ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 1 1991

ПРОБЛЕМА ПРЕДСТАВЛЕНИЯ ЗНАНИЙ И ЕСТЕСТВЕННЫЙ ЯЗЫК

КУЛАГИНА О. С.

ОБ АСПЕКТЕ МЕРЫ В ЛИНГВИСТИЧЕСКОМ ЗНАНИИ

1. Введение. В настоящей работе рассматривается аспект меры (или, иными словами, аспект измерения, количества, упорядочения по количественным критериям) в лингвистическом знании. Основная идея, которая развивается в дальнейшем изложении, состоит в следующем: знания носителей естественного языка (ЕЯ) о языке имеют не только качественный аспект, в них есть также аспект меры, порядка, который несколько условно можно назвать количественным. Первый обеспечивает широкий круг возможностей для выражения некоторой мысли (это знание «как можно сказать»). Второй регулирует выбор из этих возможностей наилучшего решения там, где выбор не детерминируется никакими другими более жесткими факторами (это знание «как лучше сказать»).

Мы будем рассматривать аспект, названный нами количественным, в связи с проблемой представления лингвистического знания для автоматизации работы с текстами на естественных языках. При этом речь будет идти не о форме представления, т. е. не о том, как представить лингвистическое знание в ЭВМ, а о его содержании, т. е. о том, какие сведения входят в состав знаний о ЕЯ у носителей языка и, соответственно, должны быть отражены в системах автоматической обработки текстов.

Практически с момента появления первых ЭВМ они начали использоваться не только как средство проведения вычислений, но и как средство переработки символьной информации, в том числе текстов, написанных на естественных языках. Со временем объем обработки символьной информации чрезвычайно возрос. В рамках данной работы представляют интерес те виды обработки текстов, где текст подвергается нетривиальным содержательным преобразованиям (в отличие от разного вида программ-редакторов). Иными словами, говоря об автоматизации работы с текстами, мы имеем в виду лингвистические процессоры.

Лингвистический процессор (ЛП) — это система, реализованная на ЭВМ, которая способна, по крайней мере, проводить анализ или синтез текстов на ЕЯ. Анализ — это переход от текста на ЕЯ к его представлению на некотором внутреннем языке, отражающему все существенные для решаемой задачи аспекты содержания и строения входного текста. Синтез — это обратный переход: от внутреннего представления — к тексту на ЕЯ. В ряде систем, относимых к ЛП, кроме анализа и синтеза происходят также преобразования внутренних представлений (поэтому выше и сделана оговорка «по крайней мере»).

Внутреннее представление, которое строит анализирующий ЛП, может быть вырожденным или полным. Под полным представлением мы имеем в виду такое, в котором никакая существенная информация не утрачивается, так что по нему в идеале можно восстановить исходный текст с точностью до синонимического преобразования. Примером такого представления может служить синтаксическая структура предложения в виде дерева зависимостей, в узлах которого стоят лексемы входного текста. Такое дерево позволяет восстановить фразу с точностью до порядка слов.

Вырожденное представление — это представление, в котором значительная часть информации утрачена (примером может служить поисковый образ документа в виде списка ключевых слов). Мы будем иметь в виду в дальнейшем только полные представления.

ЛП включает в себя три вида обеспечения: лингвистическое, математико-алгоритмическое и программное. Математико-алгоритмическое обеспечение — это формальные языки представления данных в ЭВМ и алгоритмы переработки этих данных, а программное обеспечение — набор программ, реализующих алгоритмы. На этих двух видах обеспечения мы останавливаться не будем, речь будет идти только о лингвистическом обеспечении ЛП, которое включает в себя словари и грамматические правила.

ЕЯ обладают рядом характерных особенностей, затрудняющих создание для них достаточно подробных описаний. К этим особенностям относятся сложность, неоднородность, недетерминированность и нечеткость. Поясним очень коротко, что имеется в виду.

Сложность проявляется в наличии очень большого числа очень разнородных элементов, способных вступать в большое число разнообразных отношений. Неоднородность проявляется в том, что при любых способах классификации этих элементов разброс в величине классов очень велик. Недетерминированность выражается в том, что все отображения в ЕЯ многозначны, в том числе переход от текста к смыслу, и, особенно, от смысла к тексту. Нечеткость проявляется и в нечеткости областей значений лексем и выражений, и в нечеткости границ синтаксической правильности выражений на ЕЯ. Совокупность этих особенностей создает богатство и гибкость естественных языков, но она же определяет тот факт, что проблема построения полного описания ЕЯ для систем автоматической обработки текстов столь трудна.

Системы грамматических правил, которые используются в ЛП, строятся, естественно, на основе тех грамматик, которые создавались в свое время для использования людьми. Но уже с самого начала исследований по автоматизации обработки текстов выяснилось, что для целей создания ЛП существующие грамматики мало пригодны. Прежде всего бросалось в глаза то обстоятельство, что многие правила в них не формализуемы или трудно формализуемы: они апеллируют к человеческой интуиции, к пониманию содержания соответствующего текста и т. д. Менее ясно осознавалась трудность, связанная с неполнотой лингвистических описаний. Позднее других выяснилось то обстоятельство, что существующие грамматики практически совершенно не отражают один из аспектов знания, которым располагает носитель языка. Именно об этом аспекте пойдет речь в настоящей работе.

Грамматические правила обычно формулируются в виде правил разрешающих, предписывающих или запрещающих определенные грамматические конструкции. При этом в них не находит отражения то обстоятельство, что человек, владеющий некоторым ЕЯ, не только умеет совершать

переход от текста к его смыслу и от смысла к тексту, но и ранжировать разные результаты всех переходов по определенной шкале правильности, нормативности, естественности выражения данной мысли на этом языке. Видимо, именно отсутствие владения этим аспектом лингвистического знания является причиной того, почему на иностранном языке человек может формировать фразы вполне правильные с точки зрения всех норм, но неестественные для носителя языка. Дело, по-видимому, не только в лексической идиоматичности, но и в различии в представлениях о норме для более глубоких уровней.

Аспект лингвистического знания, названный нами количественным, может проявляться очень по-разному. Прежде чем переходить к разбору его различных проявлений, сделаем одну оговорку. Для многих «количество» в применении к лингвистике ассоциируется в первую очередь с лингвистической статистикой. Как будет показано ниже, рассматриваемый аспект не сводится только к частотным характеристикам, хотя многие из показателей, о которых идет речь, действительно проявляются в частоте. Наряду с этим, рассматриваемый аспект проявляется и в размере классов при тех или иных классификациях; и в упорядоченности в круге альтернатив по степени их естественности или нормативности и, соответственно, предпочтительности одних перед другими; и в представлении о близости элементов языка, с точки зрения их значений и т. д.

2. **Уровень морфологии.** На уровне морфологии аспект меры в знании о ЕЯ проявляется в наиболее явной и простой форме, а именно в виде знания о количестве лексем, обладающих теми или иными морфологическими свойствами. Как мы уже отмечали выше, для ЕЯ характерен большой разброс в размере классов слов с одинаковыми морфологическими свойствами, будь то свойство иметь одинаковые окончания или свойство иметь одинаковые чередования в основах и т. д.

Например, в русском языке классы слов с полностью совпадающими окончаниями (т. е. классы более мелкие, чем обычные склонения и спряжения) могут насчитывать от нескольких тысяч единиц до одной единственной (известно, что слово *путь* не имеет аналогов). Как правило, в словарях и справочниках, описывающих словоизменительные возможности языка, сведения о числе элементов в соответствующих классах не приводятся.

В ЛП сведения о размерах морфологических классов могут быть полезны при решении вопроса о том, как наиболее эффективно составить словарь и правила для обработки слов на морфологическом уровне. В качестве примера можно предложить несколько разных способов для описания русского словоизменения.

Способ, наиболее близкий к принятому в обычных словарях, состоит в том, что для каждого слова в словарь включается одна основа, а в правила работы вводятся операции чередования (наборы окончаний при этом обычные). Второй способ состоит в том, что для слов, имеющих чередования в основах, в словарь включаются все разновидности основ; наборы окончаний при этом те же, что и в первом случае. Третий способ состоит в том, что в словарь включаются некоторые необычные «основы»: для каждого слова берется его графически неизменяемая часть, а все остальное, включая элементы чередования, присоединяется к окончанию. В результате получаются новые наборы окончаний и число разных наборов возрастает. Можно предложить и другие подходы. Каждый из указанных способов, выигрывая в одних показателях, проигрывает в других. При втором способе размер словаря является максимальным, но меньше наборов окончаний и работа по расчленению словоформ проще, чем при других способах.

При первом и третьем способах объем словаря меньше, но либо сложнее обработка (при первом способе), либо увеличиваются в объеме таблицы окончаний (при третьем). Легко понять, что нахождение эффективного описания словоизменения требует знания количественных оценок: насколько возрастет словарь при допущении нескольких основ для одного слова, как много новых наборов окончаний может возникнуть при введении квазиоснов и т. д. Кроме того, на работе этапов морфологического анализа и синтеза может сказаться и такая чисто количественная характеристика, как частота распределения словоформ в текстах. Так, если допускается чередование, то частота появления в текстах словоформ с чередованием может сказаться на скорости работы.

Таким образом, на уровне морфологии неоднородность ЕЯ проявляется в самой простой и явной форме: в величине классов и частоте появления их элементов в текстах. Так что в данном случае можно при желании оба эти проявления интерпретировать как частоту: частоту использования той или иной словоизменительной модели в пределах словарного запаса (которая и дает объем классов) и частоту использования той или иной словоизменительной модели в текстах. Аналогично можно говорить о неравномерности употребления словообразовательных моделей. При этом в лингвистической литературе, даже если и сообщается о различной продуктивности разных моделей, какие бы то ни было количественные оценки практически всегда отсутствуют. При создании описаний ЕЯ для построения ЛП эти сведения, как уже было отмечено, сказываются на компактности и эффективности системы. Однако это не слишком интересный случай, поскольку на результаты морфологического анализа или синтеза они не влияют.

Что касается неоднозначности переходов, то на уровне морфологии она проявляется, с одной стороны, в наличии нескольких разложений для одной словоформы (т. е. в омонимии) и, с другой стороны, в возможности нескольких словоформ для одной лексемы при одних и тех же морфологических характеристиках. Таковы, например, случаи типа *чая/чаю* для родительного падежа единственного числа слова *чай* или *лесу/лесе* для предложного падежа слова *лес*. Но последняя ситуация сравнительно редка.

3. Уровень синтаксиса. При переходе от морфологии к синтаксису свойственные ЕЯ неоднородность и неоднозначность переходов, отмеченные в п. 1, встречаются чаще. Как и в морфологии, имеется неоднородность и по величине классов слов с одинаковыми синтаксическими признаками, и по частоте употребления тех или иных синтаксических конструкций. Под синтаксическими признаками мы здесь имеем в виду и модели управления, и возможности появления определенных управляющих и др. Выражение «синтаксическая конструкция» здесь тоже употреблено в широком смысле: имеется в виду комбинация представителей определенных синтаксических классов (с заданными синтаксическими признаками и морфологическими характеристиками), связанных определенными синтаксическими отношениями.

В качестве примера рассмотрим способы оформления в русском языке подлежащего, или, в иных терминах, те конструкции, которые используются для оформления предикативного отношения. Подлежащим может быть имя (существительное, местоимение, числительное, субстантивированное прилагательное) в именительном или родительном падеже, глагол в неопределенной форме, придаточное предложение и др. Тут можно отметить разные проявления того, что мы назвали неоднородностью. С од-

ной стороны, далеко не каждый глагол допускает все эти способы. С другой стороны, очевидно и другое проявление неоднородности, именно то, что перечисленные случаи явно неравноправны с точки зрения употребимости. В грамматиках только указывается их допустимость, но сведения о частоте употребления или об их ранжировании по степени естественности отсутствуют.

Кроме разнообразия в способах оформления того или иного синтаксического отношения, русский язык предоставляет пользователю еще свободу в выборе порядка слов. Известно, что русское подлежащее может стоять как впереди, так и позади сказуемого. Однако хотя оба эти расположения возможны, они явно неравноправны: в большинстве случаев порядок прямой.

Вопроса о том, какими свойствами строения текста в целом определяется естественность выбора, мы касаться не будем. Для нас представляет интерес тот факт, что знания носителей языка о данной конструкции включают в себя если не точные оценки, то, во всяком случае, некоторую ранжированность этих возможностей по степени естественности, нормативности. Так, случай подлежащего, выраженного существительным в именительном падеже, является наиболее естественным, глагол в инфинитиве употребляется в роли подлежащего реже, чем местоимение, и т. п. Причины выбора одного из вариантов, который определяется, конечно, далеко не только естественностью, разнообразны. Для целей нашего изложения важно сейчас только следующее. В тех случаях, когда для выражения некоторого синтаксического отношения в ЕЯ имеется некоторый круг альтернативных возможностей, эти альтернативы неравноправны по употребимости; однако сведения о предпочтительности одних перед другими, которые в какой-то форме имеются у носителей языка, не зафиксированы в традиционных лингвистических описаниях, хотя для построения ЛП они нужны.

Вообще говоря, выбор одной из альтернатив при создании некоторого текста определяется целым рядом причин, например, стремлением избежать повторов, сократить текст и т. п. Это ведет к употреблению местоимения вместо существительного или вместо целой именной группы и т. д. Однако такого рода условия не детерминируют синтез полностью, у говорящего всегда есть некоторая свобода выбора, в пределах которой он решает вопрос не по строгим правилам, а на основе некоторых предпочтений. Можно предположить, что в пределах так называемой деловой прозы, где естественно считать априорными стремление к простоте и ясности изложения, автор обычно выбирает наиболее естественный нормативный вариант. Если мы сумеем отразить в системе автоматической обработки текста это представление о ранжированности по мере естественности и нормативности, то мы получим возможность улучшить результат работы системы.

Рассмотрим для начала, как проявляется отсутствие этого типа сведений.

До недавнего времени все алгоритмы синтаксического анализа базировались на понятии синтаксической правильности. Таковы и те алгоритмы, в основе которых лежат формальные грамматики Хомского, и алгоритмы фильтрового типа и другие. Нечеткость границ синтаксической правильности наряду с огромным разнообразием синтаксических конструкций, сочетаемость которых трудно проследить и описать, приводит к следующим трудностям.

Если круг синтаксических возможностей языка описан без всяких

сведений метрического характера, о которых шла речь выше, т. е. в правилах ЛП альтернативы выступают как равноправные (самый стандартный случай описывается так же, как и редчайшая возможность), то при проведении анализа, основанного только на правильности, мы сталкиваемся с одной из двух неудачных ситуаций. В одном случае в числе учитываемых возможностей мы допускаем все то, что может встретиться в языке, и тогда возрастают переборы в процессе анализа, а при многовариантном подходе возрастает и число построенных вариантов (за счет того, что в простом по строению предложении перебираются все возможности вплоть до самых экзотических). В другом случае, если мы сужаем круг допустимых ситуаций до самых стандартных, то мы не получим части правильных синтаксических структур. Выход видится в том, чтобы приблизить знания о ЕЯ, включенные в ЛП, к тем, которыми располагает носитель языка, учтя в них не только качественный аспект, но и разбираемый нами аспект меры. При этом, основываясь на постулатах общения, можно предположить, что мы сможем наилучшим образом имитировать человеческое понимание, если для каждого входного предложения сумеем построить самый естественный из правильных вариантов.

На уровне синтаксического анализа это приводит к необходимости построения такой системы правил, в которой альтернативные возможности снабжены некоторыми показателями степени нормативности или естественности, а алгоритм обеспечивает построение синтаксического представления, которое является не только правильным, но и наилучшим с точки зрения этих показателей.

4. Пример использования аспекта меры при автоматическом синтаксическом анализе. Алгоритм синтаксического анализа (САн) для русских текстов, основанный на очерченном здесь подходе, построен в системе АРТ (Анализ Русских Текстов), разработанной в Институте прикладной математики им. М. В. Келдыша АН СССР, и описан в работах [1—4]. Этот подход является обобщением фильтрового подхода, подробное описание которого содержится в [5].

Напомним кратко основные моменты фильтрового метода САн. Целью САн является построение для анализируемого предложения одной или нескольких синтаксических структур в виде размеченных деревьев зависимостей. Размеченное дерево зависимостей — это дерево, в узлах которого стоят текстовые единицы (ТЕ) входного предложения, снабженные наборами разнообразных признаков, а дуги соответствуют синтаксическим связям и снабжены пометами о типах связей. К ТЕ относятся словоформы, идиоматические словосочетания, выступающие в САн как единое целое, и знаки препинания.

Процесс построения синтаксической структуры такого вида состоит из следующих основных этапов. Сначала для анализируемого предложения строится некоторый, вообще говоря, избыточный набор синтаксических связей (ССв). При его построении учитывается только возможность двух ТЕ выступать компонентами некоторой ССв, но не учитывается ни контекст, ни сочетаемость ССв. Следующий этап состоит в сокращении исходного набора ССв путем учета ограничений на правильность сочетаний ССв. Ограничения сформулированы в виде правил, называемых фильтрами. Фильтры учитывают требования проективности, сочетаемость разных подчиненных элементов при общем управляющем и другие факторы.

Основной принцип отбрасывания лишних ССв — это принцип сохранения уникальных ССв. ССв называется уникальной, если она является

единственной, в которой ее подчиненный элемент выступает в качестве подчиненного. Поскольку мы стремимся построить связное дерево, для каждой словоформы, кроме корня дерева, надо найти ровно один управляющий элемент. Это значит, что уникальные ССв должны сохраняться при применении фильтров, а те ССв, которые с ними несовместимы, должны быть отброшены. Основываясь на этом, можно отбросить часть гипотетических ССв, но не всегда оставшийся набор оказывается единичным. Единичный набор ССв — это набор, в котором для каждой ТЕ есть ровно одна ССв и она уникальна в указанном выше смысле; но единичный набор может быть несвязным за счет наличия петель, т. е. еще не быть деревом.

Если в результате применения фильтров получился единичный набор, то это означает, что для данного предложения будет построен единственный правильный вариант анализа. Если же получен набор, который не является единичным и уже не поддается сокращениям при применении фильтров, то такой набор расщепляется на единичные, каждый из которых порождает свое дерево зависимостей.

Описанный подход имеет целый ряд положительных свойств: ясность и прозрачность, возможность лингвистически содержательного расчленения на этапы, устойчивость, застрахованность от зацикливания, ограничение переборов и т. д. (см. [5]). Однако, как уже было отмечено, его недостатком является то, что нередко правильный вариант оказывается не единственным и алгоритм строит правильные варианты, никак не ранжируя их по каким бы то ни было критериям.

Для названной выше системы АРТ разработано расширение фильтрового подхода. Расширение делается в нескольких направлениях, но здесь мы остановимся только на разбираемом нами аспекте меры (подробное описание САн в системе АРТ см. в [3. 4]).

Для учета предпочтений одних синтаксических конструкций по сравнению с другими вводится аппарат оценок ССв. Общая структура алгоритма выглядит при этом следующим образом. Как и раньше, сначала строится набор гипотетических синтаксических ССв, но при этом теперь они снабжаются некоторыми исходными оценками. Исходные оценки учитывают разнообразные факторы. Для актантных ССв, устанавливаемых по моделям управления, оценки учитывают степень необходимости заполнения данного места предиката, степень естественности каждого варианта заполнения и расположение связываемых ТЕ. Для сирконстантных ССв оценки учитывают прежде всего расположение связываемых ТЕ в анализируемом предложении: как их взаимный порядок, так и то, представителями каких синтаксических классов разделяются связываемые ТЕ.

Например, от существительного к согласованному с ним прилагательному будет устанавливаться гипотетическая определительная ССв, оценка которой будет тем ниже, чем больше ТЕ отделяют определяемое от определяющего; если эти две ТЕ соседствуют, оценка ССв будет максимальной. Таким образом, разрешенными оказываются определительные ССв и между соседними ТЕ (например, в таких сочетаниях, как *соседние слова*, *такой случай*), и между весьма далекими ТЕ, разделенными представителями разнообразных синтаксических классов, в том числе другими существительными (как, например, между *такой* и *случай* в сочетании *Такой ранее автором не рассматривавшийся случай*) Но при этом оценки они получат разные. В результате, если встретится сочетание *Такой случай автор называет...*, гипотетические определительные ССв к слову

такой пройдут и от *случай*, и от *автор*, но исходная оценка первой ССв будет выше.

Следующий этап состоит в пересчете оценок в зависимости от сочетания ССв. Если при чисто фильтровом подходе учет сочетаемости сводится к отбрасыванию каких-то ССв, то оценки дают более тонкие средства работы. Наряду с запрещением некоторых сочетаний, мы получаем возможность указать, что они допустимы, но нежелательны. Это выразится низкой оценкой данного сочетания; если наряду с ним возникнет другое, с более высокой оценкой, то нежелательное сочетание будет отброшено, а если лучшего не окажется, то оно войдет в результирующую структуру.

Здесь надо отметить, что мы описываем круг возможностей, которые при таком подходе предоставляются составителю системы правил в лингвистическом обеспечении некоторого ЛП для того, чтобы он мог отразить свои представления о степени допустимости того или иного сочетания, не предрешая вопрос о том, каковы должны быть конкретные оценки. Эти оценки он может выбирать.

При пересчете оценок они могут как увеличиваться, так и уменьшаться. Так, например, если для некоторого места предиката нашелся только один претендент на заполнение, то оценка связывающей их ССв повышается в зависимости от степени необходимости заполнения данного места, а оценки ССв, соединяющих другие ТЕ с той же ТЕ как подчиненной, соответственно понижаются. Снижение оценки до нуля означает отбрасывание данной ССв.

После пересчета оценок ССв для каждой ТЕ анализируемого предложения выбираются те (или та) ССв, которые получили в результате максимальные оценки. Если таких ССв оказывается несколько, то из них выбирается «наиболее естественная», например, учитывается направление ССв: так, для определения в русском языке естественнее препозиция: существительное в косвенном падеже, напротив, чаще стоит после своего управляющего, чем предшествует ему. В результате последнего выбора получаем единичный набор ССв, на основе которого строится дерево зависимостей, которому естественно приписать в качестве оценки сумму оценок вошедших в него ССв.

В исследовательских целях можно строить все варианты синтаксической структуры с соответствующими оценками. В частности, в применении к известному примеру *Мать любит дочь* обычный фильтровый метод выдаст два равноправных варианта анализа, на основе описанного подхода либо будет построено два варианта, но с разными оценками, либо, если будет дана установка на получение только наилучшего, будет выдан тот из них, где первое существительное является субъектом, а второе — объектом.

На одном простом примере хочется предостеречь от упрощенного использования при автоматическом САн того понятия меры, о котором идет речь. При построении синтаксической структуры чаще всего истинным управляющим для ТЕ является тот из них, который является ближайшим. Естественно, что возникает искушение строить синтаксическую структуру из кратчайших ССв, тем более, что легко количественно оценить расстояния и потом минимизировать сумму длин ССв. Однако можно показать на примерах, что выбор управляющего зависит не только от расстояния.

Рассмотрим предложение: *Он взял ближайший из примеров.* К предлогу *из* пройдут две гипотетические ССв: от глагола *взял* и от прилагательного *ближайший*. Управляющим в данном случае является прилагательное,

которое здесь соседствует с предлогом. Однако и при другой расстановке слов ССв должна идти к предлогу от прилагательного. Так, в предложении *Он взял из примеров ближайший* хотя глагол и соседствует с предлогом, все равно управляющим для предлога является прилагательное. Причем дело отнюдь не в том, что по каким-то причинам глагол *взял* не должен управлять сочетанием *из примеров*: в предложении *Он взял много интересного из примеров* предлогом *из* управляет *взял*. Дело именно в сочетании силы требований на наличие данного подчиненного: глагол *взял* может выступить в роли управляющего для предлога *из*, но если появляется более сильный претендент, он уступает эту роль ему.

Представляется, что этот пример достаточно явственно показывает, что нельзя ограничиваться выбором ближайшего из претендентов на роль управляющего. Заметим, что в рассмотренном примере требование проективности не помогает решить вопрос, определяющим является соотношение именно в силе требований на заполнение соответствующего места со стороны глагола и прилагательного.

Если пытаться обосновать выбор на чисто качественном уровне, то придется формулировать сложный набор условий, учитывающих различные возможные сочетания претендентов на роль управляющего для предлога, и указать, какое из них надо выбирать в том или ином сочетании. Представляется, что проще унифицировать этот процесс, введя в качестве средства аппарат оценок, присваиваемых ССв, как это было очерчено выше.

5. Уровень семантики. Очень коротко коснемся проявлений аспекта количества и меры при описании семантики. Здесь соображения можно высказать в основном в виде пожеланий относительно представления семантических сведений. Если мы хотим проводить семантический анализ сообщений, который позволил бы делать определенные выводы из содержания, то мы должны уметь строить соответствующие семантические структуры. В них словам должны соответствовать толкования, из которых на основе синтаксической структуры и специальных правил комбинирования толкований должно получаться толкование целого высказывания.

На уровне синтаксиса основной классифицирующий принцип — это принцип эквивалентности: синтаксические правила в основном формулируются применительно к классам слов, эквивалентных по своим синтаксическим свойствам.

В семантике полная эквивалентность (полная синонимия) практически отсутствует. На уровне семантики главным организующим принципом становится близость, а не совпадение. Соответственно желательно, чтобы семантические представления имели встроенные средства явного указания семантической близости и упорядоченности по степени близости. В семантических сетях, концептуализациях, семантических формулах и других распространенных семантических представлениях таких встроенных средств нет.

Один из возможных способов получения толкований, пригодных для явного указания наличия семантической близости, состоит в приписывании группе семантически близких слов некоторого общего набора параметров с разными значениями, определенным образом упорядоченными (это могут быть числовые значения либо значения, задаваемые словами, но с понятной упорядоченностью; в качестве примера укажем такой ряд значений: «никогда, очень редко, редко, нередко, часто, очень часто, всегда»).

Тривиальные примеры дают такие группы глаголов: группа *шептать*,

говорить. *кричать* естественным образом упорядочивается по громкости; группа *ползти*. *плестись*, *идти*. *бежать*, *нестись*. *мчаться* — по скорости; группа *уговорить*, *упросить*, *умолить* по степени трудности реализации и т. д. Параметрическое описание такого типа для группы ментальных предикатов дано в работе [6].

Представляется также, что существенными семантическими характеристиками глаголов должны быть такие, как характерная длительность, четкость границ по длительности, наличие естественного завершения. сохраняемость результата, возможность повторения без антидействия, кратность, стандартность/заурядность соответствующего действия. Каждый из этих признаков характеризуется своей шкалой значений.

Разберем в качестве примера одну из названных семантических характеристик, интересную в данном контексте проявлением количественных отношений. Речь пойдет о характерной длительности. Представляется, что для ряда действий и состояний в нашем представлении есть определенная стандартная длительность, определяемая либо самим действием, либо иногда его актантами. Она измеряется такими интервалами времени. как мгновения, минуты, часы, дни, недели, месяцы, годы и т. д. Иногда в характерной длительности объединяются несколько таких интервалов: например, для *умываться* стандартная длительность — минуты, для *болеть* — дни или недели, для *учиться в школе* — годы, для *писать записку* — минуты, *писать письмо* — минуты или часы, *писать роман* — недели, месяцы или годы. Наряду с видом глагола представление о характерном времени протекания некоторого действия и сочетание этого характерного времени с тем временным интервалом, который указан при данном действии, влияет на понимание того, достигло ли это действие завершения. Так, если временной интервал существенно больше, чем характерная длительность действия, то содержащее их высказывание даже при несовершенном виде глагола воспринимается так: в указанный интервал времени субъект начал некоторое действие, произвел его и завершил.

Рассмотрим например, высказывание: *Вчера А читал доклад*. Поскольку *вчера* — задает временной интервал длительностью в один *день*, а для *читать доклад* нормальное время — минуты или часы, то это высказывание понимается так же, как высказывание *Вчера А прочитал доклад*, т. е. «начал читать, читал и прочитал доклад». Для высказывания *Вчера А читал повесть* вопрос о завершенности остается открытым, а высказывание *Вчера А читал роман* воспринимается как сообщение о действии чтения, которое не было закончено, поскольку тут интервал, задаваемый словом *вчера*, меньше характерного времени действия. Других источников информации о завершении действия, кроме знания характерного для него времени, в этих высказываниях нет.

Соотношение характерного времени протекания действия и указанного в явном виде временного интервала может иногда быть источником информации о кратности действия. Сравним, например, высказывания *На прошлой неделе А ездил в Ленинград*, *В прошлом году А ездил в Ленинград*, *В молодости А ездил в Ленинград*. Хотя никаких явных указаний о кратности нет, первое мы воспримем скорее как указание на однократное действие, второе неопределенно с точки зрения кратности, а третье — воспринимается скорее как сообщение о многократном действии.

Заметим, что решение о завершенности или кратности действия на основе соотношения временного интервала и характерной длительности

действия происходит не абсолютно, а на уровне предпочтений. Это соотношение позволяет дополнить явно заданную информацию и выбрать наиболее естественное понимание из возможных.

Представляется, что к рассматриваемому аспекту относятся также положения о неравноправии тех или иных сочетаний слов (выражающиеся в предпочтительности одних по сравнению с другими), которые разобраны в известной работе Уилкса [7].

6. О схеме модели для описания естественных языков. В работе [8] приведена общая схема модели для описания естественных языков, в которой систематически проведена идея использования аппарата оценок для операций и конструкций ЕЯ.

Предлагается строить модель ЕЯ в виде многоуровневой лингвистической системы (МЛС), состоящей из одноуровневых лингвистических систем (ОЛС).

Каждая ОЛС — это иерархическая система, где каждый ранг иерархии включает в себя объекты, отношения трех видов (классифицирующие, структурные и ограничительные), оценочные функции и операции. Объекты нижнего уровня — атомы; они разбиты на классы классифицирующими отношениями. Атомы объединяются в неатомарные объекты при помощи структурных отношений, при этом учитываются классифицирующие отношения. Для составных объектов должны выполняться ограничительные отношения. Неатомарные объекты конструируются из простых или разлагаются на простые при помощи операций, которые делятся на объединяющие, разъединяющие и заменяющие. Применение операции влечет за собой вычисление оценок для построенных с ее помощью объектов. Благодаря этому параллельно с процессом анализа (разложения на более элементарные) или синтеза некоторого объекта идет процесс получения оценки результата в зависимости от того, какая последовательность операций была при этом применена и каковы были оценки операндов.

Различные ОЛС в пределах МЛС связаны между собой преобразующими операциями, для которых также определено вычисление оценок результата на основе оценок операндов.

Например, ОЛС, соответствующая морфологическому уровню, описывает построение словоформы из морфем с помощью объединяющих операций, разбиение словоформы на морфемы с помощью разъединяющих операций, чередование с помощью заменяющих операций.

ОЛС другого уровня служит для описания синтаксиса. Здесь атомарные объекты — это представления словоформ. Классифицирующие отношения задают деление на классы в соответствии с наличием определенных синтаксических признаков. Структурные отношения связывают словоформы в синтаксические конструкции, затем объединяют их в синтаксическую структуру всего предложения (в ней могут быть части, соответствующие простым предложениям, группам слов и т. п., а также части, соответствующие теме и реме). Ограничительные отношения, или условия, регулируют применимость операций и тем самым направляют процесс анализа или синтеза текста. Альтернативные результаты анализа получают при этом оценки в зависимости от того, каким путем (т. е. при помощи какой последовательности операций) они были получены, и, соответственно, на какие более простые объекты был разложен анализируемый.

СПИСОК ЛИТЕРАТУРЫ

1. *Кулагина О. С.* Морфологический анализ русских глаголов / Препринт ИПМ им. М. В. Келдыша АН СССР. 1985. № 195.
2. *Кулагина О. С.* Морфологический анализ русских именных словоформ Препринт ИПМ им. М. В. Келдыша АН СССР. 1986. № 10.
3. *Кулагина О. С.* Об автоматическом синтаксическом анализе русских текстов /. Препринт ИПМ им. М. В. Келдыша АН СССР. 1987. № 205.
4. *Кулагина О. С.* О синтаксическом анализе на основе предпочтений Препринт ИПМ им. М. В. Келдыша АН СССР. 1990. № 3.
5. *Кулагина О. С.* Исследования по машинному переводу. М., 1979.
6. *Кулагина О. С.* О параметрическом представлении смысла некоторых ментальных предикатов / Препринт ИПМ им. М. В. Келдыша АН СССР. 1990. № 3.
7. *Wilks Y.* Preference semantics. / Stanford A. I. Lab. Memo AIM-206. Stanford. 1973.
8. *Кулагина О. С.* О моделировании естественных языков Препринт ИПМ им. М В. Келдыша АН СССР. 1981. № 138.