OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 519.767.2

УСК МАРТЫНОВА – ТРИДЦАТЬ ЛЕТ СПУСТЯ

Ефименко И.В.[*], Хорошевский В.Ф.[**]

[*] *Центр информационно-аналитических систем ИСИЭЗ НИУ ВШЭ,*
г. Москва, Россия
iefimenko@hse.ru

[**] *Вычислительный центр им. А.А. Дородницына РАН,*
Центр информационно-аналитических систем ИСИЭЗ НИУ ВШЭ
г. Москва, Россия
khor@ccas.ru, vkhoroshevsky@hse.ru

В работе дается ретроспективный анализ систем представления знаний семейства УСК (Универсальный Семантический Код), разработанных в конце 70-х, начале 80-х годов прошлого века известным советским ученым В.В. Мартыновым. Показано, что УСК-3 Мартынова в определенном смысле опередил существующие в то время подходы к представлению лингвистических знаний и определил вектор использования методов и средств компьютерной лингвистики в практически значимых приложениях. Представлена попытка интерпретации идей В.В. Мартынова, положенных в основу УСК-3, с позиций современных достижений в области компьютерной лингвистики и искусственного интеллекта.
Ключевые слова: понимание естественного языка; семантический код; язык описания и исчисления смыслов; псевдофизическая логика; онтологическое моделирование предметной области.

Вместо введения

В 2014 году исполняется 90 лет со дня рождения Виктора Владимировича Мартынова – ученого, внесшего существенный вклад в исследования на стыке лингвистики с неклассическими логиками и в становление научного направления «представление и обработка знаний».

В.В. Мартынов окончил Одесский университет в 1948 году, в 1951 году – аспирантуру по славистике при Львовском университете, в 1952-1960 гг. был заведующим кафедрой иностранных языков Одесского университета, а в 1960 г. переехал в Минск, где много лет работал в Институте языкознания АН БССР. Доктор филологических наук, профессор В.В. Мартынов автор 20 книг и более чем 200 статей.

Три научные области были наиболее близки В. В. Мартынову. Первая – это славистика, вторая – компаративистика и, в частности, глоттогенез и онтогенез славян. Третья, которой Виктор Владимирович занимался с особым вдохновением, связана с искусственным интеллектом и информационными технологиями, где в 60-е годы XX века формировалось новое научное направление – представление и обработка знаний.

С позиций сегодняшнего дня можно сказать, что основная заслуга В.В. Мартынова как ученого состоит в том, что ему удалось, опираясь на взаимосвязь лингвистики с теорией информации, предложить новый способ «исчисления языковых смыслов», а затем и универсальный семантический код как средство снятия «неопределеннозначности» естественного языка при построении формальных моделей предметных областей.

Настоящая работа посвящена обсуждению семейства языков представления знаний УСК, а изложение организовано следующим образом. Сначала дается краткое авторское обоснование того, почему и каким образом классический филолог и специалист в области компаративистики пришел к проблеме представления и обработки знаний в парадигматике неклассических логик, затем приводится ретроспективный анализ систем представления знаний семейства УСК Мартынова (Универсальный Семантический Код), которое в значительной мере опередило существующие в то время подходы к представлению лингвистических знаний и использованию методов и средств компьютерной лингвистики в практически значимых приложениях. В заключительной части

работы представлена попытка интерпретации идей В.В. Мартынова, положенных в основу УСК-3, с позиций современных достижений в области компьютерной лингвистики и искусственного интеллекта.

1. От компаративистики к представлению и обработке знаний

В области лингвистики и филологии в сферу интересов В. В. Мартынова входил широкий набор направлений: литературоведение и культурология, сравнительно-историческое языкознание, славистика и востоковедение. В. В. Мартынов внес существенный вклад в развитие дисциплин, находящихся на стыке гуманитарных и точных / инженерных наук, в рамках исследований по формализации семантики и лингвистическим аспектам искусственного интеллекта.

Однако, как представляется, основным направлением, в контексте которого следует рассматривать историю создания универсального семантического кода, является сравнительно-историческое языкознание, или компаративная лингвистика и, в частности, исследования В. В. Мартынова в области глоттогенеза славян и славяно-неславянских контактов ([Мартынов, 1963; Мартынов, 1983; Мартынов, 2003] и др.).

В силу необходимости работы с большими объемами разнородного языкового материала компаративистика предусматривает развитие формальных методов описания, цель которых – обеспечить возможности сравнения языков и обоснования гипотез об их происхождении и родстве, а также в той или иной степени реконструировать праязык. Одним из важных результатов при этом является классификация языков.

Как отмечает С. А. Старостин, выдающийся специалист в области сравнительного языкознания, «возникшая в начале XIX в. и с тех пор неуклонно развивавшаяся как в отношении предмета изучения, так и в отношении методологии, компаративистика послужила основой для методологии всего современного языкознания, явившись, по сути, главной стимулирующей силой для выхода общей лингвистики на новый, научный этап своего развития. К настоящему времени компаративистика является полноценной научной дисциплиной, оперирующей строго формализованными методами (в том числе компьютерными) с целью проникновения вглубь истории языков и реконструкции все более и более отдаленных от современности праязыков человечества» [Старостин, 2007].

Таким образом, можно отметить две важнейшие черты компаративистики, в которых, вероятно, следует искать истоки интереса В. В. Мартынова к проблематике создания универсального семантического кода: во-первых, это поиск единой основы естественных языков (общее ядро, единый праязык, универсальные лингвистические явления) и, во-вторых, создание формализованных средств их описания.

Как представляется, исследования по компаративистике, с одной стороны, заложили основы «персональной траектории» В. В. Мартынова, результатом которой стало создание универсального семантического кода, с другой стороны – во многом определили области приложения УСК. Можно утверждать, что формальное языконезависимое представление лингвистических знаний является одной из ключевых задач для подавляющего большинства приложений, связанных с синтезом или анализом естественного языка. Однако именно для решений в области взаимодействия человека и компьютера, которые, прежде всего, интересовали В. В. Мартынова, это является наиболее очевидным. Подходы, основанные на универсальном представлении лингвистических знаний, сыграли важную роль и в развитии других типов приложений – например, в сфере машинного перевода, о чем будет сказано позднее.

2. Семейство УСК – ретроспективный анализ

2.1. Общие замечания

Как отмечается в работах [Поспелов, 1981; Поспелов, 1986], которые, по нашему мнению, сыграли значительную роль в формировании взглядов В.В. Мартынова на проблему создания универсального семантического кода, при использовании ЕЯ в качестве основы для построения языка представления знаний в нем выделяются следующие классы элементов (слов и словосочетаний), играющие функциональную роль в представлении знаний: ***понятия***, ***имена*** и ***отношения***.

При этом для любого естественного языка характерно наличие слов и словосочетаний, определяющих ***понятия-классы***, обладающие определенными свойствами (например, «статья», «лаборатория»), имена служат для идентификации элементов, входящих в понятие-класс (например, понятие-класс «вычислительный центр» в качестве элементов может содержать понятия-классы «лаборатория» с определенным номером, который играет роль имени, а лаборатории в качестве своих элементов могут содержать отделы или сектора, тоже снабженные именами. В отличие от понятий-классов ***понятия-процессы*** описывают группы однородных процессов (например, «учебная нагрузка»). По своей функциональной роли близки к ним ***понятия-состояния***, примерами которых могут быть словосочетания «нормальный режим», «компьютерная сеть работает» и т. п. Для идентификации понятий-процессов и понятий-состояний, как и в случае понятий-классов, могут использоваться имена. Важно, что для естественных

языков множества понятий и имен потенциально бесконечны.

Отношения служат для установления связей на множестве понятий или идентифицированных понятий. При этом уже сама идентификация реализуется с помощью специального отношения «называться». Существует гипотеза о конечности множества различных, не сводимых друг к другу, отношений для естественных языков (базовых отношений) и сводимости других отношений, присутствующих в ЕЯ-текстах, к комбинации базовых. Поэтому при построении моделей представления знаний для реальных предметных областей можно всегда считать, что мы имеем дело с конечным числом различных понятий, имен и отношений.

С помощью понятий, имен и отношений можно описывать ситуации, имеющие место в предметной области. Для этого в терминальный словарь (***T***) ЯПЗ вводятся три рассмотренных функциональных класса: понятия (***V***), имена (***I***) и отношения (***R***)

$$\boldsymbol{T} = \boldsymbol{V} \cup \boldsymbol{I} \cup \boldsymbol{R}, \text{ где}$$

$V = B \cup D \cup G;$
$\boldsymbol{B}=\{b_1, b_2, \ldots, b_n\};$
$\boldsymbol{D}=\{d_1, d_2, \ldots d_m\};$
$\boldsymbol{G}=\{g_1, g_2, \ldots, g_k\};$
$\boldsymbol{I}=\{i_1, i_2, \ldots i_i\};$
$\boldsymbol{R} = \{(,), \rho, r_1, \ldots, r_q\},$

и строится совокупность следующих синтаксических правил, определяющих правильно построенные формулы (ППФ) в системе представления знаний:

1. Любой элемент, кроме элементов множества ***R***, является ППФ.
2. Если ***α*** — любой элемент из ***V*** и ***β*** — любой элемент из ***I***, то тройка (αρβ) есть ППФ.
3. Если ***δ*** и ***γ*** — любые элементы из ***V***, то тройка (δrγ), где ***r*** — любой элемент из ***R***, кроме скобок и ***ρ***, есть ППФ.
4. Если ***ε*** и ***χ*** суть ППФ, то тройка (εrχ) есть ППФ.
5. Других ППФ нет.

Для примера рассмотрим следующую ЯПЗ-запись, удовлетворяющую введенным выше обозначениям, и попробуем восстановить ЕЯ-текст, соответствующий этой записи:

((«судно»ρ«№1»)r3(«причал»ρ«№8»)) & ((«груз»ρ«лес»)r12(«процесс»ρ«погрузка»)

Две тройки, входящие в первую «большую» скобку, соединенные отношением ***r3***, имеющим значение «быть в окрестности», дают фразу «Судно №1 находится около причала №8». Следующая «большая» скобка соответствует фразе «Идет погрузка леса», непосредственная расшифровка которой даст предложение «Груз по имени лес участвует в процессе (***r12***) по имени погрузка». Но смысл этой фразы тождествен смыслу ранее приведенной фразы, которая по-русски звучит естественнее. Таким образом, приведенная выше формальная запись может быть заменена одной фразой «Судно № 1 находится под погрузкой леса у причала № 8».

Не менее, а быть может более важной проблемой при создании ЯПЗ, является установление смысловой эквивалентности различных ППФ, которая обычно решается с помощью специальных ограничений на синтез ППФ, поскольку некоторые отношения не могут произвольно вводиться между понятиями произвольной природы. Проблема эта известна в структурной лингвистике как проблема построения языка семантических представлений (СЕМП), к которому предъявляются следующие требования. Если ЕЯ-фраза, по мнению его носителей, имеет смысл, то в СЕМП этой фразе должно соответствовать, по крайней мере, одно представление и обратно – если ЕЯ-фраза, по мнению его носителей, смысла не имеет, то в СЕМП для нее не должно найтись ни одного представления. Если две ЕЯ-фразы, по мнению его носителей, имеют совпадающий смысл, то в СЕМП им должно соответствовать либо одно общее представление, либо различные представления, но такие, что с помощью системы формальных преобразований фраз в СЕМП они переводятся одно в другое. Таким образом, построение СЕМП предполагает наличие эффективной процедуры перевода фраз естественного языка в СЕМП и эффективную процедуру установления эквивалентности записей в СЕМП.

И, наконец, в рамках семантического представления смысла существует еще одна важная проблема – проблема установления противоречивости описаний. По существу, ее решение сводится к построению СЕМП второго уровня, где вместо семантических правил соединения слов во фразе (и в дополнение к ним) разрабатываются семантические правила соединения фраз в тексте.

Таким образом, даже на уровне достаточно простых моделей представления смысла в те годы, когда В.В. Мартыновым создавались ЯПЗ семейства УСК, имелись серьезные научные проблемы, как в лингвистике, так и в области представления знаний. И, если учесть, что в реальных моделях представления знаний требовалось в дополнение к понятиям, именам и отношениям учитывать еще, как минимум, наличие в ЕЯ императивов, квантификаторов, модификаторов, а также модальностей и оценок, станет ясно, что задача создания универсального семантического кода для представления смысла текстов на естественном языке была чрезвычайно сложной. Справедливости ради отметим, что данная проблема не решена в общем случае и до сих пор, через 30 лет после того, как ее решением занялся В.В. Мартынов.

В процессе исследований В.В. Мартыновым предлагались различные версии УСК [Мартынов, 1977], но последним и, по-видимому, наиболее проработанным в этом семействе ЯПЗ является

УСК-3 [Мартынов, 1984]. Как представляется авторам настоящей работы, именно в процессе создания этого языка В.В. Мартынов окончательно определил основные концепции всех систем представления знаний семейства УСК, которые состоят в следующем:

- Язык представления знаний должен быть языком формального описания и исчисления смыслов и, следовательно, псевдофизическим языком и псевдофизической логикой одновременно.
- УСК должен быть языком полной экспликации смысла (т.е. каждый комбинаторный тип цепочки элементов должен иметь один и только один смысл).
- УСК должен быть языком универсальной канонизации (т.е. ограничения, накладываемые на его систему, не должны зависеть от того, какой фрагмент реального мира описывается).
- УСК должен быть языком неконвенциального представления семантики (т.е. его цепочкам семантика должна не приписывается, а выводиться из аксиом-универсалий).
- УСК должен строиться как система, способная понимать мир (т.е. формировать новые понятия и строить гипотезы о причинах и следствиях ситуаций, причем и то, и другое должны реализовываться на базе формальных преобразований цепочек).

Как следует из приведенных выше требований, уже в начале 80-х годов прошлого века В.В. Мартыновым была поставлена сверх сложная и, вместе с тем, чрезвычайно актуальная проблема создания формальной модели представления знаний, в рамках которой были бы интегрированы не только методы собственно описания смысла ЕЯ-текстов, но и средства формального описания человеческих рассуждений о мире.

2.2. УСК: «чужой среди своих»

Как указывалось выше, В. В. Мартынов внес вклад в развитие целого ряда направлений в сфере филологии и лингвистики. Основными работами В. В. Мартынова, которые изучаются лингвистами, можно считать труды по компаративистике и славистике. В частности, классической является уже упоминавшаяся выше монография «Язык в пространстве и времени. К проблеме глоттогенеза славян». Однако необходимо отметить, что непосредственно УСК не нашел широкого признания и применения среди лингвистов, чему может быть дано следующее объяснение.

Для того, чтобы описать естественноязыковые выражения, порождаемые носителями, с помощью УСК, требуется построить сложные, многокомпонентные цепочки, что ограничивает возможности прикладного применения соответствующего формализма. Кроме того, при таком описании существует риск стирания значимых семантических различий между близкими, но не идентичными фрагментами естественно-языковых текстов (см. пример в [Мартынов, 1984]) – как в части естественно-языковых цепочек, так и в части описаний на УСК: «Высказывание Главк обеспечивает завод оборудованием реально означает Главк держит оборудование на заводе. SAOŌ обозначает X держит Y (в себе). SAŌO – X держит себя в Z-е (≡ X находится в Z-е). SAŌŌ – X держит себя в себе (≡ X есть). Последняя интерпретация, вероятно, нуждается в некотором содержательном комментарии. Если мы интерпретировали X держит себя в Z-е как X находится в Z-е, то X держит себя в себе сначала интерпретируется как X находится в себе, а потом как X есть, потому что быть значит ~~н~~аходиться где-то (покоиться)", а находится где-то в любом случае означает ~~н~~аходиться в себе"».

Что касается использования УСК в области лингвистической теории, то здесь наблюдается ряд расхождений с общепринятыми методами и принципами описания языка, а также, в некотором роде, с имеющимися задачами.

С одной стороны, подход, лежащий в основе УСК, в частности, идея «разложения» языковых примеров на элементарные семантические единицы, соответствует традициям формальной семантики (см. пример в [Мартынов, 1984] – интерпретацию высказывания «Носильщик везет чемодан в тележке» как «Носильщик держит чемодан в тележке и перемещает тележку», т.е., в терминах УСК: X держит Y в Z-е и перемещает Z). Идея трансформации одних цепочек УСК в другие также перекликается с идеями перехода от глубинной структуры к поверхностной в генеративной лингвистике. Не противоречит лингвистическим традициям и тот факт, что УСК предназначен, прежде всего, для описания нарративных предложений. Так, например, известный российский лингвист, автор работ по компаративистике и происхождению языка С. А. Бурлак отмечает: «Если вы посмотрите на знаменитые лингвистические примеры, они все до единого – нарративные тексты: «Фермер убил утенка», «Бесцветные зеленые идеи яростно спят», «Глокая куздра штеко будланула бокра и курдячит (или «кудрячит») бокренка». И никто не начал свою теорию с предложений типа ~~Д~~ай, пожалуйста!" или, скажем, ~~М~~арш отсюда! ". Говорится, что такие предложения неполные, что они особые. А «правильные», «настоящие» – это именно нарративные предложения, то есть, по сути, комментарии» (см. стенограмму лекции на http://polit.ru/article/2008/11/07/lang/).

С другой стороны, задача канонизации естественно-языковых высказываний, являющаяся основой подхода к описанию языка с использованием УСК, по сути своей предполагает маргинальный характер тех выражений, которые должны быть канонизированы. Это, как и ограниченный состав атомарных элементов в УСК, изначально является препятствием для описания естественного языка на уровне, необходимом для развития лингвистической теории. Впрочем, это

отмечается самим автором: «Необходимость канонизации естественного языка для диалога человек – ЭВМ объясняется тем, что естественный язык в его полном виде нельзя сделать понятным для машины» [Мартынов, 1984]. В некотором роде, в УСК сделана попытка стереть границы между лексической и грамматической (в частности, синтаксической) составляющими языка и описать их с использованием единого формализма, а также создать «универсальное синтаксическое описание». Между тем, известно, что описание синтаксиса является одной из самых сложных задач и для компаративистики (например, при реконструкции праязыка), где, как указывалось выше, могут быть в определенном смысле найдены истоки УСК, и для прикладной лингвистики, к которой относятся работы по УСК как таковые. В частности, полномасштабное синтаксическое описание считается камнем преткновения в области извлечения информации из текстов (Information Extraction), машинного перевода и других задач автоматической обработки естественного языка.

В области анализа (и синтеза) естественного языка «драйвером» создания УСК послужила, прежде всего, задача человеко-машинного общения, диалогового взаимодействия с ЭВМ. В частности, В. В. Мартынов исследует работы А. П. Ершова, И. А. Мельчука, А. С. Нариньяни, Э. В. Попова по указанной проблематике. И несмотря на то, что УСК не стал общепринятым способом описания естественного языка в лингвистике, в т. ч. прикладной, можно констатировать, что В. В. Мартынов сформулировал целый ряд важных принципов и разработал методологию и инструментарий, позволяющие взглянуть на естественный язык с новых позиций, важных для решения целого ряда прикладных задач.

Одной из таких задач является машинный перевод естественно-языковых текстов. Детальный анализ методов и средств машинного перевода, в том числе, с использованием универсальных семантических языков (что наиболее близко к идее УСК), выходит за рамки настоящей работы. Здесь целесообразно отметить лишь то, что подходы, основанные на использовании единого формального описания, к которому приводится текст на языке-источнике (задача анализа) и на основе которого порождается представление на языке-цели, стали исторически первыми в прикладной области «машинный перевод». Так, считается, что сама идея автоматического перевода зародилась в XVII веке, когда Рене Декарт предложил универсальный язык, в котором один символ выражает эквивалентные идеи, формулируемые на различных естественных языках. В XX веке универсальные описания, промежуточные между языком-источником и языком-целью, активно использовались в лингвистических подходах, основанных на правилах (Rule-based approaches). Однако на современном этапе развития систем машинного перевода ведущими можно считать подходы, использующие статистические методы и инструменты корпусной лингвистики, в частности, подходы, которые относятся к классу Example-based и которые основаны на применении параллельных корпусов. В частности, такого рода подход используется переводчиком Google. Однако это не означает, что Rule-based подходы, опирающиеся на промежуточную формальную спецификацию, потеряли свою актуальность. Поэтому наиболее перспективными на современном этапе представляются гибридные методы, сочетающие достоинства лингвистических и статистических подходов к машинному переводу.

2.3. УСК: «свой среди чужих»

Приведенные во введении к настоящей работе краткие библиографические сведения о научном пути профессора В.В. Мартынова показывают, что по своему образовательному и научному background-у это был филолог, который в процессе работы «перешел» в лагерь специалистов по искусственному интеллекту и компьютерной лингвистике. Следует отметить, что в те годы, когда само научное направление представления и обработки знаний только формировалось, это было не единичным случаем, а скорее естественным следствием междисциплинарности новой науки. И появление в научном сообществе математиков и программистов специалистов из области филологии, психологии и других направлений воспринималось с энтузиазмом, который активно поддерживался Научным советом по искусственному интеллекту АН СССР, в рамках междисциплинарных проектов «Диалог» и «Диалог-2», а позже и в рамках международных рабочих групп РГ-18 и РГ-22. Общепризнанными «центрами притяжения» в новом научном сообществе были такие научные лидеры, как профессор Д.А. Поспелов и член-корр. РАН А.Е. Кибрик, академик Г.С. Поспелов и академик А.П. Ершов, в общении с которыми формировались и активно работали молодые коллективы из разных республик и городов нашей страны. Естественным образом «вписался» в это научное сообщество и В.В. Мартынов, который сконцентрировался на проблеме языков, с помощью которых интеллектуальные устройства могли бы вести диалог с человеком.

При этом в дискуссиях о языке, которые велись в то время (нужно ли учить ЭВМ понимать естественный язык или следует разрабатывать искусственный язык, удобный для общения ЭВМ с человеком), В.В. Мартынов четко стал на позицию применения ограниченных вариантов естественных языков с основным упором на канонизации языка общения, т. е., на спецификации того, какие синтаксические и лексические средства языка можно использовать, а какие нельзя в соответствии с некоторой теорией эффективного представления знаний.

Необходимость канонизации естественного языка для диалога человек — ЭВМ обосновывалась В.В. Мартыновым тем, что естественный язык в его

полном виде нельзя сделать понятным для машины уже в силу принципиальной семиологической эллиптичности и многозначности (омонимичности) ЕЯ-фраз, что ведет к неэксплицированности их смысла на уровне структурных описаний. При этом правильно отмечалось, что лишь незначительная часть информации, извлекаемой адресатом из ЕЯ-сообщения, содержится в самом сообщении, а большая ее часть восстанавливается на основе коллективного и индивидуального опыта человека (пресуппозиции), что подтверждалось и специалистами по искусственному интеллекту. Так, например, Р. Шейк и Р. Абельсон в своей, теперь уже классической, работе [Шенк и др., 1975] отмечали, что «исследователи понимания естественного языка в течение некоторого времени уже чувствуют, что проблема может быть решена в той мере, в какой мы способны характеризовать наши знания о мире...».

Как следствие, в качестве единственного средства преодоления барьеров на пути создания языка общения человека с ЭВМ В.В. Мартынов видел канонизацию ЕЯ, которая может быть ориентирована на проблему либо на сам язык, т. е. проводиться в соответствии с семиологической теорией эффективного представления мира без его ограничения. Следует отметить, что примерно такая же позиция была в то время и других исследователей [Ершов, 1982].

В.В. Мартынов в своей монографии [Мартынов, 1984] отмечал, что «лингвисты, работающие в области семантического синтаксиса, и кибернетики, занятые проблемами искусственного интеллекта, по существу, ведут исследования в одном направлении, но, поскольку при этом используется различная терминология, в известной мере дублируют друг друга. То, что лингвисты называют пресуппозицией в широком смысле данного понятия (без различия ее видов), в теории ИИ оказывается моделью мира, или представлением знаний. Если лингвисты сомневаются насчет того, считать ли пресуппозицию лингвистической или экстра лингвистической категорией, то кибернетики уверены в экстра лингвистичности модели мира. Вернее, такого вопроса у них просто не возникает»… В действительности это утверждение было не вполне верным даже в то время, а сейчас, по-видимому, корректнее говорить о интеграции лингвистических и внелингвистических подходов к представлению и обработке знаний.

Вместе с тем, можно согласиться с В.В. Мартыновым в том, что в большинстве систем искусственного интеллекта «лингвистический процессор действует только на входе, т. е. является устройством, выполняющим перевод с ограниченного естественного языка на внутренний язык машин, на котором информация обрабатывается внутренними процессорами. Внутренний язык машин, хотя и назван языком, в действительности рассматривается как объект не лингвистический, а логический, как система исчислений… В системах естественного интеллекта дело обстоит иначе. Язык информации, поступающей на вход, и внутренний язык совпадают. Естественный язык одновременно выполняет функции описания феномена и рассуждения о нем. Если канонизированный естественный язык или искусственный представить в виде формальной системы, то он сможет стать языком описания и исчисления смыслов, или псевдофизическим языком и псевдофизической логикой. Иными словами, отчужденное от человека кибернетическое устройство получит инструмент и с его помощью будет описывать реалии и ситуации, в которых эти реалии могут находиться (псевдофизический язык), вести рассуждения по поводу возможных изменений ситуаций (псевдофизическая логика)».

Целесообразность разработки подобного языка для В.В. Мартынова и многих других специалистов подкреплялось результатами исследований и разработок в области ИИ. Например, работами по ситуационному управлению [Поспелов, 1981], где явно указывалось, что существующие в то время «интеллектуальные системы должны, но пока не умеют: а) формировать модель мира в виде многоуровневых обобщенных знаний о классах объектов и ситуаций; б) устанавливать полные ассоциативные связи между классами объектов и ситуаций; в) строить формальную классификацию задач; г) принимать решения». А основная причина этих недостатков виделась исследователями в отсутствии эффективных способов «вложения семантических знаний в формализмы представления» [Попов и др., 1976]. При этом фактически речь шла о том, как регулярно соотносить план выражения и план содержания внутреннего языка ЭВМ, который для этого должен стать семантически мощным полифункциональным языком описания и исчисления смыслов. На разработке именно такого языка и сконцентрировался В.В. Мартынов.

Как представляется авторам настоящей работы, большое влияние на разработку семейства языков УСК оказал один из активно использовавшихся в то время в теории ситуационного управления для представления знаний язык RX-кодов [Скороходько, 1968]. В алфавите языка RX-кодов использовались два типа символов,– ***X*** (названия предметов) и ***R*** (названия отношений),– и ряд специальных обозначений. Для терминов языка было характерно, во-первых, ступенчатое кодирование, что давало возможность экономного и максимального приближения к полному выражению релевантных семантических характеристик, а, во-вторых, использование отношений, как в синтагматике, так и в парадигматике, что имитировало формирование новых понятий посредством разноструктурной модификации старых. В-третьих, сами отношения в языке RX-кодов образуются как производные путем сочетания элементарных. Эти и ряд других характеристик языка RX-кодов превратили его в весьма эффективный, по тому времени, инструмент

информационного поиска. Однако, поскольку язык создавался именно для целей информационного поиска, он не мог претендовать на роль полифункционального языка представления и обработки знаний. Именно на этом позиционировались языки семейства УСК Мартынова и, в частности, обсуждаемый ниже УСК-3.

Теории УСК предшествует логическая теория отношений, поэтому часть базовых предложений УСК формулируется в терминах последней. При этом вводятся три группы базовых предложений: о структуре отношений; о свойствах отношений; семиологические. Первые две группы объединяются как логические и, таким образом, противопоставлены третьей.

Для каждого отношения предполагается класс элементов, составляющих область отношения, и класс элементов, образующих противообласть (конверсную область) отношения. Отношения, согласно А. Тарскому, делятся по структуре на три и только три класса: первой степени (элемент-элемент) — ***XRY***; второй степени (элемент-отношение) — ***XR***(***YRZ***); третьей степени (отношение—отношение) — (***XRY***)***R***(***WRZ***).

В соответствии с числом элементов или отношений, выступающих в роли элементов, отношения делятся на бинарные и n-арные, а свойства отношений сводятся к трем типам: рефлексивность, симметричность и транзитивность. Классы отношений, упорядоченные по их свойствам, противопоставлены друг другу по

- рефлексивности—нерефлексивности — антирефлексивности,
- симметричности—несимметричности— антисимметричности,
- транзитивности—нетранзитивиости— антитранзитивности.

Дополнительно к базовым предложениям из области теории отношений вводятся базовые предложения из модальной логики и теории множеств.

Утверждения могут получать теоретико-множественную интерпретацию. Тогда утверждение с квантором общности интерпретируется посредством полного, или универсального, множества, а утверждение с квантором существования интерпретируется посредством частичного, или парциального, множества отдельных объектов, входящих в предметную область, на которую распространяется данное утверждение.

При переходе от логической базы теории УСК к семиологической логическую символику В,В, Мартынов заменяет семиологической. При этом элемент класса элементов, составляющих область отношения, обозначается как S (субъект, левая маргинальная позиция), элемент класса элементов, образующих противообласть отношения, обозначается как O (объект, правая маргинальная позиция), а отношение S к O выражается как A (акция, центральная позиция). Такого рода цепочка (SAO) называется ядерной. Кроме единичных элементов (S, A, O) в состав цепочки могут входить множественные элементы (S), (A), (O).

Таким образом, бинарное отношение между элементами S и O ядерной цепочки определяется наличием специального элемента A, задающего данное отношение, называемое эксплицитным. В n-арном отношении, сводимом к бинарным, не всякое из них эксплицитно. Отношение, определяемое отсутствием специального элемента, задающего это отношение, называется имплицитным. В соответствии с определением не может быть ядерной цепочки, состоящей из более чем одного эксплицитного отношения или из одного имплицитного отношения.

В случае незамещенности позиции ее элементом она замещается элементом S. Незамещенность позиции O в цепочке $SA\bar{O}$ (черта над позицией означает ее незамещенность) указывает на то, что позицию O замещает S, и, следовательно, элемент A задает отношение S к самому себе (рефлексивность). Рефлексивность S в ядерной цепочке означает отношение S к самому себе, что получает в семиологической терминологии название регрессивной доминации (в отличие от прогрессивной или собственно доминации). Незамещенность позиции A ($S\bar{A}O, S\bar{A}\bar{O}$) не интерпретируется в терминах теории отношений, поскольку в этом случае S замещает элемент, который сам задает отношение между S и O, а позиция S в ядерной цепочке не может быть незамещенной. Имплицитное отношение как дополнительное к эксплицитному вводится при расширении ядерной цепочки путем увеличения числа позиций одного из ее элементов (SAO=>SAOO; SAO->SSAO; SAO=>SAAO). Так как S всегда доминирует над остальными элементами, доминация сохраняется и после расширения. В этом случае S первое (S ядерной цепочки) доминирует над вторым S, возникшим в результате увеличения числа позиций.

Увеличение числа позиций в результате введения в ядерную цепочку имплицитной доминации называется мультипликацией позиции. Цепочка с тернарным отношением и мультипликацией позиции называется расширенной, а возникшие в результате позиции — вторичными. Расширенные цепочки, как и ядерные, являются отношениями первой степени, или эксплицитными отношениями элемент-элемент. Не может быть расширенных цепочек с мультипликацией обеих маргинальных позиций (*SSAOO, *SSAAOO). Приведенные выше определения специфицируют возможность следующих типов расширенных цепочек: SSAO, SAAO, SAOO, SSAAO, SAAOO и только их.

Вторичные позиции расширенной цепочки могут быть замещены цельнооформленными цепочками

(S(SAO)AO, SA(SAO)O). В результате в маргинальных позициях возникнут отношения второй степени — имплицитные отношения элемент-отношение.

Цепочками, воплощающими отношения второй степени, могут быть только расширенные. В цепочках с отношением второй степени отношения занимают только вторичные позиции. Цельнооформленные цепочки, занимающие вторичные позиции расширенных цепочек, сами подчиняются тем же правилам порождения, которым подчиняются ядерные и расширенные цепочки.

Отношения третьей степени, или отношения отношений, воплощаются в сложные цепочки, все позиции в них замещаются цельнооформленными цепочками (ядерными и расширенными). Сложные цепочки представляют собой цепочки типа (SAO)A(SAO) (SAO).

Бинарные отношения типа *(SAO)A(SAO) и *(SAO) (SAO) A (SAO) (SAO) невозможны, что объясняется их сводимостью к ядерным цепочкам и противоречит определению последних. Невозможно также сведение этих цепочек к расширенным, ибо последним не свойственно эксплицитное отношение отношений и замещение первичных позиций цельнооформленными цепочками.

Сложная цепочка не сводима к имплицитному бинарному отношению отношений типа (SAO)(SAO). Такого рода выражение не составляет цепочки. Оно не может быть ядерной цепочкой, поскольку невозможна ядерная цепочка, состоящая из одного имплицитного отношения и воплощающая бинарное отношение отношений.

Структура цепочек, замещающих позиции сложной цепочки, определяется следующими базовыми предложениями: а) цепочки в позиции субъекта реализуются всегда как цепочки типа SSAO или S(SAO)AO; б) правая часть сложной цепочки состоит из последовательности двух эквивалентных цепочек или цепочек, различающихся по знаку. На основании этого определяются следующие три типа сложных цепочек: (SSAO)A(SAO) (SAO), (SSAO)A - (SAOO) (SAOO) и (SSAO) A(SSAO) (SSAO) или соответственно с S (SAO)AO в левой части. Сложные цепочки с незамещенной позицией одного из объектов невозможны, а производные цепочки строятся на основе доминации смежных маргинальных элементов в порядке их следования слева направо.

На основании предшествующих предложений определяется порядок диффузных преобразований для цепочек типа SAOO (исключая тавтологии): $SAOO$=>$SAO\bar{O}$=>$SA\bar{O}O$=>$SA\bar{O}\bar{O}$.

Вторым (вслед за диффузией) вводится преобразование, которое представляет собой превращение позиции элемента S^2 в позицию элемента O^2, или замену второго субъекта вторым объектом при сохранении всех остальных позиций и их элементов. Такое преобразование В.В. Мартынов называет транспозицией (переносом элементов вторичных позиций вместе с самими позициями). Первичные позиции и их элементы остаются при этом неизменными. Транспозиция асимметрична. Второй субъект заменяется вторым объектом, но обратное неверно. Транспозиция с сохранением позиций невозможна.

Транспозиция включается в качестве начального преобразования в порядок диффузных преобразований. Поскольку SSAO=>SAOO асимметрично, устанавливается следующий порядок преобразований цепочек УСК: $SSAO$=>$SAO\bar{O}$ =>$SA\bar{O}O$=>$SA\bar{O}\bar{O}$.

Выше мы определили способы порождения и преобразования цепочек УСК по В.В. Мартынову и тем самым дали описание его синтаксиса в статике и динамике. Тезаурус основных понятий УСК-3 представлен на Рисунок 1.

Аналогичным образом описывается и его семантика. При этом в качестве метаязыка используется канонизированный вариант русского языка.

Не имея возможности в данной работе полностью описать УСК-3, авторы отсылают заинтересованных читателей к монографии [Мартынов, 1984], а ниже приводят несколько примеров семантических представлений ЕЯ-фраз с помощью УСК-выражений, взятых из этой работы.

Так, например, в силу рассмотренных выше свойств УСК-выражений, цепочка SAO интерпретируется как «X преобладает над Y-ом», цепочка $S\bar{A}\bar{O}$ – как «X преобладает над собой», цепочка $S\bar{A}O$ – как «X постоянно преобладает над Y-ом», а цепочка $S\overline{AO}$ – как «X постоянно преобладает над собой».

Сложнее интерпретируются расширенные цепочки. Так, например, фразе «Директор пьет кофе в кабинете» в действительности должна интерпретироваться двумя фразами: «Директор пьет кофе» & «Директор находится в кабинете» («Директор держит себя в кабинете»). УСК-цепочка, соответствующая первой фразе, очевидна. Второй же фразе соответствует цепочка $SA\bar{O}\bar{O}$, которая интерпретируется как «X держит себя в Z-е».

И, наконец, совсем сложной и далеко не всегда соответствующей ожиданиям носителей языка (в данном случае русского) является следующая интепретация цепочки $(SSAO)A(S\bar{A}\bar{O}\bar{O})(S\bar{A}\bar{O}\bar{O})$: «X посредством Y-а воздействует на Z, в результате чего сначала Z является частью V, потом Z является частью V». В.В. Мартынов в данном случае рассматривает несколько ступеней канонизации исходного текста. Так, на второй ступени представленная выше формулировка преобразуется в следующую: «X посредством Y-а воздействует на Z, в результате чего Z сохраняется как часть V => X препятствует Z-у перестать быть частью V».

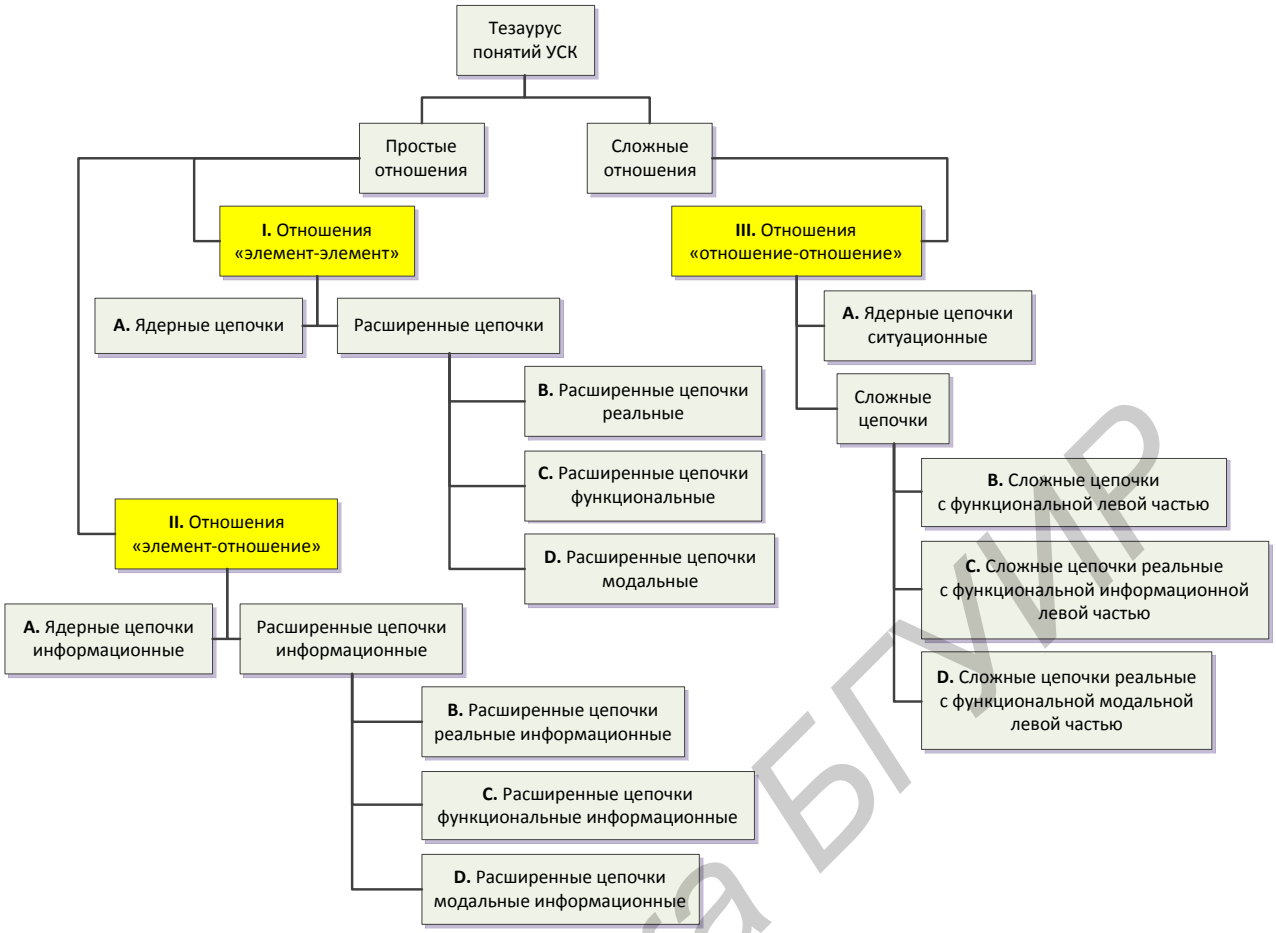

Рисунок 1 – Тезаурус основных понятий УСК-3

Как следует из приведенных выше примеров, задача преобразования ЕЯ-текста в УСК-представление является, в свою очередь, достаточно сложной. Именно это, на наш взгляд, стало еще одним препятствием на пути практического использования УСК в реальных системах того времени.

Заключение

В настоящей работе дан ретроспективный анализ систем представления знаний семейства УСК, разработанных в конце 70-х, начале 80-х годов прошлого века известным советским ученым В.В. Мартыновым.

С позиций сегодняшнего дня можно констатировать, что основные идеи, положенные в основу универсального семантического кода, актуальны до сих пор.

В первую очередь это идея единого языка и для представления, и для манипулирования знаниями. Не менее важно и то, что такой язык должен базироваться на идеях псевдофизических логик. И, наконец, идея о том, что семантика цепочек языка должна выводиться из системы аксиом.

В настоящее время часть этих идей уже воплощена в современных системах представления знаний. В частности, это идея использования псевдофизических логик в качестве базиса для представления знаний, которая предполагает формальную спецификацию семантики выражений ЯПЗ [Поспелов, 1986]. Вместе с тем, идея единого языка для представления и манипулирования знаниями в настоящее время еще не нашла своего воплощения, а технологическая целесообразность привела к тому, что эти две компоненты явно разделяются – для представления знаний используются онтологические модели и языки типа OWL, а для манипулирования знаниями – специальные языки запросов к базам знаний типа SPARQL и/или языки машин вывода на правилах типа Rule ML.

Несколько иначе в настоящее время видится и перевод с ЕЯ на язык представления знаний. Если в концепции В.В. Мартынова для этого предполагалось использовать специальные табличные формы и словари канонизированного русского языка [Мартынов, 1984], то в современных системах понимания ЕЯ акцент делается на извлечении информации из текстов и трансформации полученных результатов в системы взаимосвязанных онтологий [Хорошевский, 2008; Efimenko, et al., 2009; Хорошевский, 2009; Хорошевский, 2010; Хорошевский, 2012].

Однако общее направление универсализации семантических представлений на основе

формальных систем представления и манипулирования знаниями по-прежнему развивается и уже приносит практически значимые результаты.

Библиографический список

[Efimenko, et al., 2009] Efimenko I., Minor S., Starostin A., Drobyazko G., Khoroshevsky V. Generating Semantic Content for the Next Generation Web, Chapter in Monograph "Semantic Web", Publisher IN-TECH, 2009, ISBN 978-953-7619-33-6.

[Ершов, 1982] Ершов А. П. К методологии построения диалоговых систем: феномен деловой прозы // Вопросы кибернетики. Общение с ЭВМ на естественном языке. М., 1982.

[Мартынов, 1963] Мартынов, В.В. Славяно-германское лексическое взаимодействие древнейшей поры (к проблеме прародины славян). Минск, 1963**.**

[Мартынов, 1977] Мартынов, В.В. Универсальный семантический код (Грамматика. Словарь. Тексты). – Минск: «Наука и техника», 1977.

[Мартынов, 1983] Мартынов, В.В. Язык в пространстве и времени. К проблеме глоттогенеза славян. М., 1983.

[Мартынов, 1984] Универсальный семантический код: УСК-3 / Мартынов В.В.; – Минск: «Наука и техника», 1984.

[Мартынов, 2003] Мартынов, В.В. Кельто-славянские этноязыковые контакты // Мовазнаўства. Літаратура. Культаралогія. Фальклярыстыка: даклады беларускай дэлегацыі на XIII Міжнародным з'ездзе славістаў, Любляна, 2003 / НАН Беларусі камітэт славістаў. Мінск, 2003.

[Попов и др., 1976] Попов Э.В., Фирдман Г.Р. Алгоритмические основы интеллектуальных роботов и искусственного интеллекта. – М., 1976.

[Поспелов, 1981] Поспелов Д.А. Логико-лингвистические модели в системах управления. – М., 1981.

[Поспелов, 1986] Поспелов Д. А. Ситуационное управление: теория и практика, М., Наука,1986.

[Скороходько, 1968] Скороходько Э.Ф. Информационно-поисковая система БИТ. – Киев, 1968.

[Старостин, 2007] Старостин, С.А. Труды по языкознанию. - М., 2007. - С. 770-778.

[Хорошевский, 2008] Хорошевский, В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений. - 2008. - № 1. - С.80-97.

[Хорошевский, 2009] Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 2) // Искусственный Интеллект и Принятие решений, № 4, 2009, стр. 15-36.

[Хорошевский, 2010] Хорошевский В.Ф. Извлечение информации из текстов на конференциях серии Диалог: взгляд соседа по лестничной клетке //Труды международной конференции «Диалог–2010». Компьютерная лингвистика и интеллектуальные технологии. М.: 2010.

[Хорошевский, 2012] Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 3) // Искусственный Интеллект и Принятие решений, № 1, 2012, стр. 3-38.

[Шенк и др., 1975] Шенк Р., Абельсон Р. Сценарии, планы, знания. // В сборнике трудов IV Международной объединенной конференции по искусственному интеллекту. –М., 1975.

MARTYNOV'S USC – 30 YEARS LATER

Efimenko I.V. [*], Khoroshevsky V.F. [**]

[*] *Center for Information Intelligence Applications of Institute for Statistical Studies and Economics of Knowledge, NRU HSE*
Moscow, Russia
iefimenko@hse.ru

[**] *Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Center for Information Intelligence Applications of Institute for Statistical Studies and Economics of Knowledge, NU HSE*
Moscow, Russia
khor@ccas.ru, vkhoroshevsky@hse.ru

The paper represents a retrospective analysis of knowledge representation systems of the so called USC (i.e. universal semantic code) family. This formalism was developed in 1970-80s by a Soviet scientist Victor V. Martynov.

It is stated and shown that one of the most sophisticated versions of USC, which is USC-3, was ahead of many contemporarily existing approaches to linguistic knowledge representation in a number of ways. Martynov was among first evangelists of using methods and tools of computational linguistics in human-computer interaction applications. His approach has set a framework for following works.

The paper attempts to interpret Martunov's ideas which form the basis of the USC third version within a context of currently dominating paradigms of computational linguistics and artificial intelligence.

Keywords: natural language processing; semantic code; formal languages; calculation of meanings; pseudophysical logics; ontology engineering