УДК 004.04: 519.767
DOI: 10.15827/2311-6749.16.4.4

# *МЕТОДЫ И СИСТЕМЫ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ*

*Т.В. Батура, к.ф.-м.н., tatiana.v.batura@gmail.com*
*(Институт систем информатики им. А.П. Ершова СО РАН,*
*просп. Акад. Лаврентьева, 6, г. Новосибирск, 630090, Россия)*

Статья посвящена проблемам семантического анализа текстов. Рассмотрены различные методы: концептуальные диаграммы зависимостей и семантические сети; подходы, основанные на лексических функциях и тематических классах; фреймовые и онтологические модели; логические модели представления знаний. У каждого из них имеются свои достоинства и недостатки.

Создание новых методов семантического анализа текстов актуально в решении многих задач компьютерной лингвистики, таких как машинный перевод, автореферирование, классификация текстов и других. Не менее важна разработка новых инструментов, позволяющих автоматизировать семантический анализ.

***Ключевые слова:*** *семантический анализ, автоматическая обработка текста, извлечение информации, семантические сети, логика предикатов, представление знаний, смысл высказывания.*

Семантика – раздел лингвистики, изучающий смысловое значение единиц языка. Помимо знаний о структуре языка, семантика тесно связана с философией, психологией и другими науками, так как неизбежно затрагивает вопросы о происхождении значений слов, их отношении к бытию и мышлению. При семантическом анализе необходимо учитывать социальные и культурные особенности носителя языка. Процесс человеческого мышления, как и язык, который представляет собой инструмент выражения мыслей, является очень гибким и трудно поддается формализации. Поэтому семантический анализ по праву считается самым сложным этапом автоматической обработки текстов.

Создание новых методов семантического анализа текстов откроет новые возможности и позволит существенно продвинуться в решении многих задач компьютерной лингвистики, таких как машинный перевод, автореферирование, классификация текстов и других. Не менее актуальна разработка новых инструментов, позволяющих автоматизировать семантический анализ.

На данный момент существует несколько методов представления смысла высказываний, однако ни один из них не является универсальным. Над соотнесением смысла тексту работали многие исследователи. Так, И.А. Мельчук [1] ввел понятие лексической функции, развил понятия синтаксических и семантических валентностей и рассмотрел их в контексте толково-комбинаторного словаря, который представляет собой языковую модель. Он показал, что значения слов соотносятся не непосредственно с окружающей действительностью, а с представлениями носителя языка об этой действительности.

Большая часть исследователей склоняются к мысли, что семантический анализ должен выполняться после синтаксического. В.Ш. Рубашкин и Д.Г. Лахути [2] ввели иерархию синтаксических связей для более эффективной работы семантического анализатора. Самыми важными являются обязательные ролевые связи, далее идут связи кореференции, потом факультативные ролевые связи и только потом предметно-ассоциативные.

Известный лингвист Е.В. Падучева [3] предлагает рассматривать тематические классы слов, в частности глаголов, поскольку они несут основную смысловую нагрузку: глаголы восприятия, глаголы знания, глаголы эмоций, глаголы принятия решения, речевых действий, движения, глаголы звука, бытийные глаголы и др. Существенной в данном подходе является идея разделять понятия языка на некоторые семантические группы с учетом того, что эти понятия имеют некоторый нетривиальный общий смысловой компонент. Элементы таких групп склонны иметь один и тот же набор зависимых понятий. Однако главная проблема такого подхода заключается в том, что выделение тематических классов и составление семантических словарей чрезвычайно трудоемкий процесс, сильно зависящий от индивидуального восприятия и интерпретации понятий конкретным человеком.

Универсальный язык представления знаний должен быть удобным инструментом для вывода новых знаний из уже имеющихся, а значит, необходимо создать аппарат для проверки правильности высказываний. Здесь как раз полезны логические модели представления знаний. Например, семантический язык, предложенный В.А. Тузовым [4], содержит в себе формализмы логики предикатов, в нем присутствуют «атомарные» понятия, «функции» над этими понятиями и правила вывода, с помощью которых можно описывать новые понятия. Не исключено, что в направлении создания подобных семантических языков будет развиваться научная мысль в будущем.

Несмотря на то, что некоторые научные и технические идеи в области обработки текстов развиваются довольно быстро, многие проблемы семантического анализа остаются нерешенными. Большинство исследователей пришли к выводу, что словарь для поддержки семантического анализа должен оперировать

смыслами и, следовательно, описывать свойства и отношения понятий, а не слов. Но возникает вопрос, как правильно структурировать и представлять информацию в подобных словарях, чтобы поиск по ним был удобным и быстрым, к тому же была бы возможность учитывать изменения в естественном языке (исчезновение старых и возникновение новых понятий). В данной статье предпринята попытка систематизировать известные достижения в области семантического анализа и в какой-то мере найти ответ на этот и другие вопросы.

## Исследование семантики в рамках теории «Смысл ⇔ Текст»

При создании теории «Смысл ⇔ Текст» И.А. Мельчук [1] ввел понятие **лексической функции**. С формальной точки зрения лексическая функция – есть функция, аргументами которой являются некоторые слова или словосочетания данного языка, а значениями – множества слов и словосочетаний этого же языка. При этом представляют содержательный интерес и рассматриваются только такие лексические функции, у которых имеются фразеологически связанные значения – значения, которые возможны при одних аргументах и невозможны при других.

Иначе говоря, лексическая функция есть определенное смысловое соотношение, например, «равенство по смыслу» (*Syn*), «противоположность по смыслу» (*Anti*) и др. Пусть имеется ряд лексических единиц – слов и словосочетаний; тогда данная лексическая функция ставит в соответствие каждой из этих единиц набор лексических единиц, находящихся с исходной единицей в соответствующем смысловом соотношении.

Значения одной лексической функции от разных аргументов могут полностью или частично совпадать; могут совпадать и значения разных функций от одного аргумента. Альтернативные корреляты, входящие в значение данной лексической функции от данного аргумента, вовсе не обязаны быть взаимозаменяемыми всегда и в любом контексте. Они могут различаться по стилистическим характеристикам, по всем типам сочетаемости, по грамматическим условиям употребления и, наконец, даже по смыслу. Последнее особенно важно подчеркнуть: различные корреляты не всегда должны быть полностью синонимичны; достаточно, если у их смыслов есть общая часть, отвечающая данной лексической функции, а различия не выходят за некоторые рамки, то есть не являются «слишком значительными».

В общем случае лексическая функция определяется не для всех слов и словосочетаний. Во-первых, некоторые функции определены лишь для той или иной части речи: так, *Oper*, *Func* и *Labor* мыслимы лишь для существительных. Во-вторых, та или иная функция может определяться только для слов определенной семантики: *Magn* – для слов, смысл которых допускает градацию («больше – меньше»); *Oper*, *Func* и *Labor* определены только для названий ситуаций.

Следует иметь в виду, что и при вполне подходящем (по своим синтаксическим и семантическим свойствам) аргументе лексическая функция может не иметь значения (в данном языке). Например, синонимы, в принципе, возможны для любых слов, а имеются только у некоторых. Это связано с фразеологическим характером лексических функций.

Необходимо еще раз подчеркнуть, что изначально лексические функции были введены специально для описания лексической сочетаемости, а не для представления смысла в общем понимании, поэтому не все из них следует трактовать как семантические единицы. Соотношения лексических функций со смыслом далеко не однозначны. Одни лексические функции могут претендовать на статус смысловых элементов, другие могут вообще не иметь смысла, третьи могут покрывать весьма сложный смысл.

С нашей точки зрения, говорить о лексических функциях как о «многозначных» функциях, не совсем корректно и удобно. Удобнее говорить о лексических предикатах. Далее приведен перечень простых стандартных лексических «функций» (здесь они будут представлены в виде предикатов).

1. *Syn* (*x*, *y*), *x*, *y* – синонимы.

2. *Conv* (*x*, *y*), *x*, *y* – конверсивы.

3. *Anti* (*x*, *y*), *x*, *y* – антонимы.

4. *Der* (*x*, *y*), *y* – синтаксический дериват *x*, то есть *y* совпадает с *x* по смыслу, но принадлежит к другой части речи:

$S_0(x, y)$, *y* – существительное, производное от *x* (*x* – не существительное);

$A_0(x, y)$, *y* – прилагательное, производное от *x* (*x* – не прилагательное);

$Adv_0(x, y)$, *y* – наречие, образованное от *x* (*x* – не наречие);

$V_0(x, y)$, *y* – глагол, образованный от *x* (*x* – не глагол).

Иначе говоря, $\forall x \forall y$ (*Der* $(x, y) \leftrightarrow S_0(x, y) \vee A_0(x, y) \vee \mathrm{Adv}_0(x, y) \vee V_0(x, y)$).

5. *Gener* (*x*, *y*), *y* – обобщающее понятие по отношению к понятию, обозначенному *x* (*x* = *клубника*, *y* = *ягода*). Этот предикат зависит от лексической сочетаемости слов в данном языке: если *x* и *m* – совпадающие по смыслу слова двух разных языков, то для *Gener* (*x*, *y*) и *Gener* (*m*, *n*), соответственно, *y* и *n* могут не совпадать по смыслу.

**Ситуация** – определенное лексическое отражение (в данном языке) некоторой части действительности. Ситуации, обозначаемые отдельными лексическими единицами естественных языков (лексемами), имеют, как правило, от одного до четырех смысловых компонентов, или **семантических актантов,** обозначаемых заглавными латинскими буквами A, B, C, D. В то же время каждой такой лексеме сопоставляются **глубинно-синтаксические актанты** – ее зависимые, соответствующие подлежащему и сильным дополнениям (в случае, если данная лексема реализуется глаголом-сказуемым). Глубинно-синтаксические актанты нумеруются арабскими цифрами: 1, 2, 3, 4.

6. $S_i(x, y)$, $i = 1, \ldots, 4$, $y$ – типовое название $i$-го актанта для $x$.

7. $S_c(x, y)$, $y$ – сирконстанта, то есть типовое название второстепенной компоненты данной ситуации $x$:

$S_{loc}(x, y)$, $y$ – типовое название места осуществления данной ситуации $x$; «то, где…» ($x$ = *битва*, $y$ = *поле* (*битвы*));

$S_{instr}(x, y)$, $y$ – типовое название инструмента, используемого в данной ситуации $x$; «то, чем/посредством чего…» ($x$ = *борьба*, $x$ = *орудие* (*борьбы*));

$S_{mod}(x, y)$, $y$ – типовое название способа (манеры, характера) осуществления данной ситуации $x$; «то, как…» ($x$ = *жизнь*, $y$ = *образ* (*жизни*));

$S_{res}(x, y)$, $y$ – типовое название результата данной ситуации; «то, что получается» ($x$ = *копировать*, $y$ = *копия*).

Иначе говоря, $\forall x \forall y\ (S_c(x, y) \leftrightarrow S_{loc}(x, y) \vee S_{instr}(x, y) \vee S_{mod}(x, y) \vee S_{res}(x, y))$.

8. Соотносительные предикаты *Sign* ($x$, $y$), $y$ – типовое название одной «штуки», одного «кванта» некоторого $x$; *Mult* ($x$, $y$), $y$ – типовое название совокупности, множества.

9. *Sigur* ($x$, $y$), $y$ – метафора для $x$ ($x$ = *сон*, $y$ = *объятия* (*сна*)).

10. *Centr* ($x$, $y$), $y$ – типовое обозначение «центральной» части предмета или процесса.

11. $A_i(x, y)$, $i = 1, \ldots, 4$, $y$ – типовое определение $i$-го актанта по его реальной роли; «такой, который…»; «такой, которого…».

12. $Able_i(x, y)$, $i = 1, \ldots, 4$, $y$ – типовое определение $i$-го актанта по его потенциальной роли в ситуации; «такой, который может...»; «такой, которого можно...».

13. $Magn_0(x, y)$ и $Magn_i(x, y)$, $i = 1, \ldots, 4$, $y$ обозначает «высокую степень», «интенсивность» самой ситуации $x$ ($Magn_0$) или ее $i$-го актанта ($Magn_i$).

14. *Ver* ($x$, $y$), $y$ – «правильный», «соответствующий назначению», «какой следует» применительно к $x$.

15. *Bon* ($x$, $y$), $y$ – «хороший» применительно к $x$.

16. $Adv_{ix}(z, y)$, $i = 1, \ldots, 4$, $x$ = A, B, C, D, $y$ – имя ситуации в роли определения при глаголе, называющем другую ситуацию:

$Adv_{iA}(z, y)$, $i = 1, \ldots, 4$, $y$ – слово, образованное от $z$, которое, заменяя $z$ в тексте, требует превратить в вершину (вместо $z$) первый актант этого $z$ ($x$ = *сопровождать*, $y$ = *вместе с*).

$Adv_{iB}(z, y)$, $i = 1, \ldots, 4$, $y$ требует становиться вершиной второй актант $z$ ($x$ = *ошибаться*, $y$ = *ошибочно*).

17. *Loc* ($x$, $y$), $y$ – предлог типовой локализации (пространственной, временной или абстрактной):

$Loc_{in}(x, y)$, $y$ – «статическая» локализация ($x$ = *Москве*, $y$ = *в*);

$Loc_{ad}(x, y)$, $y$ – предлог направления ($x$ = *Москву*, $y$ = *в*);

$Loc_{ab}(x, y)$, $y$ – предлог удаления ($x$ = *Москвы*, то $y$ = *из*).

Иначе говоря, $\forall x \forall y\ (Loc(x, y) \leftrightarrow Loc_{in}(x, y) \vee Loc_{ad}(x, y) \vee Loc_{ab}(x, y))$.

Иногда *Loc* ($x$, $y$) не удается определить однозначно ($x$ = *снег*, $y$ = *на* и $y$ = *в*).

18. *Copul* ($x$, $y$), $y$ – глагол-связка «быть», «являться» ($x$ = *нападал*, $y$ = *совершал нападение*).

19. $Oper_1(x, y)$, $Oper_2(x, y)$, $y$ – глагол, связывающий название первого (соответственно второго) актанта в роли подлежащего с названием ситуации в роли первого дополнения (если $x$ = *поддержка*, то $y$ = *оказывать* для $Oper_1(x, y)$, а $y$ = *находить* или *встречать* для $Oper_2(x, y)$).

20. $Func_0(x, y)$, $Func_1(x, y)$, $Func_2(x, y)$, $y$ – глагол, имеющий название ситуации в роли подлежащего $x$ с названиями актантов (если они есть) в роли дополнения ($x$ = *дождь*, $y$ = *идти*).

21. $Labor_{12}(x, y)$, $y$ – глагол, связывающий название первого актанта в роли подлежащего, с названием второго актанта в роли первого дополнения и с названием ситуации в роли второго дополнения ($x$ = *орден*, $y$ = *награждать*; $x$ = *наказание*, $y$ = *подвергать*).

22. $Caus_{ij}(x, y)$, $y$ – действие актантов «делать так, чтобы…», «каузировать». В случае без актантных индексов *Caus* ($x$, $y$), $x$ – название неучастника ситуации ($x$ = *преступление*, $y$ = *толкать*). Отдельно выступает только при глаголах, в остальных случаях входит в состав сложных параметров.

23. *Incep* ($x$, $y$), $y$ – «начинать». Свойства те же, что и у $Caus_{ij}(x, y)$.

24. *Perf* ($x$, $y$), $y$ – «перфектив», $y$ несет завершенность действия, достижение им своего естественного предела. Отдельного самостоятельного выражения *Perf* ($x$, $y$) в русском языке не имеет; как правило, этот предикат выдает истинное значение, если $y$ имеет форму совершенного вида ($x$ = *читать*, $y$ = *прочитать*).

25. *Result* (*x*, *y*), *y* – «результатив», то есть *y* – «состояние в результате…»; используется для форм несовершенного вида (*x* = *ложиться*, *y* = *лечь* для *Perf* (*x*, *y*), *y* = *лежать* для *Result* (*x*, *y*)).

26. $Fact^{j}$ (*x*, *y*), *y* – «реализоваться», «выполниться». Верхний индекс (римские цифры) представляет, если это надо, степень осуществления подразумеваемого требования, причем меньший индекс присваивается более низкой степени (если *x* = *капкан* и *j* = *I*, то *y* = *срабатывать*; если *j* = *II*, то *y* = *поймать*).

27. $Real^{j}_{1,2}(x, y)$, *y* – «реализовать», «выполнить требование», содержащееся в *x*. Индекс *j* имеет то же значение, что и выше – степень выполнения; нижний индекс обозначает глубинно-синтаксический актант, выполняющий требование (*x* = *долг (денежный)*, *y* = *признавать* для $Real^{I}_{1,2}(x, y)$, *y* = *погашать* для $Real^{II}_{1,2}(x, y)$).

28. *Destr* (*x*, *y*), *y* – типовое название «агрессивного» действия (*x* = *оса*, *y* = *жалит*).

29. *Cap* (*x*, *y*), *y* – «начальник» (*x* = *факультет*, *y* = *декан*).

30. *Equip* (*x*, *y*), *y* – «личный состав» (*x* = *население*, *y* = *государства*).

31. *Doc* (*x*, *y*), *y* – «документ»:

$Doc_{res}$ (*x*, *y*), *y* – «документ, являющийся результатом»; «воплощающий в себе» (*x* = *отчитываться*, *y* = *отчет*);

$Doc_{perm}$ (*x*, *y*), *y* – «документ на право…» (*x* = *поезд*, *y* = (*проездной*) *билет* для $Doc_{perm}\, Oper_2$ (*x*, *y*));

$Doc_{cert}$ (*x*, *y*), *y* – «документ, удостоверяющий…» (*x* = *высшее образование*, *y* = *диплом*).

Иначе говоря, $\forall x \forall y\, (Doc\ (x, y) \leftrightarrow Doc_{res}\ (x, y) \vee Doc_{perm}\ (x, y) \vee Doc_{cert}\ (x, y))$.

Помимо перечисленных выше простых лексических предикатов, для описания лексической сочетаемости могут использоваться и их комбинации – составные предикаты:

$AntiReal_2$ (*x*, *y*): проваливать экзамен/проваливаться на экзамене;

$IncepOper_2$ (*x*, *y*): приобретать популярность, впадать в отчаяние;

$IncepOper_2$ (*x*, *y*): поступать в продажу, попадать под обстрел;

$CausOper_2$ (*x*, *y*): ставить под контроль, пускать в обращение.

Как уже отмечалось ранее, в общем случае лексическая функция определяется не для всех слов и словосочетаний. Функция может определяться только для слов определенной семантики. Например, *Cap* и *Equip* – для слов, смысл которых предполагает наличие «начальника» и «персонала», то есть для названий учреждений и организаций в самом широком понимании; *Real* – для слов, в смысл которых входит компонент «требовать» («нужно»), и т. д.

Если лексические функции представлять в виде предикатов, не возникает никаких затруднений. В случаях, когда лексические функции не определены, соответствующие им предикаты будут ложными.

Особую роль при исследовании семантики в подходе И.А. Мельчука играют **валентности слов**, то есть способности слов вступать в связи с другими словами. Валентностями обладают слова, которые задают ситуацию. Это все глаголы, некоторые существительные (отглагольные), прилагательные (обозначающие сравнение: больше, меньше, выше, ниже), некоторые предлоги и наречия.

Различают два вида валентностей слова: синтаксические и семантические. Хотя это разделение иногда бывает довольно условным. **Семантические валентности** определяются лексическим анализом ситуации, задаваемой конкретным словом. Приведем пример со словом *аренда* или *арендовать*. A *арендует* C означает, что за какое-то вознаграждение D лицо A приобретает у другого лица B право на эксплуатацию собственности C в течение времени T. Следовательно, существенными для ситуации *аренды* являются следующие «участники» или семантические актанты: субъект аренды (тот, кто арендует), первый объект аренды (то, что арендуют), контрагент (тот, у кого арендуют), второй объект (плата) и срок.

Эти актанты необходимы, так как устранение какого-либо из них изменяет смысл ситуации. Например, если убрать срок, то ситуация аренды трансформируется в ситуацию купли-продажи. С другой стороны, эти актанты достаточны, поскольку в ситуации аренды не требуется указание того, по какой причине, где, когда и с какой целью она осуществлялась. Хотя соответствующие словоформы грамматически присоединимы к глаголу *арендовать*.

Другими словами, семантическая валентность определяется числом семантических актантов. Таким образом, глагол *арендовать* имеет семантическую валентность 5, так как у него 5 семантических актантов. Формально эту ситуацию можно записать в виде предиката *P* ($x_1$, $x_2$, $x_3$, $x_4$, $x_5$), где $x_1$ – «кто», $x_2$ – «что», $x_3$ – «у кого», $x_4$ – «плата», $x_5$ – «срок».

В предложении могут быть определены не все семантические актанты, некоторые могут просто не упоминаться или вообще не иметь синтаксического выражения. **Синтаксические валентности** определяются количеством синтаксических актантов, которые представлены непосредственно в тексте (то есть присоединяемыми к глаголу подлежащими и дополнениями) и зависят от контекста.

Например, семантическая валентность глагола *промахнуться* равна 4, так как он имеет 4 актанта: кто (деятель), во что/по чему (мишень), из чего (оружие – факультативно) и чем (орган, средство). Но в большинстве контекстов синтаксически выражается лишь одна валентность, например, в предложении *«Он долго целился, но промахнулся»*. Тем не менее не совсем корректной является фраза *«Он промахнулся в окно бутылкой»*.

С формальной точки зрения мы имеем конструкцию, описанную ниже. Чтобы не связывать с каждым глаголом (и другими словами) отдельный предикат, будем рассматривать предикат, размерность которого больше на 1: $P^{val}(y, x_1, x_2, \ldots, x_n)$, при этом $y$ будет само слово, а $x_1, x_2, \ldots, x_n$ – его валентности. Чтобы отличать синтаксические и семантические актанты, можно использовать мультииндексы для указания актантов, заданных в тексте. Запись $P^{val}_{i_1 i_2 \ldots i_k}(y, x_{i_1}, x_{i_2} \ldots, x_{i_k})$ означает, что заданы актанты $i_1, i_2, \ldots, i_k$. В частности, если заданы все актанты, то получаем $P^{val}_{1\ldots n}(y, x_1, x_2, \ldots, x_n)$. Некоторые варианты (наборов мультииндексов) могут быть недопустимы в языке. Если набор $i_1, i_2, \ldots, i_k$ допустим, то имеет место импликация

$$\forall y \forall x_1 \ldots \forall x_n (P^{val}_{1\ldots n}(y, x_1, x_2, \ldots, x_n) \to P^{val}_{i_1 i_2 \ldots i_k}(y, x_{i_1}, x_{i_2}, \ldots, x_{i_k})).$$

Более того, если имеются два набора допустимых мультииндексов $<i_1, i_2, \ldots, i_k>$ и $<i_1', i_2', \ldots, i_s'>$, таких, что $\{i_1, i_2, \ldots, i_k\} \supseteq \{i_1', i_2', \ldots, i_s'\}$, то имеет место аналогичная импликация

$$\forall y \forall x_{i_1} \ldots \forall x_{i_k} \forall x_{i_1'} \ldots \forall x_{i_s'} (P^{val}_{i_1 \ldots i_k}(y, x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \to P^{val}_{i_1' i_2' \ldots i_s'}(y, x_{i_1'}, x_{i_2'}, \ldots, x_{i_s'})).$$

**Толково-комбинаторный словарь** – одно из главных теоретических изобретений И.А. Мельчука. В каком-то смысле языковая модель, предложенная И.А. Мельчуком, представляет язык как совокупность словарных статей с огромным количеством разнообразной информации; грамматические правила при таком словаре играют скорее второстепенную роль. Толково-комбинаторный словарь отражает прежде всего нетривиальную сочетаемость лексем. Можно считать, что язык – это очень большая модель, в которой определены лексические предикаты, действующие описанным выше образом.

Статья толково-комбинаторного словаря несет информацию о валентностях конкретного слова, верную не только в ее рамках, но и в рамках всего языка в целом. Валентности соответствует предикат $P^{val}(c_x, \bar{y})$, где $\bar{y} = y_1, \ldots, y_n$ – семантические актанты слова $c_x$, $n$ – валентность слова $c_x$. Например, в предложении *Петя читает книгу* будет $c_x$ = *читать*, $n$ = 2: $y_1$ = *Петя*, $y_2$ = *книга*, то есть условно можно написать $P^{val}(c_x, y_1, y_2) = 1$.

Набор статей в толково-комбинаторном словаре можно считать некоторой подмоделью исходной модели, являющейся языком. Лексические предикаты, определенные теперь на более узком множестве, будут действовать аналогично.

Обозначим Ф множество правильно построенных фраз естественного языка $L$ и $\varphi \in \Phi$ – фразу из этого множества; $c_x \prec \phi$ – слово $c_x$ входит во фразу φ, причем $c_x \in L$. Пусть $c_x$ – существительное или прилагательное. Обозначим через *Predicate* множество предикатов, определенных на $L$. Одним из элементов этого множества является введенный ранее предикат валентности $P^{val}(c_x, y_1, \ldots, y_n)$.

Аналогично можно предполагать, что имеются другие предикаты:

– предикат рода слова $Gender(c_x, \bar{G})$, где $\bar{G} \in \{g_1, g_2, g_3\}$, $g_1$ = *жен.*; $g_2$ = *муж.*; $g_3$ = *ср.*;

– предикат предлога $Preposition(c_x, \overline{Prep})$, где $\overline{Prep} \in \{pr_1, \ldots, pr_k\}$ – множество предлогов, сочетаемых с данным словом;

– предикат падежа $Cases(c_x, \overline{Case})$, где $\overline{Case}$ – падеж слова $c_x$; для разных языков число падежей различно: например, в русском языке шесть падежей, поэтому $\overline{Case} \in \{case_1, case_2, case_3, case_4, case_5, case_6\}$, $case_1$ = *им.п.*; $case_2$ = *род.п.*; $case_3$ = *дат.п.*; $case_4$ = *вин.п.*; $case_5$ = *твор.п.*; $case_6$ = *предл.п.*; в немецком – четыре падежа, поэтому $\overline{Case} \in \{case_1, case_2, case_3, case_4\}$, где $case_1$ = *Nom*; $case_2$ = *Gen*; $case_3$ = *Dat*; $case_4$ = *Akk*.

Словарная статья толково-комбинаторного словаря содержит основное слово, лексические предикаты, связанные с ним, и информацию о валентности данного слова. Информация о валентности включает в себя число, указывающее число актантов, и для каждого актанта – указание, в каких падежах и с какими предлогами используются слова, соответствующие данному актанту. В отдельных случаях может быть указан также род слова.

Сказанное выше может быть представлено посредством набора предикатов вида

$$P(x_i, \bar{G}, \overline{Prep}, \overline{Case}) = Gender(x_i, \bar{G}) \wedge Preposition(x_i, \overline{Prep}) \wedge Cases(x_i, \overline{Case}),$$

где $x_i$ – свободная переменная, соответствующая $i$-му актанту.

Теория «Смысл ⇔ Текст» с самого начала создавалась для применения в прикладной проблематике автоматического перевода. По замыслу И.А. Мельчука, с ее помощью, в отличие от традиционных нестрогих теорий, следовало обеспечить построение «действующей» модели языка. Теория «Смысл ⇔ Текст» действительно была использована в некоторых системах машинного перевода, разработанных в России, – прежде всего в системе англо-русского автоматического перевода ЭТАП, созданной группой под руководством Ю.Д. Апресяна. Все эти системы относятся к числу экспериментальных, то есть их промышленное использование не представляется возможным. Несмотря на то, что они включают много

лингвистически полезной информации, в целом ни одна из них пока не обеспечила прорыва в качестве перевода.

На взгляд автора, основная ценная идея этой теории состоит в том, что значения слов соотносятся не непосредственно с окружающей действительностью, а с представлениями носителя языка об этой действительности (иногда называемыми концептами). Природа концептов зависит от конкретной культуры; система концептов каждого языка образует так называемую «наивную картину мира», которая во многих деталях может отличаться от «научной» картины мира, являющейся универсальной. Задача семантического анализа лексики в теории «Смысл ⇔ Текст» состоит как раз в том, чтобы обнаружить «наивную картину мира» и описать ее основные категории. Другими словами, важная роль этой теории состоит в описании не только объективной, но и субъективной картины мира.

Несмотря на то, что интерес к теории И.А. Мельчука угасает, разметка синтаксического корпуса «Национальный корпус русского языка» [5] выполняется лингвистическим процессором ЭТАП-3, основанным на принципах теории «Смысл ⇔ Текст».

Как уже упоминалось выше, в разработке процессора ЭТАП участвовал Ю.Д. Апресян. Его идеи несколько отличаются от идей И.А. Мельчука. Центральное место в исследованиях Ю.Д. Апресяна занимает синонимический словарь нового типа [6]. Для этого словаря была разработана подробная схема описания **синонимических рядов**, где каждый элемент ряда характеризовался с точки зрения семантики, синтаксиса, сочетаемости и других свойств. В словаре собрано и обобщено максимальное количество информации о языковом поведении русских синонимов.

## Концептуальные диаграммы зависимостей

***Концептуальный и прецедентный анализ***

На этапе морфологического и семантико-синтаксического анализа текстов основными единицами, обозначающими понятия, являются слова. Как правило, считается, что смысл словосочетаний и фраз может быть выражен через смыслы составляющих их слов. В качестве исключений рассматривается лишь ограниченное число устойчивых словосочетаний – идиом. Такой подход опирается на предположение, что словосочетания, встречающиеся в языке, можно разделить на «свободные» и «несвободные».

Другой подход [7] основывается на том, что наиболее устойчивыми (неделимыми) единицами смысла являются категории и понятия, состоящие не из самостоятельных слов, а из словосочетаний. Такие категории и понятия называются концептами. При таком подходе «несвободными» словосочетаниями являются не только идиоматические выражения, но и все устойчивые фразеологические единицы языка и речи (в развитых языках их насчитываются сотни миллионов).

Идея концептуального анализа как неотъемлемой составляющей семантического анализа встречается также в исследованиях В.Ш. Рубашкина и Д.Г. Лахути [2, 8, 9]. В данном разделе кратко излагаются взгляды на вопрос о том, какие задачи должны решаться средствами концептуального семантического анализа.

На вход семантического компонента должен поступить синтаксически размеченный текст. В размеченном тексте должна быть представлена различная информация: идентификаторы понятий, соответствующих слову (термину); указание синтаксического хозяина (всех альтернативных хозяев) и вида синтаксической связи и др.

До передачи в семантический компонент должны быть также опознаны термины-словосочетания, унифицировано представление числовой информации, опознаны собственные имена и т.п. В реальных проектах все эти задачи решаются с той или иной степенью приближения. Можно считать, что профессиональное сообщество пришло к согласию, по крайней мере, в следующих вопросах.

Семантический анализ, с точки зрения используемых методов и средств, должен предусматривать два этапа: а) этап интерпретации грамматически выраженных (синтаксических и анафорических) связей и б) этап распознавания связей, не имеющих грамматического выражения.

Неоднозначности должны разрешаться самим процессом анализа – по критерию степени смысловой удовлетворительности получаемого в каждом варианте результата.

Ключевым пунктом системы семантического анализа является эффективная словарная поддержка. В этом смысле любая система семантического анализа является тезаурусно ориентированной. Процедуры семантического анализа во всех без исключения случаях опираются на функциональность понятийного словаря. Словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов. Это **концептуальный словарь** [2]. В некотором смысле роль концептуального словаря могут выполнять семантические сети, описание которых приведено в следующем разделе.

В семантическом интерпретаторе прежде всего следует специфицировать различаемые **типы семантических отношений** в тексте: ролевые (связи по валентности предиката), предметно-ассоциативные

(отношения между объектами, процессами, значимые в предметной области, – *быть частью, иметь местом, быть предназначенным для, быть столицей* и т.д.) и др.

Принимаются следующие основные постулаты интерпретации синтаксических связей.

1. Тип устанавливаемого семантического отношения определяется семантическими классами и в определенных случаях более детальными семантическими характеристиками синтаксического «хозяина» и «слуги».

2. Предлоги рассматриваются не как самостоятельный объект интерпретации, а как дополнительная (семантико-грамматическая) характеристика связи между синтаксическим «хозяином» предлога и управляемым им знаменательным словом.

3. Для разрешения лексической и синтаксической омонимии, фиксируемой синтаксическим анализатором, семантический интерпретатор использует систему эмпирически устанавливаемых предпочтений. Для удобства сравнения предпочтительности вариантов интерпретации им присваиваются числовые ранги. На уровне типов семантических отношений устанавливается следующий порядок предпочтений (порядок перечисления соответствует уменьшению приоритета связи):

– функциональные связи и связи, устанавливающие факт смысловой избыточности;
– ролевые связи, определяемые как обязательные, при наличии семантически согласованного актанта;
– связи кореференции;
– ролевые связи, определяемые как факультативные;
– предметно-ассоциативные связи специфицируемые;
– предметно-ассоциативные связи неспецифицируемые.

Специфицируемые синтаксические связи – это те, которые интерпретатор в состоянии лексикализовать конкретным отношением в предметной области (*портовые сооружения* → *сооружения, находящиеся в порту*); соответственно, неспецифицируемые связи – те, для которых интерпретатору не удается предложить такую конкретизацию и которые интерпретируются общим понятием *связан.*

В случае обнаружения синтаксической омонимии сочинительных связей предпочтения определяются степенью согласованности семантических характеристик участников синтаксической связи.

Лексические и локальные синтаксические неоднозначности (наличие у слова альтернативных хозяев) обрабатываются в одном переборном механизме. Глобальные варианты синтаксического разбора предложения рассматриваются в переборном механизме следующего уровня. В этом случае сравниваются суммарные веса интерпретации всех связей предложения.

При установлении разных типов отношений интерпретация определяется следующими положениями.

При установлении **ролевых отношений** значимы и должны учитываться (применительно к русскому языку) следующие грамматические характеристики участников синтаксической связи:

– семантико-синтаксический тип предиката (словарная характеристика);
– грамматическая форма предиката;
– падеж актанта, возможность адъективной формы для актанта по данной валентности;
– возможность предложного управления актантом и способность оформляющего синтаксическую связь предлога выражать отношение по данной валентности; информация о способности предлога служить указателем роли для данной валентности хранится в словарном описании предлога.

Операционально процедура определения возможной роли актанта определяется грамматикой ролевых связей, устанавливающей соответствие вида

(Rf, GFP, TSEMU) → VAL,

где Rf – имя синтаксической связи; GFP – грамматическая форма предиката; TSEMU – семантико-синтаксический тип предиката; VAL – имя возможной валентности либо отсылка к ролевой функции предлога.

Затем проверяется соответствие семантических характеристик актанта семантическому условию заполнения валентности предиката (соответствующая пара понятий проверяется на объемную совместимость).

Для установления **отношения кореференции** необходимыми и достаточными являются следующие условия:

– «хозяин» и «слуга» принадлежат семантической категории *Объект*;
– понятия, соответствующие термам «хозяина» и «слуги», находятся в отношении объемной совместимости;
– в случае предложной связи проверяется способность данного предлога выражать отношение кореференции.

Для установления **специфицируемых предметно-ассоциативных отношений** необходимыми и достаточными являются следующие условия:

– понятия, соответствующие термам «хозяина» и «слуги», находятся в отношении объемной несовместимости либо (в случае их совместимости) эти термы синтаксически связаны через предлог, не способный выражать отношение кореференции;

– с парой термов «хозяин – слуга» словарно ассоциировано некоторое предметное отношение (<*автомобиль, кузов*> → *иметь частью*) и/или (если связь предложная) предметное отношение ассоциировано с предлогом и падежом.

Для установления **неспецифицируемых предметно-ассоциативных отношений** необходимым и достаточным является истинность первого и ложность второго условия.

Анализ «по образцу» (**прецедентный анализ**) [10], основанный на использовании корпуса предварительно размеченных текстов, приобретает все большее значение. Разумно построенная система анализа должна обеспечивать не только извлечение знаний из конкретного текста, но и накопление результатов как на синтаксическом, так и на семантическом уровне – для использования их далее в качестве прецедентов.

Одним из наиболее масштабных и значимых проектов, осуществляемых в настоящее время, является создание Национального корпуса русского языка. В нем участвует большая группа лингвистов Москвы, Санкт-Петербурга, Казани, Воронежа, Саратова и других научных центров России.

Национальный корпус русского языка [5] – коллекция электронных текстов, снабженных обширной лингвистической и метатекстовой информацией. Корпус представляет все разнообразие стилей, жанров и вариантов русского языка XIX–XX вв. В Национальном корпусе русского языка в настоящее время используются пять типов разметки: метатекстовая, морфологическая (словоизменительная), синтаксическая, акцентная и семантическая. Не будем подробно рассматривать все имеющиеся виды разметок, остановимся лишь на семантической разметке.

При семантической разметке большинству слов в тексте приписываются один или несколько семантических и словообразовательных признаков, например, «лицо», «вещество», «пространство», «скорость», «движение» и пр. Разметка текстов осуществляется автоматически с помощью программы Semmarkup (автор А.Е. Поляков) в соответствии с семантическим словарем корпуса. Поскольку ручная обработка семантически размеченных текстов очень трудоемка, семантическая омонимия в корпусе не снимается: многозначным словам приписываются несколько альтернативных наборов семантических признаков.

В основу семантической разметки положена система классификации русской лексики, принятая в базе данных «Лексикограф», которая разрабатывалась с 1992 г. в Отделе лингвистических исследований ВИНИТИ РАН под руководством Е.В. Падучевой и Е.В. Рахилиной. Для корпуса был существенно увеличен словник, расширен состав и усовершенствована структура семантических классов, добавлены словообразовательные признаки.

Словник семантического словаря базируется на морфологическом словаре системы «Диалинг» (общим объемом порядка 120 тыс. слов), представляющим собой расширение грамматического словаря русского языка А.А. Зализняка. Текущая версия семантического словаря включает слова знаменательных частей речи: существительные, прилагательные, числительные, местоимения, глаголы и наречия.

Лексико-семантическая информация, приписываемая произвольному слову в тексте, состоит из трех групп помет:

– разряд (например, имя собственное, возвратное местоимение);

– собственно лексико-семантические характеристики (например, тематический класс лексемы, признаки каузативности, оценки);

– деривационные (словообразовательные) характеристики (например, «диминутив», «отадъективное наречие»).

Лексико-семантическая информация имеет различную структуру для разных частей речи. Кроме того, каждый из разрядов существительных – имена предметные, непредметные и собственные – имеет свою структуру помет.

Собственно лексико-семантические пометы сгруппированы по следующим полям:

– таксономия (тематический класс лексемы) – для имен существительных, прилагательных, глаголов и наречий;

– мереология (указание на отношения «часть – целое», «элемент – множество») – для предметных и непредметных имен;

– топология (топологический статус обозначаемого объекта) – для предметных имен;

– каузация – для глаголов;

– служебный статус – для глаголов;

– оценка – для предметных и непредметных имен, прилагательных и наречий.

***Тематические классы глаголов***

Как особое направление в изучении семантики русского языка рассматриваются также исследования Е.В. Падучевой. Наиболее интересными являются работы относительно тематических классов [3] русских глаголов. **Тематический класс** объединяет слова с общим семантическим компонентом, который занимает центральное место в их смысловой структуре. Различают, например, фазовые глаголы, глаголы

восприятия, глаголы знания, глаголы эмоций, глаголы принятия решения, речевых действий, движения, глаголы звука, бытийные глаголы и др.

Слова одного тематического класса имеют некоторый нетривиальный общий компонент в толковании. Тематический класс важен по нескольким причинам. Во-первых, тематический класс часто имеет характерные проявления в синтаксисе – например, у класса обычно есть характерный участник. Во-вторых, члены одного тематического класса склонны иметь один и тот же набор семантических дериватов, то есть зависимых от него понятий.

В статье [11] приведен наиболее полный перечень частных видовых значений глаголов несовершенного вида. Различают следующие видовые значения: актуально-длительное (процесс или состояние длится в момент наблюдения); процессное (то есть просто длящееся); постоянно-непрерывное (значение постоянного свойства или соотношения); узуальное (значение узуального, то есть общепринятого, повторяющегося действия или события); потенциальное; многократное (но не узуальное и не потенциальное); общефактическое непредельное (значение прекратившегося состояния или непредельного процесса); общефактическое результативное (действие достигло предела); общефактическое двунаправленное (результат был достигнут, но аннулирован противоположно направленным действием); общефактическое нерезультативное (неизвестно, достигло ли действие своего предела).

В работе [12] анализируются отпредикатные имена, то есть существительные, образованные от глаголов и прилагательных, такие как *борьба, приход, отчаяние, скупость*. В результате удается различить процессы, события, состояния и свойства.

Например, имена процессов допустимы в контексте глаголов со значением «протекать», «идти», то есть «иметь место» (*идет беседа, происходит забастовка, обновление*). Частная разновидность процессов – совершающиеся действия, то есть целенаправленные процессы с активным субъектом, такие как *борьба*, *проверка*, но не такие, как *купание, бегство, восстание, прогулка, сон, курение*. Имена действия допустимы в контексте глаголов со значением «производить», «вести»: с*лежку осуществляла группа агентов; они производят прием* (*замену, отбор*); мы *ведем расследование*.

Все имена процессов употребляются в контексте фазовых глаголов со значением «начинаться», «кончаться», «продолжаться»: *началась борьба* (*дождь, бой*); *кончилось преследование инакомыслящих*; *продолжается посадка* (*осада*). Имена действий допустимы в контексте фазовых глаголов со значением «начинать», «кончать», «продолжать»: *вступил в переговоры*; *закончил проверку тетрадей*; *прервал чтение*; *принялся за, приступил к, прекратил* (*выдачу*). Контекст фазового глагола является диагностическим для имен процессов, в противоположность именам событий.

Имена событий употребляются в контексте глаголов со значением «произошло», «случилось»: *произошло землетрясение.* События отличаются от процессов тем, что имеют ретроспективного наблюдателя. Наблюдатель процесса синхронный, поэтому, если имеем процесс, то глагол несовершенного вида, а если событие, то совершенного.

Мы опустим множество других деталей, касающихся различий между процессами, событиями, состояниями и свойствами, отметив лишь, что прикладной потенциал данных исследований еще предстоит раскрыть.

Ниже приведен **список глаголов восприятия**, обозначенный Е.В. Падучевой [3] как один из наиболее подробно изученных тематических классов. Казалось бы, чтобы установить принадлежность глагола к тематическому классу восприятий, достаточно убедиться в том, что его семантическая формула включает компонент «восприятие». Однако все не так просто. Дело в том, что перцептивный компонент легко входит в семантику глаголов самых разных классов. Восприятие реальное перетекает в восприятие ментальное.

1. Глаголы движения и состояния, предполагающие наблюдателя:

а) глаголы наблюдаемого движения: *мелькать, промелькнуть, проступить, проскользнуть*;

б) глаголы наблюдаемого состояния: *белеть, торчать, маячить; расстилаться, высовываться, выбиваться, раскинуться, разверзнуться, выступить*;

в) глаголы эмиссии света, запаха, звука: *блестеть, мерцать, светиться, пахнуть, вонять, звучать*.

2. Предполагает наблюдателя глагол *раздаться* (как в *раздался звонок)*, но и в следующих глаголах тоже есть перцептивный компонент: *заглохнуть, заглушить, затмить, смолкнуть, умолкнуть, стихнуть, сливаться* (как в *гимнастерка и серые штаны почти сливались с землей*).

3. Субъект восприятия (или наблюдатель) – обязательный участник ситуаций, выражаемых каузативными глаголами: *выразить, выказать (он выказал мне свое расположение); выделить, выявить, оттенить, высветить, запечатлеть, заслонить, обнажить, обозначить* (*границы*), *открыть, отметить, отобразить*; и их декаузативами (*выразиться, выявиться, выделиться, запечатлеться, обнажиться, обозначиться, открыться*).

4. Есть много глаголов, описывающих идентификацию, которая требует участия органов чувств: *идентифицировать, дифференцировать, опознать, отличить, отождествить, различить* (*очертания*), *распознать, разобрать* (как в *не разбираю второй буквы*).

5. Многие глаголы включают перцептивный компонент, но обозначают вполне специфическое действие или деятельность, для которой главное – цель, а не участие восприятия в ее достижении: *досмотреть* («произвести досмотр»), *зарегистрировать, искать, разыскать, отыскать, выискать, исследовать, изобразить, обрисовать, отследить, проследить, выследить, охранять, подкараулить, просвечивать, спрятать*(*ся*)*, скрыться, шпионить*.

6. Любой глагол передачи и получения информации, например, *писать* или *читать,* предполагает наличие сигнала, который должен быть воспринят органами чувств.

7. Глаголы *показать* и *скрыть,* поскольку в их толкование входит перцептивный компонент, также можно отнести к глаголам восприятия.

8. К глаголам восприятия, помимо прочего, относятся *ослепнуть – слепнуть* (и *ослепить* в одном из значений). Они описывают утрату органа зрения, в результате чего способность видеть утрачивается навечно. Однако сюда не относится глагол *очнуться,* который обозначает временную утрату способности воспринимать с ее последующим возвращением.

9. Некоторые стилистически окрашенные глаголы восприятия: *вериться, воззриться, вылупиться, глазеть, пялиться, узреть, застукать, засветиться.*

10. Тематическая классификация ориентируется на исходные значения слов. Между тем многие глаголы имеют перцептивное значение в качестве производного; в частности, *бдеть, столкнуться (с проблемой), проникнуть (в тайну), выступить*. Например, *Белые здания внезапно выступили из темноты*.

11. Другие аналогичные слова, у которых значение восприятия производное или контекстно обусловленное, как у *бросить (взгляд, взор), броситься (в глаза), обратить (взор, внимание), пробежать (глазами), впиться (взглядом), скользить (взглядом).*

12. Глаголы маркированных способов действия:

а) начинательные: *засквозить, забелеть, зазвучать*;

б) финитивные: *досмотреть, дослушать* и *подсмотреть, подслушать*;

в) сатуративные: *насмотреться, налюбоваться, наслушаться*;

г) глаголы полной поглощенности действием: *засмотреться – засматриваться, заглядеться – заглядываться;*

д) специально-результативные: *высмотреть – высматривать, выследить – выслеживать, отследить – отслеживать;*

е) прерывисто-смягчительные: *поглядывать, послеживать;* семельфактивные: *глянуть.*

Глаголы восприятия, как и другие тематические классы, располагают своими, свойственными именно этому классу моделями семантической деривации.

13. Характерным является семантический переход – от восприятия к ментальному значению. Производное ментальное значение развивается, например, у глаголов *видеть, смотреть, замечать, рассматривать* (*как намек;* и *мы рассматриваем ваше предложение*)*, чувствовать, казаться, обнаружить, слышать, воображать, столкнуться, следить, показаться; представляться, видеться* (та же неоднозначность у существительного *взгляд*):

а) *От прилавка ему хорошо виделось клубное крылечко* (зрительное значение);

б) *Мне видится это так* (ментальное значение).

14. Глагол *свидетельствовать* этимологически предполагает видение, но в контексте *Это свидетельствует о его незаурядном таланте* он имеет ментальное значение; *пролить свет* значит «сделать более понятным», хотя свет нужен для того, чтобы видеть. Глагол *предвкушать* вообще утратил компонент, связанный с вкусовым восприятием, и стал ментальным.

15. Производное ментальное значение возникает и у каузативных глаголов. Так, *показать* – глагол восприятия, но может иметь и значение «доказать», ментальное. Интересно, что в числе производных от *видеть* есть и глаголы знания, и глаголы мнения:

а) *я вижу, ты молчишь* (знание);

б) *он видит в этом препятствие* (мнение).

16. Глагол *оказаться* совмещает перцептивное значение (*Его там не оказалось*) с ментальным (*Оказалось, он здоров*).

17. Производное значение речи развивает глагол *заметить;* оно проявляет себя в сочетаемости с наречиями: *это ты верно заметил* («верно сказал»).

18. Глаголам *послушать, слушать, послушаться, внять* свойственна многозначность «воспринять» – «подчиниться».

19. Регулярным, то есть повторяющимся, является также семантический переход *смотреть* → *относиться*: *я на это смотрю просто* (отношусь просто); *смотреть сквозь пальцы* (потворствовать); *несмотря на* (безотносительно к).

20. Многозначность *смотреть* → *относиться* свойственна глаголу *коситься:* а) (смотреть искоса, сбоку); б) (смотреть косо, относиться с подозрением, выражать взглядом подозрительное отношение).

21. Переход *видеть* → *иметь* представлен примерами *найти, потерять*.

22. Переход от восприятия к межличностному контакту отмечается у глаголов *встречаться, заглянуть* (на огонек), *увидеться.*

23. Смысл *видеть* может выветриваться до идеи простого контакта с объектом, то есть пребывания в том же месте (*Эти стены видели многое; Крым всегда будет рад вас видеть*).

24. Глаголам *возникнуть* и *исчезнуть* свойственна неоднозначность *быть видимым – существовать*. Аналогичная неоднозначность у *обозначиться – обозначаться;* у *теряться*: например, *Тропинка терялась в кустах* (переставала быть видимой) и *Живость движений понемногу терялась* (переставала существовать); у совершенного вида *пропасть* (хотя несовершенный вид *пропадать* значит только не быть видимым: *где ты пропадал?*). В математическом языке *найдется* X значит *существует* X.

25. Семантическое понятие восприятия часто соседствует с перемещением: *столкнуться, наткнуться, напороться, нарваться; попасться* (*Мне попался белый гриб*).

Следствием перемещения может быть, наоборот, выход из поля зрения, как у *скрыться, деться, задеваться.*

Интересно, что для глаголов, выражающих основные виды восприятия – зрение, слух, обоняние, осязание, вкус, – можно выявить единую парадигму семантических дериватов исходной лексемы, причем она будет в существенной степени одинаковой для многих языков, что свидетельствует о древности данной лексики и данных конструкций.

Существенной в данном подходе является идея разделять понятия языка на некоторые семантические группы с учетом того, что эти понятия имеют некоторый нетривиальный общий смысловой компонент. Элементы таких групп склонны иметь один и тот же набор зависимых понятий. Словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов. Остается вопрос, как правильно структурировать и представлять информацию в подобных словарях, чтобы поиск по ним был удобным и быстрым, а кроме того, получалось учитывать изменения в естественном языке (исчезновение старых и возникновение новых понятий).

При обсуждении проблем семантики часто упоминают **принцип композициональности.** Он утверждает, что смысл сложного выражения определяется смыслами его составных частей и правилами, применяемыми для их объединения. Поскольку предложение состоит из слов, получается, что его смысл можно представить набором значений слов, входящих в него. Но не все так просто. Смысл предложения также опирается на порядок слов, фразирование и отношения между словами в предложении, то есть учитывает синтаксис.

Как видим, концептуальные диаграммы зависимостей позволяют утверждать, что в некоторых случаях принцип композициональности нарушается. Ошибочно утверждать, что смысл словосочетаний и фраз может быть выражен через смыслы составляющих их слов. Это не всегда верно. Однако главная проблема такого подхода заключается в том, что выделение тематических классов и составление семантических словарей чрезвычайно трудоемкий процесс, сильно зависящий от индивидуального восприятия и интерпретации понятий конкретным человеком.

## Сетевые модели представления знаний

### *Тезаурусы, семантические сети, фреймовые и онтологические модели*

**Тезаурус** – разновидность словаря общей или специальной лексики, в котором указаны семантические отношения между лексическими единицами. В отличие от толкового словаря тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами, благодаря чему может использоваться для наполнения баз знаний систем искусственного интеллекта.

В тезаурусах обычно используются следующие основные семантические отношения: синонимы, антонимы, гипонимы, гиперонимы, меронимы, холонимы и паронимы.

Синонимы – слова одной части речи, различные по звучанию и написанию, но имеющие похожее лексическое значение (*смелый – храбрый, бесстрашный*).

Антонимы – это слова одной части речи, различные по звучанию и написанию, имеющие прямо противоположные лексические значения (*добрый – злой*).

Гипоним – понятие, выражающее частную сущность по отношению к другому, более общему понятию (*животное – собака – бульдог*).

Гипероним – слово с более широким значением, выражающее общее, родовое понятие, название класса предметов, свойств или признаков (*бульдог – собака – животное*).

Гипероним является результатом логической операции обобщения, тогда как гипоним – ограничения.

Мероним – понятие, которое является составной частью другого (*автомобиль – двигатель, колесо, капот*).

Холоним – понятие, которое является целым над другими понятиями (*двигатель, колесо, капот – автомобиль*).

Меронимия и холонимия как семантические отношения являются взаимно обратными друг другу, так же, как гипонимия и гиперонимия.

Паронимы – слова, сходные по форме, но различающиеся по смыслу (*индеец – индиец*).

Примером тезауруса является WordNet [13]. Базовой словарной единицей WordNet является синонимический ряд (синсет), объединяющий слова со схожим значением. Синсеты состоят из слов, принадлежащих той же самой части речи, что и исходное слово. Каждый синсет сопровождается небольшой формулировкой (дефиницией), разъясняющей его значение. Синсеты связаны между собой различными семантическими отношениями, например, гипонимии, гиперонимии и др. Пример со словом *pen (ручка)* приведен на рисунке 1. Видно, что в словаре для этого слова имеются пять различных значений, оно относится к разряду письменных принадлежностей и имеет семь родственных слов: карандаш, маркер, мел для доски, восковой мелок и др.


*Рис. 1. Описание слова pen в базе WordNet*

WordNet содержит приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч синсетов. Вся БД разбита на три части: существительные, глаголы и прилагательные/наречия. Слово или словосочетание может находиться более чем в одном синсете и принадлежать более чем одной категории части речи. Более подробная информация о количестве уникальных слов, синсетов и пар «слово-синсет» в базе WordNet дана в таблице 1.

Преимущества WordNet перед остальными похожими ресурсами – его открытость, доступность, наличие большого количества различных семантических связей между синсетами. Доступ к WordNet выполняется непосредственно с помощью браузера (локально или через Интернет) или библиотек на C.

*Таблица 1*

**Информация о количестве уникальных слов, синсетов и пар «слово-синсет» в WordNet**

| Части речи | Уникальные слова | Синсеты | Общее количество пар «слово-синсет» |
|---|---|---|---|
| Существительные | 117798 | 82115 | 146312 |
| Глаголы | 11529 | 13767 | 25047 |
| Прилагательные | 21479 | 18156 | 30002 |
| Наречия | 4481 | 3621 | 5580 |
| **Всего** | **155287** | **117659** | **146312** |

Существуют реализации WordNet для других языков (около 16). Например, для европейских языков создан EuroWordNet, связь между различными языковыми версиями в котором осуществляется через специальный межъязыковой индекс. Ведутся разработки WordNet и для русского языка. Необходимо отметить, что существуют методы предметной классификации синсетов WordNet, то есть определение областей знаний, в которых они употребляются. Подобная информация может служить впоследствии для сокращения количества возможных значений слов, если известна тематика обрабатываемого документа, тем самым позволяя уменьшить значение ошибки при принятии неверного значения слова.

**Семантическая сеть** – модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (ребра) задают отношения между ними [14]. Объектами могут быть понятия, события, свойства, процессы. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений. Причем в качестве понятий могут быть как экземпляры объектов, так и их множества.

Семантические сети возникли как попытка визуализации математических формул. За визуальным представлением семантической сети в виде графа стоит математическая модель, в которой каждая вершина соответствует элементу предметного множества, а дуга – предикату. На рисунке 2 представлен пример семантической сети, взятый из Википедии [14].

Терминология, использующаяся в этой области, разнообразна. Чтобы добиться некоторой однородности, узлы, соединенные дугами, принято называть графами, а структуру, где имеется целое гнездо из узлов или где существуют отношения различного порядка между графами, называть сетью. Помимо терминологии, использующейся для пояснения, также различаются способы изображения. Некоторые используют кружки вместо прямоугольников; некоторые пишут типы отношений над или под дугами, заключая или не заключая их в овалы; некоторые используют аббревиатуры вида O или A для обозначения агента или объекта; некоторые используют различные типы стрелок.

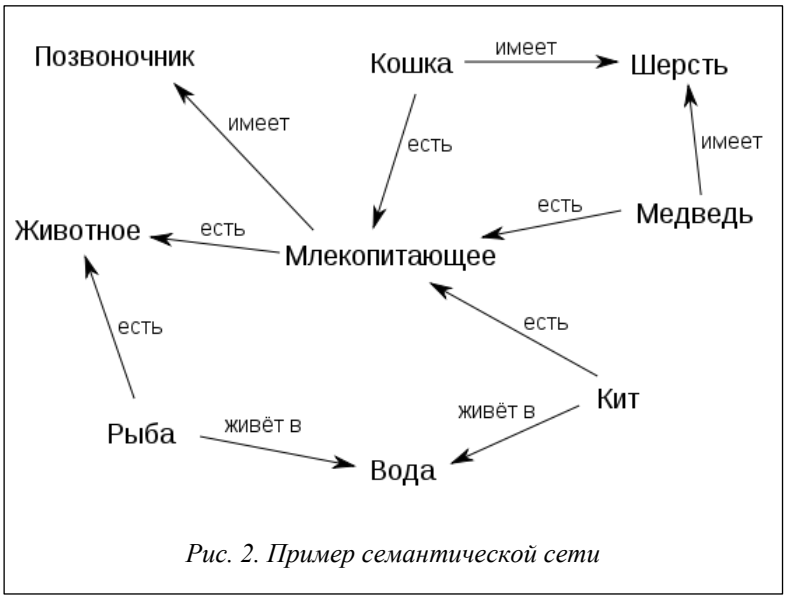


*Рис. 2. Пример семантической сети*

Самые первые семантические сети были разработаны в качестве языка-посредника для систем машинного перевода. Последние версии семантических сетей становятся более мощными и гибкими и составляют конкуренцию фреймовым системам, логическому программированию и другим языкам представления знаний.

Несмотря на различную терминологию, разнообразие методов представления кванторов общности и существования и логических операторов, разные способы манипулирования сетями и правила вывода, можно выделить существенные сходства, присущие почти всем семантическим сетям:

– узлы семантических сетей представляют собой концепты предметов, событий, состояний;

– различные узлы одного концепта относятся к различным значениям, если для них не помечено, что они относятся к одному концепту;

– дуги семантических сетей создают отношения между узлами-концептами, пометки над дугами указывают на тип отношения;

– некоторые отношения между концептами представляют собой семантические роли, такие как «агент», «объект», «реципиент» и «инструмент»; другие означают временные, пространственные, логические отношения и отношения между отдельными предложениями;

– концепты организованы по уровням в соответствии со степенью обобщенности, наподобие иерархии гиперонимов в WordNet, например, *сущность → живое существо → животное → плотоядное*.

Заметим, что среди семантических отношений, используемых для описания сетей, могут быть не только семантические отношения, используемые в тезаурусах, но и другие виды связей: функциональные (определяемые обычно глаголами *производит*, *влияет*, …), количественные (*больше*, *меньше*, *равно*, …), пространственные (*далеко от*, *близко от*, *под*, *над*, …), временные (*раньше*, *позже*, *в течение*, …), атрибутивные (*иметь свойство*, *иметь значение*), логические (*И*, *ИЛИ*, *НЕ*) и пр.

Например, семантика предложения *У Иванова есть черный BMW* может быть представлена в виде семантической сети, приведенной на рисунке 3.

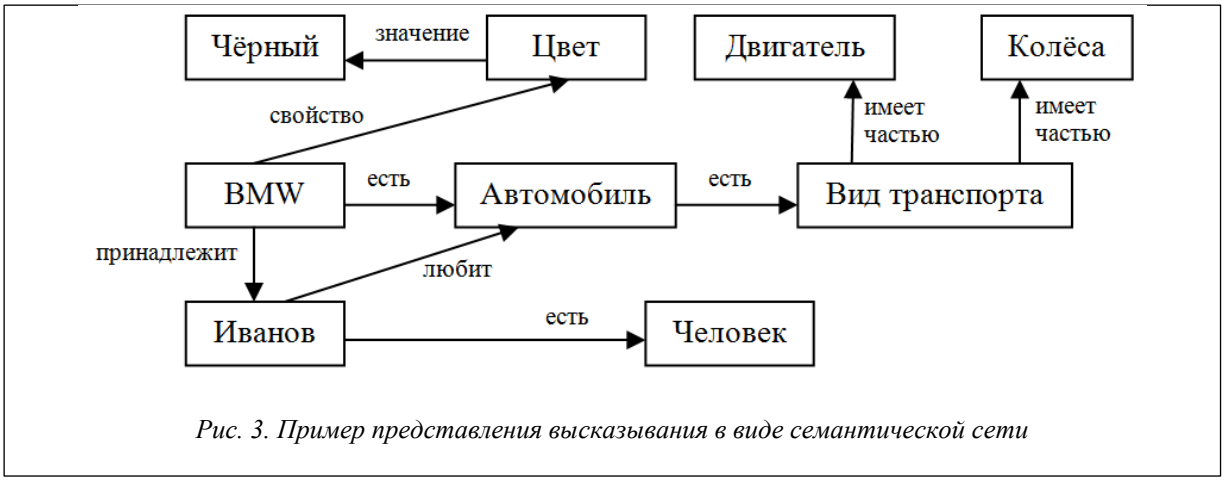


*Рис. 3. Пример представления высказывания в виде семантической сети*

Несмотря не некоторые различия, сети удобны для чтения и обработки компьютером, являются наглядным и достаточно универсальным средством представления семантики естественного языка. Однако их формализация в конкретных моделях представления, использования и модификации знаний оказывается достаточно трудоемкой, особенно при наличии множественных отношений между ее элементами.

Рассмотрим, например, некоторую сеть, описывающую высказывание *Настя попросила книгу у Даши*. Допустим, можно приписать приведенным объектам свойства: *Настя* – «*старательная*», *Даша* – «*любознательная*». Между этими объектами есть связь (через книгу). Но, кроме нее, есть много других связей, которые существуют в реальном мире: социальный статус (студентки, подруги – необязательно между собой), родственные отношения (у каждой есть родители и/или другие родственники) и т.д. Получается, что даже для такого простого примера сеть способна разрастись до большого размера и, как следствие, поиск вывода в ней будет слишком сложным.

В сложных семантических сетях, включающих множество понятий, процесс обновления узлов и контроль связей между ними, как видим, усложняют процедуру обработки информации. Стремление устранить эти недостатки послужило причиной появления особых типов семантических сетей, таких как фреймовые модели.

**Фреймовые модели** представления знаний были предложены М. Минским [15].

Фреймом называется структура для описания понятия или ситуации, состоящая из характеристик этой ситуации и их значений [16]. Фрейм можно рассматривать как фрагмент семантической сети, предназначенный для описания понятий со всей совокупностью присущих им свойств. Особенность фреймовых моделей представления знаний состоит в том, что все понятия, описываемые в каждом из узлов модели, определяются набором атрибутов и их значениями, которые содержатся в слотах фрейма < *имя фрейма*, *слот* 1, *слот* 2, …, *слот N* >. Графически это выглядит аналогично семантической сети, но принципиальное отличие заключается в том, что каждый узел во фреймовой модели имеет обобщенную структуру, состоящую из множества слотов, каждый из которых имеет имя, указатель наследования, указатель типа данных и значение.

Слот – это атрибут, связанный с узлом в модели, основанной на фреймах, он является составляющей фрейма. *Имя слота* должно быть уникальным в пределах фрейма. *Указатель наследования* показывает, какую информацию об атрибутах слотов во фрейме верхнего уровня наследуют слоты с теми же именами во фрейме более низкого уровня. *Указатель типа данных* содержит информацию о типе данных, включаемых в слот. Обычно используются следующие типы данных: указатель на имя фрейма верхнего уровня, вещественное число, целое число, текст, список, таблица, присоединенная процедура и др. *Значением слота* может быть экземпляр атрибута, другой фрейм или фасет, оно должно соответствовать указанному типу данных и условию наследования. Помимо конкретного значения, в слоте могут храниться процедуры и правила, которые вызываются при необходимости вычисления этого значения. Таким образом, слот может содержать не только конкретное значение, но и имя процедуры, позволяющей вычислить его по заданному алгоритму, а также одну или несколько продукций, с помощью которых это значение определяется. В слот может входить не одно, а несколько значений. Иногда этот слот включает компонент, называемый фасетом, который задает диапазон или перечень его возможных значений. Фасет указывает также граничные значения заполнителя слота. Чаще всего со слотами связываются процедуры добавления и удаления информации, они могут следить за приписыванием информации к данному узлу и проверять, что при изменении значения производятся соответствующие действия.

Различают фреймы-образцы (прототипы), хранящиеся в базе знаний, и фреймы-экземпляры, которые создаются для отображения реальных ситуаций на основе поступающих данных. Фреймовые модели являются достаточно универсальными, поскольку позволяют отразить все многообразие знаний о мире через фреймы-структуры (для обозначения объектов и понятий: *заем, залог, вексель*), фреймы-роли (*менеджер, кассир, клиент*), фреймы-сценарии (*банкротство, собрание акционеров, празднование именин*), фреймы-ситуации (*тревога, авария, рабочий режим устройства*) и др. Для представления знаний в виде сети фреймов существуют специальные языки и программные средства: FRL (Frame Representation Language), KRL (Knowledge Representation Language), фреймовая оболочка Kappa, PILOT/2 и другие.

Важнейшим свойством теории фреймов является заимствованное из теории семантических сетей наследование свойств. И во фреймах, и в семантических сетях наследование происходит по *ISA*. Слот *ISA* указывает на фрейм более высокого уровня иерархии, откуда неявно наследуются, то есть переносятся, значения аналогичных слотов.

Основным преимуществом фреймов как модели представления знаний является соответствие современным представлениям об организации долговременной памяти человека, а также ее гибкость и наглядность. Достоинства фреймовых моделей представления знаний проявляются в том случае, если родовидовые связи изменяются нечасто и предметная область насчитывает немного исключений.

К недостаткам фреймовых моделей относят их относительно высокую сложность [17], что проявляется в снижении скорости работы механизма вывода и увеличении трудоемкости внесения изменений в сформированной иерархии. Поэтому при разработке фреймовых систем большое внимание уделяется наглядным способам отображения и эффективным средствам редактирования фреймовых структур.

Можно заметить, что объектно-ориентированный подход является развитием фреймового представления. В этом случае шаблон фрейма можно рассматривать как класс, экземпляр фрейма – как объект. Языки объектно-ориентированного программирования предоставляют средства создания классов и объектов, а также средства для описания процедур обработки объектов (методы). Однако фреймовые модели не позволяют организовать гибкий механизм логического вывода, поэтому фреймовые системы либо представляют собой объектно-ориентированные БД, либо требуют интеграции с другими средствами обработки знаний, например, логическими моделями.

В инженерии знаний под **онтологической моделью** понимается детальное описание некоторой предметной или проблемной области, которое используется для формулирования утверждений общего характера. Онтологии позволяют представить понятия в таком виде, пригодном для машинной обработки.

В центре большинства онтологий находятся классы, описывающие понятия предметной области. Атрибуты описывают свойства классов и экземпляров. Здесь прослеживаются аналогии с фреймовым подходом к формализации знаний. Многие понятия и принципы реализации, а также графическая форма представления на начальном этапе структуризации являются в онтологиях сходными с семантическими сетями. Основным отличием является ориентация онтологий на использование непосредственно компьютером, то есть структуры данных описаны не на естественном языке (как это принято в семантических сетях и тезаурусах), а на специальном формальном языке. С тезаурусами онтологии тоже имеют много общего. Но в отличие от них для онтологических моделей необходимыми требованиями являются внутренняя полнота, логическая взаимосвязь и непротиворечивость используемых понятий. В тезаурусах эти требования могут не выполняться. Для описания онтологий используются такие формальные языки, как RDF, OWL, KIF, CycL, OCML и другие.

Обычно выделяют [18] следующие основные элементы онтологий:

– экземпляры;

– классы объектов (понятий);
– атрибуты (описывают свойства классов и экземпляров);
– функции (описывают зависимости между классами и экземплярами);
– аксиомы (дополнительные ограничения).

**Специализированные (предметно-ориентированные) онтологии** – это представление какой-либо области знаний или части реального мира. В такой онтологии содержатся специальные для этой области значения терминов. К примеру, слово *поле* в сельском хозяйстве означает участок земли, в физике – один из видов материи, в математике – класс алгебраических систем.

**Общие онтологии** используются для представления понятий, общих для большого числа областей. Такие онтологии содержат базовый набор терминов, глоссарий или тезаурус, используемый для описания терминов предметных областей.

Современные онтологические модели являются модульными, то есть состоят из множества связанных между собой онтологий, каждая из которых описывает отдельную предметную область или задачу. Онтологические модели не являются статичными, они постоянно меняются.

Если использующая специализированные онтологии система развивается, то может потребоваться объединение онтологий. Преимущественным недостатком онтологических моделей является сложность их объединения. Онтологии даже близких областей могут быть несовместимы друг с другом. Разница может появляться из-за особенностей местной культуры, идеологии или вследствие использования другого языка описания. Объединение онтологий выполняют как вручную, так и в полуавтоматическом режиме. В целом это трудоемкий, медленный и дорогостоящий процесс.

Онтологические модели широко применяются в системах, основанных на знаниях: экспертных системах и системах поддержки принятия решений. Интересный способ представления знаний о времени с учетом неопределенности в онтологиях описан в работе А.Ф. Тузовского [19].

В настоящее время довольно перспективными и широко применяемыми на практике технологиями представления знаний являются технологии Semantic Web. Центральным понятием Semantic Web является онтология – модель предметной области, состоящая из множества понятий, **множества экземпляров понятий** и множества отношений (свойств). Множество понятий и отношений между ними определяет общую схему хранения данных, представленных как множество утверждений об экземплярах понятий, или аксиом онтологии. Такие простые утверждения, называемые триплетами, имеют вид «субъект-предикат-объект». Набор правил, задаваемых пользователем, загружается в систему логического вывода, которая на основе содержащихся в онтологии утверждений создает согласно этим правилам новые экземпляры понятий и отношений онтологии.

Одной из наиболее существенных проблем как для представления знаний в контексте времени, так и для представления знаний в целом является представление знаний о времени и об изменениях знаний с его течением. Однако большинство применяемых на практике языков описания знаний основываются на логике предикатов первого порядка и используют унарные или бинарные отношения. К таким языкам, например, относятся OWL и RDF. В этом случае для описания бинарных отношений с учетом времени требуется вводить в отношения дополнительный параметр, соответствующий времени. При этом бинарные отношения превращаются в тернарные и выходят за рамки описательных возможностей языка.

Еще одной важной задачей является описание знаний о времени с учетом возможной неполноты этих знаний. Например, описание высказываний вида: «событие А произойдет когда-нибудь в будущем». Эта задача обычно решается в рамках модальных темпоральных логик, например LTL, при помощи определенных модальных операторов. Но, поскольку язык описания знаний OWL основан на дескриптивной логике, воспользоваться таким решением для OWL-онтологий становится невозможно.

В своей работе [19] А.Ф. Тузовский предлагает представить модель описания знаний о времени в следующем виде:

$< T^U, V^U, T^P, F, Rul >$, где

1) $T^U$ – множество моментов времени $T^U = \{T \cup \{t_\emptyset\}\}$, где $T$ – линейно упорядоченное множество, имеющее мощность континуума, на котором задана бинарная операция вычитания $T \times T \to R^+$, а $t_\emptyset$ – особый элемент, соответствующий «неопределенному моменту времени»;

2) $V^U$ – множество переменных, обозначающих элементы множества $T^U$, а также особая переменная $t_N$, соответствующая текущему моменту времени; значение переменной $t_N$ постоянно меняется, отражая ход времени в некоторой системе, для описания временного контекста которой используется предлагаемый подход;

3) $T^P$ – множество промежутков времени; промежуток времени соответствует упорядоченной паре $\tau = < t_{i1}, t_{i2} >$, где $t_{i1}$ и $t_{i2}$ – такие элементы множества $V^U$, что $((t_{i1} \le t_{i2}) \wedge (t_{i1} \ne t_\emptyset) \wedge (t_{i2} \ne t_\emptyset)) \vee (t_{i1} = t_\emptyset) \vee (t_{i2} = t_\emptyset)$; таким образом, промежуток времени соответствует некоторому участку на временной шкале, причем его границей может быть некоторый момент времени, текущий момент времени (переменная $t_N$) либо неопределенный момент времени $t_\emptyset$, при этом промежуток времени с совпадающими границами ($t_{i1} = t_{i2}$) соответствует некоторому моменту времени;

4) $F$ – множество предикатов, описывающих свойства промежутков времени, а также качественные отношения между ними;

5) *Rul* – множество правил вида ($G \to H$) и ($G \leftrightarrow H$), описывающих базовые механизмы логического вывода, в том числе ограничения на значения предикатов $F$, а также определенность границ промежутков времени.

Понятие промежутка времени требуется для описания некоторых интервалов времени, точные границы которых неизвестны до наступления определенного состояния модели. Можно сказать, что каждый промежуток времени описывает некоторый интервал времени, точные границы которого пока неизвестны. При этом может быть доступна информация о том, в каких пределах этот интервал гарантированно расположен на временной шкале, а точные границы интервала, описываемого промежутком времени, могут стать известны в будущем. Поэтому вводятся две разновидности границ промежутка времени: точная и гарантированная [19]. Для определения двух видов границ служат предикаты *EL* (*exactleft*), *ER* (*exactright*), *GL* (*guaranteedleft*) и *GR* (*guaranteedright*), определяющие точную левую/правую и гарантированную левую/правую границы промежутка времени, соответственно. Например, предикат *EL* ($\tau_i$ , $t_{i1}$) соответствует утверждению «точная левая граница промежутка $\tau_i$ – момент времени $t_{i1}$». Для простоты вид границы промежутка времени можно обозначать при помощи различных скобок: промежуток [$t_i$ , $t_j$] полностью определенный (обе его границы являются точными); промежуток [$t_i$ , $t_j$) – определен слева (левая граница промежутка точная, а правая гарантированная), а промежуток ($t_i$ , $t_j$) – полностью неопределенный (обе границы промежутка гарантированные). Полностью определенный промежуток времени соответствует некоторому интервалу времени, то есть отрезку на временной шкале, точные границы которого известны.

Множество предикатов $F$ содержит также предикат *duration* ($\tau_i$ , $l_k$) (сокращенная запись *duration* ($\tau_i$ ) = $l_k$), который сопоставляет промежутку $\tau_i$ некоторое значение $l_k$ , соответствующее длительности этого промежутка. Например, промежуток времени «два часа днем 20 мая» можно описать как $\tau_i$ = ($t_{i1}$ , $t_{i2}$), *duration* ($\tau_i$ ) = 2, где $t_{i1}$ = «12:00 20 мая», $t_{i1}$ = «18:00 20 мая».

Особые переменные $t_N$ и $t_\varnothing$, соответствующие текущему и неопределенному моментам времени, также служат для описания неопределенных границ промежутка времени, для которых неизвестно даже гарантированное значение границы. К примеру, промежуток < $t_N$, $t_\varnothing$ > будет соответствовать понятию «будущее», а < $t_\varnothing$, $t_N$ > – «прошлое». В предлагаемой краткой нотации промежуток времени «двое суток в будущем» можно обозначить как $\tau_i$ = ( $t_N$, $t_\varnothing$), *duration* ($\tau_i$ ) = 48.

Множество $F$ может включать также качественные отношения, связывающие промежутки времени между собой. В их роли выступают отношения интервальной алгебры Аллена [20]. На основе этих отношений в системе, использующей описываемый подход, определяются элементы множества *Rul* – правила логического вывода.

Таким образом, в предлагаемой онтологии описания времени понятие *TimePeriod* соответствует промежутку времени с длительностью *duration* и границами *PeriodBorder*, связанными с ним отношениями *hasLeftBorder* и *hasRightBorder*. Определенность границы описывается булевым свойством *defined*, а значение границы – свойством *presentedBy*. Момент времени *TimePoint* может иметь вид *TimePointType* – «точный» (*Numeric*, в этом случае свойство *value* указывает на конкретную точку на временной шкале), «неопределенный» (*Undefined*, неопределенный момент времени) и «текущий» (*Now*, в двух последних случаях значение свойства *value* не задано).

Предлагаемый подход позволяет использовать дескриптивные логики и OWL-онтологии для описания знаний, для которых обычно требуется применение модальных логик, ситуационного исчисления или исчисления событий. Кроме того, предлагаемый подход несложно использовать в существующих системах управления знаниями, не учитывающих неполноту знаний о времени, дополняя их возможностью описывать содержащиеся в них знания с учетом временного контекста.

***Семантические роли и семантические ограничения***

Семантические сети позволяют представлять семантику отдельно взятого слова согласно его внутренней структуре. Если вместе с этой структурой учитывать грамматические особенности слов (как это обычно бывает в предложениях на естественном языке), то смысл высказывания может быть представлен в терминах семантических ролей и связанных с ними семантических ограничений.

Ясно, что с точки зрения синтаксиса глагол задает синтаксические рамки. Рассмотрим теперь семантические роли и семантические ограничения на эти роли, задаваемые глаголом.

Возьмем, например, глагол *want*. Если используется его активная форма (*wanting*), то перед ним находится аргумент, который выполняет роль действующего субъекта – агенса. Если используется пассивная форма (*wanted*), то после глагола находится аргумент, который выполняет роль желаемого объекта. Заметив эти закономерности, мы можем ассоциировать поверхностные аргументы глагола со множеством дискретных ролей. Другими словами, можно сказать, что некоторые шаблоны подкатегоризации глаголов позволяют связывать аргументы поверхностной структуры с семантическими ролями, выполняемы-

ми этими аргументами. Исследование ролей, ассоциированных с определенными глаголами и классами глаголов, обычно ссылается на анализ тематических ролей (или кейс-ролей).

Понятие семантических ограничений возникает напрямую из этих семантических ролей. Например, первым аргументом словосочетания будет субъект (*wanter*), так как глагол *want* ограничивает компоненты и подразумевает, что первому аргументу требуется использование активного залога (*wanting*). Традиционно это явление называют ограничением выбора. Таким образом, глаголы могут задавать семантические ограничения на их аргументы.

Следует заметить, что не только глаголы имеют предикатно-аргументную структуру. Предлог может быть представлен двухместным предикатом, где первым аргументом является объект, который связан некоторым отношением со вторым аргументом. В следующем примере приведен четырехместный предикат, полученный на основе существительного:

*Make a reservation for this evening for a table for two persons at 8*

*Reservation (Hearer, Today, 8 PM, 2).*

Как видим, представление смысла текстовой информации может быть организовано в виде семантической предикатно-аргументной структуры.

Помимо термина «семантические роли», в различной литературе используются также понятия «тематические роли», «тета-роли», «глубинные падежи». Основоположниками данного направления исследований семантики принято считать Дж. Грубера и Ч. Филлмора. По своей сути эти понятия близки к семантическим и глубинно-синтаксическим актантам, исследованием которых занимался И.А. Мельчук. Приведем некоторые семантические роли, рассмотренные в работах [21, 22].

**Агенс** – одушевленный инициатор и контролер действия.

**Адресат** – получатель сообщения (может объединяться с Бенефактивом).

**Бенефактив (Реципиент, Посессор)** – участник, чьи интересы косвенно затронуты в процессе ситуации (получает пользу или вред).

**Инструмент** – стимул эмоции или участник, с помощью которого осуществляется действие.

**Источник** – место, из которого осуществляется движение.

**Контрагент** – сила или сопротивляющаяся среда, против которой осуществляется действие.

**Объект** – участник, который передвигается или изменяется в ходе события.

**Пациенс** – участник, претерпевающий существенные изменения.

**Результат** – участник, который появляется в результате события.

**Стимул** – внешняя причина или объект, вызывающие это состояние.

**Цель** – место, в которое осуществляется движение.

**Экспериенцер** – участник, испытывающий внутреннее состояние, не приводящее к внешним изменениям (носитель чувств и восприятий).

**Эффектор** – неодушевленный участник, часто природная сила, вызвавший изменение в состоянии Пациенса.

В соответствии с числом аргументов и их семантическими свойствами множество глагольных лексем можно разбить на классы. Например, рассмотрим следующие ролевые типы глаголов: глаголы физического воздействия (*рубить*, *пилить*, *резать*); глаголы восприятия (*видеть*, *слышать*, *чувствовать*); глаголы способа речи (*кричать*, *шептать*, *бормотать*). Внутри каждого класса существует более точное деление. Среди глаголов физического воздействия похожую семантическую предикатно-аргументную структуру имеют глаголы вида **глагол (Агенс, Инструмент, Объект):** *break – разбить, bend – согнуть, fold – загнуть, shatter – разбить вдребезги, crack – расколоть* и др. Другая предикатно-аргументная структура характерна для глаголов вида **глагол (Агенс, Инструмент, Цель):** *hit – ударить, slap – шлепнуть, strike – ударить, bump – ударить (обо что-то), stroke – погладить* и др.

Можно заметить, что существуют корреляции между морфологическими падежами, предлогами, синтаксическими ролями, с одной стороны, и семантическими ролями, с другой стороны, например, cut *with* a knife, work *with* John, spray *with* paint. Кроме того, следует учитывать, что у одного предикатного слова не может быть двух актантов с одной и той же семантической ролью. Различия в наборах ролей затрагивают в основном периферийные семантические роли (Контрагент, Стимул, Источник) или сводятся к объединению/фрагментации ядерных ролей (Агенс vs. Агенс и Эффектор; Адресат vs. Адресат, Реципиент и Бенефактив; Пациенс/Тема/Объект vs. Пациенс, Тема и Результат).

В своей работе [21] Ч. Филлмор даже предложил правило опосредованного отображения семантических ролей в синтаксические: если среди аргументов имеется Агенс, он становится подлежащим; в отсутствие Агенса, если есть Инструмент, он становится подлежащим; в отсутствие Агенса и Инструмента подлежащим становится Объект. Отсюда естественным образом возникает иерархия семантических ролей. Наиболее известные иерархии семантических ролей: Агенс > Инструмент > Пациенс; Агенс > Источник > Цель > Инструмент > Тема > Место; Агенс > Бенефактив > Экспериенцер > Инструмент > Тема > Место и некоторые др. Иерархия семантических ролей строится таким образом, чтобы можно было отразить степень тематической принадлежности аргументов (топикальность) так, что на левом конце

иерархии располагаются прагматически наиболее важные семантические роли, а на правом – семантические роли, которым не свойственна высокая топикальность.

Первоначально семантические роли предполагалось считать примитивами, не подверженными дальнейшему анализу, который мог бы выявить их внутреннюю структуру. Однако в таком случае возникает ряд проблем. Во-первых, в результате все более тщательного семантического и синтаксического анализа происходит ничем не ограниченное увеличение списка ролей [23]. Во-вторых, неструктурированные списки ролей не позволяют делать предсказания о возможных ролевых типах глаголов и объяснять отсутствие незасвидетельствованных типов. Поэтому в теории семантических ролей было предложено определять роли в терминах различительных признаков или проторолей. Например, Д. Даути [23] предлагает выделить следующие свойства протороли Агенс: волитивно вовлечен в событие или состояние; является сознательным и/или воспринимающим участником; инициирует событие или изменения состояния другого участника; движется (по отношению к точке пространства или другому участнику); существует независимо от события, обозначенного глаголом.

К сожалению, в настоящий момент не удается установить взаимно-однозначное соответствие между семантическими ролями и падежами, то есть с функциональной точки зрения категория падежа является неоднородной. Ситуация осложняется еще и тем, что сами роли нетривиально связаны между собой, а в естественных языках распространены такие генеративные приемы, как метафора и метонимия, которые порождают множество новых смыслов и в принципе не могут быть отражены в статическом лексиконе.

## Логические модели представления знаний

Основная идея подхода при построении логических моделей представления знаний состоит в том, что вся информация, необходимая для решения прикладных задач, рассматривается как совокупность фактов и утверждений, которые представляются в виде формул в некоторой логике. Знания отображаются совокупностью таких формул, а получение новых знаний сводится к реализации процедур логического вывода. В основе логических моделей представления знаний лежит понятие формальной теории, задаваемое кортежем $S = < B, F, A, R>$, где $B$ – счетное множество базовых символов (алфавит); $F$ – множество, называемое формулами; $A$ – выделенное подмножество априори истинных формул (аксиом); $R$ – конечное множество отношений между формулами, называемое правилами вывода.

Основной подход к представлению смысла в компьютерной лингвистике включает в себя создание представления смысла в формальном виде. Такое представление можно назвать языком представления смысла. Язык представления необходим для того, чтобы ликвидировать разрыв между естественным языком и общесмысловыми знаниями о мире. А поскольку предполагается использовать этот язык для автоматической обработки текстов и при создании систем искусственного интеллекта, необходимо учитывать вычислительные требования семантической обработки, такие как необходимость определять истинность высказываний, поддерживать однозначность представления, представлять высказывания в канонической форме, обеспечивать логический вывод и быть выразительными.

В естественных языках существует большое разнообразие приемов, которые используются для передачи смысла. Среди наиболее важных – способность передавать предикатно-аргументную структуру. Учитывая вышесказанное, получаем, что в качестве инструмента для представления смысла высказываний хорошо подходит логика предикатов первого порядка. С одной стороны, она относительно легко понимается человеком, с другой – хорошо поддается (вычислительной) обработке. При помощи логики первого порядка могут быть описаны важные смысловые классы, включающие события, время и другие категории. Однако следует помнить, что высказывания, соответствующие таким понятиям, как убеждения и желания, требуют выражений, включающих модальные операторы.

Семантические сети и фреймы, которые обсуждались в предыдущем разделе, могут быть рассмотрены в рамках логики предикатов первого порядка. Например, смысл предложения *У меня есть книга* можно записать четырьмя различными способами, используя четыре различных языка представления смысла (см. рис. 4, нумерация соответствует порядку на рисунке): 1) концептуальная диаграмма зависимостей; 2) представление, основанное на фреймах; 3) семантическая сеть; 4) исчисление предикатов первого порядка.

Несмотря на то, что все эти четыре подхода различны, на абстрактном уровне они представляют собой общепринятое фундаментальное обозначение того, что представление смысла состоит из структур, составленных из множества символов. Эти символьные структуры соответствуют объектам и отношениям между объектами. Все четыре представления состоят из символов, соответствующих «говорящему», «книге» и набору отношений, обозначающих обладание одного другим. Важным здесь является то, что все эти четыре представления позволяют связать, с одной стороны, выразительные особенности естественного языка и, с другой стороны, реальное состояние дел в мире.

Логические модели представления знаний обладают рядом преимуществ. Во-первых, в качестве «фундамента» здесь используется классический аппарат математической логики, методы которой доста-

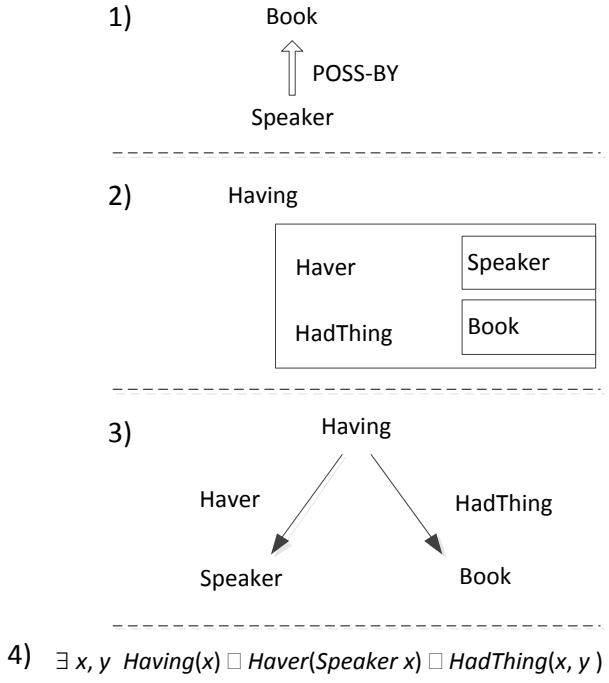


*Рис. 4. Четыре языка представления смысла высказывания*

$$
\begin{aligned}
\textit{Formula} &\rightarrow \textit{AtomicFormula} \\
&\mid \textit{Formula Connective Formula} \\
&\mid \textit{Quantifier Variable, ... Formula} \\
&\mid \neg\, \textit{Formula} \\
&\mid (\textit{Formula}) \\
\textit{AtomicFormula} &\rightarrow \textit{Predicate}(\textit{Term, ...}) \\
\textit{Term} &\rightarrow \textit{Function}(\textit{Term, ...}) \\
&\mid \textit{Constant} \\
&\mid \textit{Variable} \\
\textit{Connective} &\rightarrow \wedge \mid \vee \mid \Rightarrow \\
\textit{Quantifier} &\rightarrow \forall \mid \exists \\
\textit{Constant} &\rightarrow \textit{ChildrensBook} \mid \textit{HarryPotter} \mid ... \\
\textit{Variable} &\rightarrow x \mid y \mid ... \\
\textit{Predicate} &\rightarrow \textit{Have} \mid \textit{Read} \mid ... \\
\textit{Function} &\rightarrow \textit{AuthorOf} \mid \textit{GenreOf} \mid ...
\end{aligned}
$$

*Рис. 5. Описание контекстно-свободной грамматики языка исчисления предикатов первого порядка*

точно хорошо изучены и формально обоснованы. Во-вторых, существуют достаточно эффективные процедуры вывода синтаксически правильных высказываний. В-третьих, такой подход позволяет хранить в базах знаний лишь множество аксиом, а все остальные знания (в том числе факты и сведения о людях, предметах, событиях и процессах) получать из этих аксиом по правилам вывода.

Язык представления смысла, как и любой язык, имеет свой синтаксис и семантику. На рисунке 5 дано описание контекстно-свободной грамматики для исчисления предикатов первого порядка, предложенное в работе [24].

Рассмотрим представление смысла категорий, событий, времени, аспектов и убеждений, приведенное в книге [25].

**Представление категорий.** Под категорией понимается группа слов, объединенная общим признаком наподобие того, как это организовано в тезаурусах. Слова с предикатно-подобной семантикой часто содержат ограничения выбора, которые обычно выражаются в форме семантических категорий, где каждый представитель категории обладает набором подходящих признаков.

Простейший способ представления категорий – создать одноместный предикат для каждой категории. Однако тогда будет сложно делать утверждения о самих категориях. Рассмотрим следующий пример. Допустим, при помощи языка логики предикатов первого порядка требуется представить смысл высказывания: *«Harry Potter» is the most popular children's book*. То есть нужно найти наиболее часто встречающийся объект категории в виде *MostPopular (HarryPotter, ChildrensBook)*. Эта формула не является настоящей формулой логики первого порядка, так как аргументами в предикатах по определению должны быть термы, а не другие предикаты. Чтобы решить эту проблему, все понятия, которые участвуют в высказывании, можно представить в качестве полноценных объектов, то есть категорию *ChildrensBook* представим как объект наравне с *HarryPotter*. Принадлежность к такой категории будет обозначена посредством отношения *ISA (HarryPotter, ChildrensBook).* Отношение *ISA (is a)* указывает на связь между объектами и категориями, к которым эти объекты принадлежат. Такой прием можно применить для создания иерархий категорий. Например, используем отношение *AKO (ChildrensBook, Book)*. Здесь отношение *AKO (a kind of)* обозначает включение одной категории в другую. Конечно, для большей достоверности категории должны характеризоваться большим множеством фактов, то есть категории должны быть определены как множества.

**Представление событий.** Чтобы представить смысл события, достаточно рассмотреть его в виде предиката от множества аргументов, выполняющих определенные роли и необходимых для описания ситуации. Примеры таких предикатов приведены в первом разделе (там они получены из лексических функций, предложенных И.А. Мельчуком). Другой пример: *Reservation (Hearer, Today, 8 PM, 2)*. Здесь аргументами являются такие объекты, как человек, ресторан, день, время и количество мест для резервирования в ресторане. Для глаголов такое представление можно получить, если считать, что аргументы соответствуют синтаксическим актантам. При таком подходе существуют четыре проблемы:

– определение корректного количества ролей для каждого события;

– представление фактов о ролях, ассоциированное с событием;

– необходимость убедиться, что все корректные выводы могут быть получены напрямую из такого представления события;

– необходимость убедиться, что ни один некорректный вывод не может быть получен из представления события.

Например, глагол «есть» может иметь от одного до четырех актантов в зависимости от ситуации, описанной предложением. Поэтому заранее непонятно, какова должна быть местность предиката. Ведь в исчислении предикатов первого порядка количество аргументов должно быть фиксированным.

Одно из решений – предположить, что подобные ситуации обрабатываются на синтаксическом уровне. Можно рассматривать отдельные подкатегории для каждой из конфигураций аргументов. Семантический аналог данного метода – создать как можно больше предикатов, каждый из которых будет соответствовать отдельным ситуациям. Название предиката одно, а количество аргументов разное: $Eating_1(w)$ – *I ate*; $Eating_2(w, x)$ – *I ate a sandwich*; $Eating_3(w, x, y)$ – *I ate a sandwich for lunch*; $Eating_4(w, x, y, z)$ – *I ate a sandwich for lunch at home*. Поэтому они считаются разными. Этот подход позволит обойти проблему количества аргументов, но он неэффективный. Кроме предложенных имен предикатов, ничто не объединяет их в одно событие, хотя очевидна их логическая взаимосвязь. Выходит, что некоторые логические связи не могут быть получены на основе предложенных предикатов. Более того, придется искать эти логические связи в базе знаний.

Решить эту проблему можно, воспользовавшись **смысловыми постулатами**. Они связывают семантику предикатов в явном виде. Например, $\forall\ w, x, y, z\ \ Eating_4(w, x, y, z) \Rightarrow Eating_3(w, x, y)$.

Предикаты могут отражать морфологическую, синтаксическую и семантическую информацию. Примерами таких смысловых постулатов являются формулы, содержащие некоторые лексические предикаты из первого раздела. Смысловые постулаты, содержащие морфологические и синтаксические особенности построения слов и предложений русского языка, приведены в [26]. Примеры смысловых постулатов, несущих семантическую нагрузку, встречаются в предыдущем разделе.

Заметим, что не следует путать семантику высказывания на естественном языке и семантику предиката, который мы вводим для того, чтобы отражать семантику высказывания. Смысловые постулаты отражают семантику предикатов, то есть семантические связи между введенными нами предикатами.

Ясно, что такой подход для обнаружения семантических связей между предикатами годится для небольших областей и имеет проблемы с масштабируемостью. Более удобно было бы сказать, что эти предикаты ссылаются на один предикат с аргументами, отсутствующими на некоторых позициях. В этом случае можно обойтись без смысловых постулатов. Но этот метод также имеет недостаток. Например, если рассмотреть предикат *Eating* (*w, x, y, z*) и считать, что в качестве третьего аргумента должно присутствовать одно из слов из множества {*Breakfast, Lunch, Dinner*}, то квантор существования, навешенный на другую переменную, будет означать существование конкретной пищи, ассоциированной с каждым приемом, что не соответствует действительности.

Рассмотрим подходящий пример. Запишем три высказывания (*I ate a lunch*, *I ate at home* и *I ate a sandwich for lunch at home*), используя логику первого порядка:

$\exists\ w, x$ *Eating* (*Speaker*, *w*, *Lunch*, *x*)

$\exists\ w, x$ *Eating* (*Speaker*, *w*, *x*, *Home*)

$\exists\ w$ *Eating* (*Speaker*, *w*, *Lunch*, *Home*).

Допустим, что необходимо получить из двух первых формул, относящихся к одному событию, третью формулу. Независимые события *I ate lunch* и *I ate at home* не позволяют заключить, что *I ate lunch at home*. Как и в случае с представлением категорий, можно разрешить эту проблему, рассматривая события наравне с объектами так, чтобы на них можно было навешивать кванторы и связывать с другими объектами с помощью наборов заданных отношений. Теперь, согласно такому подходу, будет получено следующее представление.

Для предложения *I ate lunch*

$\exists\ w$ *ISA* (*w*, *Eating*) $\wedge$ *Eater* (*w*, *Speaker*) $\wedge$ *Eater* (*w*, *Lunch*).

Для предложения *I ate at home*

$\exists\ w$ *ISA* (*w*, *Eating*) $\wedge$ *Eater* (*w*, *Speaker*) $\wedge$ *Place* (*w*, *Home*).

Для предложения *I ate a sandwich for lunch at home*

$\exists$ *w ISA* (*w, Eating*) $\wedge$ *Eater (w, Speaker*) $\wedge$ *Eaten* (*w, Sandwich*) $\wedge$ *MealEaten* (*w, Lunch*) $\wedge$ *Place* (*w, Home*)

Изложенный подход позволяет избавиться от необходимости задавать фиксированное количество аргументов в предикате независимо от ролей и других актантов. Другие роли, которые не упоминаются в предложении, отсутствуют, при этом логические связи между семантически связанными предикатами не требуют использования смысловых постулатов.

**Представление времени.** Для описания последовательностей событий и их взаимосвязи во временной шкале используется **темпоральная логика**. В естественных языках таким инструментом является время глагола. Можно считать, что одно событие предшествует другому, если поток времени ведет от первого события ко второму. Отсюда возникают привычные нам понятия прошлого, настоящего и будущего.

В темпоральной логике используются два вида операторов: логические и модальные. В качестве логических операторов используются обычные операторы логики исчисления высказываний: конъюнкция, дизъюнкция, отрицание и импликация. Модальные операторы определяются следующим образом [27].

**N** φ – **Next:** φ должно быть истинным в состоянии, непосредственно следующим за данным.

**F** φ – **Future:** φ должно стать истинным хотя бы в одном состоянии в будущем.

**G** φ – **Globally:** φ должно быть истинно во всех будущих состояниях.

**A** φ – **All:** φ должно выполняться на всех ветвях, начинающихся с данной.

**E** φ – **Exists:** существует хотя бы одна ветвь, на которой φ выполняется.

φ **U** ψ – **Until** (strong): ψ должно выполниться в некотором состоянии в будущем (возможно, в текущем), свойство φ обязано выполняться во всех состояниях до обозначенного (невключительно).

φ **R** ψ – **Release:** φ освобождает ψ, если ψ истинно, пока не наступит момент, когда φ первый раз станет истинно (или всегда, если такого момента не наступит). Иначе, φ должно хотя бы раз стать истинным, пока ψ не стало истинным первый раз.

**Представление аспектов.** Для описания действий в естественных языках применяются глаголы. Американский философ З. Вендлер в 1957 году предложил модель разделения глаголов по лексическим аспектам [28]. Он выделил четыре класса:

– **стативы (states)** – глаголы, описывающие статические состояния, которые не имеют конечной точки (например, «знать», «любить»);

– **деятельности (activities)** – глаголы, описывающие состояния, которые являются динамическими и не имеют конечной точки (например, «бежать», «ехать»);

– **совершения (accomplishments)** – глаголы, описывающие события, которые имеют конечную точку и являются постепенными (например, «рисовать картину», «строить дом»);

– **достижения (achievements)** – глаголы, описывающие события, которые имеют конечную точку и происходят мгновенно (например, «узнавать», «замечать»).

В таблице 2 приведена сравнительная таблица классов Вендлера для английских глаголов, взятая из работы [29].

Как видно, продолженность действия характерна деятельностям и совершениям и отсутствует у стативов и достижений. Можно сказать *It was boiling* (деятельность) и *I was writing a letter* (совершение), но нельзя сказать *It was existing* (статив) и *I was finding a book* (достижение). Достижения не сочетаются с обстоятельствами длительности. Можно сказать *It existed for two hours* (статив), но нельзя *I found it for two hours* (достижение).

*Таблица 2*

**Классификация Вендлера для английских глаголов**

| Класс глагола | Продолженность действия | Обстоятельство длительности | Обстоятельство срока завершения |
|---|---|---|---|
| Стативы (States) | – | + | – |
| Деятельности (Activities) | + | + | – |
| Совершения (Accomplishments) | + | + | + |
| Достижения (Achievements) | – | – | (+) |

Совершения и достижения описывают целенаправленные действия, они сочетаются с обстоятельствами срока завершения в отличие от стативов и деятельностей. Можно сказать *I wrote a letter in two hours* (совершение), но нельзя сказать *I walked in two hours* (деятельность).

**Представление убеждений, желаний и намерений**. Для выражения отношения говорящего к сообщаемой информации в высказываниях на естественном языке используются такие слова, как *верить, хотеть, полагать, представлять* и т.д. Такие высказывания описывают не объективную картину мира, а особенности личностного восприятия говорящего, его «внутренние» представления о мире. Рассмотрим высказывание *I belive that John read «Harry Potter»*. Неверно пытаться представить его смысл при помощи логики предикатов: *Believing* (*Speaker, Reading* (*John, HarryPotter*). Здесь второй аргумент должен быть термом, а не формулой. Эта синтаксическая ошибка влечет за собой семантическую. В логике первого порядка предикаты связывают объекты, а не отношения между ними. Стандартный способ преодолеть эту проблему – добавить операторы, которые позволяют делать нужные нам утверждения. Если ввести оператор *Believes*, имеющий в качестве аргументов формулы, тогда получим следующее представление:

*Believes* (*Speaker*, $\exists$ *x ISA* (*x, Reading*) $\wedge$ *Reader* (*x, John*) $\wedge$ *Read* (*x, HarryPotter*)).

Нельзя сказать, что такое представление записано в терминах исчисления предикатов первого порядка, но оно является подтверждением того, что в языке существует группа глаголов, выполняющая особую роль в семантическом анализе. В системах автоматического анализа иногда необходимо отслеживать убеждения и намерения пользователей. Ситуация осложняется тем, что убеждения, желания и намерения могут меняться в процессе диалога.

Введенный оператор называется модальным. Бывают различные модальные операторы. О временной модальности уже упоминалось немного ранее, когда говорилось о представлении времени в высказываниях. Помимо временной, существует пространственная модальность, логика знания («известно, что»), логика доказуемости («можно доказать, что») и другие. Логика, расширенная модальными операторами, называется **модальной логикой**. В настоящее время в этой области остается множество сложных неизученных вопросов. Как логический вывод работает в присутствии специфических модальных операторов? К каким видам формул могут применяться те или иные операторы? Как модальные операторы взаимодействуют с кванторами и логическими связками? Эти и другие вопросы еще предстоит исследовать. Здесь останавливаться на них не будем.

Вывод синтаксически правильных высказываний в логических моделях представления знаний опирается на **правило резолюций**, разработанное Дж. Робинсоном в 1965 году. Оно утверждает, что если группа выражений, образующая посылку, является истинной, то применение правила вывода гарантированно обеспечит получение истинного выражения в качестве заключения. Результат применения правила резолюций называют **резольвентой.**

Метод резолюций (или правило устранения противоречий) позволяет проводить доказательство истинности или ложности выдвинутого предположения методом от противного. В методе резолюций множество предложений обычно рассматривается как составной предикат, который содержит несколько предикатов, соединенных логическими функциями и кванторами существования и всеобщности. Так как одинаковые по смыслу предикаты могут иметь разный вид, предложения сначала необходимо привести к унифицированному виду (дизъюнктивной или конъюнктивной нормальной форме), в которой удалены кванторы существования, всеобщности, символы импликации, эквивалентности и др. Правило резолюций содержит в левой части конъюнкцию дизъюнктов. Поэтому приведение посылок, используемых для доказательства, к виду, представляющему собой конъюнкции дизъюнктов, является необходимым этапом практически любого алгоритма, реализующего логический вывод на базе метода резолюций [30]. В процессе логического вывода с применением правила резолюций выполняются следующие шаги [17].

1. Устраняются операции эквивалентности и импликации:

$A \leftrightarrow B = (A \rightarrow B) \wedge (B \rightarrow A)$;

$A \rightarrow B = \neg A \vee B$.

2. Операция отрицания продвигается внутрь формул с помощью законов де Моргана:

$\neg (A \wedge B) = \neg A \vee \neg B$;

$\neg (A \vee B) = \neg A \wedge \neg B$.

3. Логические формулы приводятся к дизъюнктивной форме:

$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$.

В логике предикатов для применения правила резолюций необходимо осуществить более сложное преобразование логических формул для приведения их к системе дизъюнктов. Это связано с наличием дополнительных элементов синтаксиса, в основном кванторов, переменных, предикатов и функций. Алгоритм унификации предикатных логических формул состоит из следующих шагов [17].

1. Исключение операций эквивалентности.

2. Исключение операций импликации.

3. Внесение операций отрицания внутрь формул.

4. Исключение кванторов существования. Это может произойти на третьем шаге вследствие применения законов де Моргана, а именно, отрицание $\exists$ меняется на $\forall$, но при этом может произойти и обратная замена. Тогда для исключения $\exists$ поступают следующим образом: все вхождения некоторой перемен-

ной, связанной квантором существования, например ($\exists x$), заменяются в формуле на новую константу, например $a$. Эта константа представляет собой некоторое (неизвестное) значение переменной $x$, для которого утверждение, записанное данной формулой, истинно. При этом важно то, что на все места, где присутствует $x$, будет подставлено одно и то же значение $a$, пусть оно и является неизвестным в данный момент.

5. Кванторы общности выносятся на первые места в формулах. Это также не всегда является простой операцией, иногда при этом приходится делать переименование переменных.

6. Раскрытие конъюнкций, попавших внутрь дизъюнкций.

После выполнения всех шагов описанного алгоритма унификации можно применять правило резолюций.

Именно правило резолюций послужило основой для создания языка программирования Prolog. В языке Prolog факты описываются в форме логических предикатов с конкретными значениями. Правила вывода описываются логическими предикатами с определением правил вывода в виде списка предикатов над базами знаний и процедурами обработки информации. Интерпретатор языка Prolog самостоятельно реализует вывод, подобный вышеописанному. Для того, чтобы инициировать вычисления, выполняется специальный запрос к базе знаний, на который система логического программирования генерирует ответы «истина» и «ложь».

Метод резолюций легко программируется, это одно из важнейших его достоинств, однако он применим только для ограниченного числа случаев, так как для его применения доказательство не должно иметь большую глубину, а число потенциальных резолюций не должно быть большим.

Чтобы инструмент исчисления предикатов первого порядка был более гибким, его можно расширить лямда-исчислением. **Лямбда-исчисление** – это язык более высокого порядка, чем исчисление предикатов первого порядка. В нем в качестве аргументов лямбда-функция может работать не только с переменными, но и с предикатами. Тем не менее, использование лямбда-выражений формально не увеличивает выразительную мощность логики первого порядка, поскольку любую конструкцию, содержащую лямбда-выражение, можно преобразовать к эквивалентному виду без него [24].

После того, как язык Prolog приобрел большую популярность, в начале 80-х годов прошлого века появился термин «компьютеры пятого поколения». В то время ожидалось создание следующего поколения компьютеров, ориентированного на распределенные вычисления. Вместе с этим считалось, что пятое поколение станет основой для создания устройств, способных имитировать процесс человеческого мышления. Тогда же возникла идея создания аппаратной поддержки параллельных реляционных БД Grace и Delta [31, 32] и параллельного логического вывода (Parallel Inference Engine, PIE), опирающийся на принципы языка Prolog. Каждый блок логического вывода сигнализировал свою текущую рабочую нагрузку таким образом, чтобы работа могла быть передана в блок логического вывода с наименьшей нагрузкой [33]. Но, как известно, подобные попытки не позволили создать искусственный интеллект, а лишь послужили очередным подтверждением того, что человеческое мышление еще недостаточно изучено.

Логические модели представления знаний позволяют проверить синтаксическую правильность высказывания. Однако с помощью правил, задающих синтаксис языка, нельзя установить истинность или ложность того или иного высказывания. Высказывание может быть построено синтаксически правильно, но оказаться совершенно бессмысленным. Помимо этого, логические модели сложно использовать при доказательстве рассуждений, отражающих специфику конкретной предметной проблемы, из-за высокой степени единообразия. [17]

### Системы с компонентами семантического анализа

***Open Cognition***

В рамках проекта Open Cognition [34] разрабатывается анализатор Link Grammar Parser, который отвечает за обработку естественного языка. Link Grammar Parser начал разрабатываться в 1990-е гг. в университете Карнеги–Меллона [35]. Данный подход отличается от классической теории синтаксиса. Система приписывает предложению синтаксическую структуру, которая состоит из множества помеченных связей (коннекторов), соединяющих пары слов. Link Grammar Parser использует информацию о типах связей между словами.

Анализатор имеет словари, включающие около 60 000 словарных форм. Он позволяет разбирать большое число синтаксических конструкций, включая многочисленные редкие выражения и идиомы. Link Grammar Parser довольно устойчив, он может пропустить часть предложения, которая ему непонятна, и определить некоторую структуру оставшейся части предложения. Анализатор способен работать с неизвестной лексикой и делать разумные предположения (на основе контекста и написания) о синтаксической категории неизвестных слов. У него есть данные о собственных именах, числовых выражениях и различных знаках препинания.

Анализ в системе проходит в два этапа.

1. Построение множества синтаксических представлений одного предложения. На этом этапе рассматриваются все варианты связей между словами и выбираются среди них те, которые удовлетворяют **критерию проективности** (связи не должны пересекаться) и **критерию минимальной связности** (получившийся граф должен содержать наименьшее число компонент связности; компонента связности графа – некоторое множество вершин графа такое, что для любых двух вершин из этого множества существует путь из одной в другую, и не существует пути из вершины этого множества в вершину не из этого множества).

2. Постобработка. Предназначена для работы с уже построенными альтернативными структурами предложения.

Получаемые диаграммы по сути являются аналогом деревьев подчинения. В деревьях подчинения от главного слова в предложении можно задать вопрос к второстепенному. Таким образом, слова выстраиваются в древовидную структуру. Синтаксический анализатор может выдать две или более схемы разбора одного и того же предложения. Это явление называется синтаксической синонимией.

Главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей (около 100 основных, причем некоторые из них имеют 3-4 варианта). В некоторых случаях тщательная работа над разными контекстами привела авторов системы к переходу к почти семантическим классификациям, построенным исключительно на синтаксических принципах. Так, выделяются следующие классы английских наречий: ситуационные наречия, которые относятся ко всему предложению в целом (clausal adverb); наречия времени (time adverbs); вводные наречия, стоящие в начале предложения и отделенные запятой (openers); наречия, модифицирующие прилагательные, и т.д.

Из достоинств системы нужно отметить, что организация самой процедуры нахождения вариантов синтаксического представления очень эффективна. Построение идет не сверху вниз (top-down) и не снизу вверх (bottom-up), а все гипотезы отношений рассматриваются параллельно: сначала строятся все возможные связи по словарным формулам, а потом выделяются возможные подмножества этих связей. Это, во-первых, приводит к алгоритмической непрозрачности системы, поскольку очень трудно проследить за всеми отношениями сразу, а во-вторых – не к линейной зависимости скорости алгоритма от количества слов, а к экспоненциальной, поскольку множество всех вариантов синтаксических структур на предложении из слов в худшем случае равномощно множеству всех основных деревьев полного графа с вершинами.

Последняя особенность алгоритма заставляет разработчиков использовать таймер для того, чтобы вовремя останавливать процедуру, которая работает слишком долго. Однако все эти недостатки с лихвой компенсируются лингвистической прозрачностью системы, в которой с одинаковой легкостью прописываются валентности слова, причем порядок сбора валентностей внутри алгоритма принципиально не задается, связи строятся как бы параллельно, что полностью соответствует нашей языковой интуиции.

Для каждого слова в словаре записывается, какими коннекторами оно может быть связано с другими словами предложения. Коннектор состоит из имени типа связи, в которую может вступать рассматриваемая единица анализа. Только основных, наиболее важных связей имеется более 100. Для обозначения направления связи справа к коннектору присоединяется знак «+», слева – знак «–». Левонаправленный и правонаправленный коннекторы одного типа образуют связь (link). Одному слову может быть приписана формула коннекторов, составленная с помощью определенных связок.

Отметим также недостатки Link Grammar Parser.

1. Практическое тестирование системы показывает, что при анализе сложных предложений, длина которых превышает 25–30 слов, возможен комбинаторный взрыв. В этом случае результатом работы анализатора становится «панический» граф, как правило, случайный вариант синтаксической структуры, с лингвистической точки зрения неадекватной.

2. Применение описанных выше идей затруднено для флективных языков типа русского ввиду значительно возрастающего объема словарей, которые возникают в силу морфологической развитости флективных языков. Каждая морфологическая форма должна описываться отдельной формулой, где нижний индекс входящего в нее коннектора должен обеспечивать процедуру согласования. Это приводит к усложнению набора коннекторов и к увеличению их количества.

Проект Open Cognition, в рамках которого развивается Link Grammar Parser, открытый и бесплатный, что является большим преимуществом для проведения исследований. Довольно подробное описание и исходный код можно найти на сайте [36]. Open Cognition продолжает развиваться в настоящее время, что также важно, поскольку есть возможность взаимодействовать с разработчиками. Наравне с Link Grammar ведется разработка анализатора RelEx [37], который позволяет извлекать отношения семантической зависимости в высказываниях на естественном языке, и в результате представляются предложения в виде деревьев зависимостей. Он использует несколько наборов правил для перестроения графа с учетом синтаксических связей между словами. После каждого шага, согласно набору правил сопоставления, в полученном графе добавляются теги структурных характеристик и отношений между словами. Однако неко-

торые правила, наоборот, могут сокращать граф. Таким образом происходит преобразование графа. Этот процесс применения последовательности правил напоминает метод, используемый в ограничительных грамматиках. Главное отличие состоит в том, что RelEx работает с графовым представлением, а не с простыми наборами тегов (обозначающими отношения). Эта особенность позволяет применять более абстрактные преобразования при анализе текстов. Другими словами, основная идея состоит в том, чтобы использовать распознавание образов для преобразования графов. В отличие от других анализаторов, которые полностью опираются на синтаксическую структуру предложения, RelEx больше ориентирован на представление семантики, в частности, это касается сущностей, сравнений, вопросов, разрешения анафор и лексической многозначности слов.

***Система «Диалинг»***

«Диалинг» – автоматическая система русско-английского перевода, которая разрабатывалась с 1999 по 2002 гг. в рамках проекта «Автоматическая обработка текста» (АОТ). В разное время в работе над системой принимали участие 22 специалиста, большинство из которых – известные ученые-лингвисты. За основу системы «Диалинг» была взята система *французско-русского автоматического перевода* (ФРАП), разработанная в ВЦП совместно с МГПИИЯ им. М. Тореза в 1976–1986 гг., и система анализа политических текстов на русском языке «Политекст», разработанная в центре информационных исследований в 1991–1997 гг.

Система «Политекст» была направлена на анализ официальных документов на русском языке и содержала полную цепочку анализаторов текста: графематический, морфологический, синтаксический и частично семантический. В системе «Диалинг» был частично заимствован графематический анализ, но адаптирован под новые стандарты программирования. Программа морфологического анализа была написана заново, поскольку скорость работы была низкой, но сам морфологический аппарат не изменился [38].

На графематическом уровне константами являются графематические дескрипторы. Например, ЛЕ (лексема) – присваивается последовательностям, состоящим из кириллических символов; ИЛЕ (иностранная лексема) – присваивается последовательностям из латинских символов; ЦК (цифровой комплекс) – присваивается последовательностям, состоящим из цифр; ЦБК (цифробуквенный комплекс) – присваивается последовательностям, состоящим из цифр и букв, и т.д.

На морфологическом уровне для обозначений используются граммемы – грамматические характеристики, относящие словоформу к определенному морфологическому классу. Различные граммемы одной категории исключают друг друга и не могут быть выражены в одном слове. Например, *жр* – женский род, *тв* – творительный падеж, *мн* – множественное число, *но* – неодушевленность, *св* – совершенный вид, *дст* – действительный залог, *пе* – переходность глагола, *пвл* – повелительная форма глагола, *нст* – настоящее время глагола и т. д.

Фрагментационный анализ ставит своей целью деление предложения на неразрывные фрагменты (синтаксические единства), бóльшие словосочетания или равные ему (синтаксической группе), и установление частичной иерархии на множестве этих единств. Возможные типы фрагментов: главные предложения, придаточные предложения в составе сложного, причастные, деепричастные и другие обособленные обороты. Про каждый фрагмент известно, какие фрагменты в него непосредственно вложены и в какие он непосредственно вложен.

Система ФРАП содержала полную цепочку анализа текста вплоть до семантического, который был реализован только частично. В системе ФРАП был разработан и опробован семантический аппарат, на основе которого в системе «Диалинг» был создан особый метод семантического анализа – **метод полных вариантов**. ФРАП не содержала механизмов структурных оценок семантического представления, то есть методов не просто одного вхождения текстового элемента, а всей структуры в целом. Идея метода полных вариантов состоит в том, что в анализе должны быть четко разделены варианты анализа, возникающие на разных этапах, и декларативные лингвистические правила (частичные модели), которые строят и оценивают отдельные варианты. Такой подход, ранее применяемый только для предсемантических анализаторов, теперь, ввиду развития компьютерных мощностей, стало возможным перенести на семантику, тем самым повысив уровень разделения процедурной и декларативной частей системы [38]. Процедурная часть семантического анализа в идеальном случае сводится к циклам, перебирающим разные лингвистические варианты. Таким образом, стало возможно упростить лингвистические модели благодаря увеличившейся скорости компьютеров.

Основными составляющими применяемого в «Диалинге» семантического аппарата являются семантические отношения (СО) и семантические характеристики (СХ). Примеры семантических отношений: ИНСТР – «инструмент», ЛОК – «локация, местоположение», ПРИНАДЛ – «принадлежность», РЕЗЛТ – «результат» и пр. Они довольно универсальны и имеют сходства с предикатами, рассмотренными в первом разделе, и семантическими ролями, упоминаемыми в третьем разделе. Семантические характеристики позволяют строить формулы с использованием логических связок «и» и «или». Каждому слову приписывается некоторая формула, составленная из семантических характеристик. В семантическом словаре

«Диалинга» содержится около 40 семантических характеристик. Примеры семантических характеристик: АБСТР – абстрактное существительное или прилагательное, ВЕЩВО – название химического вещества или того, что можно отмерять по весу или объему; ГЕОГР – географический объект; ДВИЖ – глаголы движения; ИНТЕЛ – действия, связанные с мыслительной деятельностью; КОММУНИК – глаголы речи; НОСИНФ – носители информации; ОРГ – организация; СОБИР – все, что обозначает множество однотипных объектов; ЭМОЦ – прилагательные, которые выражают эмоции и т.д. Некоторые характеристики являются составными, так как их можно выразить через другие. Есть характеристики, которые являются антонимами. Использование их в одной конъюнкции запрещено. Существуют характеристики, которые являются разновидностями других. Семантические характеристики наравне с грамматическими характеристиками обеспечивают проверку согласования слов при интерпретации связей в тексте.

В данный момент все инструменты, разработанные в рамках проекта АОТ (в том числе система «Диалинг»), являются свободным кроссплатформенным программным обеспечением. Демоверсия и подробная документация доступны на сайте [39].

**Системы извлечения информации и представления знаний**

Существуют и другие системы, содержащие компоненты семантического анализа. Однако они имеют существенные недостатки для исследований: сложно найти описания, которые не являются бесплатными и свободно распространяемыми или не работают с текстами на русском языке. К ним относятся OpenCalais (http://www.opencalais.com/opencalais-api/), RCO (http://www.rco.ru/?page_id=3554), Abbyy Compreno (https://www.abbyy.com/ru-ru/isearch/compreno/), SemSin (http://www.dialog-21.ru/media/1394/kanevsky.pdf), DictaScope (http://dictum.ru/) и др.

Следует упомянуть систему извлечения данных из неструктурированных текстов Pullenti (http://semantick.ru/). Она заняла первое место на дорожках T1, T2, T2-m и второе место на T1-l на конференции Диалог-2016 в соревновании FactRuEval. На сайте разработчиков системы Pullenti есть также демоверсия семантического анализатора, позволяющего по предложению строить семантическую сеть.

Инструментальная среда «ДЕКЛ» (http://ipiranlogos.com/) разработана в конце 90-х годов и использована для построения *экспертных систем* (ЭС), оболочек для ЭС, *логико-аналитических систем* (ЛАС), *лингвистических процессоров* (ЛП), обеспечивающих обработку и автоматическое извлечение знаний из потоков неформализованных документов на естественном языке.

Система машинного перевода «ЭТАП-3» предназначена для анализа и перевода текстов на русском и английском языках. Система использует преобразование текстов на естественном языке в их семантическое представление на языке Universal Networking Language. Как уже говорилось ранее, разметка синтаксического корпуса «Национальный корпус русского языка» [5] выполняется лингвистическим процессором ЭТАП-3, основанным на принципах теории «Смысл ⇔ Текст».

В последнее время появляется все больше систем представления баз знаний в виде графов. Поскольку объемы информации постоянно увеличиваются с невероятной скоростью, такие системы должны поддерживать построение и пополнение баз знаний в автоматическом режиме. Автоматическое построение баз знаний может осуществляться на основе структурированных источников данных.

Примеры таких систем: Yago (http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/), DBpedia (http://wiki.dbpedia.org/), Freebase (https://developers.google.com/freebase/), Google's Knowledge Graph (https://developers.google.com/knowledge-graph/), OpenCyc (http://www.opencyc.org/). Другой подход позволяет извлекать информацию из открытых ресурсов в Интернете без участия человека: ReadTheWeb (http://rtw.ml.cmu.edu/rtw/), OpenIE (http://nlp.stanford.edu/software/openie.html), Google Knowledge Vault (https://www.cs.ubc.ca/~murphyk/Papers/kv-kdd14.pdf). Подобные системы являются экспериментальными, каждая из них имеет свои особенности. Например, Knowledge Vault пытается учитывать неопределенности, каждому факту ставится в соответствие коэффициент доверия и происхождения информации. Таким образом, все утверждения делятся на те, которые имеют высокую вероятность быть истинными, и те, которые могут быть менее вероятными. Предсказание фактов и их свойств осуществляется методами машинного обучения на основе очень большого количества текстов и уже имеющихся фактов. В данный момент Knowledge Vault содержит 1,6 млрд фактов. Система NELL, разрабатываемая в рамках проекта ReadTheWeb университетом Карнеги–Меллона, содержит более 50 млн утверждений с разными степенями доверия. Около 2 млн 800 тыс. фактов имеют высокую степень доверия. Процесс обучения NELL также еще не завершен.

Сделаем следующие выводы. С развитием компьютерных технологий и постоянным приростом объемов текстовой информации исследования в области автоматической обработки текстов сфокусировались на прикладном аспекте. Возможности большинства инструментов ограничиваются морфологическим и синтаксическим анализом в сочетании с методами из теории вероятностей и математической статистики. Таким образом, лишь выбранная часть наиболее простых задач оказалась решенной. Остальные проблемы по-прежнему еще предстоит решить.

Как мы убедились, причин этому много. Например, существует мнение, что каждое правило в синтаксисе имеет свой аналог в семантике. Этот постулат называют гипотезой «правило к правилу» (rule-to-

rule hypothesis [40]). На самом деле это соответствие не является взаимно-однозначным, и в этом состоит главная сложность. Действительно, каждому синтаксическому правилу (дереву разбора) можно сопоставить семантическое правило (дерево разбора), но оно не будет единственным. В обратную сторону аналогично семантическому правилу сопоставляется синтаксическое правило, но необязательно единственное. Именно эта неоднозначность приводит к неразрешимым на сегодняшний день проблемам в области автоматической обработки текстов. В связи с этим рассуждением возникает вопрос выбора нужного сопоставления из большого количества возможных вариантов.

Из всего вышесказанного можно сделать еще один очень важный вывод. Не следует рассматривать процессы генерации и интерпретации высказывания по отдельности, они неразрывно связаны между собой. Выражая свою мысль, человек ориентируется на то, поймет ли его собеседник. В процессе генерации высказывания человек как бы «перепроверяет» себя, моделируя, как собеседник воспримет информацию. Похожий механизм присутствует при интерпретации высказывания. Когда мы осмысливаем услышанное или прочитанное, мы опять же «сверяемся» с нашими знаниями и представлениями о мире. Только благодаря этому нам удается выбрать подходящий смысл.

Современные исследователи склоняются к мысли, что нужный выбор возможно сделать, имея дополнительную базу знаний о мире. Такая база знаний должна содержать общесмысловую информацию о понятиях и отношениях между ними, чтобы при обращении к ней можно было определить подходящий контекст высказывания в автоматическом режиме. Она помогла бы учитывать накопленные знания о мире, которые в явном виде не присутствуют в конкретном высказывании, но непосредственно влияют на его смысл.

### *Литература*


1. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». М.: Языки русской культуры, 1999. 346 с.
2. Лахути Д.Г., Рубашкин В.Ш. Семантический (концептуальный) словарь для информационных технологий // Научно-техническая информация. 2000. № 7. С. 1–9.
3. Падучева Е.В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004. 608 с.
4. Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во СПбГУ, 2003. 391 с.
5. Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата обращения: 22.08.2016).
6. Апресян В.Ю. и др. Новый объяснительный словарь синонимов русского языка. М.–Вена: Языки славянской культуры–Венский славистический альманах, 2004. 1488 с.
7. Хорошилов А.А. Методы автоматического установления смысловой близости документов на основе их концептуального анализа // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: тр. XV Всерос. науч. конф. RCDL' 2013. Ярославль: Изд-во ЯрГУ, 2013. С. 369–376.
8. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. М.: Наука, 1989. 189 с.
9. Лахути Д.Г., Рубашкин В.Ш. Средства и процедура концептуальной интерпретации входных сообщений на естественном языке // Изв. АН СССР. Сер. Технич. киберн. 1987. № 2. С. 49–59.
10. Рубашкин В.Ш. Семантический компонент в системах понимания текста // КИИ-2006. Тр. 10 национ. конф. по искусствен. интеллекту с междунар. участ. 2006. URL: http://www.raai.org/resurs/papers/kii-2006/#dokladi (дата обращения: 23.08.2016).
11. Падучева Е.В. Семантика вида и точка отсчета // Изв. АН СССР: Сер. лит. и яз. 1986. Т. 45. № 5. С. 18–25.
12. Падучева Е.В. Отпредикатные имена в лексикографическом аспекте // Науч.-технич. инф. 1991. Сер. 2. № 5. С. 21–31.
13. WordNet. A lexical database for English. URL: http://wordnet.princeton.edu/ (дата обращения: 23.08.2016).
14. Семантическая сеть. URL: https://ru.wikipedia.org/wiki/Семантическая_сеть (дата обращения: 23.08.2016).
15. Minsky M. Minsky's frame system theory // Proceedings of the workshop on Theoretical issues in natural language processing (TINLAP '75). 1975, pp. 104–116.
16. Хабаров С.П. Представление знаний в информационных системах: конспекты лекций. URL: http://www.habarov.spb.ru/bz/bz07.htm (дата обращения: 23.08.2016).
17. Луценко Е.В. Представление знаний в информационных системах: электр. учеб. пособие для студентов. Краснодар: Изд-во КубГАУ, 2010. 428 с.
18. Константинова И.С., Митрофанова О.А. Онтологии как системы хранения знаний // Всерос. кон-

курсн. отбор стат. по приорит. направл. «Информационно-телекоммуникационные системы». 2008. 54 с.

19. Разин В.В., Тузовский А.Ф. Представление знаний о времени с учетом неопределенности в онтологиях Semantic WEB // Докл. Томского гос. ун-та систем управления и радиоэлектроники. 2013. № 2 (28). С. 157–162.
20. Patel-Schneider P.F., Horrocks I. et al. SWRL: A Semantic Web Rule Language Combining OWL and RuleML // World Wide Web Consortium (W3C). 2004. URL: http://www.w3.org/Submission/SWRL (дата обращения: 18.08.2016).
21. Fillmore Ch. The Case for Case. Proc. Texas Symp. on Language Universals, 1967, 134 p.
22. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. М.: Прогресс, 1981. С. 369–495.
23. Dowty D. Thematic Proto-Roles and Argument Selection // Language, 1991, vol. 67, no. 3, pp. 547–619.
24. Норвиг П., Рассел С. Искусственный интеллект: современный подход. М.: Вильямс, 2007. 1408 с.
25. Jurafsky D., Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2008, 1024 p.
26. Батура Т.В., Мурзин Ф.А. Машинно-ориентированные логические методы отображения семантики текста на естественном языке: монография. Новосибирск: Изд-во НГТУ, 2008. 248 с.
27. Темпоральная логика. URL: https://ru.wikipedia.org/wiki/Темпоральная_логика (дата обращения: 23.08.2016).
28. Vendler Z. Verbs and times. The Philosophical Review, 1957, vol. 66, no. 2, pp. 143–160.
29. Падучева Е.В. Лексическая аспектуальность и классификация предикатов по Маслову–Вендлеру // Вопросы языкознания. 2009. № 6. С. 3–21.
30. Вывод в логических моделях. Метод резолюций. URL: http://www.aiportal.ru/articles/knowledge-models/method-resolution.html (дата обращения: 11.08.2016).
31. Boral H., Redfield S. Database Machine Morphology. Proc. 11th Intern. Conf. Very Large Data Bases, 1985, pp. 59–71.
32. Fushimi S., Kitsuregawa M., Tanaka H. An overview of the system of a parallel relational database machine GRACE. Proc. 12th Intern. Conf. Very Large Data Bases, 1986, pp. 209–219.
33. Tanaka H. Parallel Inference Engine. IOS Press Publ., 2000, 296 p.
34. Open Cognition. URL: http://opencog.org/ (дата обращения: 23.08.2016).
35. Link Grammar Parser. AbiWord, 2014. URL: http://www.abisource.com/projects/link-grammar/ (дата обращения: 20.08.2016).
36. The CMU Link Grammar natural language parser. URL: https://github.com/opencog/link-grammar/ (дата обращения: 22.08.2016).
37. RelEx Dependency Relationship Extractor. OpenCog. URL: http://wiki.opencog.org/wikihome/index.php/Relex (дата обращения: 22.08.2016).
38. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ). Дисс. … канд. тех. наук. М.: МГПИИЯ, 2001. 120 с.
39. Автоматическая обработка текста. URL: http://www.aot.ruhttp://aot.ru/ (дата обращения: 23.08.2016).
40. Prószéky G. Machine Translation and the rule-to-rule hypothesis. New Trends in Translation Studies (In Honour of Kinga Klaudy). Budapest: Akadémiai Kiadó, 2005, pp. 207–218.