OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 811

ВЕРОЯТНОСТНО-АЛГОРИТМИЧЕСКОЕ МОДЕЛИРОВАНИЕ СТАТИЧЕСКОЙ И ДИНАМИЧЕСКОЙ СОСТАВЛЯЮЩИХ СОДЕРЖАНИЯ ТЕКСТА

А.В. Зубов (*proscien@mslu.by*)
Минский государственный лингвистический университет,
г. Минск, Республика Беларусь

В статье обосновывается необходимость выделения в содержании текста статической и динамической составляющих. Предлагается метод формального выделения этих составляющих, а также семантико-синтаксический язык для представления этих составляющих в памяти компьютера.

Ключевые слова: динамический, содержание, статический, текст, формальный язык.

Введение

Современные интеллектуальные системы исходят из того, эмпирически подтвержденного положения, что эффективность такой системы определяется не столько её способностью к сложной логической переработке некоторой информации, сколько полнотой представления предметного знания профессионалов-экспертов [Концепции…, 1988]. Если речь идет об автоматической обработке текстов, то в качестве такого эксперта выступает лингвист, хорошо знающий некоторый естественный язык и умеющий представить свои знания о текстах в формализованном виде. Но для этого такой лингвист должен иметь некоторый формальный язык, позволяющий представить его знания о языке в памяти компьютера.

Были многочисленные попытки создания такого формального языка. Однако, они не были ориентированы на семантическое представление содержания всего текста и связь этого содержания с первоначальным замыслом. Предлагаемый в статье семантико-синтаксический язык для записи в памяти компьютера содержания текста может быть использован для извлечения из текстов на иностранных языках основной информации и для порождения текстов для обучения и для общения в Интернет.

Основной текст

Многоаспектные исследования различных текстов показывают, что каждый текст по своей структуре является многоплановым и противоречивым образованием. В нем можно обнаружить социальное и индивидуальное, детерминированное и случайное, обязательное и факультативное, содержательное и формальное, глубинное и поверхностное, языковое и речевое [Проблемы, 1983]. К тем же противоречивым особенностям текстов относится и наличие в тексте статического и динамического аспектов.

Термины «статический» и «динамический» по отношению к тексту употребляется обычно в смысле «покой» и «движение» [Гальперин, 1981]. Текст как последовательность лексических единиц, как некоторый результат, продукт речемыслительной деятельности находится в статическом состоянии или в состоянии покоя. Текст же в процессе его порождения, восприятия и понимания считается находящимся в движении.

Рассмотрим проблему статики и динамики текста с несколько иной точки зрения. Как отмечают многие исследователи, при воспроизведении предложений, запоминаются две группы данных о предложении: информация о его семантическом содержании и информация о

его синтаксической структуре. При этом семантический аспект запоминается в качестве первого шага, а синтаксический – в качестве второго. В дальнейшем это положение было развито по отношению к тексту и свелось к констатации того, что в каждом тексте есть свой словарь и свой синтаксис. По-разному понимаются при этом такие составляющие как «словарь» и «синтаксис». Ближе всего нам та точка зрения, когда под словарем понимаются так называемые слова «содержания» или «несомые слова», а под синтаксисом имеется в виду грамматический строй вместе с «несущими» или служебными словами [Проблемы, 1983]. Таким образом, можно представить, что слова «содержания» представляют статику текста и являются отражением в тексте некоторого множества предметов, явлений, фактов реальной действительности, а синтаксис – динамику текста, отражающую те отношения между этими предметами, фактами, явлениями, которые устанавливает автор текста в зависимости от целевой установки, типа текста, речевого опыта и целого ряда других факторов. Наличие этих двух составляющих в тексте подтверждают и эксперименты по анализу процесса понимания текста.

Любая ситуация, отраженная в тексте, может быть рассмотрена с той или иной позиции. Помимо этого, различный отбор элементов ситуации связан с тем, что человек воспринимает действительность в условиях некоторой вероятности, зависящей от его жизненного и языкового опыта. Все это приводит к тому, что в процессе развертывания содержания отбираются узловые элементы системы референтов, являющиеся наиболее существенными для построения данного текста. Каждый из этих элементов позволяет задавать порядок дальнейшего отбора необходимых элементов в подмножестве связанных с ним референтов [Новиков, 1983]. А так как референтам в тексте соответствуют различные имена, то формальная структура содержания текста сводится к формальной структуре линейной последовательности материальных знаков – слов.

Анализ большого числа работ по смысловой организации текста показывает, что смысловое единство текста связано с наличием в нем некоторых «опорных слов» («топиков», «смысловых вех», «дескрипторов», «ключевых слов», «фокальных точек», «ядер» и т.д.) [Вейнрейх, 1981], которые пронизывают весь текст от его начала и до конца.

Существуют достаточно интересные попытки формального представления смысла текста через такие опорные слова и словосочетания, связанные с построением различного рода графов. Однако, как утверждают авторы подобных построений, такая формальная связь не позволяет полностью решить задачу построения графа денотативной структуры, соответствующего всему тексту, так как его окончательная связь всех вершин, представляющих текст, осуществляется на содержательном уровне с привлечением не содержащейся в тексте информации [Новиков, 1983].

Известно, что в структуре «опорных слов» или «смысловых вех» существует определенная иерархия. Часть из них является «главными», «важными» словами, «номенклатурными дескрипторами», «ключевыми» словами, «выступающими точками» всего содержания [Новиков, 1983]. Они определяют главный предмет сообщения и по существу являются свернутым замыслом текста. При этом один и тот же референт может выражаться различными словами – контекстуальными синонимами в тех пределах, при которых возможно их взаимное приравнивание. Ввиду линейности означающего текста такие слова оказываются разнесенными по разным абзацам текста.

Анализ работ по выделению опорных слов с учетом вышесказанного позволяет предложить два формальных критерия определения главных опорных слов (ГОС): наибольшая частота употребления слова (включая все его контекстуальные синонимы и местоименные замены) и максимальное число абзацев, в которых встречается слово. Такие слова в дальнейшем становятся своеобразным центром, вокруг которого формируются другие элементы, отражающие составляющие микроситуации (глаголы, прилагательные, наречия и т.п.). Главные опорные слова являются, как правило, именами, отражающими главные действующие лица или главные предметы описания, повествования или рассуждения.

Главные референты, которым в тексте соответствуют главные опорные слова, в рамках той или иной ситуации могут быть связаны с другими референтами, также являющимися относительно важными для той же ситуации, но которые являются центром не всей ситуации, а некоторых микроситуаций общей ситуации. Соответствующие им слова текста назовем второстепенными опорными словами (ВОС). Они встречаются с меньшей частотой, чем ГОС, и в меньшем числе абзацев всего текста.

Существует достаточно большое число способов определения «степени важности» для содержания текста того или иного слова или словосочетания. Их можно разделить на: 1) анкетные, 2) структурные, 3) словарные, 4) частотные, 5) синтаксические, 6) диалоговые.

Все эти методы имеют один существенный недостаток: они не исходят из текста как связной единицы. Например, с помощью структурных методов выделяются опорные слова на основе их вхождения в заголовок, первое предложение или в несколько отдельных предложений. Частотные методы исходят из частоты встречаемости слов в тексте. При таком подходе по существу учитывается не смысл текста, а свойства плана выражения текста [Гиндин, 1977].

Для выделения из текста ГОС и ВОС по каждому тексту строится частотно-алфавитный словарь и на его основе для каждого слова вычисляется коэффициент важности по формуле:

$$К_{важ.} = \frac{F \cdot m}{N \cdot n}$$

В этой формуле: F – абсолютная частота слова в тексте (в неё входит суммарная частота всех типов синонимов этого слова его ассоциативных и местоименных замен); m – число абзацев, в которых встретилось слово; N – общее число слов в тексте; n – общее число абзацев в тексте. Слова текста, имеющие наибольшие значения Кваж., относят к числу ГОС и ВОС [Зубов, 1986]. ГОС и ВОС текста оформляются в специальную таблицу основного статического содержания текста.

Рассмотрим теперь несколько подробнее проблему смысловой организации абзаца. Отмечая, что эта проблема является в целом семантико-синтаксической, исследователи в то же время подчеркивают, что синтаксические особенности этой единицы плана содержания не могут быть поняты без детального проникновения в её лексическую структуру. Для микротемы, так же как и для всей темы характерны иерархия и повторяемость опорных слов и словосочетаний. При этом выделяются два именных ряда соотнесённости: ряд смыслового субъекта и ряд смыслового предиката. Действительно, как отмечалось выше, абзац содержит отражение некоторого фрагмента психической ситуации. По существу, в абзацах отражаются те связи и отношения, которые зафиксированы у главных и второстепенных референтов текста с другими, менее значимыми элементами ситуации: различные типы действий и состояний ГОС и ВОС, различные места действий, время действий, цели, причины действий и состояний и т.п. [Новиков и др., 1981].

Как было отмечено выше, динамический аспект содержания текста должен характеризовать те отношения между объектами реальной действительности, которые зафиксированы в конкретной, описанной в тексте ситуации. Точнее, речь идет о фиксации тех событий (процессов, действий, состояний), которые происходят с референтами, имена которых представлены в тексте опорными словами. В связи с ограниченностью оперативной памяти человека и невозможностью полного охвата событий, человеческое сознание в процессе познания вычленяет некоторый фрагмент этих событий. Это позволяет последовательно рассмотреть различные признаки одного предмета при подведении каждого из них под общую категорию, известную ранее и зафиксированную в сознании. Такое членение действительности на фрагменты проводится не произвольно, а в соответствии с социально отработанными моделями, сформировавшимися в ходе длительного исторического развития человечества [Чесноков, 1982].

Ввиду неразрывности мышления и языка каждому отраженному в сознании фрагменту действительности в некоторое соответствие ставятся определенные языковые формы. Одни исследователи отмечают, что разным элементам реальной действительности в мышлении соответствуют фиксированные слова, другие ставят им в соответствие словосочетания, предложения и целые тексты. Третья группа авторов, не отрицая определенной роли указанных единиц, ведущую роль в отражении фрагментов реальной действительности отдаёт логическому аппарату, позволяющему получать выводные знания из известных индивиду обобщенных эмпирических фактов.

Ученые самых различных направлений отмечают, что в процессе человеческого познания наиболее важные и часто повторяющиеся отношения реальной действительности закрепляются в форме синтаксических структур. Эти структуры усваиваются нами с детства вместе со словами и звуками родного языка. Когда мы говорим, то очень часто не задумываясь, выбираем один из наиболее привычных для нас в данной ситуации языковых шаблонов. Такие шаблоны («синтаксические модели», «структурные схемы предложений», «конструкции», «фразовые стереотипы», «синтаксические формы» и т.д. и т.п.) являются психическими реальностями, существующими в нашем сознании вместе с единицами других уровней языка, участвующими в процессе создания текста.

Вместе с тем, по данным современной психологии и лингвистики внеязыковое содержание чаще всего передается не одним, а несколькими взаимосвязанными предложениями, т.е. в языке существуют определенные структурно-смысловые модели для выражения отношений между мыслями в структурах, находящихся за пределами предложения [Лосева, 1980]. Как правило, цепочку таких взаимосвязанных мыслей называют высказыванием. Основная особенность таких единиц заключается в том, что они обладают определенными и относительно устойчивыми типическими формами построения целого. Такими единицами мы уверенно пользуемся, но теоретически можем и не знать об их существовании. Они приходят к нам в процессе обучения языку и жизни, соотносятся с определенными фрагментами действительности и несут в себе не только собственный опыт автора, но и социальный опыт общества.

Более глубокие исследования позволяют прийти к выводу, что наличие таких синтаксических шаблонов, охватывающих несколько предложений, связано с общим положением об опережающем отражении действительности, разработанным учеными Д.Н. Узнадзе и П.К. Анохиным [Узнадзе, 1961]. В соответствии с этой теорией человек «обладает способностью на основе врожденного и приобретенного опыта предвидеть развитие ситуативных обстоятельств и настраиваться на определенную реакцию с некоторым упреждением вероятностного хода событий» [Колшанский, 1983]. В процессе создания текста это положение реализуется в виде составления некоторой общей схемы развития мысли и, следовательно, её речевого оформления. Таким образом, подобные синтаксические структуры есть не что иное, как некоторые устоявшиеся, отложившиеся в памяти логические образования, отразившиеся в человеческом мышлении типические связи и отношения объективной действительности. Наличие речевых шаблонов–высказываний отмечается в исследованиях лингвистов, психологов и психолингвистов и специалистов по обучению языкам. Причем набор таких шаблонов конечен, относительно невелик и специфичен для текстов одного автора или одной достаточно узкой предметной области.

Как было отмечено выше, такие синтаксически единые объединения, включающие несколько предложений, были названы «сверхфразовыми единствами» или «сложными синтаксическими целыми». Исследования последних лет позволяют допустить, что и абзац письменного текста также является «логическим слепком» с действительности [Лосева, 1980]. Это подтверждают не только лингвистические исследования, но и эксперименты психологов и психолингвистов. Так, в одной серии опытов испытуемым предъявлялись одни и те же тексты, но подготовленные по-разному: с разбивкой на авторские абзацы, без абзацев и с абзацами, не учитывающими логическую структуру текста. В итоге первые тексты воспринимались уверенно, легко и быстро. Во 2-ом и 3-ем случаях на усвоение таких текстов было потрачено гораздо больше времени и текст усваивался плохо или совсем не усваивался.

Таким образом, приведенные факты показывают, что абзацы автором выделяются не произвольно. Они строятся по определенным правилам, более строгим и стереотипным для научных текстов и более свободным и творческим в текстах художественных, в то же время в особенностях композиции абзаца отражается семантическая динамика текста. Исследователи отмечают специфичность абзацев каждого автора. Можно предположить, что и тексты разных авторов, относящиеся к достаточно узкой предметной области, можно описать в виде определенной последовательности конечного числа специализированных абзацев.

Как же расположены абзацы прозаического текста или строфы поэтических текстов в структуре текста?

Известно, что в процессе познания жизни на нервную систему человека одновременно воздействует огромное число различных нервных импульсов, сигнализирующих о влиянии на организм бесчисленных факторов окружающей среды и процессов внутри самого организма. Поэтому трудно сказать с полной определенностью, какова будет реакция организма на то или иное действие среды. Такую реакцию можно предсказать лишь с некоторой вероятностью, индивидуальной для каждого члена общества [Петунин, 1971]. Однако речевая деятельность человека имеет принципиально ту же психологическую структуру, как и всякая другая форма психической деятельности. Поэтому и в речевом поведении человека можно обнаружить такие же вероятностные составляющие, как и в поведении вообще, т.е. индивидуальный речевой опыт имеет вероятностную структуру. Но «речи для себя» не существует: она кому-то адресована и связана с определенной жизненной ситуацией. Как показывают многочисленные исследования, модель ситуации формируется говорящим в эталонах восприятия, которые, с одной стороны, зависят от прошлого (социального) опыта человека, а с другой – от задач и мотива совершаемой деятельности. Иными словами, в виду того, что психика личности социальна по своей сути, язык представляет собой средство опосредованного овладения социальным опытом. Таким образом, парадокс языка и идиолекта решается в современной лингвистике таким образом, что все индивидуальное, входя в речь, не перестает быть языком до тех пор пока оно принадлежит и другому. Именно поэтому индивидуальный язык детерминирован социально. Эти, отмеченные психологами и психолингвистами, особенности речевого поведения, были зафиксированы и лингвистами. Так, Л.В. Щерба говорил, что индивидуальная речевая система является лишь конкретным проявлением языковой системы, а А.Л. Потебня полностью соглашался с мыслью В. Гумбольдта: Sprechen heisst sein besonderes Denken an das allgemeine ankupfen – «говорить – значит связывать свою личную узкую мысль с мышлением своего племени, народа, человечества» [Потебня, 1976].

Первым из лингвистов, кто свел все подобные факты в единое целое, был Г. Хердан, выдвинувший теорию языка как выбора и случайности [Herdan, 1965]. Интерпретируя статистически основную дихотомию Ф. де Соссюра la langue – la parole, Г. Хердан на самом деле под la langue понимает le langage и соотносит le langage и la parole как совокупность и выборку, что и позволяет le langage изучать статистическими методами. Г. Хердан отмечает, что основным законом языковой коммуникации является стабильность распределения языковых элементов. Эта стабильность не зависит от воли говорящего и объясняется тем, что людям свойственно приобретать речевые навыки построения фраз путем подражания другим носителям языка. Он при этом допускает индивидуальность в языке и объясняет ее правом выбора со стороны носителя языка. Дихотомия языка как выбора и случайности дает возможность проводить исследование независимо от означаемого, ибо закону случайных чисел подчиняется все, что связано с означающим. Означаемое же предполагает наличие выбора и поэтому язык как случайность имеет дело с означающим, а язык как выбор – с означаемым. Это, по Г. Хердану, делает возможным логическое описание языка на всех его уровнях, описание формы независимо от содержания. Близкие по содержанию идеи содержатся и в работах французского лингвиста П. Гиро.

Большой цикл работ по обоснованию объективной значимости вероятностно-статистических отношений в языке и речи выполнен Р. Г. Пиотровским и его учениками [Пиотровский и др., 1977]. Эти исследования показали, что в языке существует некоторый

«эталон» статистического построения и упорядочения текста. Таким эталоном является норма, находящаяся между системой языка и речью (текстом).

Было выполнено большое число исследований по изучению строения текста и распределению в нем отдельных единиц. В итоге выяснилось, что порождение единиц заполнения в тексте описывается математическими моделями – распределениями Гаусса, Пуассона, Фукса, логнормальным распределением, кривыми Пирсона и т.п., а ситуативные ключевые элементы таким моделям не подчиняются. Так, например, при анализе английских технических текстов выяснилось, что нормальному закону подчиняется распределение в тексте артиклей, предлогов, частиц, союзов, местоимений и наречий. Терминологическая же лексика этому закону не подчиняется. Высказывается предположение, что существуют определенные функции распределения не только для означающих языкового знака, но и для значений этого знака. Так В. В. Налимов пишет: «... как в обыденном языке, так и во многих других языках с каждым знаком вероятностным образом связано множество смысловых значений. Можно говорить об априорной функции распределения смысловых значений знака» [Налимов, 1974].

Сравнивая и оценивая изложенные выше вероятностно-статистические концепции Г. Хердана, П. Гиро, Р. Г. Пиотровского и других исследователей, ряд авторов отмечают, что в процессе функционирования языка действуют вероятностно-статистические законы. Этот факт ставит перед исследователями много новых вопросов: «На каких уровнях действуют эти законы?», «Какой механизм обеспечивает их функционирование?», «Как обеспечивается необходимое соотношение детерминированных и вероятностных процессов?» и т.д. Пока можно лишь предположить, что в сложных многоуровневых и многофункциональных системах (а язык и является именно такой системой) узкая специализация отдельных уровней, функций и органов может сосуществовать с недетерминированной гибкостью других уровней и частей этих систем. Можно надеяться, что анализ конкретных текстов различного типа позволит в определенной степени ответить на поставленные выше вопросы.

Текст, являясь конечным результатом речевой деятельности, должен, очевидно, содержать в себе какие-то следы действия случайных факторов и детерминированных правил. Являясь результатом взаимодействия порождающей системы языка и нормы, а также внешней ситуации, текст включает два типа языковых элементов – элементов заполнения, которые порождаются системой и нормой языка, и ключевых единиц, которые стимулируются немодулируемой и непредсказуемой ситуацией.

Какие же именно вероятности зафиксированы в тексте?

Для ответа на этот вопрос необходимо вернуться к проблеме вероятностной структуры среды и вероятностному восприятию человеком элементов этой среды. В каждой жизненной ситуации устанавливается определенная степень соответствия между субъективной вероятностью наступления некоторого события и его объективной вероятностью, и это помогает человеку прогнозировать своё будущее поведение. Отмеченный факт лежит в основе так называемого вероятностного прогнозирования – способности человека сопоставлять поступающую информацию о наличной ситуации с хранящейся в памяти информацией о прошлом опыте и использовать эти данные для построения гипотез и предстоящих событиях, приписывая им ту или иную вероятность. Эта общефизиологическая закономерность справедлива и для речевой деятельности, т.е. человек владеет определенными субъективными представлениями о вероятностях употребления тех или иных элементов речи, может на основе этого строить субъективную модель вероятностной структуры предстоящей речевой ситуации, осуществлять на основе этой модели субъективный прогноз и строить свою речевую деятельность в соответствии с этим прогнозом. Но как было показано выше, психика человека социальна по своей сути, поэтому речевая способность каждого индивидуума так же социальна. Ученые отмечают, что в каждом тексте может быть выделено нечто «внеличное, общее, надиндивидуальное». Эта социальность практически проявляется в том, что человек в процессе порождения текста использует слова, словосочетания, предложения и даже более крупные единицы не в бесконечных вариантах, а в пределах некоторых границ. Эти границы для каждого человека различны и зависят от большого числа факторов, которые действуют на

протяжении всей жизни человека. Ядро, основа такой совокупности единиц общения формируется в детстве под влиянием подражания взрослым и активного воздействия со стороны взрослых на речь ребенка. В тексте это ядро проявляется в явлении стереотипии, проявляющемся на разных уровнях. Стереотипичность или стандартность речевого действия экономит умственные усилия как автора, так и получателя речи. Малая вариантность, характерная для стереотипичности, оставляет все меньшие возможности для свободного выбора нужного языкового средства и это способствует автоматичности выбора. Отмечается что стереотип – это прагматическое средство наиболее целесообразного исторического применения языка, существующее как автоматизированный и всеобщий переносчик смысла. Стереотип усреднен исторически из индивидуальных решений и находится в рамках нормативной грамматики.

Различают фонематический, словесный, фразовый и текстовый стереотип. Их объединяет то, что все они являются своеобразным пусковым звеном в развертывании соответствующего фрагмента речевой программы. Различие между ними заключается в том, что с повышением уровня стереотип становится менее «жестким», получает большее разнообразие форм воплощения. Как отмечает профессор Р.Г. Пиотровский текст строится на сочетании плана (детерминированная составляющая), действующего на коротких текстовых цепочках, и вероятностного прогнозирования – на длинных текстовых дистанциях [Piotrowski, 1975].

Опуская фонематический стереотип, относящийся к устному тексту, отметим, что суть словесного стереотипа заключается в том, что для выражения одного и того же понятия в пределах достаточно узкой предметной области и одной и той же ситуации используются одни и те же слова. Это связано с тем, что современная массовая коммуникация породила тенденцию языкового развития – концентрирование речевых средств, основывающееся на принципе экономной затраты сил. Концентрация речевых средств при этом заключается в накоплении и систематизации лексических единиц в закрепленных значениях и речевых отрезков вокруг определенных тематических линий. На этом принципе строятся лексические минимумы для чтения специальной литературы, отбираются лексические единицы в базовые языки.

Проблема фразового стереотипа подробно рассматривалась выше в ходе анализа динамических аспектов текста. Такие стереотипичные фразовые построения были названы «шаблонами», «синтаксическими моделями» и т.п. Как показывают исследования психологов и лингвистов, с возрастом запоминание материала приходит во все большей зависимости от грамматической структуры. Такие структуры оказываются в сознании наиболее устойчивыми при различного рода оговорках и нарушениях нормальной речевой деятельности. С точки зрения речевой деятельности фразовая стереотипичность – это не что иное, как вызов из памяти некоторой общей схемы развития мысли, а, следовательно, и его речевого оформления. Стереотипичность на фразовом уровне приводит к тому, что в каждой фразе текста актуализация одних лексических элементов неизбежно сопровождается автоматизацией выбора других элементов.

Приведенные выше рассуждения о сути абзаца, его роли в общей структуре текста, специфике абзацев являются подтверждением того, что можно говорить и о сверхфразовой стереотипичности. В процессе обучения жизни и языку в нашем сознании откладываются вероятностные модели, представления о синтаксических структурах самых различных текстов: научных, публицистических, художественных, деловых, поэтических и т.д. Стереотипичность на фразовом и сверхфразовом уровнях приводит к тому, что внутри одного вида письменной речи индивидуальные различия авторов становятся несравненно менее значимыми, чем различия между разными видами письменных текстов, созданных одним и тем же автором. Можно полагать, что такая сверхфразовая стереотипичность в определенной степени связана с фактом, доказанным специалистами различного профиля – порождение текста и его восприятия осуществляется не отдельными элементами, а определенными «блоками», «квантами», в которых сохранена связь с предшествующим и заложено на некоторое семантическое расстояние последующее. Так как основу речевой деятельности составляет, прежде всего, вещественное содержание речи, то сама коммуникация дискретна и деление текста на какие-то «куски» основано на «логическом слепке» с действительности на основе обобщения

человеческого опыта научного познания. Так как внешняя и внутренняя деятельность человека имеют одинаковое общее строение и вторая из них происходит из первой, то указанное деление текста на составляющие вполне согласуется с данными биологии и психологии вообще. Показано, что поведение животных слагается из таких элементарных актов как «убегание», «преследование», «поиск пищи», «ухаживание» и т.д. Притом каждый вид животных обладает определенным набором стандартных двигательных актов, которые, комбинируясь в различных сочетаниях, производят разнообразные формы сложного поведения. Показано также, что сложные формы умственной деятельности слагаются из элементарных информационных процессов, различные комбинации которых дают различные формы деятельности: игру в шахматы, решение задач, распознавание образов, перевод текстов и т.п. Объединение таких «кирпичиков» информации в единое целое возможно в самом общем виде на основе трех принципов: вероятностном, когда элементы объединяются в целое на основе вероятностных распределений; детерминированном, когда порядок следования «кирпичиков» однозначен и строго задан; и, наконец, вероятностно-алгоритмическом, когда в процессе объединения учитываются связи как случайные, так и детерминированные (алгоритмические). Вероятностно–алгоритмические процессы объединения составляющих в единое целое выступают как основной тип процессов объективного мира. Как уже не раз отмечалось, наличие вероятностных составляющих в сложных структурах объективного мира позволяет таким системам меняться, развиваться во времени.

Возвращаясь к проблеме организации текста из блоков, «квантов» смысла, следует отметить, что такое деление зависит от типа текста (научный, художественный и т.п.), его направленности, намерений автора, его языковых знаний и целого ряда других факторов. В процессе изучения семантической структуры текстов проводилось выделение различных смысловых «кусков» в тексте. Это – «текстовые элементы» и «блоки-мотивы», и «фрагменты», и «аспекты», и «сгустки», и абзацы. Как было отмечено выше, абзац, как смысловой «квант» и как «застывший» синтаксический шаблон, может быть принят за минимальный логико-семантический компонент текста. Наличие абзацев в тексте помогает лучшему усвоению текста, позволяет строить различные варианты одного и того же содержания в зависимости от намерений автора и цели создания текста.

Пока остается неясным, каким образом вероятность используется при формировании единого текста из абзацев. Для научных текстов, например, в силу преобладания в них логических составляющих «выстраивание» абзацев происходит стереотипично, по очень ограниченному числу схем. Художественные тексты характеризуются более свободным расположением абзацев. В то же время высказываются предположения, что каждому автору в пределах определенного функционального стиля свойственны определенные структуры абзацев и типы их взаимного расположения. Конечно, в каждом тексте, кроме стереотипичных элементов, обусловленных социолингвистически, можно обнаружить и индивидуальные авторские особенности. Для писателя и поэта они являются средством выражения их сугубо индивидуальной мысли, то есть аффективного и эстетического содержания, а ученый использует их как индивидуальное средство усиления связности и логичности текста. В научных текстах такая индивидуальность «авторского почерка» может заключаться в умении автора владеть слогом в рамках определенного научного стиля (использование образных, субъективно-оценочных средств, разговорной лексики и конструкций и т.п.), в индивидуальном предпочтении того или иного синонимичного варианта, предъявляемого стилем, в особенностях слога, связанных с возрастом, региональными различиями и т.д. Однако основная особенность этого индивидуального в речи заключается в том, что оно ограничено нормами стиля и в целом за каждым текстом стоит система языка.

Что касается текстового стереотипа, то его можно понимать как особый самоорганизующийся психологический механизм, способствующий «укладыванию» поступающей из текста разнообразной информации в русло разветвленной структурно-смысловой схемы в условиях взаимодействия детерминированных и вероятностных правил. Число таких формул текста, выбираемых для передачи одного и того же содержания, зависит

как от типа текста, так и от языковых знаний автора текста. Последние, как уже неоднократно отмечалось, зависят в конечном счете от социального опыта личности.

Как же можно построить такие формулы текста, чтобы в них были зафиксированы как детерминированные, так и вероятностные составляющие этого текста?

Для этого в нашей модели используется специальный семантико-синтаксический язык СЕМСИНТ [Зубов, 1990].

СЕМСИНТ включает в себя следующие основные составляющие: 1) алфавит языка, 2) систему средств для записи семантических отношений между членами отдельного предложения, 3) систему средств для записи синтаксических отношений между членами предложения, 4) систему элементов для фиксации логико-семантических отношений между предложениями отдельного абзаца и между абзацами, 5) систему средств для фиксации темы текста.

Рассмотрим подробнее эти составляющие.

Алфавит языка СЕМСИНТ составляют:

а) все русские и латинские буквы;
б) все десятичные цифры;
в) все орфографические знаки;
г) знаки арифметических действий;
д) знаки логических действий.

С помощью этих букв, цифр и знаков описываются все остальные составляющие языка СЕМСИНТ.

Для фиксации семантических отношений между членами отдельного предложения текста используются 20 семантических функций, подобных семантическим падежам Ч. Филлмора [Филлмор, 1981]. К числу их мы относим следующие: Субъект (AAG), Предмет (AH1), Понятие (AS2), Н-понятие (научное понятие – AS1), Процесс (AS3), П-деятель (активный природный или стихийный деятель – AEL), Объект (AP1), Получатель (AB1), Адресат (AB2), Н-объект (неодушевленный объект – AO), Место (ALK), Средство (AMD), Инструмент (AIN), Состав (AKM), Определитель (AD1), Время (ATM), Способ (AAD), Условие (AIF), Причина (ACS), Цель (AAM), Вставка (IN). При этом в скобках даются принятые в языке СЕМСИНТ формальные коды семантических функций.

Второй семантической составляющей языка СЕМСИНТ является алфавитный словарь достаточно узкой предметной области, где каждое слово имеет код класса слова (существительное, глагол и т.д.) и код принадлежности к определенному семантическому подклассу (например, запись N21 говорит о том, что это – существительное, обозначающее человека и части тела человека).

Отмеченные выше семантические функции относятся не к отдельному слову предложения, а к аргументной группе – группе слов, включающих в себя имя существительное (именное словосочетание) и все относящиеся к нему определители, выраженные именем прилагательным, причастием, числительным, местоимением. В семантической формуле предложения аргументная группа состоит из цепочки кодов семантических подклассов определителей и существительного, соединенных знаком «*», ограниченных слева знаком «<», а справа – знаком «>». Перед такой цепочкой ставится один из кодов семантических падежей.

Для обозначения в семантической формуле предложения места глагола-сказуемого используется код R, за которым ставится цифра, указывающая семантическую валентность глагола, затем знак «<», код семантического подкласса соответствующего глагола и знак «>». Наречия в такой формуле представляются отдельно от глагола в виде соответствующего кода семантического подкласса.

Все эти составные части семантической формулы предложения соединяются между собой знаком «+» («плюс») в строгом соответствии с порядком следования слов в предложении.

Синтаксические отношения между членами предложения в такой семантико-синтаксической формуле предложения (СЕСФП) выражаются по-разному.

Основной способ – «поверхностные значения» перечисленных выше семантических функций (табл. 1). Так каждому из 82 предлогов русского языка дан определенный код (KIJ) и указаны падежи, которыми предлог управляет. Предлоги включаются в состав аргументной группы с помощью знака «*».

Таблица 1 – «Поверхностные значения» семантических падежей языка СЕМСИНТ (русский язык)

Код семантических функций	Поверхностное значение семантических функций	Код семантических функций	Поверхностное значение семантических функций
AAG	именительный	A03	предложный
AH1	именительный	A04	творительный
AS2	именительный	A05	дательный
AS1	именительный	ALK	зависит от предлога
AS3	именительный	AMD	творительный
AEL	именительный	AIN	творительный
AP1	творительный	AKM	зависит от предлога
AP2	винительный	ADI	зависит от предлога
AP4	предложный	ATM	зависит от предлога
AP5	родительный	AAD	зависит от предлога
AB1	дательный	AIF	зависит от предлога
AB2	дательный	ACS	зависит от предлога
A01	винительный	AAM	зависит от предлога
A02	родительный		

Для выражения синтаксической связи между некоторыми словами в предложении и между простыми предложениями в пределах сложных в языке СЕМСИНТ используются союзы и союзные слова (коды CIJ). В русском языке выделено 73 такие единицы. Перед ними в СЕСФП всегда ставится знак «+». Если союз «и» связывает однородные члены предложения, то в СЕСФП он изображается знаком «&». Этим же знаком соединяются и аргументные группы, являющиеся распространенными однородными членами.

Синтаксическая связь в субстантивно-субстантивных словосочетаниях (независимо от выражаемых ими отношений) фиксируется с помощью знака «/» («наклонная черта»). Например:

«нарушение речи» – N13/N09

«поражение лобных долей мозга» – N05/J06 * N20/N23.

Помимо этого язык СЕМСИНТ содержит ряд менее значимых средств для выражения синтаксических отношений между словами предложения.

Для возможности полного отражения в памяти ЭВМ любого текста в язык СЕМСИНТ включаются все частицы языка. Так для русского языка (коды TIJ), их число составляет 60 единиц. Они являются важными элементами предложения, вносящими в смысл предложения дополнительные логико-смысловые, модальные или экспрессивные оценки.

Специалистами по лингвистике текста показано, что между предложениями абзаца, как основной единицы письменного текста, существуют различные логико-семантические связи. В языке СЕМСИНТ для описания русских текстов, например, используются 12 типов логико-смысловых скрепов (код LIJ). Фиксируются следующие виды связей между предложениями: предшествование приведению примера, факта, детализации или некоторой аргументации; субъективное отношение автора к высказываемому в тексте; причинно-следственные; перечисления; возражения; результат, итог; обстоятельственные; оценка достоверности

сообщения, подтверждение, положительная аргументация; сопутствующие обстоятельства; перифрастические конструкции; отсылки при повторениях слов и словосочетаний; связи, выражаемые обобщающими местоимениями, заменяющими части предложений.

Многие исследователи структуры текста отмечают, что средства логико-семантической связи между абзацами текстов в основном совпадают с теми формальными средствами связи, которые соединяют предложения в абзаце. Поэтому для выражения связей между абзацами используются слова указанных выше типов логико-смысловых скрепов (в русском языке 105 слов и словосочетаний).

Тема текста, как было отмечено выше, задается в виде таблицы основного статического содержания. Входящие в неё ГОС и ВОС отмечаются в СЕСФП знаком «**», который ставится над кодом семантического подкласса соответствующего слова текста.

Тогда, например, предложение

«Маленький мальчик быстро открыл белую дверь»,

где слова «мальчик» и «дверь» относятся к ГОС, будет представлено на языке СЕМСИНТ в виде следующей семантико-синтаксической формулы:

** **

AAG<J04*N26>+D11+R3<V13>+A01<J04*N31>.

При заполнении такой формулы конкретными словами вместо N26** и N31** вставляются конкретные слова из таблицы основного статического содержания всего текста, откуда взято это предложение, а вместо кодов других слов берется по датчику случайных чисел из словаря предметной области одно из слов, имеющих соответствующий код принадлежности к определенному семантическому подклассу.

Имея в памяти компьютера текст представленный на языке СЕМСИНТ, можно автоматически извлекать из него любую информацию («кто действует?», «где действует?», «когда действует?» и т.п.). Вся семантико-синтаксическая формула текста может служить основой для порождения текста по заданному содержанию в виде задаваемого набора ГОС и ВОС.

Язык СЕМСИНТ использован для автоматического создания табличного реферата группы английских публицистических текстов, освещающих в Интернет поездки Дж. Буша. С его помощью порождены 4 типа электронных англоязычных деловых писем, а также английские и французские сказки и французские стихотворения.

Библиографический список

[Вейнрейх, 1981] Вейнрейх, У. Опыт семантической теории / У. Вейнрейх // Новое в зарубежной лингвистике. Вып. X. Лингвистическая семантика. – М.: Прогресс, 1981. – С. 50–176.

[Гальперин, 1981] Гальперин, И.О. Текст как объект лингвистического исследования / И.О. Гальперин. – М.: Наука, 1981.

[Гиндин, 1977] Гиндин, С.И. Семантика текста и различные теории информации / С.И. Гиндин // НТИ. сер. 2, 1977. № 10. – С. 10–15.

[Зубов, 1986] Зубов, А.В. Статический аспект содержания текста и его формальное представление / А.В. Зубов // Уч. зап. Тартуского ун-та. Вып. 745. Квантитативная лингвистика и автоматический анализ текстов. Tartu, 1986. – С. 75–94.

[Зубов, 1990] Зубов, А.В. Семантико-синтаксический язык для записи текстов в памяти ЭВМ / А.В. Зубов // Функционирование и развитие языковых систем. – Минск: Высшая школа, 1990. – С. 110–117.

[Колшанский, 1983] Колшанский, Г.В. О языковом механизме порождения текста / Г.В. Колшанский // Вопросы языкознания. 1983. № 4. – С. 44–51.

[Концепции…, 1988] Концепции интеллектуальных систем. Научно-аналитический обзор. – М.: ИНИпоОН, 1988. – 55 с.

[Лекомцев, 1973] Лекомцев, Ю.К. Психическая ситуация, предложения и семантический признак / Ю.К. Лекомцев // Труды по знаковым системам, VI. Уч. зап. Тартусского ун-та. Вып. 308. – Тарту, 1973. – С. 5–44.

[Лосева, 1980] Лосева, Л.М. Как строится текст. Пособие для учителей / Л.М. Лосева. – М.: Просвещение, 1980.

[Налимов, 1974] Налимов, В.В. Вероятностная модель языка / В.В. Налимов. – М.: Наука, 1974.

[Новиков и др., 1981] Новиков, А.И. К вопросу о теме и денотате / А.И. Новиков, Г.Д. Чистякова // Известия АН СССР. Сер. Лит. и языка. 1981. Т. 40. № 1. – С. 48–56.

[Новиков, 1983] Новиков, А.И. Семантика текста и её формализация / А.И. Новиков. – М.: Наука, 1983.

[Падучева, 1965] Падучева, Е.В. О структуре абзаца / Е.В. Падучева // Труды по знаковым системам. Учен. зап. Тартусского гос. ун-та. Вып. 181. – Тарту, 1965. – С. 184–292.

[Петунин, 1971] Петунин, Ю.И. Мозг как вероятностная система / Ю.И. Петунин // Философия и естествознание. Вып. 3. – Воронеж: ВГУ, 1971. – С. 154–165.

[Пиотровский и др., 1977] Пиотровский, Р.Г. Математическая лингвистика / Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская // Математическая лингвистика. – М.: Высшая школа. 1977.

[Потебня, 1976] Потебня, А.А. Эстетика и поэтика / А.А. Потебня. – М.: Искусство, 1976.

[Проблемы, 1983] Проблемы текстуальной лингвистики. – Киев: КГУ, 1983.

[Скороходько, 1983] Скороходько, Э.Ф. Семантические сети и автоматическая обработка текста / Э.Ф. Скороходько. – Киев: Навукова думка, 1983.

[Узнадзе, 1961] Узнадзе, Д.Н. Экспериментальные основы психологии установки / Д.Н. Узнадзе. – Тбилиси: Мецниереба, 1961.

[Филлмор, 1981] Филлмор, Ч. Дело о падеже / Ч. Филлмор // Новое в зарубежной лингвистике. Вып. X. Лингвистическая семантика. – М.: Прогресс, 1981.

[Чесноков, 1982] Чесноков, П.В. Об отношении между речевыми и мыслительными процессами с точки зрения единства языка и мышления / П.В. Чесноков // Синтаксическая семантика и прагматика. – Калини: КГУ, 1982. – С. 38–47.

[Herdan, 1965] Herdan, G. Language as Choice and Chance / G/ Herdan. – Gronigen, 1965.

[Piotrowski, 1975] Pijtrowski, R. The Linguistics of the Text and Machine Translation / R. Piotrowski // American Journal of Computational Linguistics, 1975. V.12. № 3. – P. 55–56.