины и являются частью ее содержания. Такого рода

тексты художники использовали еще в XIX в., а в живописи XX в. прием широко распространился, причем не только на плакатах, но и в произведениях, причисляемых к живописи как высокому искусству. Существуют два вида внедрения утилитарного текста в живописное полотно. В первом случае текст, исходно не бывший утилитарным, изображается в таком качестве, например как указатель. Так случилось на полотне В.М. Васнецова «Витязь на распутье», где на камне написано «Как пряму ехати – живу не бывати – нет пути ни прохожему, ни проезжему, ни пролетному». Эти слова, по уверению самого художника, сделанному 20 сентября 1898 г. в письме к критику В. В. Стасову, взяты из летописи. Во втором случае становится непосредственным объектом художественного изображения текст, используемый в утилитарном дискурсе. Примером может служить картина русского художника-нонконформиста Э. Булатова «Вход – входа нет», где текст «вход» устремлен вперед, а текст «входа нет» преграждает путь. Разумеется, в обоих случаях текст используется в аллегорическом смысле, а орнаментальная функция в нем преобладает. Но как переносное значение невозможно без понимания прямого, так и аллегорический смысл вырастает из утилитарного, которым было указание пути и запрет движения.

В результате изучения поля утилитарного дискурса, мы прояснили его границы и его внутреннюю структуру, которая представлена в виде концентрической схемы на рисунке, где центр изображает ядро дискурса, а внешние области – соответственно три диффузные зоны.

В заключение отметим, что поле утилитарного дискурса представляет собой информационную среду, саморазвивающуюся под влиянием изменяющихся потребностей общества. Именно благодаря самопроизвольному развитию утилитарный дискурс, отделенный от рекламы, может быть использован как образец навигации. Во-первых, информационное поле утилитарного дискурса может быть полезно как алгоритм для программы информационных терминалов в учреждениях, в частности в государственных службах и торговых центрах. Во-вторых, строение утилитарного дискурса может быть использовано при разработке сайтов интернет магазинов. В-третьих, на основе утилитарного дискурса можно разработать тематическую классификацию используемых человеком вещей, которая могла бы лечь в основу тематической поисковой системы, что оказалось бы принципиально ново, так как все существующие системы основаны на поиске заданного текстового фрагмента. В-четвертых, знание поля утилитарного дискурса могло бы быть использовано для разработки помощника для слепых, способного не только читать, но и выбирать актуальные для навигации надписи и озвучивать их.

Рисунок

СПИСОК ЛИТЕРАТУРЫ

1. Букчина Б.З., Золотова Г.А. Слово на вывеске // Русская речь. – 1968. – №3. – С. 49 – 56.
2. Шмелева Т.В. Современная городская эпиграфика. // Язык и культура. Третья междунар. конф.: докл. и тез. – Киев, 1994. – С. 107.
3. Подберезкина Л. З. Современная городская среда и языковая политика // Русский язык в его функционировании: Тез. докл. междунар. научно-методич. конф. – СПб.: Изд-во “Питер”, 1998. – С. 26 – 29.
4. Романова Т.П. Проблемы современной эргонимии. // Вестник СамГУ Филология. – 1998. – №1. – С. 82 – 90.
5. Китайгородская М.В., Розанова Н.Н. Малые письменные жанры в городском общении: на пути к диалогу с. 256 – 262 // Жизнь языка: Сб. ст. к 80-летию М.В. Панова. – М.: Языки славянской культуры, 2001. – 544 с.
6. Уэллс У., Бернет Дж., Мориарти С. У. Реклама: принципы и практика / пер. с англ. – СПб.: Изд-во «Питер», 1999. – 736 с.
7. Бабайцева В.В. Односоставные предложения в современном русском языке. – М.: Просвещение, 1968 – 160 с.

Материал поступил в редакцию 11.10.13.

Сведения об авторе

ЛИХАЧЕВ Сергей Владимирович – кандидат филологических наук, доцент кафедры филологических дисциплин и методики их преподавания в начальной школе ИППО Московского городского педагогического университета
e-mail: lihachovSV@rambler.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [004 : 61] – 047.44

Т.И. Булдакова, Д.А. Миков

Анализ информационных процессов виртуального центра охраны здоровья

Предложено решение задачи выявления информационных рисков виртуальной системы здравоохранения на основе функционального моделирования с использованием методологии IDEF0. Разработана функциональная IDEF0-модель виртуального центра охраны здоровья населения и проведён анализ информационных потоков на основе построенных диаграмм. Выявлены существующие уязвимости в функционировании информационной системы. Предложен комплекс контрмер для снижения уровня информационного риска и способы их реализации.

Ключевые слова: *информационные риски, здравоохранение, виртуальные системы, методология IDEF0*

Процесс модернизации системы здравоохранения сопровождается активным внедрением информационно-коммуникационных технологий, созданием виртуальных инфраструктур для развития телемедицинских услуг: функционирования электронной регистратуры, обеспечения доступа к электронным медицинским картам пациентов и т.д. Примером такой инфраструктуры являются виртуальные центры охраны здоровья (ВЦОЗ) населения [1-3]. При этом в предлагаемых вариантах реализации концепции виртуального здравоохранения обходится стороной важный аспект –информационная безопасность.

Главная особенность защиты конфиденциальной информации медицинских учреждений вообще и ВЦОЗ в частности заключается в необходимости более детального анализа бизнес-процессов для выявления информационных рисков, характера информационных потоков в медицинской информационной системе (МИС) и их взаимосвязей. Такое требование можно осуществить, используя различные подходы – статистические методы, методы экспертных оценок, методы моделирования [4].

Однако наиболее подходящим является функциональное моделирование с применением методологии IDEF0 [5]. Для построения функциональной модели ВЦОЗ сначала необходимо исследовать его информационные потоки.

1. ХАРАКТЕРИСТИКА ОСНОВНЫХ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ ВИРТУАЛЬНОГО ЦЕНТРА ОХРАНЫ ЗДОРОВЬЯ

Развитие микроэлектроники и телекоммуникаций позволяет включить пациента в единое информационное пространство системы здравоохранения, независимо от его местоположения [6]. Сделать это можно, например, регистрацией биосигналов сердечно-сосудистой системы с помощью датчиков, вмонтированных в нательную одежду [7-9]. Биосигналы должны передаваться по каналам связи в медицинские центры мониторинга и обработки информации (рис. 1), где посредством математических моделей элементов и подсистем организма создаётся виртуальный физиологический образ (ВФО) пациента, описывающий физиологическую деятельность подсистем человека [1, 3].

Макет виртуального центра представляет собой распределённую биомедицинскую информационно-аналитическую систему, содержащую устройства удалённого доступа, web-сервер, клиентские приложения и сервер обработки (рис. 2) [3].

Начальный этап построения модели – сбор и первичная обработка данных. В роли данных в этом случае выступают персональные данные и биоинформация пациента, которые проходят сбор и первичную обработку с помощью устройств удалённого доступа – мобильных терминалов (МТ) или информационно-измерительных систем лечебно-профилактических учреждений (ЛПУ). Устройства удалённого доступа формируют регистрационный номер пациента и структуру модели виртуального физиологического образа на основе лингвистических данных (классификационные признаки, необходимые для формирования структуры моделей, базы знаний и ВФО пациента).

Далее регистрационный номер и структура модели пересылаются на Web-сервер ВЦОЗ. Информационно-измерительные системы ЛПУ выполняют передачу через ADSL (англ. Asymmetric Digital Subscriber Line – «асимметричная цифровая абонентская линия»), а мобильный терминал – через GPRS (англ. General Packet Radio Service – «пакетная радиосвязь общего пользования»).

Регистрационный номер пополняет базу данных, а структура моделей – базу знаний. В базе данных хранятся регистрационные номера пациентов и сотрудников ВЦОЗ, каждый из которых принадлежит к определённой группе пользователей МИС ВЦОЗ, наделённой соответствующими полномочиями и правом доступа. База знаний содержит информацию о структуре моделей и знания, накопленные врачами и исследователями. База знаний, в свою очередь, связана с базой электронных карт, куда поступают все личные данные пациента, включая информацию о госпитализации, историю болезни, результаты проведённых исследований. Электронные карты пациентов периодически пополняют архив, который предназначен для ведения статистических данных. База данных, база знаний, база электронных карт и архив располагаются на Web-сервере и управляются единой системой управления базами данных (СУБД).

Необходимая информация из базы знаний поступает на сервер обработки, который выполняет функции провайдера виртуального центра [3]. Это набор программно-аппаратных средств, осуществляющих обработку и накопление знаний о пациентах. В основу этой подсистемы положена технология web-сервисов, оперирующая тремя основными понятиями:

1) набор сервисов (анализ, моделирование, прогнозирование и др.);

2) генератор сценариев;

3) исполняемый сценарий.

Фактически каждый исполняемый сценарий является приложением, разработанным для определённой группы пользователей. Набор сервисов представляет совокупность базовых функций, которые могут потребоваться для работы приложений. Логика формирования сценариев основывается на предопределённых правилах, адекватных конкретному случаю.

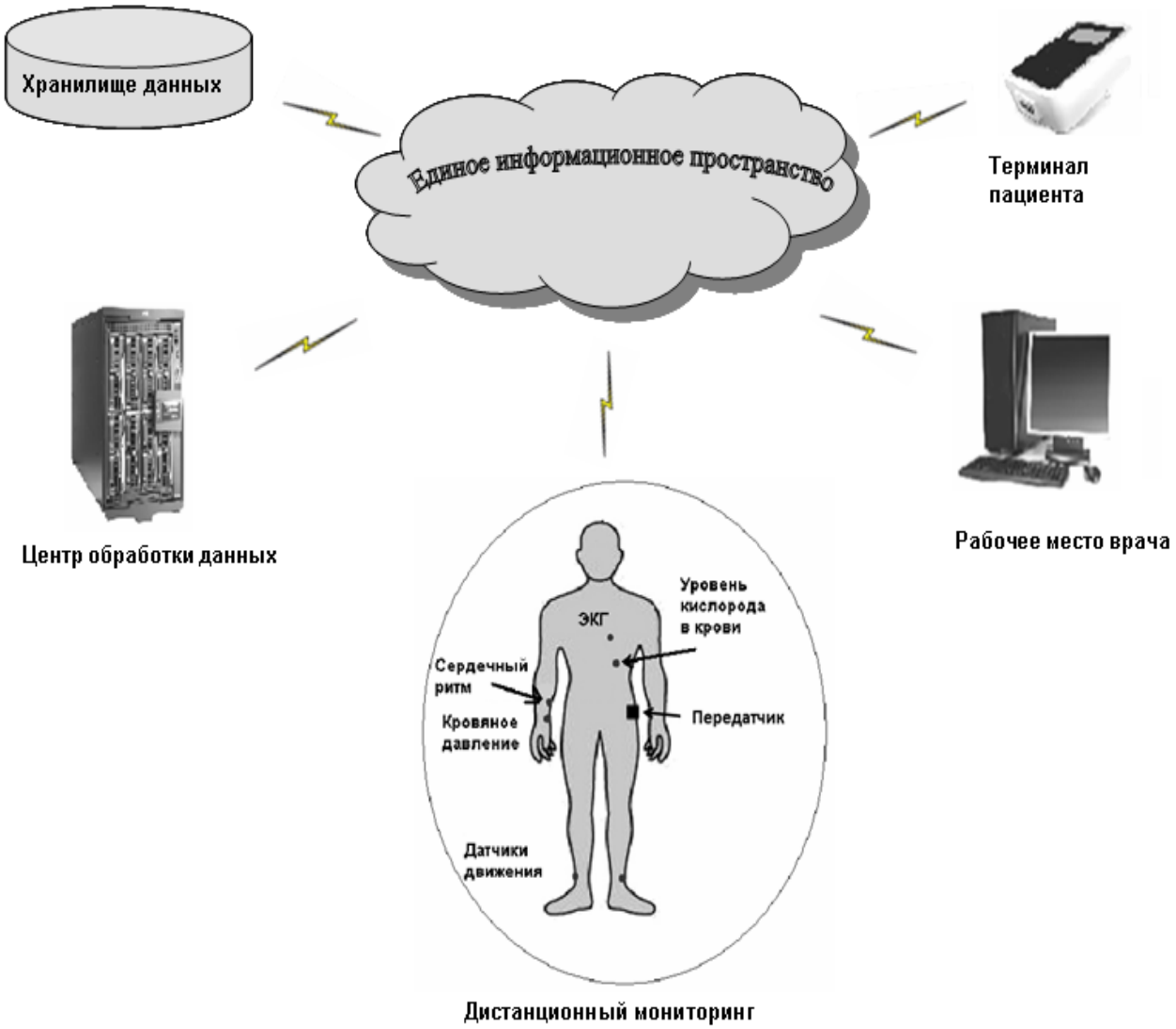

Рис. 1. Дистанционный мониторинг состояния человека на основе единого информационного пространства

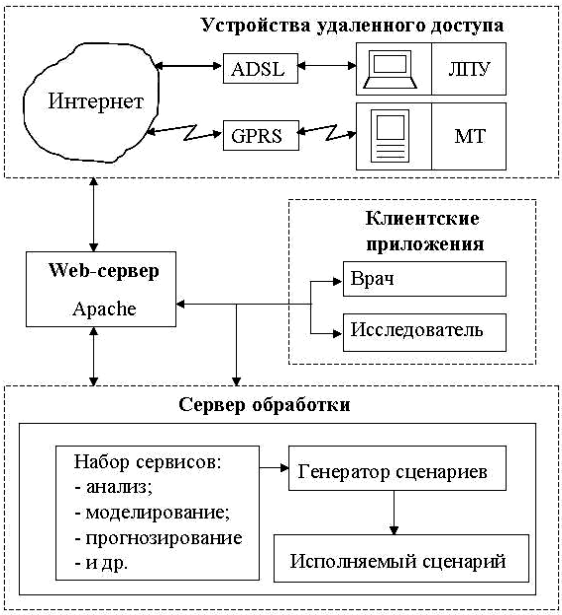

Рис. 2. Структура прототипа ВЦОЗ

Сервер обработки на основе информации, полученной из базы знаний, производит выбор структуры модели ВФО для соответствующей электронной карты пациента. На основе выбранной структуры ВФО и данных, содержащихся в электронной карте, формируется конкретный ВФО пациента. Сервер обработки производит также статистическую обработку данных, поступающих к нему из архива, которые необходимы для создания типовых ВФО (эталонов), с которыми будет происходить сравнение ВФО пациента. Результаты сравнения с помощью клиентских приложений используются персоналом для определения диагноза и прогнозирования изменения функционального состояния организма (ФСО) пациента. После этого персонал выдает рекомендации пациенту по профилактике и лечению.

Далее необходимо осуществить анализ функционирования ВЦОЗ с применением методологии IDEF0.

2. ПОСТРОЕНИЕ IDEF0-МОДЕЛИ ВИРТУАЛЬНОГО ЦЕНТРА ОХРАНЫ ЗДОРОВЬЯ «КАК ЕСТЬ»

Процесс моделирования начинается с определения субъекта моделирования, цели и точки зрения на модель. В нашем исследовании субъект моделирования – ВЦОЗ, цель моделирования – анализ информационных рисков, точка зрения – специалист по информационной безопасности. Сначала строится модель «как есть». Контекстная диаграмма этой модели представлена на рис. 3. (Каждая диаграмма IDEF0-модели имеет свой порядковый номер).

Далее необходимо провести функциональную декомпозицию контекстной диаграммы. При этом все интерфейсные дуги с внешних сторон диаграмм декомпозиции нижних уровней должны соответствовать интерфейсным дугам диаграммы декомпозиции верхнего уровня, а интерфейсные дуги диаграммы декомпозиции первого (самого верхнего) уровня – интерфейсным дугам контекстной диаграммы. Диаграмма декомпозиции первого уровня представлена на рис. 4.

Процесс сбора и первичной обработки данных должен происходить с учётом требований Федерального закона от 27 июля 2006 г. №152-ФЗ «О персональных данных».

Типы структур моделей и алгоритм их формирования должны быть определены во внутренних нормативных актах ВЦОЗ, как и правила и особенности формирования и ведения базы данных, базы знаний, базы электронных карт и архива.

Таким образом, на выходе процесса структурирования данных ключевыми элементами, необходимыми для дальнейшего функционирования ВЦОЗ, являются база знаний, электронная карта, архив, регистрационный номер. Эти элементы используются в качестве входных интерфейсных дуг в процессе обработки данных.

Функциональные диаграммы декомпозиции процессов обработки данных и анализа и извлечения данных строятся аналогично.

Обработка данных состоит из следующих процессов:

1) выбор структуры модели,

2) формирование виртуального физиологического образа (ВФО),

3) статистическая обработка данных,

4) формирование отчётов,

5) организация доступа к данным,

6) обмен данными с клиентами.

Все виды формируемых отчётов, принципы организации доступа к данным и правила обмена данными с клиентами должны быть отражены во внутренних нормативных актах ВЦОЗ.

Рис. 3. Начальная контекстная диаграмма виртуального центра охраны здоровья

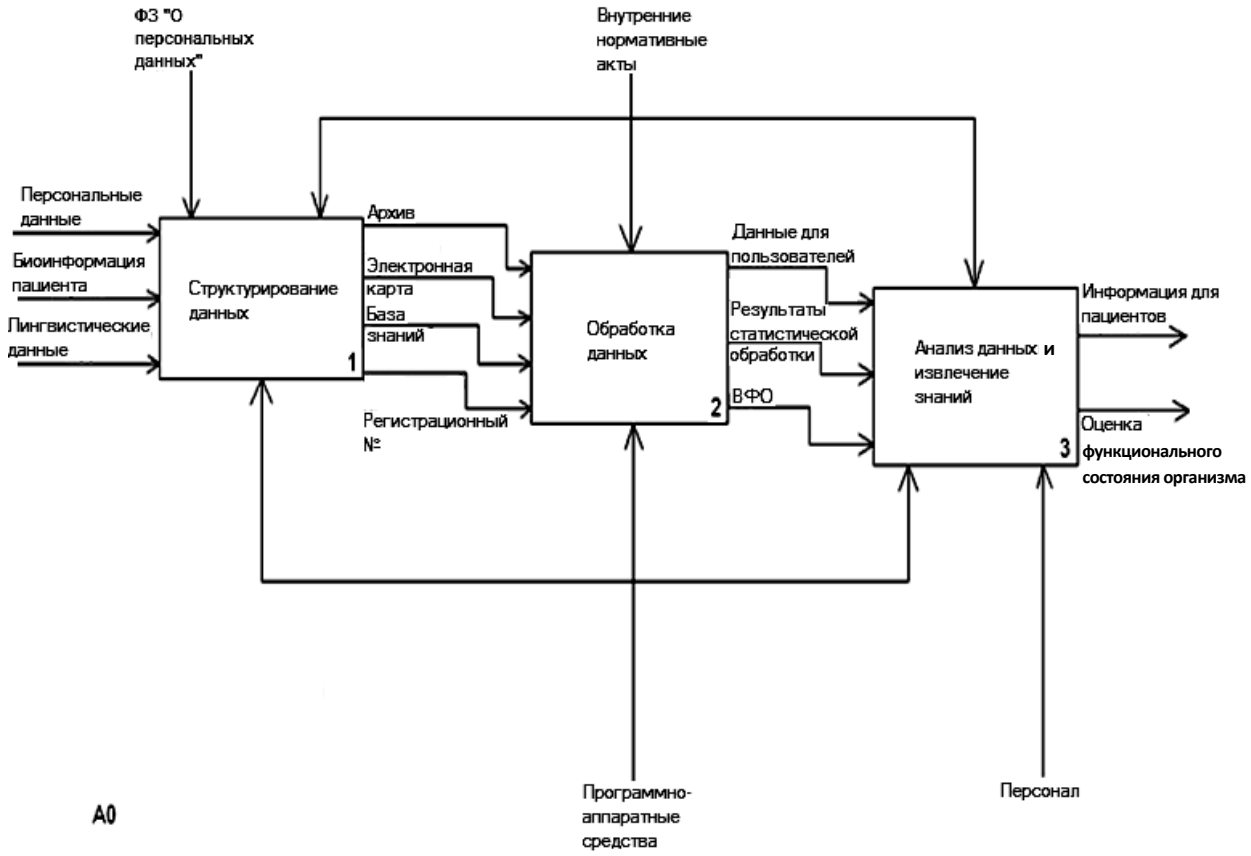

Рис. 4. Диаграмма декомпозиции первого уровня

На выходе процесса обработки данных имеются результаты статистической обработки, ВФО, данные для пользователей. Эти элементы используются в качестве входных интерфейсных дуг для процесса анализа данных и извлечения знаний, который состоит из следующих этапов:

1) формирование эталонов (типовых ВФО);

2) моделирование и визуализация состояния системы «сердце – сосуды – легкие»;

3) сопоставление модели с эталонами;

4) диагностика и прогнозирование;

5) выработка рекомендаций;

6) информирование пациента о функциональном состоянии организма.

IDEF0-модель «как есть» можно считать завершённой. Достигнутый уровень детализации является достаточным для анализа системы с точки зрения специалиста по защите данных. Следовательно, можно провести анализ информационной безопасности, определить уязвимости в МИС и выявить существующие информационные риски.

3. ВЫЯВЛЕНИЕ ИНФОРМАЦИОННЫХ РИСКОВ И ВЫРАБОТКА КОНТРМЕР

В соответствии со статьёй 6 «Условия обработки персональных данных» Федерального закона №152-ФЗ, обработка персональных данных осуществляется только с согласия субъекта персональных данных. Поэтому необходимо разобраться, в чём должно заключаться согласие пациента на обработку его персональных данных.

Статья 10 «Специальные категории персональных данных» этого закона указывает, что обработка специальных категорий персональных данных, касающихся расовой, национальной принадлежности, политических взглядов, религиозных или философских убеждений, состояния здоровья, интимной жизни, не допускается, за исключением случаев, предусмотренных частью 2 настоящей статьи. То есть персональные данные, обрабатываемые МИС ВЦОЗ, относятся к специальным категориям персональных данных, так как касаются состояния здоровья пациентов.

Во 2 части этой статьи Федерального закона говорится, что обработка указанных специальных категорий персональных данных допускается в случае, если субъект персональных данных дал согласие в письменной форме на обработку своих персональных данных. Таким образом, для обработки персональных данных необходимо письменное согласие пациента.

В статье 9 «Согласие субъекта персональных данных на обработку его персональных данных» сказано, что равнозначным содержащему собственноручную подпись субъекта персональных данных согласию в письменной форме на бумажном носителе признаётся согласие в форме электронного документа, подписанного в соответствии с федеральным законом электронной подписью.

Следовательно, эффективным решением для выполнения требования Федерального закона о письменном согласии пациента на обработку его персональных данных может быть документ, подписанный электронной подписью.

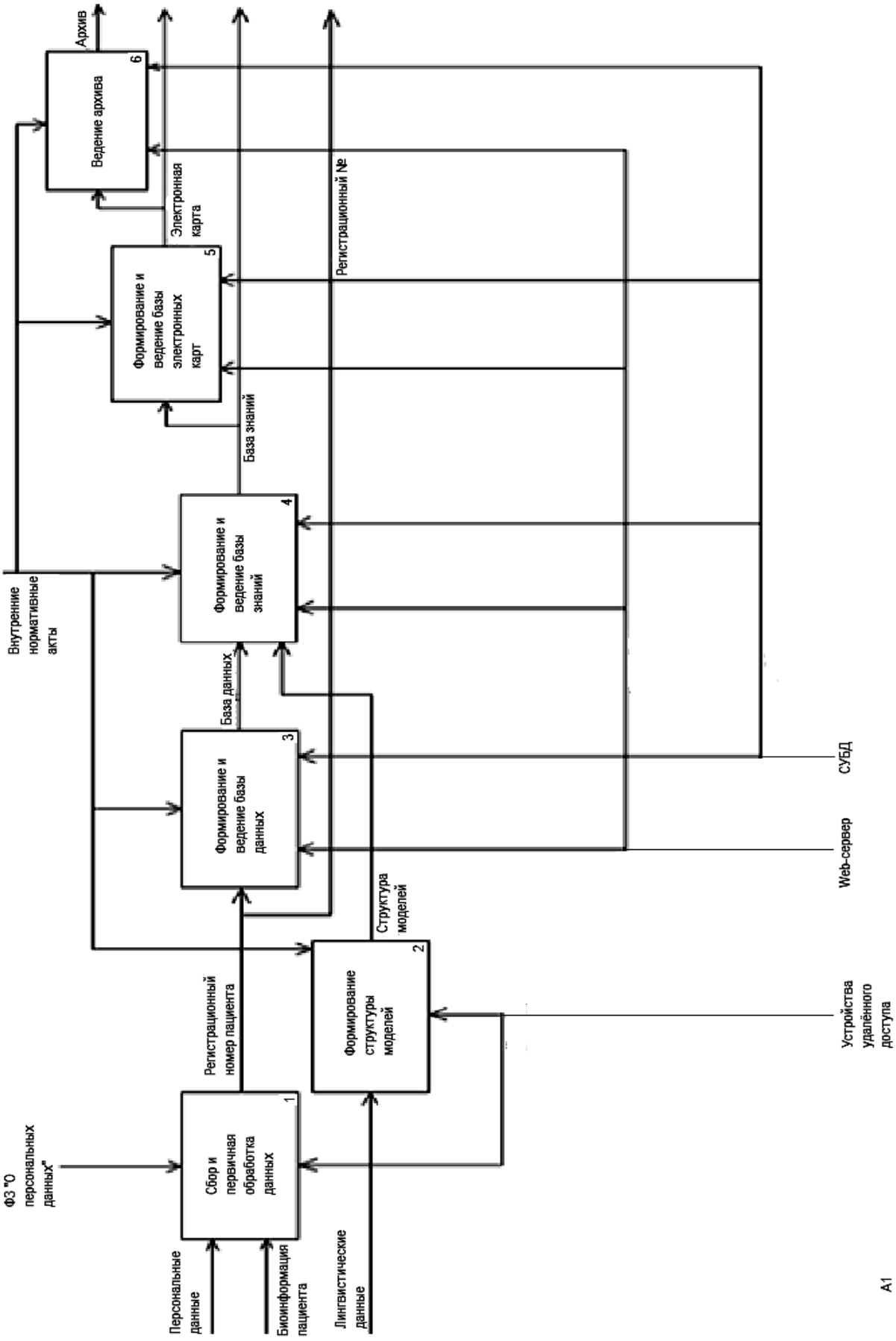

Рис. 5. Структурирование данных

В подсистеме структурирования данных проблемой является управление базой данных, базой знаний, базой электронных карт и архивом с помощью единой СУБД, то есть необходимо обеспечить информационную совместимость при организации информационного обмена. Также в подсистеме структурирования данных необходимо гарантировать проверку целостности, резервное копирование и восстановление данных для оперативного восстановления всей подсистемы либо любой её части.

Представляется, что естественным решением указанных проблем может быть использование CALS-технологий. Подтверждением этого является, например, широкое использование языка XML для представления, хранения и использования различными пользователями биотехнических и данных о состоянии здоровья [10].

Наиболее подходящей XML-СУБД для ВЦОЗ является Sedna, так как данная система обеспечивает политику безопасности и предоставляет возможность резервирования данных, что позволяет проводить проверку целостности и восстановление данных в случае повреждений, возникающих вследствие нарушений в работе подсистемы структурирования данных. Sedna поддерживает древовидную модель данных (хранимых в двоичном виде), которые загружаются и извлекаются в виде XML-документов. Данные оптимизируются и индексируются для рационального хранения и быстрого доступа [11].

Далее необходимо исследовать подсистему обработки данных на наличие уязвимостей. Наибольшее внимание следует обратить на процессы организации доступа к данным и обмена данными с клиентами. Предоставление данных с высокой доступностью и поддержанием их защиты от несанкционированного доступа является сложной задачей. Контроль доступа в МИС ВЦОЗ сопряжён с рядом проблем, которые не возникали в других сферах безопасности. Перечислим их.

Поскольку защищаемая информация составляет персональные данные, то нарушение безопасности может привести к необратимым последствиям для заинтересованных лиц. При этом существует необходимость в полном доступе ко всей информации, относящейся к пациентам (например, аллергия на определённые лекарства, ВИЧ-статус и т.д.).

Однако самым сложным аспектом является объём полномочий, которые необходимо предоставить пациентам. Согласие пациентов и конфиденциальность являются ключевыми вопросами. Это предоставляет пациентам возможность указать степень ограничения контроля доступа к их медицинским данным, что потенциально ведёт к сложной политике безопасности [12].

Распределение прав доступа к данным обеспечит всем участникам процесса обработки защищаемой медицинской информации присвоение специальной метки («секретно», «для служебного пользования» и т.д.), именуемой уровнем безопасности или уровнем доступа. Все уровни безопасности должны быть основаны на принципе доминирования. Контроль доступа должен осуществляться в зависимости от уровня безопасности взаимодействующих сторон на основании двух правил:

1) уполномоченное лицо (субъект) имеет право читать только те документы, уровень безопасности которых не превышает его собственный уровень безопасности (допуск);

2) уполномоченное лицо (субъект) имеет право заносить информацию только в те документы, уровень безопасности которых не ниже его собственного уровня безопасности.

Синхронизация доступа клиентов к информационным ресурсам ВЦОЗ позволит предотвратить совместный доступ к ним, одновременное использование ресурсов несколькими клиентами может привести к порче данных и нарушениям в работе системы.

Добиться выполнения требований распределения прав доступа к данным и синхронизации доступа клиентов к информационным ресурсам можно, используя управление доступом на основе ролей.

Управление доступом на основе ролей (англ. Role Based Access Control, RBAC) – это развитие политики избирательного управления доступом, при этом права доступа субъектов системы на объекты группируются с учётом специфики их применения, образуя роли [12, 13].

Формирование ролей дает возможность определить чёткие и понятные для пользователей правила разграничения доступа. Ролевое разграничение доступа позволяет реализовать гибкие, изменяющиеся динамически (в процессе функционирования системы) правила разграничения доступа.

Так как привилегии не назначаются пользователям непосредственно и приобретаются ими только через свою роль (или роли), управление индивидуальными правами пользователя, по сути, сводится к назначению ему ролей. Это упрощает такие операции как добавление пользователя или смена подразделения пользователем.

Ещё одной проблемой подсистемы обработки данных является большой объём архивной информации, передаваемой Web-сервером серверу обработки для статистической обработки и формирования отчётности. Решить этот вопрос можно с помощью сжатия данных.

И, наконец, в подсистеме анализа данных и извлечения знаний наиболее уязвим процесс взаимодействия сервера обработки с клиентскими приложениями при обмене данными, эталонами ВФО и результатами моделирования. Для нейтрализации этой уязвимости понадобится реализация следующих решений [14]:

1) введение протокола взаимодействия «клиент-сервер» позволит отследить действия каждого клиента в произвольный момент времени, а также выявить пользователя, действия которого привели к порче данных или к возникновению спорных ситуаций из-за ошибочно введённой информации;

2) автоматическое обновление клиентских модулей обеспечит своевременную установку обновлений и исправлений клиентских модулей, значительно упрощая работу, связанную с администрированием системы;

3) автономная работа при потере связи с сервером позволит сохранить работоспособность клиентской части в случае временной потери связи с сервером.

Таким образом, выявлены уязвимости во всех подсистемах МИС ВЦОЗ и предложены способы их устранения (см. таблицу).

Далее необходимо построить функциональную модель «как должно быть» на основе предложенных механизмов реализации выработанных контрмер.

4. ПОСТРОЕНИЕ IDEF0-МОДЕЛИ «КАК ДОЛЖНО БЫТЬ»

В подсистему структурирования данных (рис. 6) введено применение электронной подписи на этапе сбора и первичной обработки данных, что позволит обеспечить выполнение требования о получении согласия пациента на обработку его персональных данных, также уточнён формат хранения информации в базах и архиве (XML), и использована XML-СУБД Sedna для обеспечения целостности, резервного копирования и оперативного восстановления хранящейся информации.

В подсистему обработки данных (рис. 7) введено сжатие архивных данных, передаваемых web-сервером серверу обработки для формирования статистики и отчётности, что позволит уменьшить расход дисковой памяти сервера и минимизировать сетевой трафик. В процессах организации доступа к данным и обмена данными с клиентами задействована технология управления доступом на основе ролей (RBAC-модель), что обеспечит распределение прав доступа к данным через назначение пользователям соответствующих ролей трёх типов (исследователь, лечащий врач, пациент) и синхронизацию доступа клиентов к информационным ресурсам.

Таблица

Уязвимости, контрмеры и их реализации

Подсистема	Выявленная уязвимость	Контрмера	Предлагаемая реализация
Подсистема структурирования данных	1) получение письменного согласия пациента на обработку его персональных данных;	1) использование электронной подписи на этапе сбора и первичной обработки данных;	1) получение согласия пациента на обработку его персональных данных в виде цифрового документа, подписанного электронной подписью;
	2) обеспечение информационной совместимости при обмене данными;	2) хранение информации в формате XML;	2) применение CALS-технологий;
	3) организация единого управления подсистемой и обеспечение целостности хранимой информации;	3) использование XML-СУБД, обеспечивающей проверку целостности, резервное копирование и восстановление данных;	3) применение XML-СУБД Sedna, обладающей собственной политикой безопасности;
Подсистема обработки данных	4) обеспечение надёжного контроля доступа к данным без нанесения ущерба доступности информации;	4) распределение прав доступа к данным и синхронизация доступа клиентов к информационным ресурсам;	4) использование модели RBAC (управление доступом на основе ролей);
	5) большой объём архивной информации, передаваемой Web-сервером серверу обработки;	5) сжатие архивных данных после их передачи на сервер обработки;	5) применение сжатия для уменьшения расхода дисковой памяти сервера и минимизации сетевого трафика;
Подсистема анализа данных и извлечения знаний	6) отслеживание действий каждого клиента в произвольный момент времени для выявления потенциального нарушителя;	6) ведение протокола взаимодействия «клиент-сервер»;	6) использование открытого сетевого протокола Sedna Client-Server Protocol;
	7) своевременная установка обновлений и исправлений клиентских модулей;	7) автоматическое обновление клиентских модулей;	7) использование модуля обновления XUpdate;
	8) сохранение работоспособности клиентской части в случае временной потери связи с сервером.	8) обеспечение автономной работы при потере связи с сервером.	8) использование 64-разрядного диспетчера памяти, адресации и подкачки.

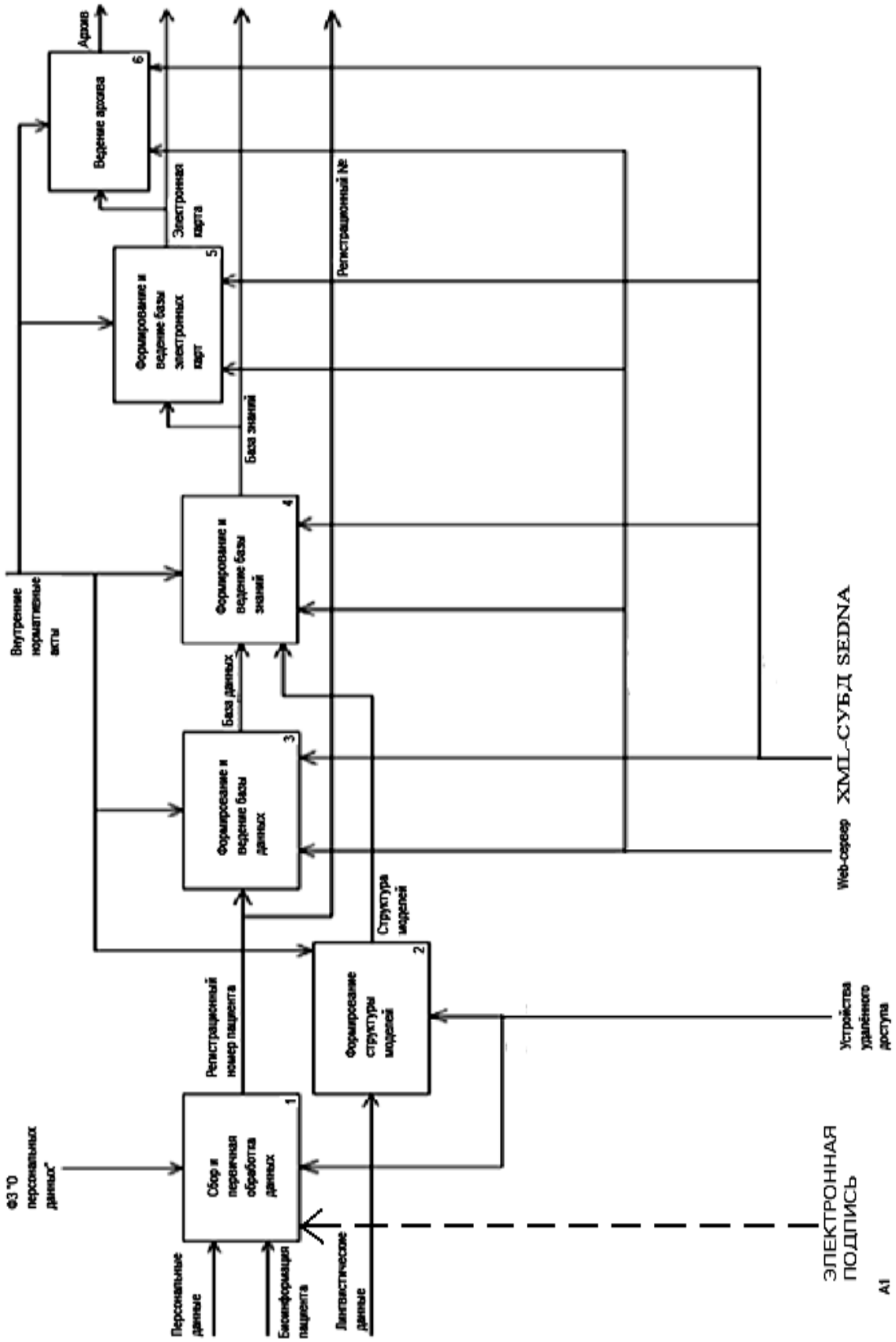

Рис. 6. Подсистема структурирования данных

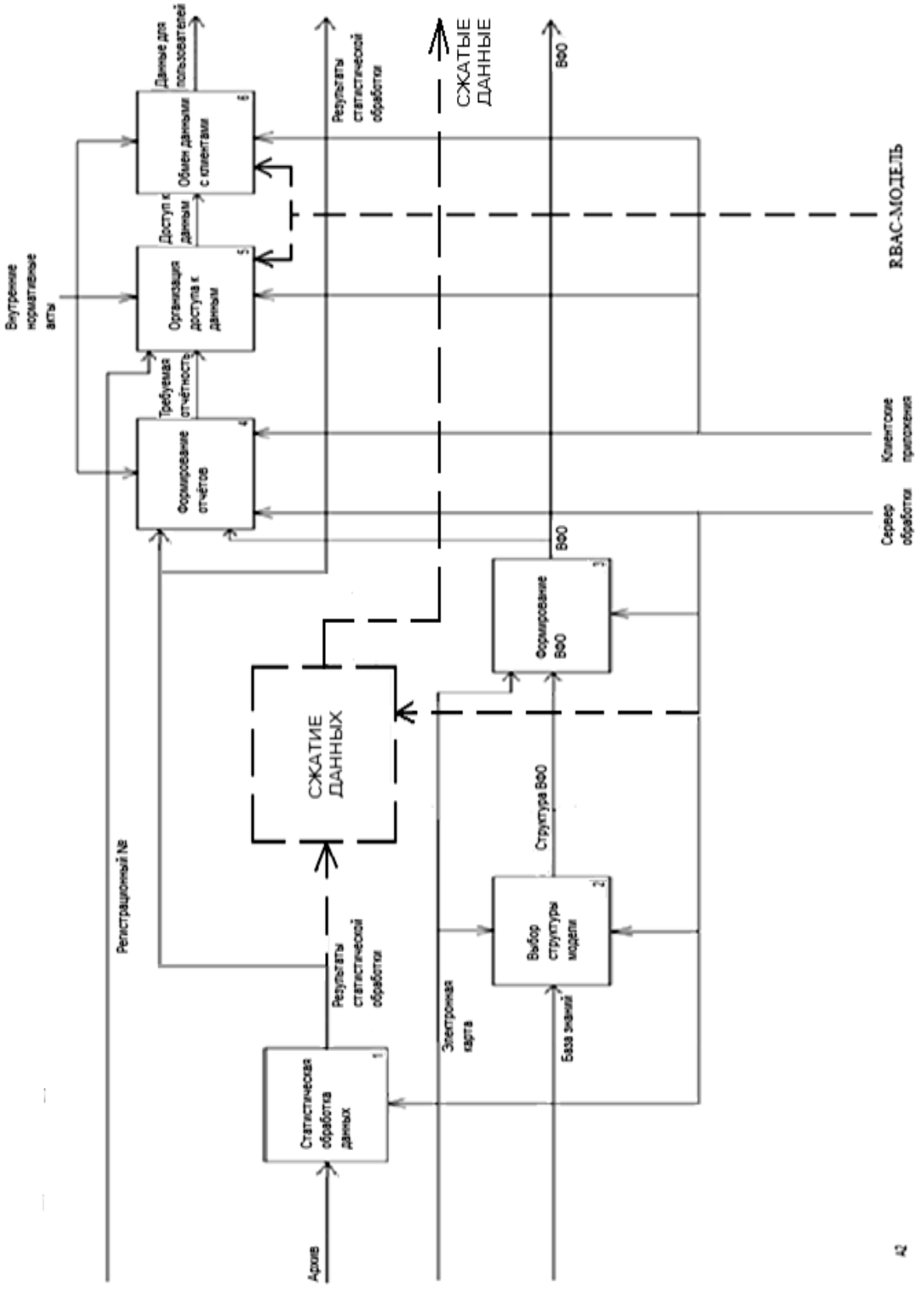

Рис. 7. Подсистема обработки данных

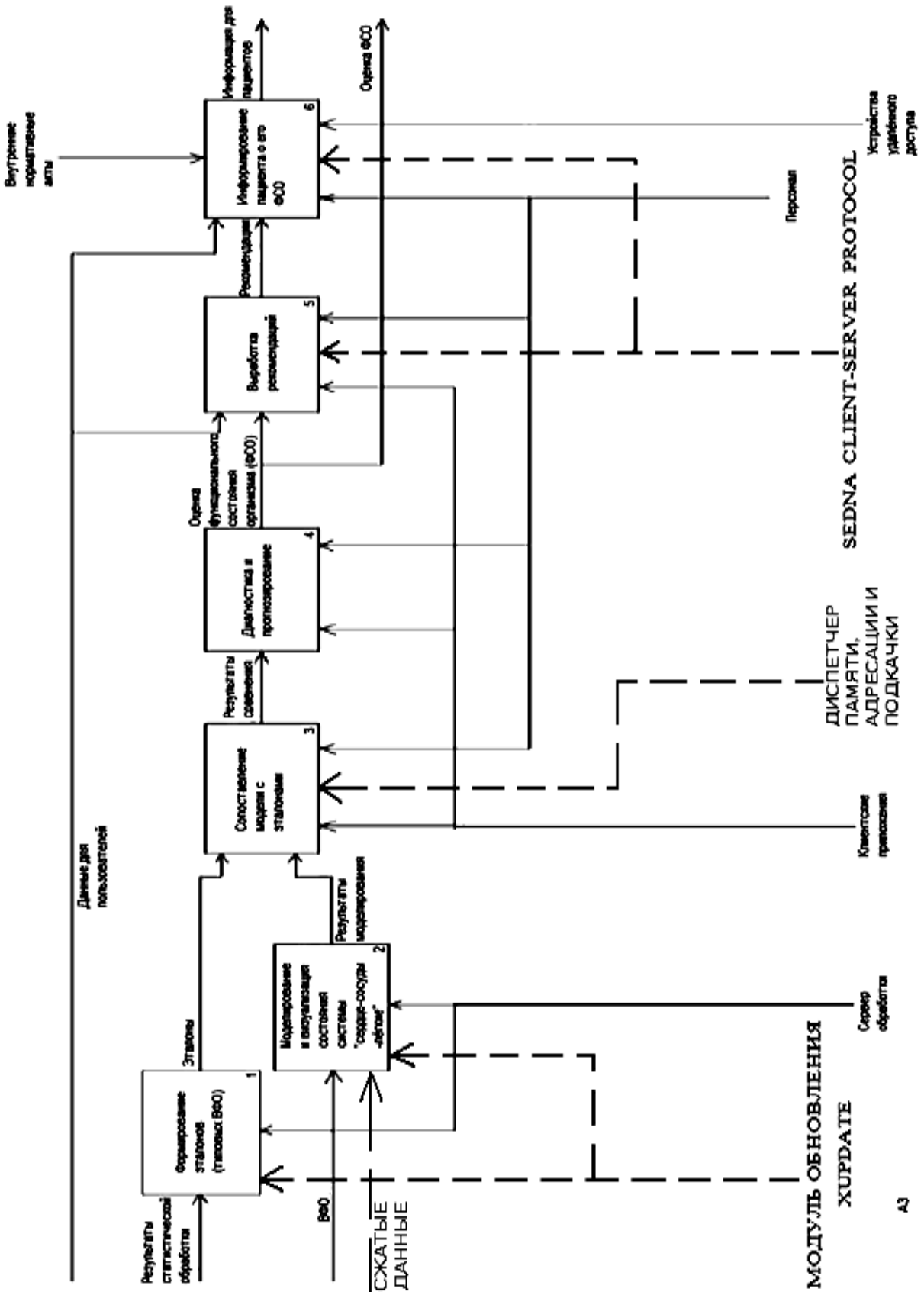

Рис. 8. Подсистема анализа данных и извлечения знаний

В подсистему анализа данных и извлечения знаний (рис. 8) внедрено ведение протокола Sedna Client-Server Protocol, который отслеживает действия каждого клиента в процессах выработки рекомендаций и информирования пациента о его ФСО. Модуль обновления на языке XUpdate обеспечит своевременную установку обновлений и исправлений клиентских модулей, значительно упрощая работу, связанную с администрированием системы. Диспетчер памяти, адресации и подкачки сохранит работоспособность клиентской части в случае временной потери связи с сервером.

Таким образом, разработка функциональной IDEF0-модели «как должно быть» завершена. Проведённое исследование позволило осуществить анализ информационных рисков ВЦОЗ, выявить уязвимости в его информационной системе и разработать комплекс контрмер, направленных на снижение уровня риска.

ЗАКЛЮЧЕНИЕ

В настоящей работе был проведён анализ функционирования ВЦОЗ с целью выявления информационных рисков. Анализ осуществлялся при помощи функционального моделирования на основе методологии IDEF0, которая является наиболее эффективным средством для решения поставленной задачи ввиду простоты построения и наглядности описания функциональных диаграмм.

Информационная система ВЦОЗ была представлена на функциональных IDEF0-диаграммах в виде трёх подсистем – структурирования данных, обработки данных и анализа данных и извлечения знаний (модель AS-IS – «как есть»). Для каждой из подсистем были выявлены риски информационной безопасности, определены уязвимостями и выработаны соответствующие контрмеры для их устранения. Способы реализации выработанных контрмер были отображены в виде дополнительных интерфейсных дуг и функциональных блоков на диаграммах модели TO-BE – «как должно быть».

Предложенный подход может быть использован для анализа информационных рисков при других вариантах построения виртуальных систем здравоохранения.

СПИСОК ЛИТЕРАТУРЫ

1. Prado M., Roa L., Reina-Tosina J. Virtual Center for Renal Support: Technological Approach to Patient Physiological Image // IEEE Transaction on biomedical engineering. –2002. – Vol. 49, № 12. – P. 1420-1430.
2. Paradiso R., Loriga G., Taccini N. A Wearable Health Care System Based on Knitted Integrated Sensors // IEEE Transactions on Information Technology in Biomedicine. – 2005. – Vol. 9, № 3. – P. 337-344.
3. Анищенко В.С., Булдакова Т.И., Довгалевский П.Я. и др. Концептуальная модель виртуального центра охраны здоровья населения // Информационные технологии. –2009. – №12. – С. 59-64.
4. Булдакова Т.И., Миков Д.А. Метод повышения адекватности оценок информационных рисков // Вестник Моск. гос. тех. университета им. Н.Э. Баумана. Серия «Приборостроение». Спец. выпуск «Информатика и системы управления». – 2012. – С. 261-271.
5. Калянов Г.Н. CASE-технологии. Консалтинг при автоматизации бизнес-процессов. – М.: Горячая линия – Телеком, 2002. – 320 с.
6. Гончаров Н.Г., Гулиев Я.И., Гуляев Ю.В. и др. Вопросы создания Единого информационного пространства в системе здравоохранения РАН // Информационные технологии и вычислительные системы. – 2006. – № 4. – С. 83-94.
7. Winters J., Wang Y. Wearable Sensors and Telerehabilitation // IEEE Engineering in Medicine and Biology Magazine. – 2003. – № 3. – P. 56-65.
8. Булдакова Т.И., Коблов А.В., Суятинов С.И. Информационно-измерительный комплекс совместной регистрации и обработки биосигналов // Приборы и системы. Управление, контроль, диагностика. – 2008. – № 6. – С. 41-46.
9. Булдакова Т.И., Гриднев В.И., Кириллов К.И. и др. Программно-аналитический комплекс модельной обработки биосигналов // Биомедицинская радиоэлектроника. –2009. – № 1. – С. 71-77.
10. Manickam S, Abidi A. Extracting clinical cases from XML-cased electronic patient records for use in web-based medical case based reasoning systems // Proc. Medinfo. – 2001. – P. 643-647.
11. Sedna XML Database. – URL: http://www.sedna.org/.
12. Eyers D.M., Bacon J., Moody K. Oasis role-based access control for electronic health records // IEE Proc.-Softw. – 2006. – Vol. 153, № 1. – P. 16–23.
13. Sandhu R., Coyne E.J., Feinstein H.L., Youman C.E. Role-Based Access Control Models // IEEE Computer. – 1996. – № 29 (2). – P. 38–47.
14. Бодин О.Н., Логинов Д.С., Семенкин М.А. Организация информационного обеспечения медицинской компьютерной диагностической системы // Информационно-измерительные и управляющие системы. – 2008. – № 12. – С. 110-114.

Материал поступил в редакцию 21.08.13.

Сведения об авторах

БУЛДАКОВА Татьяна Ивановна – доктор технических наук, профессор кафедры «Информационная безопасность» Московского государственного технического университета имени Н.Э. Баумана.
e-mail*:* buldakova@bmstu.ru

МИКОВ Дмитрий Александрович – магистрант кафедры «Информационная безопасность» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: MikovDA@yandex.ru

ИНФОРМАЦИОННЫЕ ЯЗЫКИ

УДК 007.52 : 004.934.5

А.Ю. Кирков, В.Е. Павловский

Акустический мультичастотный язык коммуникации роботов

Рассматривается задача коммуникации роботов на основе языка мультичастотных акустических сигналов. Приведена формальная модель языка, в котором каждому символу языка соответствует последовательность акустических мультичастотных сигналов, рассмотрен алгоритм распознавания таких характерных звуков.

Ключевые слова: *акустическая коммуникация, разговор роботов, распознавание речи*

ВВЕДЕНИЕ

Описывается искусственный язык для осуществления «разговора» роботов в обычной воздушной среде (или других средах) на акустических частотах, близких к воспринимаемым человеком. Этот язык основан на простых по своей структуре малочастотных сигналах, являющихся искусственно синтезированными фонемами языка.

Настоящая работа является продолжением публикаций [1-4], описывающих создание системы тональной акустической коммуникации роботов, где был представлен язык коммуникации, в котором каждому символу соответствует последовательность одночастотных акустических сигналов. В настоящей работе, в отличие от указанных публикаций, описывается язык, фонема которого представляет собой последовательность не одночастотных, а многочастотных сигналов. Язык усложняется для более устойчивой работы системы в условиях посторонних шумов, более сложные конструкции языка должны уменьшить процент ложного распознавания.

В нашей работе введен также словарь роботов, представляющий собой специализированную базу данных, состоящую из слов, которыми пользуются роботы. Цель введения словаря – исправление ошибок, возникших в результате коммуникации роботов.

Представлен также алгоритм распознавания конструкций указанного языка. Реализована программная среда для визуализации и анализа работы построенного алгоритма, приведены результаты ряда экспериментов на основе созданных компьютерной и аппаратной моделей.

СТРУКТУРА ЯЗЫКА РОБОТОВ

В нашей работе акустический язык состоит из алфавита и набора фонем. Фонемы делятся на фрагменты. В отличие от одночастотной системы [1], фрагмент в настоящей работе представляет собой не одночастотный, а многочастотный сигнал, определенной длительности и амплитуды. Будем рассматривать фрагменты, состоящие из одного и того же количества частот.

Введем некоторый параметр N_L, который будет определять количество частот каждого фрагмента языка. Параметр N_L назовем частотностью языка роботов.

Рассмотрим некоторое подмножество множества из N_L различных неупорядоченных положительных чисел, которое будем называть множеством частот. Каждый элемент этого множества представляет собой некоторый набор частот.

$$Frq \subset \{(frq_1, frq_2, .., frq_{N_L}) \mid frq_i \in N, i = \overline{1, N_L}, \\ frq_i \neq frq_j \ \forall i, j = \overline{1, N_L}, i \neq j\}.$$

Фрагмент будет представлять элемент следующего множества:

$$Frg = \{(amp, frq, dur) \mid amp \in N, dur \in N, frq \in Frq\}.$$

Отметим, что в отличие от одночастотной системы, фрагмент содержит не одну частоту, а некоторый набор частот, который характеризуется элементом множества Frq. Поскольку для рассматриваемого языка фрагменты играют роль неделимых частиц, систему будем называть многочастотной.

Любое конечное подмножество

$$F \subset \{(frg^1, frg^2, ..., frg^n) \mid n \in N, frg^i \in Frg, i = \overline{1, n}\},$$

множества конечных упорядоченных наборов элементов из Frg будем называть множеством фонем.

Таким образом, любая фонема представляет собой последовательность многочастотных сигналов. Пример аудиосигнала, соответствующего фонеме, графически изображен на рис 1.

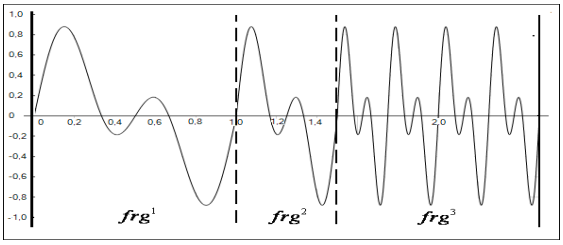

Рис.1. Пример аудиосигнала, соответствующего фонеме двухчастотной системы

Введем параметр D_f, который будем называть общей длительностью фонемы. В настоящей работе будем рассматривать фонемы, имеющие одинаковую длительность D_f

$$D_f \mid f \mid = D_f, \forall f \in F$$

Аналогично одночастотной системе рассмотрим алфавит B, соответствие R между алфавитом и множеством фонем и пробел $S^* = \{S, dur_S\}$. Набор $L = (B, F, R, S^*)$ алфавита, множества фонем, соответствия между ними и пробела будем называть языком роботов.

Любой упорядоченный набор букв из множества B будем называть словом. Множество слов будем обозначать W.

$$W = \{(b_1, b_2, .., b_n), b_i \in B, i = \overline{1,n}, n \in N\}$$

Длиной слова $w \in W$ будем называть количество букв в этом слове. Введем обозначение $|w|$ для длины слова.

Будем называть слово $w' \in W$ подсловом слова $w = (b_1, .., b_n) \in W$, если существуют такие индексы $i, j, 1 \le i \le j \le n$, что $w' = (b_i, .., b_j)$. Введем обозначение $w' \subset w$. Подслово $w' = (b_i, .., b_j) \in W$ будем называть началом слова $w = (b_1, .., b_n)$, если $i = 1$ и концом этого слова, если $j = 1$.

СЛОВАРЬ И ШАБЛОН СЛОВА

Множество слов W – бесконечно, однако, если провести аналогию с человеческим языком, набор слов, которые использует человек в своей речи, ограничен. Например, словарный запас русского языка насчитывает примерно 200 тыс. слов, из которых 3000 являются наиболее употребительными.

Введем ограничения для словарного запаса языка роботов. Для этого рассмотрим некоторое конечное подмножество множества слов

$$D \subset W, \mid D \mid < \infty .$$

Множество D будем называть словарем роботов, он будет состоять из тех слов, которыми будут пользоваться роботы. Будем считать, что роботы могут проигрывать слова исключительно из словаря. В настоящей работе будем предполагать, что все роботы пользуются одним и тем же словарем.

Поиск слова в словаре будем вести по шаблону. Шаблоном слова назовём произвольный упорядоченный набор подмножеств множества B, в котором первый и последний элемент не являются пустыми множествами, т.е. элемент множества

$$W_T^B = \{(B_1, B_2, .., B_n), B_i \subset B, B_1, B_n \ne \varnothing, i = \overline{1,n}, n \in N\} .$$

Длиной шаблона слова $w_t^b \in W_T^B$ будем называть количество подмножеств B_i в этом шаблоне, обозначим её $|w_t^b|$. Количество непустых множеств B_i шаблона w_t^b обозначим $\| w_t^b \|$. Множества $B_i, i = \overline{1,n}$ будем называть буквенными множествами шаблона w_t^b.

Шаблон слова $w_t^{b'} \in W_T^B$ будем называть составной частью шаблона $w_t^b = (B_1, .., B_n)$, если $w_t^{b'} = (B_i, .., B_j)$ для некоторых $i, j, 1 \le i < j \le n$. Для составной части шаблона введем обозначение $w_t^{b'} \subset w_t^b$.

Рассмотрим также множество, элемент которого будет представлять упорядоченный набор подмножеств множества фонем F. При этом первый и последний элемент этого набора должны быть непустыми множествами:

$$W_T^F = \{(F_1, F_2, .., F_n), F_i \subset F, F_1, F_n \ne \varnothing, i = \overline{1,n}, n \in N\} .$$

Элемент этого множества будем называть фонемным шаблоном слова. Множества $F_i, i = \overline{1,n}$ будем называть фонемными множествами шаблона w_t^f.

В силу того, что язык роботов $L = (B, F, R, S^*)$ задает однозначное соответствие R между алфавитом B и множеством фонем F, каждому элементу множества W_T^B можно поставить в соответствие элемент множества W_T^F и – наоборот. Обозначим это соответствие R_T^W.

Определенный выше словарь роботов D, а также множества W_T^B, W_T^F и соответствие между ними R_T^W будут использоваться в алгоритме распознавания.

АЛГОРИТМ РАСПОЗНАВАНИЯ ШАБЛОНА СЛОВА

В работе [1] был рассмотрен алгоритм поиска фонем в полученном роботом аудиосигнале для одночастотного языка роботов. Задача этого алгоритма заключалась в поиске всех отрезков времени, на которых была проиграна некоторая фонема. Отметим, что указанный алгоритм не использовал тот факт, что фонемы могут проигрываться не обособленно, а в составе некоторого слова. Однако, учитывая это свойство языка, можно извлечь дополнительную информацию, которую потом будем применять в алгоритме распознавания.

Согласно введенным в настоящей работе ограничениям, роботы могут говорить слова только из словаря роботов, что сильно ограничивает варианты слов для распознавания. Этот факт тоже будет полезно использовать в алгоритме распознавания.

В настоящей работе рассматривается алгоритм, цель которого поиск не фонем языка, а слов из словаря роботов. Алгоритм состоит из двух частей: первая часть – поиск отрезков времени, в течение которых было проиграно некоторое слово, и в определении шаблона, соответствующего этому слову; задача второй части – поиск слов из словаря роботов по полученному шаблону.

Такое разделение алгоритма распознавания обусловлено тем, что в процессе распознавания некоторые буквы слова могут быть не распознаны или, наоборот, вместо одной буквы в результате процесса распознавания могут быть получены два или более символов. Например, при проигрывании слова «привет» на выходе алгоритма можно получить слово «првет» без буквы «и» в середине слова. При этом между интервалами, на которых были распознаны буквы «р» и «в», будет «пропуск» длиной, равной длительности фонемы. Совершенно очевидно, что на этом месте была пропущена какая-то буква. И слово «првет» на самом деле является некоторым словом «пр?вет», где знаком «?» помечен некоторый символ языка роботов. Учитывая использование словаря, из подобных шаблонов несложно получить исходное слово.

Общая схема распознавания

Аудиосигнал будем рассматривать в качестве функции $s(t) \in L_2[t_1, t_2]$, заданной на некотором отрезке времени $[t_1, t_2]$. Задача алгоритма заключается в определении всех интервалов $[t', t''] \subset [t_1, t_2]$, на которых было проиграно некоторое слово, и шаблон $w_t^b \subset W_T^B$, соответствующий этому слову.

Поскольку между множествами W_T^B и W_T^F существует однозначное соответствие, обусловленное языком роботом, поиск шаблона слова $w_t^b \subset W_T^B$ сводится к поиску соответствующего фонемного шаблона $w_t^f \subset W_T^F$.

Алгоритм работает по разбиению отрезка $[t_1, t_2]$ на малые интервалы, которые будем называть операционными. Длительность операционного интервала задается значением τ, которое является параметром алгоритма.

Обработка аудиосигнала на каждом операционном интервале заключается в определении самых громких частот языка роботов. Громкость частот определяется согласно спектральным коэффициентам. Выделение частот производится по двухпороговой схеме. Частоту считаем выделенной, если её громкость выше двух порогов: абсолютного и относительного. Абсолютный порог задается некоторым значением $a_{\min}$. Относительный порог равен произведению некоторого значения $\sigma \in (0,1)$ на наименьшее из N_L наибольших значений амплитуды сигнала, соответствующих частотам языка роботов. Значения $a_{\min}$ и σ рассматриваются в качестве параметров алгоритма.

В том случае, если для фонемы $f \in F$ найдется отрезок, равный длительности этой фонемы, на каждом операционном интервале которого частоты соответствующих фрагментов окажутся выделенными, фонему будем считать распознанной на этом отрезке.

Фонемы проигрываются не обособленно, а в составе некоторого слова. Поэтому после распознавания первой фонемы слова будем говорить, что система переходит в режим распознавания слова. В этом режиме она будет находиться до тех пор, пока в течение промежутка, равного длительности пробела, ни одна из фонем не будет распознана.

В режиме распознавания слова интервалы, на которых могут быть распознаны фонемы, определены точно. Эти интервалы следуют за отрезком времени, на котором была распознана первая фонема слова, и равны общей длительности фонемы D_f. Такие отрезки времени будем называть фонемными интервалами слова.

После выхода системы из режима распознавания слова отрезок времени, на котором было проиграно слово, определяется от начала первого фонемного интервала до конца последнего фонемного интервала, на котором была распознана хотя бы одна фонема.

Фонемный шаблон $w_t^r = (F_1, .., F_n) \in W_T^F$, соответствующий слову, проигранному на этом отрезке, будет состоять из тех множеств фонем, которые были распознаны на соответствующих фонемных интервалах.

Вспомогательные алгоритмы

В алгоритме распознавания шаблона слова используется ряд вспомогательных алгоритмов. Представим краткое описание каждого из них, подробно с этими алгоритмами можно ознакомиться в работе [5].

Алгоритм определения начала фонемы. Необходимость алгоритма обусловлена тем, что начало проигрывания фонемы может и не совпадать с началом операционных интервалов.

В работе [1] представлен алгоритм определения начала проигрывания фонемы для одночастотной системы. В этом алгоритме вводится функция $\delta_i(f)$, которая для каждого операционного интервала определяет смещение начала проигрывания фонемы $f \in F$ относительно начала этого интервала. Расчет этой функции осуществляется в начале проигрывания фонемы. При этом используется тот факт, что фонема начинает проигрываться частотой первого фрагмента. В указанной работе приводится алгоритм определения начала проигрывания частоты.

В силу того, что фрагмент в настоящей работе может состоять из двух и более частот, при расчете функции $\delta_i(f)$ берется минимальное смещение, рассчитанное для частот первого фрагмента. Кроме этого, расчет этой функции осуществляется только в начале проигрывания слова. В режиме распознавания слова в качестве $\delta_i(f)$ для каждой фонемы $f \in F$ берется значение, рассчитанное для первой фонемы этого слова.

Алгоритм подавления эха. Во время проигрывания фрагмента амплитуда, соответствующая каждой его частоте, резко увеличивается. Однако после окончания проигрывания фрагмента, она резко не падает. Как показывают эксперименты, в течение 100-250 мс амплитуда, соответствующая частотам проигранного фрагмента, остается достаточно высокой. Подобное является следствием отражения звуковых волн.

На рис. 2 слева изображен график изменения амплитуды, соответствующей частоте проигрываемого фрагмента. Пунктирными линиями график разделен на две области. Левая область соответствует времени, в течение которого была проиграна фонема. Правая область соответствует эху этого сигнала.

Эффект эха негативно влияет на процесс распознавания. К примеру, если фонема, соответствующая букве «а», состоит из одного фрагмента, то при проигрывании только буквы «а» можно получить две-три буквы в результате распознавания.

На рис. 2 справа изображены графики амплитуд двух фрагментов, которые были проиграны один за другим. Фрагмент, который был проигран вторым, частично оказался менее громким, чем эхо от первого фрагмента.

Вследствие этих проблем был создан алгоритм подавления эха. Целью этого алгоритма является поиск частот, аудиосигнал соответствующий которым на операционном интервале является эхом только что проигранных фрагментов. Указанные частоты впоследствии не используются при определении самых громких частот языка роботов на операционном интервале.

Алгоритм подавления шума. Вместе с полезным сигналом микрофон принимает различные шумы – разговор людей, шум бытовой техники, машин, шелест листьев и т.д. Подобные звуки оказывают отрицательное воздействие на качество работы алгоритма распознавания.

Задача рассматриваемого алгоритма на каждом операционном интервале заключается в поиске частот, которые не соответствуют проигранным фонемам языка робота. Алгоритм работает в предположении, что длительность фрагмента не меньше длительности трех операционных интервалов.

Указанные частоты определяются путем выявления либо непродолжительных интервалов (не более 2τ), на которых амплитуда, соответствующая этим частотам, выше некоторого уровня шума a_n, либо интервалов, длительностью 3τ или 4τ в том случае, если этот интервал попадает на стык проигрывания двух фрагментов. Значение a_n рассматривается в качестве параметра алгоритма.

Алгоритм корректировки смещения. При определении начала проигрывания фонемы рассчитывается функция смещения $\delta_i(f)$. Если фонема f входит в аудиосигнал, на следующем шаге значение $\delta_{i+1}(f)$ для этой фонемы берется из предыдущего.

Таким образом, если в начале проигрывания фонемы значение $\delta_i(f)$ было рассчитано неправильно, оно же и будет использоваться в течение последующего периода. Эта погрешность в расчете функции $\delta_i(f)$ может негативно повлиять на процесс распознавания.

В связи с этим был построен алгоритма, который корректирует значение функции $\delta_i(f)$ в случае необходимости. В работе рассмотрено два подобных алгоритма: корректировки смещения влево и корректировки смещения вправо. Оба алгоритма корректируют значение функции смещения на стыке фрагментов и фонем.

АЛГОРИТМ ПОИСКА СЛОВА В СЛОВАРЕ ПО ШАБЛОНУ

Цель алгоритма – определить минимально возможный набор слов $D^* \subset D$ из словаря, который будет соответствовать рассматриваемому шаблону слова $w_t^b \in W_T^B$, полученному в результате работы предыдущего алгоритма. В том случае, если множество D^* окажется непустым, будем считать, что одно из слов этого множества было проиграно некоторым роботом.

Критерии соответствия. Для определения множества D^* для произвольного шаблона $w_t^b \in W_T^B$ рассмотрим критерии, по которым будем определять соответствие слова $w \in D$ из словаря шаблону w_t^b. Для этого введем ряд следующих определений.

Будем считать, слово $w \in D$ будет соответствовать шаблону $w_t^b = (B_1,..,B_n) \in W_T^B$, если для некоторого подслова $w' = (b_1,..,b_n) \subset w, |w'| = |w_t^b|$, длина которого будет совпадать с длиной шаблона w_t^b, каждая его буква будет содержаться в соответствующем множестве B_i шаблона w_t^b, либо это множество будет пустым, т.е. будет выполняться одно из следующих условий

$$1.\ b_i \in B_i$$
$$2.\ B_i = \varnothing \qquad \textbf{(1)}$$

Множество слов из словаря D, соответствующих шаблону w_t^b, обозначим $D_T^0(w_t^b)$.

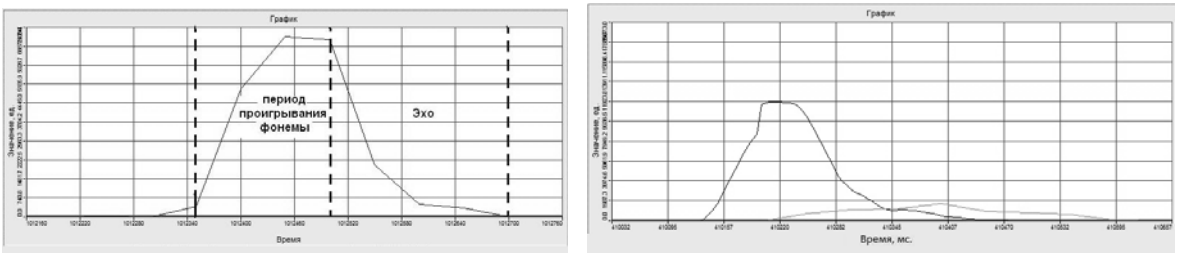

Рис.2. Демонстрация эффекта эхо

Если существует подслово w', удовлетворяющее условиям определения **(1)**, которое является началом слова w, будем говорить, что слово w соответствует шаблону w_t^b с начала. Множество слов, соответствующих шаблону w_t^b с начала, обозначим $D_T^B(w_t^b)$.

В том случае, если w' будет концом слова w, будем говорить, что оно соответствует шаблону w_t^b до конца. Множество слов, соответствующих шаблону w_t^b до конца, обозначим $D_T^E(w_t^b)$.

Если само слово w удовлетворяет условиям определения **(1)**, будем говорить, что слово w полностью соответствует шаблону w_t^b. Множество слов, полностью соответствующих шаблону w_t^b, обозначим $D_T^F(w_t^b)$.

Оператор соответствия. Рассмотрим оператор $D_T(w_t^b)$, который каждому шаблону $w_t^b \in W_T^F$ будет ставить в соответствие некоторое подмножество слов из словаря следующим образом. $D_T(w_t^b)$ будет возвращать слова, которые полностью соответствуют шаблону w_t^b. Если таких слов не будет и длина шаблона не меньше трех, $D_T(w_t^b)$ будет возвращать слова, начало которых соответствует шаблону, потом конец. Если и таких слов не будет, $D_T(w_t^b)$ вернет слова, которые просто соответствуют шаблону.

$$D_T(w_t^b)=\begin{cases} D_T^F(w_t^b),\ D_T^F(w_t^b)\neq\varnothing \\ D_T^B(w_t^b),\ |w_t^b|\geq 3,\ D_T^B(w_t^b)\neq\varnothing \\ D_T^E(w_t^b),\ |w_t^b|\geq 3,\ D_T^B(w_t^b)=\varnothing, \\ D_T^E(w_t^b)\neq\varnothing \\ D_T^0(w_t^b),\ |w_t^b|\geq 3,\ D_T^B(w_t^b)=\varnothing,\ D_T^E(w_t^b)=\varnothing, \\ D_T^0(w_t^b)\neq\varnothing \\ \varnothing,\ \textit{в остальных случаях} \end{cases}.$$

Оператор сопряжения. Для шаблона w_t^b определим сопряженный шаблон $\hat{w}_t^b \in W_T^B$. Шаблон w_t^b соответствует некоторому фонемному шаблону $w_t^f=(F_1,..,F_n)\in W_T^F$, который является распознанным на некотором отрезке $[t',t'']$.

Согласно определению распознанного шаблона существует разбиение отрезка $[t',t'']$ точками $t_0',..,t_n',t_i'\in T$, $t_0'=t',t_n'=t''$, такое, что на каждом отрезке $\Delta t_i'=[t_{i-1}',t_i']$ фонемы из множества F_i являются распознанными. На каждом отрезке $\Delta t_i'=[t_{i-1}',t_i']$ этого разбиения определим сопряженное множество фонем F_i^*.

В том случае, если множества $F_i\neq\varnothing$ непустые, сопряженным множеством F_i^* будем считать само множество F_i. Иначе, в качестве F_i^* будем рассматривать множество фонем, которые чаще всех остальных были выделенными на операционных интервалах соответствующего отрезка $\Delta t_i'$.

Фонемный шаблон $\hat{w}_t^f=(F_1^*,..,F_n^*)$, состоящий из сопряженных множеств F_i^*, будем называть сопряженным фонемным шаблоном для w_t^f. Поскольку между множествами W_T^F и W_T^B существует однозначное соответствие R_T^W, обусловленное языком роботов, для исходного шаблона $w_t^b=(B_1,..,B_n)$ определим сопряженный шаблон $\hat{w}_t^b=(B_1^*,..,B_n^*)$.

Рассмотрим множество $W_T^{B'}\subset W_T^B$, состоящее из шаблонов, буквенные множества которых совпадают либо с B_i, $i=\overline{1,n}$, либо с пустыми множествами:

$$W_T^{B'}=\{w_t^{b'}=(B_1',..,B_{n'}'),\ B_i'\in\{\varnothing,B_1,..,B_n\},$$
$$i=\overline{1,n'},\ B_1'\neq\varnothing,\ B_{n'}'\neq\varnothing\}.$$

Над этим множеством рассмотрим оператор сопряжения $Q^*(w_t^{b'})$, который для каждого шаблона $w_t^{b'}\in W_T^{B'}$, $w_t^{b'}=(B_1',..,B_{n'}')$ будет возвращать шаблон $\hat{w}_t^{b'}=(B_1'^*,..,B_{n'}'^*)$, где $B_i'^*=B_j^*$, если $B_i'=B_j$ для некоторого j и $B_i'^*=\varnothing$, если $B_i'=\varnothing$.

Сопряженный оператор соответствия. Используя операторы соответствия $D_T(w_t^{b'})$ и сопряжения $Q^*(w_t^{b'})$, определим сопряженный оператор соответствия $D_T^*(w_t^{b'})$ над множеством $W_T^{B'}$. Этот оператор будет возвращать множество слов $D_T(w_t^{b'})$, в том случае, если оно непустое. Иначе, если $\|w_t^{b'}\|\geq 3$, он будет возвращать $D_T(\hat{w}_t^{b'})$, $\hat{w}_t^{b'}=Q^*(w_t^{b'})$. В противном случае он будет возвращать пустое множество.

$$D_T^*(w_t^{b'})=\begin{cases} D_T(w_t^{b'}),\ D_T(w_t^{b'})\neq\varnothing \\ D_T(\hat{w}_t^{b'}),\ \hat{w}_t^{b'}=Q^*(w_t^{b'}),\ \|w_t^{b'}\|\geq 3, \\ D_T(w_t^{b'})\neq\varnothing \\ \varnothing,\ \textit{в остальных случаях} \end{cases}.$$

Базовый алгоритм. В том случае, если оператор $D_T^*(w_t^b)$ для исходного фонемного шаблона w_t^b возвращает непустое множество, это множество рассматривается в качестве искомого множества D^* и работа алгоритма на этом прекращается. В противном случае рассмотрим алгоритм разделения слова.

Алгоритм разделения слова. Необходимость этого алгоритма обусловлена тем, что иногда два подряд проигранных слова распознаются в виде одного шаблона. Такое может возникнуть, например, если между этими словами будет распознан какой-либо лишний символ.

Алгоритм будет делить исходный шаблон на две части и пытаться определить множества слов D_T^L и D_T^R, соответствующие левой и правой частям исход-

ного шаблона. Если объединение множеств $D_T^L \bigcup D_T^R$ будет непустым, это множество будем рассматривать в качестве искомого множества D^*.

В противном случае, если $\| w_t^b \| < 4$ считаем $D^* = \varnothing$. Иначе переходим к алгоритму исправления одиночной ошибки.

Алгоритм исправления одиночной ошибки. Необходимость алгоритма обусловлена тем, что иногда возникают ситуации, когда шаблон слова распознается с ошибками.

Алгоритм для каждого непустого буквенного множества $B_i \neq \varnothing$, $i = \overline{2, n-1}$ исходного шаблона слова w_t^b строит шаблон w_t^H, в котором вместо буквенного множества B_i берется пустое множество. Для этого шаблона w_t^H применением сопряженного оператора соответствия $D_T^*(w_t^H)$ получаем множество слов D_T^H. Если множество $D_T^H \neq \varnothing$ – непустое, берем его в качестве искомого множества D^*.

В том случае, если для всех непустых буквенных множеств $B_i \neq \varnothing$, $i = \overline{2, n-1}$ исходного шаблона слова полученные множества D_T^H будут пустыми, в качестве D^* берем пустое множество.

ЭКСПЕРИМЕНТАЛЬНЫЙ АНАЛИЗ РАБОТЫ АЛГОРИТМА

Для реализации описанных алгоритмов разработана программная среда. Она позволяет осуществлять и анализировать работу алгоритмов для системы, состоящей из двух роботов.

Тестирование проводилось для одночастотной, двухчастотной и трехчастотной систем. В качестве алфавита было рассмотрено множество букв русского алфавита. Каждой букве поставлено в соответствие множество, состоящее из одной фонемы, которая, в свою очередь, состоит из одного фрагмента, длительностью 150 мс и с амплитудой 100. Частоты фрагментов, которые использовались для каждой буквы, приведены в таблице, размер словаря роботов составлял 150 слов.

Параметры алгоритма: $\tau = 50$, $a_{\min} = 50$, $\sigma = 0.9$, $\alpha = 0.7$, $\gamma = 0.5$, $a_n = 50$. Эти параметры вместе с выбором языка отвечали наилучшему качеству распознавания.

Эксперименты проводились в помещении общей площадью 20 кв.м. в условиях тишины и с шумом. В качестве шума использовался человеческий голос средней громкости. На разных расстояниях между микрофоном "распознающей" машины и динамиком "проигрывающей" машины десять раз проигрывался звуковой сигнал, соответствующий фразе «я робот я умею разговаривать на тональном акустическом языке я общаюсь с другими роботами используя мультичастотные аудиосигналы», после этого результат распознавания сравнивался с исходной фразой. Критерием качества работы алгоритма служило количество правильно распознанных слов (в %). Результаты эксперимента представлены на рис. 3.

Соответствие между буквами алфавита и частотами фрагментов

Буква	Одно-частотная система, Гц	Двух-частотная система, Гц	Трех-частотная система, Гц
а	400	400, 1000	400,800,1100
б	450	400, 1100	400,800,1200
в	500	400, 1200	400,800,1300
г	550	400, 1300	400,900,1100
д	600	400, 1400	400,900,1200
е	650	400, 1500	400,900,1300
ж	700	500, 1000	400,1000,1100
з	750	500, 1100	400,1000,1200
и	800	500, 1200	400,1000,1300
й	850	500, 1300	500,800,1100
к	900	500, 1400	500,800,1200
л	950	500, 1500	500,800,1300
м	1000	600, 1000	500,900,1100
н	1050	600, 1100	500,900,1200
о	1100	600, 1200	500,900,1300
п	1150	600, 1300	500,1000,1100
р	1200	600, 1400	500,1000,1200
с	1250	600, 1500	500,1000,1300
т	1300	700, 1000	600,800,1100
у	1350	700, 1100	600,800,1200
ф	1400	700, 1200	600,800,1300
х	1450	700, 1300	600,900,1100
ц	1500	700, 1400	600,900,1200
ч	1550	700, 1500	600,900,1300
ш	1600	800, 1000	600,1000,1100
щ	1650	800, 1100	600,1000,1200
ъ	1700	800, 1200	600,1000,1300
ы	1750	800, 1300	700,1000,1100
ь	1800	800, 1400	700,1000,1200
э	1850	800, 1500	700,1000,1300
ю	1900	900, 1000	700,800,1100
я	1950	900, 1100	700,800,1200

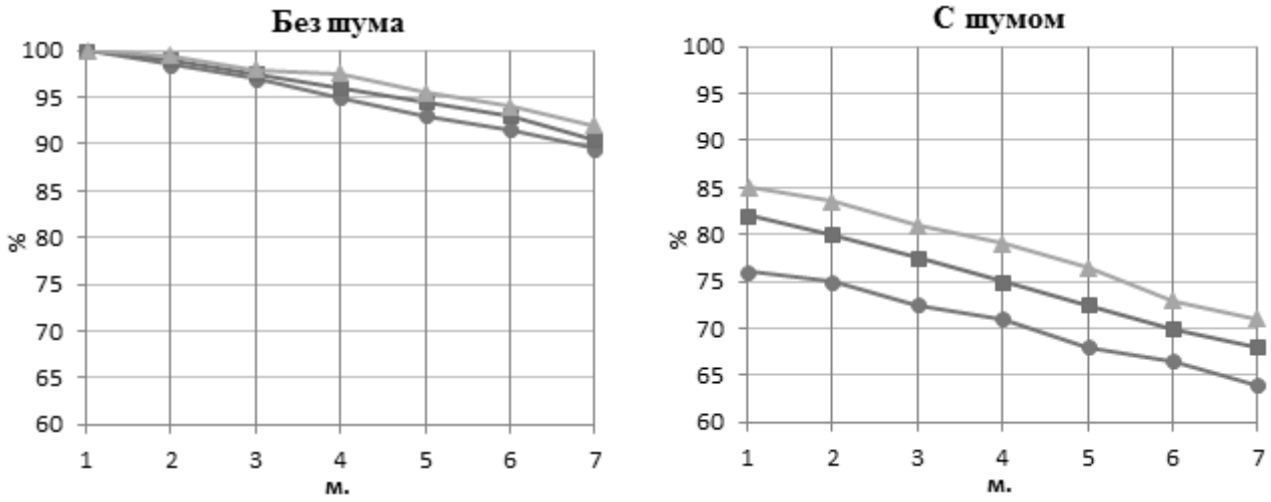

Рис.3. Результаты экспериментов. Круглый маркер – одночастотная система, квадратный – двухчастотная система, треугольный – трехчастотная система

Из результатов видно, что в условиях тишины качество распознавания систем с различной частотностью примерно одинаковое. В условиях посторонних шумов мультичастотная система оказалась надежнее.

ЗАКЛЮЧЕНИЕ

В работе описано построение модели акустического тонального мультичастотного языка роботов со словарём и представлен алгоритм выделения звуков, соответствующих фонемам этого языка.

Выполнена реализация программной среды моделирования. Проведенные эксперименты подтвердили более устойчивую работу системы в условиях посторонних шумов по сравнению с одночастотной системой. Словарь роботов позволил заметно улучшить качество распознавания. При этом благодаря использованию мультичастотных фрагментов удалось уменьшить процент ложного распознавания, что привело к более качественной работе системы при посторонних шумах. Так, качество распознавания у трехчастотной системы оказалось выше аналогичного показателя для двухчастотной системы, которая в свою очередь опережает одночастотную систему по этому показателю.

Передача акустических сигналов по надежности и скорости уступает передаче данных, например, по радиоканалу. Однако предложенная система может использоваться в тех случаях, где неприменимы радиосигналы, например, в водной среде. Были проведены первые эксперименты, полностью это подтвердившие.

В дальнейшем планируется создать версии акустического языка, которые будут работать на частотах ультразвука, создание таких версий языков, которые будут более устойчивы к помехам. Также планируется ввести некоторый понятийный словарь (тезаурус), которым будут пользоваться роботы. Для этого словаря будут созданы правила составления предложений, которые уже будут иметь специфический смысл.

СПИСОК ЛИТЕРАТУРЫ

1. Кирков А.Ю., Павловский В.Е. Акустический тональный язык коммуникации роботов // Научно-техническая информация. Сер. 2. - № 2. – 2013. – С. 8-15; Kirkov A. Yu., Pavlovskii V.E. The acoustic tonal robot communication language. // Automatic documentation and mathematical linguistics. – 2013. – Vol. 47, № 1. – P. 27-35.
2. Кирков А.Ю., Павловский В.Е. Искусственный тональный язык акустической коммуникации роботов // Искусственный интеллект. – 2012. – № 4. – С. 440-458.
3. Павловский В.Е., Кирков А.Ю. Тональная акустическая коммуникация роботов // Препринты ИПМ им. М.В.Келдыша. – 2012. – № 85. – 24 с. – URL: http://library.keldysh.ru/preprint.asp?id= 2012-85
4. Кирков А.Ю., Павловский В.Е. Искусственный мультичастотный язык акустической коммуникации роботов // Мультиконференция по управлению (МКПУ-2013), с. Дивноморское, Краснодарский край, 30 сентября – 5 октября 2013 г. – Ростов-на-Дону, 2013.
5. Павловский В.Е., Кирков А.Ю. Тональная мультичастотная акустическая коммуникация роботов // Препринты ИПМ им. М.В.Келдыша. – 2013. – № 2. – 32 с. (в печати).

Материал поступил в редакцию 23.09.13.

Сведения об авторах

ПАВЛОВСКИЙ Владимир Евгеньевич – доктор физико-математических наук, профессор, главный научный сотрудник Института прикладной математики им. М.В.Келдыша РАН, Москва
e-mail: vlpavl@mail.ru

КИРКОВ Андрей Юрьевич – аспирант Института прикладной математики им. М.В.Келдыша РАН, Москва
e-mail: kirkov.andrey@gmail.com

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’367’37

Д.А. Ильвовский

Применение семантически связанных деревьев синтаксического разбора в задаче поиска ответов на вопросы, состоящие из нескольких предложений

Рассматривается актуальная для многих прикладных областей проблема нахождения релевантных ответов на вопросы, состоящие из нескольких предложений. В частности, она возникает в промышленных системах, ориентированных на предоставление товаров и услуг. Один из основных подходов к данной проблеме заключается в том, что множество потенциальных ответов, полученное с помощью поиска по ключевым словам, повторно упорядочивается с помощью сопоставления деревьев синтаксического разбора ответов с деревом разбора, составленным для вопроса. В этой работе подход, основанный на применении деревьев разбора, был модифицирован и улучшен за счет перехода к более точному представлению семантико-синтаксической структуры текста: рассмотрению абзацев текста в качестве единицы анализируемой информации. Данный подход был программно реализован, а результаты реализации размещены в открытом доступе в виде надстройки для поисковой машины Apache SOLR, благодаря чему предлагаемая технология может быть легко интегрирована с промышленными поисковыми системами.

Ключевые слова: *чаща разбора, текстовый абзац, поиск ответов на вопросы, семантические отношения*

ВВЕДЕНИЕ

Современные поисковые машины недостаточно хорошо обрабатывают запросы, состоящие из нескольких предложений. Они находят либо очень похожие документы (если таковые имеются), либо документы, сильно отличающиеся от ожидаемого результата, что делает результаты поиска не слишком полезными для пользователя. Это обусловлено тем, что для запросов, состоящих из нескольких предложений, довольно трудно построить ранжирование, основанное на данных о пользовательских «кликах» по результатам, так как число такого рода запросов практически не ограничено. Поэтому необходима лингвистическая технология, которая бы переупорядочивала потенциальные ответы, используя структурное сходство между вопросом и ответом. В данном исследовании предлагается представление текстового абзаца, позволяющее отслеживать упомянутые структурные различия, используя не только информацию, содержащуюся в деревьях разбора, но и семантическую информацию, характеризующую абзац как лингвистическую структуру. Исследование ориентировано на обработку текстов на английском языке.

Использование абзацев текста в качестве запросов применяется, например, в основанных на поиске рекомендательных системах [1–3]. Рекомендательные агенты отслеживают действия пользователей чатов, блогов и форумов, комментарии пользователей на торговых сайтах и предлагают веб-документы и их фрагменты, относящиеся к решениям о покупке товара. Для формирования рекомендации агенты должны взять части текста, построить запрос для поисковой системы, запустить его с помощью API поисковой системы, такой как Yahoo или Bing, и отфильтровать нерелевантные по отношению к решению о покупке результаты поиска. Последний шаг имеет решающее значение для разумного функционирования агента, поскольку низкая релевантность приведет к утрате доверия по отношению к механизму рекомендаций. Поэтому нахождение точной оценки сходства между двумя частями текста имеет решающее значение для успешного использования рекомендательных агентов.

Деревья синтаксического разбора являются стандартной формой представления синтаксической структуры предложений [4–6]. В нашем исследовании для представления лингвистической структуры абзаца текста используются деревья разбора, конструируемые для каждого предложения абзаца. В рабо

те [7] было введено понятие Чащи Разбора (Parse Thicket), описывающей структуру текстового абзаца.

Определение 1. ***Чащей разбора*** *текстового абзаца называется совокупность множества деревьев разбора предложений абзаца и связей нескольких типов, устанавливаемых между вершинами деревьев. Каждая связь – это упорядоченная пара вершин деревьев разбора, причем связаны между собой могут быть как вершины одного дерева, так и вершины разных деревьев.*

Со структурной точки зрения чаща представляет собой ориентированный граф, который включает в себя деревья разбора, а также дуги, соответствующие несинтаксическим связям.

Определим операцию обобщения абзацев текста через обобщение соответствующих им ЧР. Использование обобщения для оценки сходства продолжает линию структурного подхода к машинному обучению [8–11], альтернативой которому является измерение статистического сходства как расстояния в пространстве признаков [12–15]. Применяемая в данной работе идея состоит в расширении понятия «наименее общего обобщения» (примером может служить антиунификация логических формул [16, 17] и т.д.) в направлении структурного представления текстовых абзацев и последующем использовании этой операции для вычисления сходства между состоящими из нескольких предложений вопросами и возможными ответами на них.

Обобщение абзацев текста основано на операции обобщения двух предложений [18, 19]. В дополнение к построению обобщений для отдельных предложений предпринимается попытка определить, как несинтаксические связи между словами в предложениях могут быть использованы для вычисления сходства между текстами. Для этого применяются специально построенные формализации семантических теорий, в частности, теории риторических структур [20].

1. ПРИМЕНЕНИЕ ЧАЩ РАЗБОРА ДЛЯ НАХОЖДЕНИЯ ОТВЕТОВ НА ВОПРОСЫ

Если мы построили последовательность деревьев разбора для вопроса и для ответа, как мы можем сопоставить их между собой? Существует ряд исследований, посвященных вопросу вычисления попарного сходства между деревьями разбора [4, 21, 22]. Тем не менее для того чтобы использовать связи внутри абзаца и избежать зависимости от распределения содержания по нескольким предложениям ответа, будем рассматривать абзац в целом (т.е. чащу разбора этого абзаца), а не просто отдельные предложения, входящие в этот абзац. В данной концепции для определения того, насколько удачным является ответ на вопрос, достаточно сопоставить чащи ответа и вопроса.

1.1. Расширенные группы

Для построения структуры абзаца синтаксические отношения, зафиксированные в деревьях разбора, дополним с помощью несинтаксических связей. В качестве таких связей используются:

- Кореферентные и таксономические связи:
  - ✓ анафора
  - ✓ «та же сущность»
  - ✓ «частный случай»
  - ✓ «более общий случай» и т.д.
- Связи, полученные с помощью применения семантических теорий.

Используя несинтаксические связи, мы можем расширить понятие синтаксической группы на случай нескольких предложений. При поиске сходства между отдельными предложениями сопоставляются именные, глагольные группы и другие виды групп, фигурирующие в предложениях. Несинтаксические связи между вершинами деревьев разбора позволяют объединять несколько групп из разных предложений или из одного предложения между собой. Таким образом, мы можем расширить понятие группы, допустив включение в группу одной или нескольких несинтаксических связей. Такие связи при обходе группы условно позволяют «перескакивать» с одного дерева разбора на другое. В данной работе рассматриваются следующие типы групп:

- Синтаксические или регулярные группы;
- Группы, включающие кореферентные (см., например, [23]) и таксономические связи. Также будем называть их *чащевыми* группами.
- RST-группы. Две группы (каждая из них может быть и чащевой, и синтаксической), соединенные RST-отношением.
- CA-группы. Здесь возможны два случая:
  - ✓ Синтаксическая или обычная группа с выделенным в ней коммуникативным действием.
  - ✓ Две группы (каждая из них может быть и чащевой, и синтаксической), объединенные связью между двумя коммуникативными действиями.

Для удобства все объединенные несинтаксическими связями синтаксические группы (чащевые, RST, CA) будем называть *расширенными группами*.

Рассмотрим пример, в котором добавление дополнительных кореферентных связей помогает правильно сопоставить ответ с вопросом:

Индексируемый текст 1: … *Tuberculosis is usually a lung disease. It is cured by doctors specializing in pulmonology.*
Индексируемый текст 2: … *Tuberculosis is a lung disease… Pulmonology specialist Jones was awarded a prize for curing a special form of disease.*
Question: *Which specialist doctor should treat my tuberculosis?*

В обоих случаях тексты содержат ключевые слова из вопроса. Но настоящим ответом является только первый текст. Понять это помогает установление связи *Tuberculosis → disease → is cured by doctors pulmonologists*.

Другие примеры расширенных групп приведены в разделах 1.3.

1.2. Различные подходы к выявлению сходства между текстовыми абзацами

Существуют различные подходы к оценке сходства между двумя абзацами текста (в рассматриваемых приложениях – вопросом и ответом):

- Применение ключевых слов: базовый подход, в котором тексты представляются в виде «мешка слов», а затем вычисляется набор общих ключевых слов / N-грамм и их частот [15].
- Попарное сравнение предложений: применяются синтаксические обобщения для каждой пары предложений, полученные результаты суммируются [19, 24].
- Попарное сопоставление абзацев текста [1, 19, 24].

Первый подход наиболее характерен для промышленного применения в современной компьютерной лингвистике. Второй подход был использован, например, в [19]. Ко второму подходу также относятся применение ядер деревьев разбора [21, 25] и ядер последовательностей деревьев [25] в алгоритмах классификации типа Метода Опорных Векторов (SVM) [26].

Рассмотрим и сравним перечисленные выше подходы на примере пары коротких текстов (статей). Первый текст можно рассматривать в качестве поискового запроса (причем он необязательно должен быть сформулирован в виде предложения в вопросительной форме), а второй текст – как потенциальный ответ на него. При этом необходимо помнить, что релевантный ответ должен быть тесно связан с текстом запроса, который в то же время не является копией запроса или его фрагмента.

Примечание. "^" в следующем примере и далее означает операцию обобщения двух абзацев. При описании деревьев разбора используется стандартная нотация, принятая для деревьев составляющих: [...] обозначает синтаксическую группу, NN, JJ, NP и т.д. – части речи и типы групп (существительное, прилагательное, именная группа и т.д.), * используется для обозначения произвольных вершин дерева. "Communicative action" обозначает коммуникативное действие, <leads to> – связь между коммуникативными действиями, "RST-evidence" – тип риторической связи (см. раздел 1.3.1).

"Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons",
"UN nuclear watchdog passes a resolution condemning Iran for developing a second uranium enrichment site in secret",
"A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons",
"Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US",
^
"UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claims that its nuclear research is for peaceful purpose",
"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret",
"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"

Список общих ключевых слов позволяет определить, что оба документа относятся к ядерной программе Ирана, однако понять на его основе что-то более конкретное весьма затруднительно.

Iran, UN, proposal, dispute, nuclear, weapons, passes, resolution, developing, enrichment, site, secret, condemning, second, uranium

Попарное обобщение предложений дает чуть более полную картину.

[NN-work IN-* IN-on JJ-nuclear NNS-weapons],

[DT-the NN-dispute IN-over JJ-nuclear NNS-*],

[VBZ-passes DT-a NN-resolution],

[VBG-condemning NNP-iran IN-*],

[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret],

[DT-* JJ-second NN-uranium NN-enrichment NN-site],

[VBZ-is IN-for JJ-peaceful NN-purpose],

[DT-the NN-evidence IN-* PRP-it],

[VBN-* VBN-fabricated IN-by DT-the NNP-us]

Обобщение с помощью чащ разбора дает существенно более детальную картину, чем результаты, полученные с помощью первых двух подходов (см. также далее рис. 3).

[NN-Iran VBG-developing DT-* NN-*enrichment* NN-site IN-in NN-secret]

[NN-*generalization*-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]

[NN-*generalization*-<Iran/envoy of Iran> *Communicative_action* DT-the NN-dispute IN-over JJ-nuclear NNS-*]

[Communicative_action - NN-work IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]

[NN-generalization <Iran/envoy to UN> *Communicative_action* NN-Iran NN-nuclear NN-* VBZ-is IN-for JJ-peaceful NN-purpose],

[*Communicative_actio*n - NN-*generalization* <work/develop> IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]

[NN-*generalization* <Iran/envoy to UN> *Communicative_action* NN-evidence IN-against NN Iran NN-nuclear VBN-fabricated IN-by DT-the NNP-us]

NN-Iran JJ-nuclear NN-weapon NN-* – *RST-evidence* – VBN-fabricated IN-by DT-the NNP-US

condemn^proceed [enrichment site] <leads to> suggest^condemn [work Iran nuclear weapon]

1.3. Несинтаксические связи, получаемые из семантических теорий

Для получения дополнительных несинтаксических связей были использованы и (частично) реализованы в виде программных компонент методы следующих семантических теорий, описывающих отношения внутри абзаца:

- Теория риторических структур (Rhetorical Structure Theory, сокр. RST) [20];
- Теория речевых актов (Speech Act Theory, сокр. SpActT) [27].

Хотя обе эти теории построены на психологических наблюдениях и имеют в основном невычислительный характер, для них были построены конкретные вычислительные реализации [19]. Для RST из текста извлекаются RST-отношения (риторические отношения). В случае SpActT для нахождения связей используется словарь так называемых коммуникативных действий (*communicative actions*) [18].

1.3.1. Пример использования риторической структуры

Рассмотрим представленный на рис. 1 пример обобщения на основе риторического отношения «evidence» (доказательство) [27]. Это соотношение имеет место между группами (перед группами указана их роль в риторическом отношении) *«доказательство-чего [Iran's nuclear weapon program]»* и *«что-происходит-с-доказательством [Fabricated by USA]»*, а также между фразами «*свидетельство-чего [against Iran's nuclear development]»* и «*что-происходит-с-доказательством [Fabricated by the USA]»*.

Нужно отметить, что в последнем случае необходимо объединить (путем разрешения анафоры) группу *«its nuclear development»* с группой *«evidence against it»,* чтобы получить *«evidence against its nuclear development»*. Анафорой в данном случае является связь *«it – development»*. *«Evidence»* удаляется из фразы, поскольку это индикатор риторического отношения. Чтобы получить итоговую фразу, необходимо разрешить еще одну анафору: *«its – Iran»*.

После обобщения двух групп, построенных на базе риторического отношения RST-evidence, мы получаем RST-группу «*Iran nuclear NNP – RST-evidence – fabricated by USA»*.

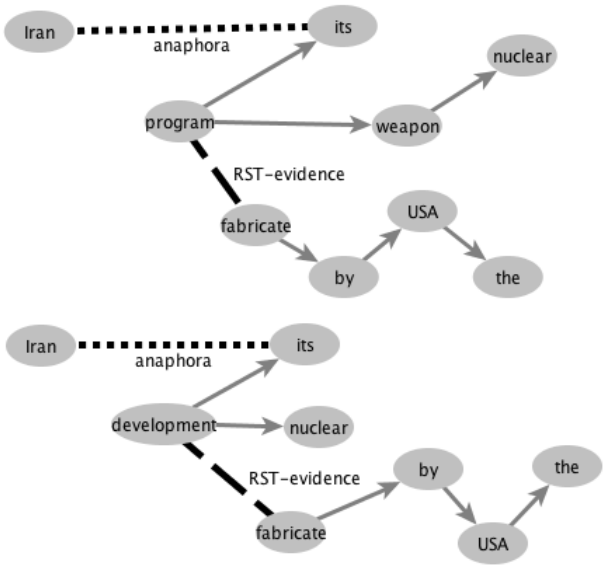

Рис. 1. Пример обобщения на основе риторического отношения RST-evidence.

1.3.2. Обобщение расширенных групп, использующих коммуникативные действия

Инструментарий глаголов – коммуникативных действий используется авторами текстов, для того чтобы показать структуру диалога или конфликта [28]. Поэтому добавление в чащу самих коммуникативных действий и связей, устанавливаемых между ними, позволяет отыскивать неявное сходство между текстами. При выполнении операции обобщения в этом случае применяются следующие правила:

1. Одно коммуникативное действие (глагол) и его субъект (подчиненная группа) из чащи T_1 можно обобщить с другим коммуникативным действием (глаголом) и его субъектом из чащи T_2. Дуга между коммуникативными действиями в этом обобщении не участвует.

2. Пару коммуникативных действий с их субъектами можно обобщить с другой парой коммуникативных действий и их субъектами из второй чащи. Связь между коммуникативными действиями включается в результат обобщения. Пример такого обобщения приведен на рис. 2.

3. При обобщении двух групп, построенных для коммуникативных действий, в первую очередь обобщаются их субъекты, затем – сами коммуникативные действия. Результат обобщения коммуникативных действий «прикрепляется» к результату обобщения их субъектов, представляющему собой множество наибольших общих поддеревьев. При этом сами коммуникативные действия всегда можно обобщить, но если результат обобщения субъектов является пустым множеством, то и соответствующие им расширенные группы тоже не обобщаются.

1.3.3. Пример использования коммуникативных действий

В примере, приведенном на рис. 2, мы имеем совпадающие коммуникативные действия с практически не совпадающими субъектами:

condemn [Iran for developing second enrichment site in secret]
vs
condemn [the work of Iran on nuclear weapon] ,

а также несовпадающие коммуникативные действия с очень похожими субъектами:

suggest [Iran was secretly working on nuclear weapons]
vs
condemn [the work of Iran on nuclear weapon]

Результатом обобщения в первом случае будет пустое множество, поскольку субъекты не обобщаются (см. правило 3). Во втором случае мы получим *suggest^condemn* [*work Iran nuclear weapon*].

Теперь, используя полученные результаты, попробуем обобщить приведенные выше пары коммуникативных действий между собой:

condemn [second uranium enrichment site] ↔ condemn [the work of Iran on nuclear weapon]
↓ *communicative action arcs* ↓
suggest [Iran is secretly working on nuclear weapon] ↔ proceed [develop an enrichment site in secret]

Такое обобщение дает пустое множество, поскольку, как было показано выше, *condemn [Iran for developing second enrichment site in secret]* и *condemn [the work of Iran on nuclear weapon]* не обобщаются.

T_1	T_2
condemn [second uranium enrichment site] ↔	proceed [develop an enrichment site in secret]
↓ *communicative action arcs* ↓	
suggest [Iran is secretly working on nuclear weapon] ↔ condemn [the work of Iran on nuclear weapon]	

Здесь результатом будет *condemn^proceed* [*enrichment site*] <*leads to*> *suggest^condemn* [*work Iran nuclear weapon*].

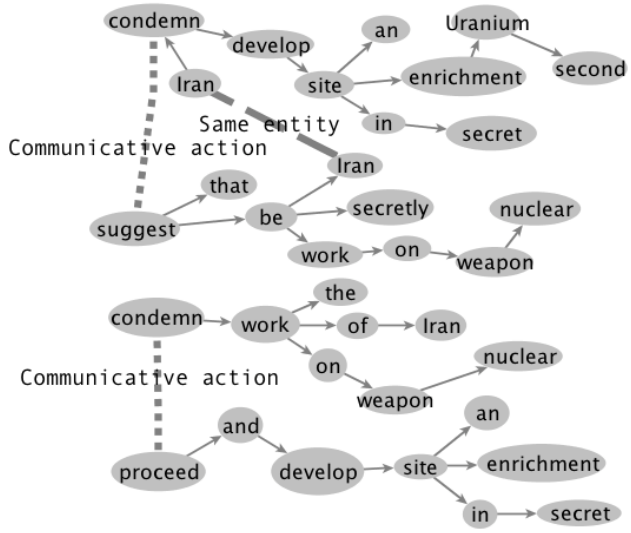

Рис. 2. Пример обобщения пар коммуникативных действий и их субъектов.

2. УЗОРНЫЕ СТРУКТУРЫ И ИХ ПРОЕКЦИИ

2.1. Анализ формальных понятий

Анализ формальных понятий (АФП) – это область прикладной теории решеток, методы которой используются для решения различных задач анализа и разработки данных [29]. Приведем основные определения АФП согласно [29].

Определение 2. ***Формальный контекст*** *– это тройка, в которой G это множество объектов, M – множество признаков, $I \subseteq G \times M$ – бинарное отношение между G и M .*

Между множествами подмножеств объектов и признаков можно задать соответствие Галуа, задаваемое с помощью двух отображений:

$$A' = \{m \in M \mid \forall g \in A, (g,m) \in I\}, \text{ для } A \subseteq G$$
$$B' = \{g \in A \mid \forall m \in M, (g,m) \in I\}, \text{ для } B \subseteq M$$

Данное соответствие сопоставляет множеству объектов максимальное множество признаков, каждый из которых находится в отношении с каждым объектом. Аналогично – для множества признаков. Например, $\{g_1, g_2\}' = \{m_4\}$, в то время как $\{m_4\}' = \{g_1, g_2, g_4\}$.

Соответствие Галуа лежит в основе определения формальных понятий и соответствующей им решетки формальных понятий.

Определение 3. ***Формальное понятие*** *– это пара (A,B), где A – это подмножество объектов, $A \subseteq G$; B – признаков, $B \subseteq M$, причем $A' = B$, а $A = B'$. Множество объектов A называют объемом, а множество признаков B – содержанием формального понятия (A,B).*

2.2. Узорные структуры

АФП преобразует формальный контекст, представленный как бинарное отношение, в решетку формальных понятий, но во многих случаях исследуемые «объекты» могут иметь более сложное описание, чем множество некоторых наперед заданных признаков. Например, исследуя множество объектов, возможно ли их исследовать без выделения специальных бинарных признаков? Узорные структуры дают ответ на этот вопрос, являясь расширением АФП для работы со сложными данными [30], такими как данные, описываемые численными значениями, множествами последовательностей или графов.

Определение 4. ***Узорная структура*** *– это тройка $(G,(D,\Pi),\delta)$, где G – множество объектов, (D,Π) – полная полурешетка всевозможных описаний, а $\delta : G \to D$ – функция, которая сопоставляет каждому объекту из множества G его описание из D .*

Полурешеточная операция $\sqcap$ соответствует операции сходства между двумя описаниями.

2.3. Проекции узорных структур

Поскольку размер решетки узорных понятий может быть существенным, а сама решеточная операция – вычислительно сложной (например, на узорной структуре на графах необходимо вычислять изоморфизм подграфов), построение решетки узорных понятий может занимать существенное время. Для сокращения времени работы алгоритмов построения узорных решеток были введены проекции узорных структур [30]. Проекция может быть рассмотрена как некоторый фильтр полурешетки описания с определенными математическими свойствами. Эти свойства позволяют доказать, что для любого понятия в спроецированной решетке существует соответствующее понятие в исходной решетке.

Определение 5. ***Проекция узорной структуры*** *– это функция* $\psi : D \to D$, *которая является монотонной* ($x \sqsubseteq y \Rightarrow \psi(x) \sqsubseteq \psi(y)$),, *сжимающей* ($\psi(x) \sqsubseteq x$) *и идемпотентной* ($\psi(\psi(x))\psi(x)$) [30].

Узорная структура проецируется следующим образом. Мы должны спроецировать функцию – описание объектов, а также полурешетку описаний:

$$\psi((G,(D,\Pi),\delta)) = (G,(D_\psi,\Pi_\psi),\psi \circ \delta \ ,$$

где $D_\psi = \psi(D) = \{d \in D \mid \exists d^* \in D : \psi(d^*) = d\}$

и $\forall x, y \in D, x \Pi_\psi y := \psi(x \Pi y)$.

3. АЛГОРИТМ ВЫЧИСЛЕНИЯ ПРИБЛИЖЕННОГО ОБОБЩЕНИЯ ЧАЩ РАЗБОРА

3.1. Проекции на чащах

Для того чтобы оценить структурное сходство двух текстовых абзацев, нам необходимо выполнить операцию обобщения на соответствующих им чащах разбора. Воспользуемся приведенными выше определениями из теории решеток и узорных структур, для того чтобы задать операцию обобщения.

Если рассматривать абзацы как *объекты*, а чащи разбора как их *описания*, то операция *обобщения* или сходства – это полурешеточная *операция пересечения*. Далее, если представить чащу в виде графа, то пересечение двух чащ наиболее естественным образом определяется как *множество наибольших общих подграфов* для соответствующих им графов. Выполнение данной операции является NP-трудной задачей [31], поэтому для эффективного вычисления с сохранением свойств операции мы можем воспользоваться введенным выше механизмом *проекций*.

Зададим проекцию чащи как *множество всех максимальных по вложению синтаксических и расширенных групп*, вычисленных для данного абзаца. Со структурной точки зрения такая проекция – это максимальные по вложению поддеревья графа с дополнительными свойствами. Пример проекций для двух абзацев приведен в разделе 1.2.

Операция пересечения двух групп определяется внутри каждого типа групп [19]. Пересечение на проекциях заключается в попарном пересечении групп для каждого типа из двух множеств и выборе наибольших по вложению подгрупп. Примеры пересечения для RST-групп и CA-групп приведены в разделах 1.3.2 и 1.3.3 соответственно. Работа с проекциями позволяет добиться экономии по сложности (переход к работе с деревьями) без ущерба для качества результата (группы учитывают все необходимые связи внутри абзаца).

3.2. Построение множества расширенных групп

Рассмотрим подробнее алгоритм нахождения проекции, т.е. множества расширенных групп, для чащи разбора.

Для каждого предложения S в абзаце P:

1. Сформировать список предшествующих предложений в абзаце S_{prev}
2. Для каждого слова в текущем предложении:

2.1 Если это местоимение (*pronoun)*: с помощью разрешения анафоры найти все существительные и именные группы в S_{prev}, которые связаны с данным словом отношением «та же сущность».

2.2 Если это существительное (*noun)*: найти все существительные и именные группы в S_{prev}, которые связаны с данным словом:

- ✓ отношением «та же сущность» (через разрешение анафоры)
- ✓ отношением синонимии
- ✓ отношением «более общий случай»
- ✓ отношением «частный случай»
- ✓ имеют общую родительскую (связанную отношением «более общий случай») сущность

2.3 Если это глагол (*verb*):

2.3.1 Если он выражает собой коммуникативное действие (*communicative action*):

2.3.1.1 Сформировать группу $VBCA_{phrase}$, включающую в себя глагольную группу данного слова VB_{phrase}.

2.3.1.2 Найти предыдущее коммуникативное действие с его субъектом $VBCA_{phrase0}$ в $S_{prev.}$

2.3.1.3 Построить расширенную группу [$VBCA_{phrase0}$, $VBCA_{phrase}$].

2.3.2 Если он указывает на риторическое отношение (*RST relation*):

2.3.2.1 Сформировать из фраз – субъектов отношения расширенную группу [$VBRST_{phrase1}$, $VBRST_{phrase2}$] ; фраза $VBRST_{phrase1}$ относится к S_{prev}.

3.3. Обобщение чащ на проекциях

Для того чтобы обобщить пару абзацев, необходимо:

1. Выполнить их фрагментацию и извлечь все синтаксические группы из каждого предложения.
2. Найти семантические связи внутри абзаца.
3. Используя семантические связи, построить на основе синтаксических групп расширенные группы.
4. Провести обобщение для каждого из четырех типов (см. раздел 1.1) групп, заключающееся в поиске множества наибольших общих подгрупп для каждой пары групп одного и того же типа.

5. Полученные на уровне групп обобщения можно интерпретировать как набор путей в результирующих общих поддеревьях [19].

4. ОПИСАНИЕ ЭКСПЕРИМЕНТОВ

Попробуем оценить, как обобщение чащ разбора может улучшить поиск в ситуации, когда сформулированный в виде поискового запроса вопрос и потенциальный ответ на него представляют собой текстовые абзацы. Количественным результатом оценки является точность в процентах, вычисленная как среднее по 100 поисковым запросам. В данной работе использовались параметры проведения эксперимента, описанные в [19].

Оценка основана на повторном ранжировании поисковых результатов, полученных с помощью API поисковой системы Bing, осуществляемом на базе меры сходства чащ разбора. Сходство определяется как общее число вершин в наибольшем общем подграфе для двух чащ. Приближенное значение этой оценки было получено путем подсчета количества слов в наибольших общих подгруппах с учетом весов для частей речи [19].

Полученные результаты приведены в таблице.

Для оценки были смоделированы следующие ситуации в трех прикладных областях:

- Выдача рекомендаций по товарам: агент рекомендательной системы читает чаты о продуктах и находит в сети релевантную информацию о каждом продукте.
- Выдача рекомендаций по путешествиям: агент рекомендательной системы читает чаты с описанием путешествий и находит в сети релевантную информацию об отелях, курортах и т.д.
- Выдача рекомендаций в социальных сетях (на примере Facebook): агент рекомендательной системы читает чаты и записи на «стене» пользователей социальной сети и отбирает информацию, которая может быть интересна друзьям этих пользователей.

Результаты экспериментов

Тип запроса	Сложность запроса	Релевантность исходного поиска в Bing, %,	Релевантность поиска с использованием обобщений для отдельных предложений, %,	Релевантность поиска с помощью чащ, построенных на **фрагментах**, %,	Релевантность поиска с помощью чащ, построенных на **оригинальных абзацах**, %,
Поиск рекомендаций по товарам	1 составное предложение	62.3	69.1	72.4	72.9
	2 предложения	61.5	70.5	71.9	72.8
	3 предложения	59.9	66.2	72.0	73.4
	4 предложения	60.4	66	68.5	69.2
Поиск рекомендаций по путешествиям	1 составное предложение	64.8	68	72.6	74.7
	2 предложения	60.6	65.8	73.1	76.9
	3 предложения	62.3	66.1	70.9	70.8
	4 предложения	58.7	65.9	72.5	73.9
Поиск рекомендаций контента на Facebook	1 составное предложение	54.5	63.2	65.3	68.1
	2 предложения	52.3	60.9	62.1	63.7
	3 предложения	49.7	57	61.7	63.0
	4 предложения	50.9	58.3	62.0	64.6
Средние показатели		**58.15**	**64.75**	**68.75**	**70.33**

В каждой из этих областей на основе Интернет-источников выбирался кусок текста, далее формировался запрос, а затем осуществлялась фильтрация результатов поиска, полученных с помощью API поисковой системы Bing.

Точность исходной поисковой выдачи составила 58,2%, применение операции обобщения на уровне предложений дало улучшение в 6,5%. Использование чащ для выдаваемых поисковой системой фрагментов (сниппетов) позволило увеличить точность еще на 4%. Наконец, применение чащ, построенных для абзацев, взятых непосредственно из оригинальных документов, дало прибавку еще в 1,5%. Нетрудно также видеть, что с ростом сложности запроса увеличивался эффект от применения технологии обобщения.

5. ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ И МАСШТАБИРУЕМОСТЬ ПРЕДЛАГАЕМОГО ПОДХОДА

Операция обобщения на деревьях разбора и чащах разбора определяется как нахождение всех наибольших общих поддеревьев и подчащ соответственно. Хотя для деревьев эта проблема решается за *O(N)*, для графа общего вида она является NP-трудной [31].

Один из подходов к обучению на деревьях разбора основан на так называемых ядрах, определенных для дерева (*tree kernel*). Авторы этого подхода предлагают технику, ориентированную специально на деревья разбора, уменьшая тем самым размерность пространства всех возможных поддеревьев. Существует несколько специальных разновидностей ядер, направленных на более эффективную обработку деревьев. Частичные ядра (*partial tree kernels*) задают правила частичного соответствия, игнорирующие некоторые дочерние узлы [21]. Ядра последовательностей на деревьях (*tree sequence kernels*) используют в качестве подструктуры не просто поддеревья, а последовательности поддеревьев [32].

Подход, основанный на сопоставлении синтаксических групп для предложений и расширенных групп для чащ разбора, с вычислительной точки зрения оказывается гораздо более эффективным, чем подходы, использующие ядра на графах разного вида, включая деревья. Вместо того чтобы рассматривать пространство всех возможных подграфов, рассматриваются пути в деревьях и графах, которые соответствуют синтаксическим и расширенным группам.

Для того чтобы оценить сложность обобщения двух чащ разбора, рассмотрим абзац, состоящий из 5 предложений, каждое из которых имеет длину в 15 слов. В таких чащах в среднем содержится 10 синтаксических групп в каждом предложении и 10 дуг между предложениями, которые дают нам до 40 расширенных групп. Поэтому для сопоставления таких чащ разбора необходимо попарно обобщить около 50 синтаксических групп и 40 расширенных групп из одной чащи с таким же множеством групп для другой. С учетом обобщения отдельных существительных и глагольных групп это составляет порядка 2 * 45 * 45 обобщений, сопровождаемых проверкой вхождения результатов друг в друга. Каждое обобщение состоит не более чем из 12 сравнений строк, если принять средний размер группы за 5 слов. Следовательно, в итоге среднее обобщение двух чащ включает в себя 2 * 45 * 45 * 12 * 5 операций. Так как сравнение строк занимает несколько микросекунд, обобщение занимает в среднем 100 миллисекунд без использования индекса. Однако в промышленной поисковой системе, где группы хранятся в обратном индексе, операция обобщения может быть выполнена за фиксированное время, не зависящее от размера индекса [33].

ЗАКЛЮЧЕНИЕ

В работах [19, 34] было показано, как использование богатого набора лингвистической информации – синтаксических связей между словами – улучшает релевантность поиска. Для того чтобы, помимо синтаксической информации, воспользоваться и семантической, было использовано понятие чащи разбора и предложен способ вычисления сходства между текстами, основанный на обобщении соответствующих им чащ разбора. Также была построена и применена к задаче повторного ранжирования результатов поиска по ключевым словам технология обобщения чащ разбора, представляемых в виде наборов групп.

Для построения чащ использовались различные виды связей между словами в предложениях: кореферентные связи, таксономические отношения, такие как «быть частным случаем», «быть обобщением», «та же сущность» и т.д., а также семантические связи, полученные на базе теории риторических структур и теории речевых актов. Было показано, что если ответ содержится в нескольких предложениях, то применение чащи позволяет повысить релевантность поиска.

Также было показано, что операция сходства или обобщения абзацев текста может быть естественным образом определена с помощью математического аппарата узорных структур. При этом обобщение с использованием общих подграфов точно описывается в терминах пересечения описаний объектов, а синтаксические и расширенные группы соответствуют взятию проекций от исходных описаний.

Традиционное машинное обучение на языковых структурах ограничено работой с формами и частотами ключевых слов. В то же время большинство семантических теорий не являются вычислительными, они моделируют определенный набор отношений между последовательными состояниями. В данной работе была предпринята попытка совместить два подхода: использовать всю информацию, полученную из дерева синтаксического разбора, дополнив ее сведениями из семантических теорий, допускающих вычислительную обработку.

Архитектура системы в целом была построена на базе OpenNLP [35], а компонента, отвечающая за вычисление обобщения, является отдельным Apache Software Foundation проектом и принимает на вход данные в форматах OpenNLP и Stanford NLP [36]. Код и библиотеки системы доступны на http://code.google.com/p/ rele-

vance-based-on-parse-trees и http://svn.apache.org/repos/asf/ opennlp/sandbox/opennlp-similarity/.

Система готова для подключения к библиотеке Lucene. Кроме того, в состав системы включен обработчик запросов SOLR, и программные инженеры могут переключиться на поиск по нескольким предложениям с использованием чащ для проверки роста релевантности.

СПИСОК ЛИТЕРАТУРЫ

1. Bhasker B., Srikumar K. Recommender Systems in E-Commerce // CUP. – 2010.
2. Thorsten H., Marchand A., Marx P. Can Automated Group Recommender Systems Help Consumers Make Better Choices? // Journal of Marketing. – 2012. – Vol. 76 (5). – P. 89–109.
3. Montaner M. Lopez B. de la Rosa J. L. A Taxonomy of Recommender Agents on the Internet // Artificial Intelligence Review. – 2003. – Vol. 19 (4). – P. 285–330.
4. Punyakanok V., Roth D., Yih W. The Necessity of Syntactic Parsing for Semantic Role Labeling // IJCAI-05. – 2005.
5. Domingos P., Poon H. Unsupervised Semantic Parsing // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. – 2009. – Singapore, ACL.
6. Abney S. Parsing by Chunks // Principle-Based Parsing. Kluwer Academic Publishers. – 1991. – P. 257–278.
7. Galitsky B., Usikov D., Kuznetsov S. O. Parse Thicket Representations for Answering Multi-sentence questions // 20th International Conference on Conceptual Structures, ICCS 2013. – 2013.
8. Mill J.S. A system of logic, ratiocinative and inductive. – London, 1843.
9. Finn V.K. On the synthesis of cognitive procedures and the problem of induction // NTI. Series 2. – 1999. – № 1–2. – P. 8–45.
10. Mitchell T. Machine Learning // McGraw Hill. – 1997.
11. Furukawa K. From Deduction to Induction: Logical Perspective. The Logic Programming Paradigm / eds. K. R. Apt, V. W. Marek, M. Truszczynski, D. S. Warren. – Springer, 1998.
12. Fukunaga K. Introduction to statistical pattern recognition (2nd ed.) // Academic Press Professional Inc. – San Diego, CA, 1990.
13. Jurafsky D., Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 2008.
14. Byun H., Lee S. Applications of Support Vector Machines for Pattern Recognition: A Survey // Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines (SVM '02), Seong-Whan Lee and Alessandro Verri (Eds.). – 2002. – Springer-Verlag. London, UK. – P. 213–236.
15. Manning C., Schütze H. Foundations of Statistical Natural Language Processing // MIT Press. – 1999. – Cambridge, MA.
16. Robinson J.A. A machine-oriented logic based on the resolution principle // Journal of the Association for Computing Machinery. – 1965. – Vol. 12. – P. 23–41.
17. Plotkin G.D. A note on inductive generalization // B. Meltzer and D. Michie (Eds.). Machine Intelligence. – 1970. – Vol. 5. Elsevier North-Holland, New York. – P. 153–163.
18. Galitsky B., Kuznetsov S. Learning communicative actions of conflicting human agents // J. Exp. Theor. Artif. Intell. – 2008. – Vol. 20(4). – P. 277–317.
19. Galitsky B., de la Rosa J., Dobrocsi G. Inferring the semantic properties of sentences by mining syntactic parse trees // Data & Knowledge Engineering. – 2012. – Vol. 81–82. – P. 21–45.
20. Mann W., Matthiessen C., Thompson S. Rhetorical Structure Theory and Text Analysis // Discourse Description: Diverse linguistic analyses of a fund-raising text / ed. by W. C. Mann and S. A. Thompson. – Amsterdam, 1992. – P. 39–78.
21. Moschitti A. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees // Proceedings of the 17th European Conference on Machine Learning. – Berlin, 2006.
22. Collins M., Duffy N. Convolution kernels for natural language // Proceedings of NIPS. – 2002. – P. 625–632.
23. Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky, D. Deterministic coreference resolution based on entity-centric, precision-ranked rules // Computational Linguistics. – 2013. – Vol. 39(4).
24. Galitsky B., Ilvovsky D., Kuznetsov S. O., Strok F. Matching sets of parse trees for answering multi-sentence questions // Proceedings of the Recent Advances in Natural Language Processing, RANLP 2013. – INCOMA Ltd., Shoumen, Bulgaria. – 2013. – P. 285–294.
25. Zhang M., Che W., Zhou G., Aw A., Tan C., Liu T., Li S. Semantic role labeling using a grammar-driven convolution tree kernel // IEEE transactions on audio, speech, and language processing. – 2008. – Vol. 16 (7). – P. 1315–1329.
26. Vapnik V. The Nature of Statistical Learning Theory. – Springer-Verlag, 1995.
27. Searle J. Speech acts: An essay in the philosophy of language. – Cambridge: Cambridge University, 1969.
28. Marcu D. From Discourse Structures to Text Summaries // Proceedings of ACL Workshop on Intelligent Scalable Text Summarization / eds. I. Mani and M. Maybury. – Madrid, 1997. – P. 82–88.
29. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. 1st edn. – Secaucus, NJ, USA: Springer, 1997. – P. I–X, 1–284.

30. Ganter B., Kuznetsov S. O. Pattern Structures and Their Projections ICCS '01. – 2001. – P. 129–142.
31. Kann V. On the Approximability of the Maximum Common Subgraph Problem // In (STACS '92) / eds. Alain Finkel and Matthias Jantzen. – 1992. – Springer-Verlag, London, UK. – P. 377–388.
32. Sun J., Zhang M., Lim Tan C. Tree Sequence Kernel for Natural Language // AAAI-25. – 2011.
33. Dean J. Challenges in Building Large-Scale Information Retrieval Systems // URL: research.google.com/people/jeff/WSDM09-keynote.pdf .
34. Galitsky B. Machine Learning of Syntactic Parse Trees for Search and Classification of Text // Engineering Application of AI. // URL: http://dx.doi.org/10.1016/j.engappai.2012.09.017.
35. Kottmann J., Ingersoll G., Kosin J., Galitsky B. // The Apache OpenNLP library // URL: http://opennlp.apache.org/documentation/ 1.5.3/ manual/opennlp.html.
36. The Stanford Natural Language Processing Group // URL: http://nlp.stanford.edu/.

Материал поступил в редакцию 27.09.13.

Сведения об авторе

ИЛЬВОВСКИЙ Дмитрий Алексеевич – аспирант Национального исследовательского университета – Высшая школа экономики (НИУ ВШЭ), младший научный сотрудник научно-учебной лаборатории интеллектуальных систем и структурного анализа НИУ ВШЭ, Москва.
e-mail: dilv_ru@yahoo.com

УДК 81’42

Н. Ю. Муравьева

«Образ автора» в художественном тексте с позиций лингвистического анализа

Рассматривается зависимость интерпретации художественного текста от последовательного отражения (фиксации) модусной составляющей значения слов в толковых словарях.

Ключевые слова: *«образ автора», субъект речи, лексическая семантика, модусный компонент лексического значения*

Выдающийся русский филолог XX в. академик В. В. Виноградов в работе «О теории художественной речи» ввел одно из базовых понятий литературоведения – «образ автора», определив его как «концентрированное воплощение сути произведения, объединяющее всю систему речевых структур персонажей в их соотношении с повествователем рассказчиком или рассказчиками и через них являющееся идейно-стилистическим средоточием, фокусом целого» [1, с. 181]. При этом введенный для удобства анализа и интерпретации литературных текстов «образ» выделяется В. В. Виноградовым на основании языковых характеристик: это центр «системы речевых структур», т.е. определенный субъект речи, или позиция, точка зрения, с которой ведется описание. Все остальные черты, включая номинации, являются дополнительными и необязательными.

В связи с проблемой разграничения субъектов, разных точек зрения (героев – автора – читателя) в филологических работах XX в. использовались такие термины, как «сказ» [2, с. 154–155 и др.], «рассказчик», «повествователь» (например: [3, с. 254; 4, с. 126; 5, с. 33–35 и др.]), «лирический герой» [6, с. 119 и др.]. Были выделены приемы «остранения» (подчеркнутого несовпадения разных сознаний в повествовательном тексте [7, с. 109]), «опрокинутых фраз» (связанных с ироническим эффектом подобного несовпадения [8, с. 126–127]). В работах М. М. Бахтина встречаем термины «чужое сознание», «чужое слово», «двуголосие» [9, с. 215, 220–231], «зона героя». Последняя «образуется из полуречей… из различных форм скрытой передачи чужого слова, из рассеянных слов и словечек чужой речи, из вторжений в авторскую речь чужих экспрессивных моментов (многоточий, вопросов, восклицаний)» [4, с. 129–130]. Г. А. Гуковский, Ю. М. Лотман, Б. А. Успенский рассматривали проблему точки зрения как важный аспект структуры художественного текста.

Итак, «образ автора» предполагает наличие в художественном тексте системы «точек зрения». Причем сменяющие друг друга субъекты речи могут быть охарактеризованы (или противопоставлены) на уровне психологии, идеологии, пространственной и временной позиции, а также фразеологии (в понимании Б. А. Успенского [10]). В настоящей статье мы сосредоточим внимание не столько на поэтике художественного текста, сколько на его лингвистическом анализе и рассмотрим, с помощью каких языковых средств может быть обозначен субъект речи в конкретном художественном тексте.

Материалом исследования служит повесть В. Г. Короленко «Слепой музыкант». В литературоведческом понимании «образа автора» в этом тексте нет: повествователь не назван, он не входит в систему персонажей. Однако нельзя говорить и о полном отсутствии автора в повести. Внимательное прочтение помогает обнаружить ряд фрагментов, в которых автор оказывается в кругу своих героев, в одном времени и пространстве с ними. Например, в приводимом далее отрывке В. Г. Короленко упоминает четырех персонажей: трех слепых (в том числе Кандыбу и слепого музыканта), которые направляются за Чудом исцеления в Почаевскую обитель, и господина, проезжающего мимо них в бричке, который не назван, но описан так, что читатели узнают в нем дядю главного героя Максима (в данном случае он тайно оберегает своего воспитанника в первом самостоятельном путешествии). Однако есть и пятый присутствующий, формально никакого имени не имеющий: тот, кто видит этого проезжающего господина. Тот, кто описывает его как *виднеющуюся фигуру, квадратную фигуру, седого господина с торчащими сбоку костылями.* Таким образом, можно говорить о присутствии в пространственно-временном мире героев еще одного субъекта, который повествует о непосредственно им самим наблюдаемых событиях. Ведь глаголы *виднеться* и *торчать* указывают на наличие расстояния между тем, кто видит, и объектом восприятия, а глагол *появиться* – на временную характеристику, последовательность сиюминутно наблюдаемых действий и событий. И тот, кого видят, и тот, кто видит, оказываются в одном (художественном) мире. Это позволяет говорить о существовании в тексте «образа автора» не только как субъекта речи и творца композиционной структуры текста, но и как субъекта пространственно-временной «точки зрения». Средства

обозначения этого субъекта – в первую очередь лексические: слова разных частей речи, содержащие в своем значении сему восприятия. Причем лексемы с семантикой определенного типа восприятия оказываются в данном тексте маркерами субъекта речи (автора или героя). В приведенном ниже отрывке **полужирным** шрифтом выделены показатели слухового восприятия, **полужирным с подчеркиванием** – зрительного, только подчеркиванием – остальные случаи (синтез или неразличение слуховых и зрительных анализаторов, а также внутренние ощущения):

<…> Часам к десяти они ушли далеко. Лес остался ***синей полосой на горизонте****. Кругом была степь, и* ***впереди слышался звон*** *разогреваемой солнцем проволоки на шоссе, пересекавшем пыльный шлях. Слепцы вышли на него и повернули вправо, когда сзади* ***послышался топот*** *лошадей и* ***сухой стук кованых*** *колес* ***по щебню****. Слепцы выстроились у края дороги. Опять* ***зажужжало*** *деревянное колесо* ***по струнам****, и* ***старческий голос затянул****:*

– Под-дайте сли-пеньким… – К ***жужжанию*** *колеса* ***присоединился тихий перебор струн*** *под пальцами юноши.*

Монета ***зазвенела*** *у самых ног старого Кандыбы.* ***Стук*** *колес* ***смолк,*** *видимо, проезжающие остановились, чтобы посмотреть, найдут ли слепые монеты. Кандыба сразу нашел ее, и на лице* ***появилось довольное выражение****.*

– ***Бог спасет,*** *– сказал он по направлению к бричке, в сиденье которой* ***виднелась квадратная фигура седого*** *господина, и два костыля* ***торчали сбоку****. <…>*

Наличие в этом фрагменте зрительных номинаций помогает нам обнаружить автора, замечающего и фиксирующего то, что физически не могут увидеть герои. Он оказывается еще одним, не названным участником описанной сцены. В то же время все слуховые лексемы свидетельствуют о смене речевого субъекта, «передаче» повествования («права» голоса и мировосприятия) главному герою – слепому музыканту, остро воспринимающему весь новый для него окружающий мир на слух.

Слова, помогающие зафиксировать в тексте определенного субъекта речи, в лингвистике принято связывать с понятиями эгоцентризма, субъективности, языковой прагматики, модусной* информации в семантике. В данном случае речь идет о семантике лексической, которая должна быть отражена в словарях. Однако существующая лексикографическая практика свидетельствует о ряде трудностей, возникающих при толковании лексем с модусным компонентом значения. Так, глагол *виднеться* толкуется как *‘быть видным, заметным для зрения’* [13], *‘быть видным, быть доступным зрению’* [14]. Оба толкования не имеют указания на наличие помехи (например, расстояния) между воспринимаемым объектом и воспринимающим субъектом, в то время как знание об этом условии обязательно для употребления слов. Толкование глагола *слышаться – ‘быть слышным, звучать, восприниматься’* [13, 14] – также не имеет указания на наличие помехи между воспринимаемым объектом и воспринимающим субъектом. В толковании существительного *горизонт – ‘линия кажущегося соприкосновения неба с земной или водной поверхностью’* [14] – содержится указание на субъективность восприятия, однако отсутствует отсылка именно к визуальной характеристике положения вещей.

Некорректность толкования подобных слов свидетельствует о том, что семантическая природа лексем с модусным компонентом в значении различна. В приведенном выше фрагменте текста можно выделить как минимум три типа лексем с модусным компонентом в семантике, причем в существующих словарных толкованиях наличие (или корректность) описания этого компонента последовательно уменьшается. Первую группу образуют слова, в чьей семантике модусная, субъективная, ситуативная информация явно доминирует над объективной (диктумной), а именно глаголы и существительные, в лексическом значении которых есть модусно-перцептивные компоненты *‘звук’*, *‘шум’*, *‘звучание’*, *‘стать видным’*, являющиеся показателями сиюминутного наблюдения ситуации: *(по)слышаться, звон, топот, зажужжать, голос, затянуть, зазвенеть, стук, смолкнуть, появилось (выражение на лице).* Именно эти лексемы оказываются маркирующими перцептивный модус в текстовом фрагменте, а значит, доказывающими наличие конкретного наблюдателя. Лексической семантике помогает грамматика: во-первых, приставка *за-* в употребленных лексемах добавляет значение *‘начало’*, т.е. речь идет о начале восприятия звука (и таким образом исключается возможность обозначения постоянного свойства предмета, потенциально возможного в ином контексте). Поэтому не только лексическое значение корня, но и семантику модифицирующей приставки в данных глаголах можно считать модусной. Во-вторых, все глаголы характеризуются совершенным видом, что свидетельствует о чьем-то личном восприятии: категория вида передает не только последовательно сменяющиеся ситуации и действия героев, но и соприсутствие наблюдающего персонажа.

Вторую группу образуют слова, в семантике которых модусная информация составляет меньшую часть по сравнению с диктумной и в большинстве случаев сводится к эгоцентрическому элементу: *далеко, впереди, сзади, довольное выражение (на лице), квадратная фигура, седой, синий, сбоку.* Пространственные наречия характеризуют ситуацию относительно говорящего-наблюдателя: именно от него *далеко*, по отношению к его движению *впереди* или *сзади*. Определения *довольный* и *квадратный*,

* Термин «модус» ввел Ш. Балли: предмет нашей речи, или действительность, о которой мы говорим, вслед за логиками, была названа диктумом; а то, как мы о ней говорим, как ее преподносим, какое к ней отношение имеем – модусом. «Модус и диктум дополняют друг друга. <…> …уверенность предполагает наличие объекта уверенности, и наоборот: не может быть объекта уверенности без самого акта веры» [11, с. 44, 46]. Далее в лингвистике была предложена типология модусной информации, выделен, в том числе, модус перцепции, восприятия [12, с. 109 и далее].

будучи изначально зрительными характеристиками, передают впечатление конкретного соприсутствующего наблюдателя, поскольку первое характеризует лицо, а второе фигуру и ни для того, ни для другого такой признак не может быть постоянным. Цветовые прилагательные указывают на частноперцептивные характеристики: цвет (подобно свету, звуку, запаху и т. д.) может быть воспринят исключительно одним конкретным органом восприятия – глазом, что снова приводит нас к конкретному внутритекстовому наблюдателю.

Третья группа состоит из слов, употребленных в данном фрагменте в своих переносных, неосновных значениях, имеющих дополнительные контекстные прочтения, связанные с непосредственно воспринимаемыми кем-либо (на слух или зрительно) характеристиками:

(лес) остался полосой (на горизонте) – это указание на большое расстояние до леса и его конфигурацию (растянутость, значительную площадь);

(шоссе), пересекающее (пыльный шлях) – это определенный рисунок дорог;

(сказал он) по направлению к (бричке) – определенный разворот головы и тела героя;

(к жужжанию колеса) присоединился (тихий) перебор струн – звуковая характеристика;

сухой (стук) кованых колес по щебню, старческий (голос) – здесь, оказавшись пояснением к звуковому существительному, изначально «нейтральные» существительные и прилагательные начинают характеризовать различные типы звуковых оттенков.

Для слов третьей группы в словарях редко отмечается отдельное переносное значение, так как их модусная семантика приравнивается к контекстному употреблению. Тем не менее переносные и контекстные значения (и употребления) в большинстве случаев связаны с модификацией уже имеющихся компонентов семантики. Следовательно, оказывается важным выявить потенциальную предрасположенность слова к выражению модусных смыслов. И лексикографически ее зафиксировать[**].

В целом, знание о модусных компонентах значения является необходимым и при употреблении слова в речи, и при анализе и интерпретации художественного произведения, в особенности при переводе, а также при изучении языка как иностранного.

* * *

«Образ автора» присутствует в любом художественном тексте: в каждом произведении есть «некоторая точка зрения на излагаемое, точка зрения психологическая, идеологическая и попросту географическая», поскольку «невозможно описать ниоткуда и не может быть описания без писателя» [16, с. 200]. Но если литературоведение интересует в первую очередь эксплицитно представленный (названный, введенный в систему персонажей) «образ автора», то лингвистика готова обнаруживать и исследовать имплицитные средства создания этого «образа» (субъекта речи).

[**] Модель учета модусной информации в лексикографическом описании для двух первых типов лексем уже разработана и была применена в «Новом объяснительном словаре синонимов русского языка» [15] – издании, рассчитанном на профессионально подготовленного читателя и предлагающем анализ порядка 300 синонимических рядов на более чем 1500 (!) страницах.

СПИСОК ЛИТЕРАТУРЫ

1. Виноградов В. В. О теории художественной речи. – М. : Высшая школа, 1971. – 240 с.
2. Эйхенбаум Б.М. Иллюзия сказа // Эйхенбаум Б. М. Сквозь литературу. – Л. : Academia, 1924. – С. 153–156.
3. Бахтин М. М. Автор и герой в эстетической деятельности // Бахтин М. М. Работы 20-х годов. – Киев, 1994. – С. 69–255.
4. Бахтин М. М. Слово в романе // Бахтин М. М. Вопросы литературы и эстетики. – М. : Худож. лит-ра, 1975. – С. 72–233.
5. Корман Б. О. Изучение текста художественного произведения. – М. : Просвещение, 1972. – 113 с.
6. Тынянов Ю. Н. Блок // Тынянов Ю. Н. Поэтика. История литературы. Кино. – М. : Наука, 1977. – С. 118–123.
7. Шкловский В. Б. Материал и стиль в романе Толстого «Война и мир». – М. : Федерация, 1928. – 249 с.
8. Чичерин А. В. Идеи и стиль. О природе поэтического слова. – М. : Советский писатель, 1965. – 374 с.
9. Бахтин М. М. Проблемы поэтики Достоевского. – М. : Советская Россия, 1979. – 317 с.
10. Успенский Б. А. Поэтика композиции // Успенский Б. А. Семиотика искусства. – М. : Языки русской культуры, 1995. – С. 9–218.
11. Балли Ш. Общая лингвистика и вопросы французского языка. – М. : Изд-во иностранной литературы, 1955. – 416 с.
12. Арутюнова Н. Д. Типы языковых значений. Оценка. Событие. Факт. – М. : Наука, 1988. – 341 с.
13. Ожегов С. И. Словарь русского языка. – М. : Русский язык, 1986.
14. Словарь русского языка : в 4 т. – М. : Русский язык, 1980.
15. Новый объяснительный словарь синонимов русского языка. Вып. 1–3. – М., 1999 – 2003.
16. Гуковский Г. А. Реализм Гоголя. – М. : Гос. изд-во худож. лит-ры, 1959. – 532 с.

Материал поступил в редакцию 24.12.2013 г.

Сведения об авторе

МУРАВЬЕВА Наталия Юрьевна – кандидат филологических наук, доцент кафедры русского языка Института гуманитарных наук Московского городского педагогического университета (МГПУ), Москва
e-mail: natasha1000@mail.ru