УДК 681.3

# ОБЩЕНИЕ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ КАК ПРОЦЕСС ПРЕОБРАЗОВАНИЯ «ТЕКСТ-СМЫСЛ»

Омар А.Х. Авадала
(представил д.т.н., проф. Е.И. Бобыр)

*Рассматривается процесс использования естественного языка как средства коммуникации. Предлагается представлять тексты общения в плане выражения (поверхностная структура текста) и в плане содержания (семантика текста – смысл), а достижение взаимопонимания взаимодействующих партнеров трактовать как переход от смысла к тексту и обратно. Обосновываются формализмы, применяемые для представления модели естественного языка, используемого как средства общения.*

**Введение**. Одним из основных направлений создания экспертных систем является разработка методов, обеспечивающих реализацию процесса общения с ЭВМ на естественном языке (ЕЯ). В этой связи представляют несомненный интерес вопросы формализации ЕЯ. Рассмотрим процесс использования естественного языка человеком как средства коммуникации. С этой целью обратимся к рис. 1, на котором изображены взаимодействующие партнеры A и B, каждый из которых обладает языковой системой и совокупностью знаний о внешней среде – моделью среды.

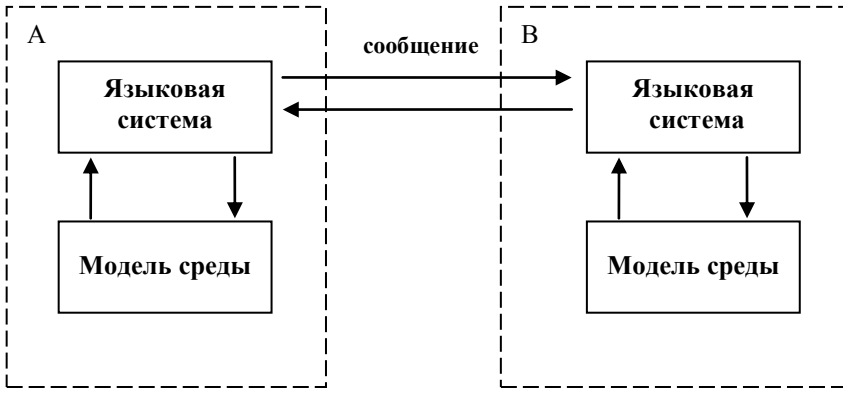


Рис. 1. Процесс взаимодействия

Пусть A и B обмениваются сообщениями на естественном языке.

Согласно существующим на сегодняшний день представлениям о мыслительной деятельности человека процесс понимания сводится к интерпретации сообщения партнера в собственной языковой системе, соотнесении его с моделью среды с последующей выработкой ответного сообщения [1]. При этом считается, что естественный язык представляет собой сложную многоуровневую структуру, в которой языками крайних уровней применительно к обозначениям рис. 1 являются язык сообщений и внутренний язык представлений человека. На внутреннем языке человек представляет смысл информации, использует его структуры для постановки целей и формирования стратегий их достижения. Язык сообщений служит для общения: смысл, представленный на внутреннем языке, переводится на язык сообщений и обратно, поверхностные структуры языка сообщений позволяют формировать структуры внутреннего языка. Таким образом, процесс общения состоит из следующих основных компонентов: информации, подлежащей передаче и восприятию, которая представлена смыслами; физических носителей информации (текстов), которые в зависимости от носителя могут представлять собой устную речь, печатный текст, сообщения, набираемые на пульте терминала и т.п.; структуры, заполняющей пространство между смыслами и текстами и представляющей вместе с ними язык общения [2, 3].

Исходя из этого, процесс общения можно представить как переход от текста к смыслу и обратно. *Для реализации взаимодействия пользователей (операторов) с ЭВМ необходимо построить модель языка, реализующую подобный переход, что и является* **целью данной статьи**.

**Обоснование модели естественного языка, используемого как средство общения.** Учитывая вышесказанное, каждый текст взаимодействия будем рассматривать в двух аспектах: в плане выражения (поверхностная структура текста, именуемая для краткости текстом) и в плане содержания (семантика текста – смысл), а достижение взаимопонимания взаимодействующих партнеров можно трактовать как переход от смысла к тексту и обратно.

Существенным для естественного языка является наличие синонимии и омонимии, в силу чего отображение множества текстов на множество смыслов не является взаимно однозначным. Поэтому непосредственный переход от текста к смыслу в общем случае крайне затруднен. Исходя из этого, в лингвистике используется многоуровневое описание языка, причем не существует единой точки зрения на количество уровней, необходимое для его адекватного представления. В процессе построения модели языка взаимодействия (ЯВ) необходимо определить число уровней и выбрать формализмы для их описания.

Практически все способы, используемые в настоящее время для формального задания поверхностного уровня языка, могут быть представлены описанием с помощью *систем составляющих* или описанием с использованием *деревьев синтаксического подчинения*.

Суть первого способа состоит в следующем. Пусть $a = x_{i1}\ x_{i2}\ \ldots\ x_{ik}$ – некоторая цепочка слов. Для подцепочки $b = x_{j1}\ x_{j2}\ \ldots\ x_{jp}$ выполняется вхождение (b, r) в цепочку a, если $x_{ir} = x_{j1}\ x_{ir+1} = x_{j2},\ \ldots\ x_{ir+p-1} = x_{jp}$, где $i < r$, $r + p - 1 < k$. Пусть, далее $d = x_{t1}\ x_{t2}\ \ldots x_{tg}$ – подцепочка a , для которой выполняется вхождение (d, s) в a. Вхождения (b, r) и (d, s) перекрываются, если $r < s < r + p - 1$ или $s < r < s + g - 1$.

Рассмотрим множество σ (a ) вхождений всех подцепочек в цепочку a, причем $(a,_1) \in \sigma$ (a), где $(a,_1)$ – вхождение a самой в себя. Система $C \in \sigma$ (a) называется системой составляющих цепочки a, если:

1) C содержит вхождение $(a,_1)$, а также вхождение всех символов, из которых состоит a;

2*)* из любых вхождений ά, β Є C либо одно вложено в другое, либо вхождения ά и β не перекрываются.

С описанием поверхностной (синтаксической) структуры цепочек в терминах их составляющих тесно связан класс грамматик непосредственно составляющих – НС-грамматик G(A.Ан, I, C), продукции которых имеют вид $u_1 \psi\ u_2 \rightarrow u_1\ z\ u_2$ (I), где A и Ан – конечные неперекрывающиеся множества терминального и нетерминального словарей; I Є Ан – начальный символ, указывающий совокупность языковых объектов, на порождение которых ориентирована грамматика; C – схема грамматики – совокупность продукций типа (I); z Є F(ν) – непустая цепочка; ψ Є Ан – нетерминальный символ; $u_1$, $u_2$ Є F(ν) – цепочки, часто называемые контекстами; F (ν) – свободная полугруппа над словарем $\nu = A \cup A_н$.

Если $u_1 = u_2 = Ф$ для всех продукций схемы C, то получим частный случай НС-грамматик – контекстно-свободные или КС -грамматики, широко используемые при задании формальных языков.

Существует естественная связь между выводом произвольной цепочки в НС-грамматике и системой составляющих данной цепочки. Нетрудно показать, что каждому способу вывода некоторой цепочки в грамматике однозначно соответствует дерево составляющих.

В качестве примера, иллюстрирующего вышесказанное, для КС-грамматики, продукции которой имеют следующий вид:

1. I → <ИГм.ед.им> <Гн.з.ед.> < ИГср.ед.вин. > ;
2. <ИГм.ед.им._> → < Пм.ед.им.> < См.ед.им. > ;
3. <ИГср.ед.вин> → <С ср.ед.вин.> ;
4. <Пн.ед.им.>→ < коммерческий > ;

5. <См.ед.им.> → <директор > ;

6. <Гн.з ед. > →<проводит > ;

7. <С ср.ед.вин.>→ <совещание>

на рис. 2 изображено дерево вывода некоторого текста, составляющие которого отмечены скобками (I – начальный символ – текст; ИГм.ед.им. – именная группа мужского рода единственного числа именительного падежа; ИГср.ед.вин. – именная группа среднего рода единственного числа винительного падежа; П и С – прилагательное и существительное именной группы; Гн.з.ед. – глагол настоящего времени третьего лица единственного числа).

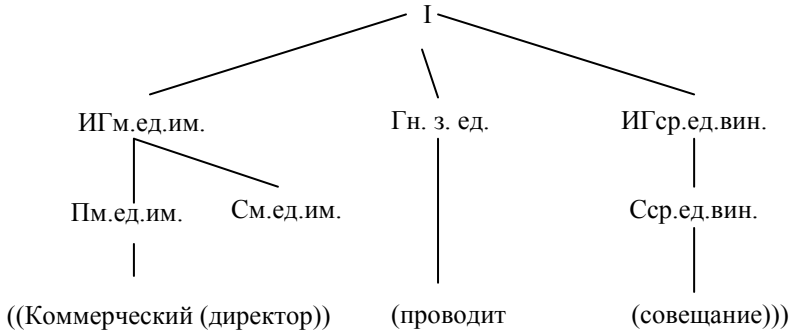


Рис. 2. Дерево вывода текста с использованием КС-грамматики

Естественный язык имеет большое количество разнообразных форм предложения: повествовательное, вопросительное, утвердительное, отрицательное, активная, пассивная и т.д. В НС-грамматике такие предложения будут порождаться более или менее независимо и, следовательно, не будут находиться в каких-либо явных отношениях друг к другу.

Неспособность НС-грамматик учитывать подобные "оттенки" предложения привела к созданию Хомским теории трансформационных грамматик [4]. Последние, однако, не нашли широкого применения в силу низкой эффективности алгоритмов синтаксического анализа, построенных на их основе.

В определенной степени преодолеть указанные ограничения удалось созданием системы, описывающей ЕЯ с использованием расширенной сети переходов. В этом случае некоторая КС-грамматика представляется сетью из узлов и связанных с ними дуг, помеченных символами, ввод которых в систему может вызвать переход вдоль данной дуги. Расширение мощности КС-грамматики по описанию ЕЯ осуществляется путем

введения множества регистров, ассоциированных с данной сетью, при этом переходы на сети могут осуществляться а зависимости от состояния этих регистров [2, 3]. Подобные системы позволяют обрабатывать довольно сложные предложения естественного языка. При этом достижение "человеческой разговорности" осуществляется за счет довольно сложных алгоритмов обработки текста. Несмотря на различие используемых средств, общим для данного способа описания языка является то, что входному тексту сопоставляется система составляющих, а язык рассматривается как множество цепочек (текстов), порождаемых формальной грамматикой. Описание ЕЯ с использованием аппарата формальных грамматик наряду с наличием ряда достоинств: наглядностью, достаточно хорошо разработанными формализмами – имеет существенные недостатки, связанные с трудностью учета контекстных ограничений, а также с тем, что теории порождающих грамматик не создали однозначной детализированной процедуры для обнаружения глубинных структур языка.

Рассмотрим сущность метода ***представления текста с использованием деревьев синтаксического подчинения***.

Пусть x – произвольная непустая цепочка над словарем A, а X – множество всех точек (слов) в x.

Произвольное бинарное отношение → на X, при котором граф {X; →} является деревом, будем называть отношением синтаксического подчинения или поверхностно-синтаксическим отношением для x, а {X; →} – деревом синтаксического подчинения. Свяжем с цепочкой x размеченное дерево синтаксического подчинения {X, →,Z,ψ}, где {X; →} – дерево синтаксического подчинения; Z – конечное множество меток; ψ – отображение множества дуг дерева {X; →} в Z.

Представление вышерассмотренного текста с использованием данного формализма дано на рис. 3.

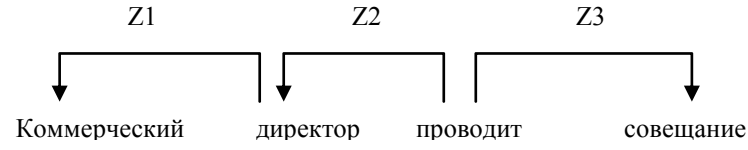


Рис. 3. Представление текста с использованием деревьев синтаксического подчинения

Здесь метки Zi, i = 1, 2, 3 – имена отношений синтаксического подчинения, определяемые функцией тех или иных частей речи в предложении. В настоящее время для русского языка составлен экспериментальный список, включающий 42 поверхностно-синтаксических отношения. По мнению ав-

тора [1], список этот является открытым и зависит от того, где проведена граница между семантикой и синтаксисом при анализе языка.

При таком подходе к описанию текста цепочка слов $x_i$ разбивается на пары $x_{ij} \rightarrow x_{ik}$ ($x_{ij}$, $x_{ik}$ – точки цепочки), связанные отношением синтаксического подчинения. При этом $x_{ij}$ называют хозяином; $x_{ik}$ – слугой, причем у каждого слова текста может быть только один хозяин, слуг же может быть несколько. В полном предложении все слова $x_i$ должны иметь хозяина, кроме одного (предиката), которое всегда остается независимым и считается вершиной предложения (“проводит” на рис. 3).

Изображение синтаксических связей с помощью отношений синтаксического подчинения в основном совпадает с принятыми представлениями о грамматической структуре предложения (дополнения зависят от сказуемого, определение от определяемого и т.д.). Вследствие этого любые предложения русского языка могут быть представлены древесной структурой.

**Выводы:** 1. Системы составляющих и деревья синтаксического подчинения характеризуют синтаксическую структуру текста в разных аспектах. С помощью первых описываются в явном виде словосочетания, но игнорируется ориентация связей (т.е. не различаются главные и зависимые элементы); вторые дают возможность рассматривать направленные связи между словами.

2. Язык описания смыслов и способ представления и использования знаний взаимообусловлены. Они должны иметь единую формальную основу, а также выразительные средства, достаточные для описания понятий, отношений и ситуаций предметной области.

## ЛИТЕРАТУРА


1. *Поспелов Д.Н. Прикладная семиотика и искусственный интеллект. – 1996. – № 3. – С. 10 – 13.*
2. *Кургаев А.Ф. Метаязык представления знаний // Управляющие системы и машины. – 1998. – № 4. – С. 79 – 86.*
3. *Вагин В.Н. Некоторые базовые принципы построения интеллектуальных систем поддержки принятия решений реального времени // Изв. РАН. Теория и системы управления. – 2001. – № 6. – С. 114 – 123.*
4. *Бондарев В.Н., Аде Ф.Г. Искусственный интеллект. – Севастополь: Изд-во Сев. НТУ, 2002. – 615 с.*


*Поступила 3.04.2003*

***Омар А.Х. Авадала**, аспирант НТУ «ХПИ». В 1996 году окончил НТУ «ХПИ». Область научных интересов – обработка языковой информации.*