СТРУКТУРНЫЙ АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1984 год Выпуск 101

УДК 519.766

ОБ ОДНОМ ПРИМЕНЕНИИ МЕТОДА АВТОМАТИЧЕСКОЙ РЕКОНСТРУКЦИИ ГРАММАТИК

М.К.Тимофеева

В [1] предложен метод автоматической реконструкции грамматики произвольной языковой подсистемы*) по конечному набору порожденных этой подсистемой текстов. В [2] исследовались числовые характеристики реконструированных грамматик и показывалась возможность использования этих характеристик для классификации текстов, но различия между языковыми подсистемами не исчерпываются их числовыми признаками. Анализ содержательных свойств подсистем позволяет выявить специфические черты отражения этими подсистемами объективной действительности. Цель данной статьи – демонстрация возможности применения метода автоматической реконструкции грамматик как средства установления содержательных особенностей грамматик языковых подсистем. Рассматриваемые содержательные характеристики восстанавливаются на основе интерпретации результата реконструкции как фрагмента традиционной грамматики анализируемого языка**). Оценивается длина входного текста, достаточная для восстановления рассматриваемых содержательных характеристик языковых подсистем.

*) Под языковой подсистемой здесь понимается совокупность языковых средств, используемых в некоторой ограниченной области применения языка.

**) Под выражением "традиционная грамматика" понимается традиционно-лингвистическая концепция, сложившаяся в 50-х – 60-х годах этого века и отраженная, например, в [3].

I. Содержательные характеристики

К числу основных функций традиционной грамматики Г относятся следующие: 1) наименование явлений объективной действительности (семантические средства Г) и 2) формирование способов наименования явлений объективной действительности (формальные средства Г). Поясним каждую из этих функций.

Наименование явлений объективной действительности словами некоторого языка всегда опосредованно его грамматикой [4]. Нельзя назвать ни одно явление, не подведя его под какую-либо из имеющихся в языке грамматических категорий (для русского языка - это категории существительного, прилагательного, глагола, единственного или множественного числа, настоящего, прошедшего или будущего времени и т.д.). Каждая из таких категорий имеет определенное значение, т.е. именует некоторый признак рассматриваемого явления (например, категория существительного именует признак "является предметом"). Любое средство описания действительности представляет собой результат комбинирования морфем - единиц грамматики Г, обладающих своим собственным значением. Совокупность значений грамматических категорий и морфем, участвующих в процессе наименования явлений действительности, относится к семантическим средствам Г.

Введем некоторые определения.

Пусть A - алфавит, T - текст в алфавите A, r - подмножество символов-разделителей из A. С е г м е н т - подцепочка текста, заключенная в нем между двумя разделителями или началом текста и разделителем и не содержащая разделителей внутри себя. Для естественного языка понятие сегмента совпадает с понятием с л о в о ф о р м ы в традиционной грамматике.

В грамматике Г каждая словоформа u представляется в виде конкатенации двух цепочек u = gh: основы g и окончания h. Множество словоформ, различающихся только своей грамматической формой, называется л е к с е м о й*), а упорядоченный набор окончаний, используемых для образования словоформ одной и той же лексемы, - п а р а д и г м о й P_i. В Г все лексемы с одинаковыми парадигмами

*) В одну лексему могут быть объединены а) словоформы с одной и той же основой; б) словоформы, содержащие чередующиеся символы в основе ("кусок" - "куска"); в) словоформы, образованные от разных основ ("идти" - "шел"). Лексемы последнего типа при данном методе реконструкции не рассматриваются.

объединяются в один парадигматический тип $П_1$. Произвольная лексема ω называется *изменяемой в тексте Т*, если в Т имеются хотя бы две разные словоформы из ω. Словоформы, входящие в одну и ту же лексему, связаны между собой *словоизменительными отношениями*.

Основа g делится на *морфемы* g_i: $g=g_1 \dots g_n$. Лексема ω, имеющая основу $g = p_1p_2$, называется *производной в тексте Т*, если существует лексема ω', отличная от ω, такая, что а) в Т присутствуют словоформы из ω и ω'; б) основа g' лексемы ω' представляется в виде p_1p_3 (или p_3p_2), p_1, p_2, p_3 - последовательности морфем. Лексемы ω и ω' в этом случае связаны *словообразовательными отношениями*. Последовательность морфем p_1 (или соответственно p_2) называется *словообразующей* для ω и ω'.

Совокупность правил изменения и производства лексем в текстах рассматриваемого языка относится к формальным средствам Г.

Каждая языковая подсистема использует свой комплекс семантических и формальных средств Г. Исследование этих комплексов дает возможность оценить значимость тех или иных аспектов описываемых языком явлений для пользующихся соответствующей подсистемой. Проиллюстрируем это на примере отражения языком процессов.

Возможности подведения явлений действительности под имеющиеся грамматические категории Г, как правило, неоднозначны. При наименовании некоторого процесса одним словом могут учитываться разные его признаки. По тому, какие из признаков регулярно включаются в наименование процесса, можно судить о значимости этих признаков для пользующихся языковой подсистемой. Так, если для делающего высказывание большее значение представляют предметы, участвующие в процессе, а не сам процесс как таковой, то он описывается как атрибут этих предметов и выражается в форме причастия. Если для говорящего не важны ни время осуществления процесса, ни его участники, то он выражается в форме инфинитива или отглагольного существительного (например, "бегание"). Средством указания на завершенность процесса служит вид глагола, а на отношение к говорящим - формы 1-го и 2-го лица глагола. Рефлексивность процесса (совпадение субъекта и объекта процесса) отражают возвратные формы, например, "двигаться", "умываться".

Описанные семантические и формальные средства относятся к числу содержательных характеристик языковых подсистем. В имеющей-

ся лингвистической практике процесс выявления таких средств, как правило, опирается на знание самой грамматики Г. При автоматизации этого процесса возникает необходимость хотя бы частичной формализации Г, представляющей собой чрезвычайно сложную систему. Опыт реконструкции показал, что те достаточно простые формальные критерии, на основе которых построен применяемый метод реконструкции (опирающиеся только на самые общие свойства грамматических систем), позволяют выявить семантические и формальные средства Г, наиболее активно используемые в рассматриваемой языковой подсистеме. Для выявления таких средств результат реконструкции интерпретируется как подсистема традиционной грамматики Г.

2. Интерпретация реконструированных грамматик

Процесс реконструкции грамматики языка L по представляющему этот язык тексту T базируется на совокупности формальных предположений π, определяющих наиболее общие признаки грамматических систем [I]. Конструируемая на основе предположений π результирующая грамматика G_T является формальным объектом, интерпретация которого зависит от его практического использования. Интерпретацию G_T как подсистемы традиционной грамматики Г будем обозначать через I_T.

Распечатка грамматики G_T на устройстве печати ЭВМ состоит из трех частей.

Первая часть имеет следующую структуру:

$$u_1;u_{11};u_{12};\ldots;u_{1n_1}$$
$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$
$$u_t;u_{t1};u_{t2};\ldots;u_{tn_t}$$

Каждая i-я строка является объединением n_i сегментов: u_iu_{i1}, $u_iu_{i2},\ldots,u_iu_{n_i}$. Обозначим через W_i множество $\{u_{ij} \mid j=1,\ldots,n_i\}$, $U_i = \{u_iu_{ij} \mid u_{ij} \in W_i\}$, $U = \bigcup_{i=1}^{t} U_i$. Один и тот же сегмент u может присутствовать в нескольких разных множествах U_i. В каждом из множеств U_i, как правило, объединяются словоформы, связанные словоизменительными отношениями или относящиеся к лексемам, связанным словообразовательными отношениями. (В грамматике, реконструированной по общественно-политическому тексту, около 97% множеств U_i обладает этим свойством, в грамматиках, реконструирован-

ных по художественному тексту и тексту научно-технических рефератов - около 93%.)

Цепочки $u_{i_1 j_1}$ и $u_{i_2 j_2}$ называются в з а и м о з а м е щ а е м ы м и, если существует не менее S разных множеств W_i, каждое из которых содержит обе эти цепочки (S - заданное целое число). Во второй части распечатки указываются отношения взаимозамещаемости на множестве цепочек $\{u_{ij}\}$. Третья часть распечатки содержит значения формальных характеристик результирующей грамматики G_T.

Интерпретация I_T грамматики G_T как подсистемы Г состоит в сопоставлении каждой лексеме ω из Г, имеющей словоформы в U, пары $\langle F_1^\omega, F_2^\omega \rangle$, где F_1^ω - отображение, сопоставляющее ω множество значений окончаний, используемых в словоформах ω, присутствующих в U; F_2^ω - отображение, сопоставляющее ω множество значений словообразующих последовательностей морфем для лексемы ω.

Интерпретация I_T строится неформально на основе знания грамматики Г. Процесс формирования множеств U_i проводится автоматически и опирается только на наиболее общие сведения о Г, образующие множество предположений π.

3. Пример анализа содержательных характеристик

Рассмотрим результаты реконструкции грамматики по текстам трех языковых подсистем разной степени специализированности*): общественно-политической (P), художественной (H), научно-технических рефератов (R). Наименее специализированной является подсистема художественных текстов, наиболее - подсистема реферативных текстов. Каждая из подсистем представлена текстом объемом около 100 тысяч символов.

Реконструкция грамматики для каждой из перечисленных подсистем проводилась в два этапа. На первом этапе реконструировалась грамматика G_T^{15} начального участка текста (длиной 15 тысяч символов). Назначением второго этапа являлась проверка полноты построенной грамматики. На этом этапе по выборке из всего текста реконструировалась грамматика G_T^{100}. Выборка строилась путем отбора из текста всех словоформ, начинающихся с выделенных символов, общее число разных словоформ в каждой выборке не менее 500.

*) Под специализированностью здесь понимается степень ограниченности сферы использования языковой подсистемы.

В обозначениях длина входного текста указывается как верхний индекс, а тип текста - как нижний индекс, например, G_R^{15} - результирующая грамматики реферативного текста длиной 15 тысяч символов, I_p^{15} - интерпретация G_p^{15} как подсистемы Г. Через I_R, I_p, I_H обозначим объединение соответственно I_R^{15} и I_R^{100}, I_p^{15} и I_p^{100}, I_H^{15} и I_H^{100}.

Для примера рассмотрим способ отражения процессов в I_R, I_p, I_H.

В I_H средства обозначения процессов имеют наибольшую употребительность (см. табл.I), причем названия этих процессов, как правило, включают указания на время их совершения и на отношение к говорящим. I_H характеризуется значительно большим, чем I_R и I_p, развитием средств отражения признаков процессов (формы деепричастий). В словообразовательной подсистеме I_H также наиболее употребительны морфемы, служащие для образования глаголов.

Таблица I

Распределение лексем по классам

Грамматика	Всего изменяемых лексем	Существительные	Прилагательные	Глаголы (без причастий)	Причастия
I_R	278	II6 (42%)	70 (25%)	37 (I3%)	55 (20%)
I_p	345	I5I (44%)	83 (24%)	69 (20%)	42 (I2%)
I_H	278	94 (34%)	35 (I3%)	I34 (48%)	I5 (5%)

Чем более специализирована языковая подсистема, тем более внимание пользующихся ею смещается с процессов и их признаков на предметы и их признаки*). Процессы все чаще оформляются лишь как атрибуты предметов, в них участвующих (т.е. называются посредством причастий). Среди указывавшихся выше признаков процессов все большее значение приобретает рефлексивность. Время совершения процесса и его отношение к говорящим, как правило, в название этого процесса не включается.

Для пользующихся реферативной языковой подсистемой наибольшее значение представляют рефлексивные процессы; вневременные процессы, безотносительные к участвующим в них предметам (называются

*) Подтверждением этой тенденции является то, что искусственно создаваемые специализированные языковые подсистемы в информатике часто строятся без использования глагольных форм отражения процессов (см., например, [6]).

посредством отглагольных существительных); процессы как атрибуты некоторых предметов (называются посредством причастий). Процессы, как правило, рассматриваются безотносительно к говорящим.

Кроме указанных выше тенденций, с возрастанием специализированности языковой подсистемы а) уменьшается число используемых парадигматических типов; б) возрастает число неиспользуемых окончаний в глагольных парадигматических типах, т.е. уменьшаются возможности локализации высказывания во времени (табл.2).

Таблица 2

Употребительность глагольных форм

Грамматика	I_R	I_p	I_H
Настоящее время	65 (94%)	93 (64%)	I06 (32%)
Прошедшее время	3 (4%)	40 (27%)	I87 (56%)
Повелительное наклонение	0	I (2%)	II (3%)
Деепричастие	I (2%)	II (7%)	29 (9%)

Опыт исследования результирующих грамматик G_A английского текста показал, что для отнесения сегмента G_A к той или иной части речи часто бывает необходимо рассмотреть непосредственное окружение этого сегмента в исходном тексте. Результирующая же грамматика не содержит информации для анализа контекстов сегментов. Поэтому для установления содержательной специфики подсистем английского языка должен использоваться способ интерпретации результирующей грамматики, отличный от того, который использовался для результирующих грамматик русских текстов. Такой способ интерпретации может состоять, например, в сопоставлении каждой лексеме, имеющей словоформы в G_A, семантического класса, к которому она относится в традиционной грамматике английского языка.

4. Полнота выявления грамматических средств

При практическом исследовании содержательных характеристик языковых подсистем важно знать, каков приблизительный объем текста, достаточный для выявления этих характеристик. Оказывается, что

уже небольшой объем текста (около I5 тысяч символов или около 5 страниц) позволяет исследовать наиболее активно используемые в рассматриваемой подсистеме семантические и формальные средства Г.

Рассмотрим грамматику I_T некоторого текста Т. Пусть $П_1,\ldots,П_m$ - парадигматические типы, используемые в изменяемых лексемах грамматики I_T; $P_1,\ldots,P_m$ - соответствующие им парадигмы; k_i - число элементов в P_i; V_T - множество всех словообразующих последовательностей морфем в лексемах из U, $k = \max k_i$. Правила изменения лексем из U изображаются в виде матрицы Q_T, содержащей m столбцов, поименованных парадигмами P_i, и k строк, поименованных номерами окончаний. В каждой позиции q_{ij} матрицы Q_T стоит число изменяемых лексем из U, относящихся к парадигматическому типу $П_j$ и имеющих в U словоформу с окончанием, входящим в P_j под номером i. Парадигмы P_i в общем случае состоят из разного числа элементов, поэтому в матрице Q_T может заполняться не более $N = k \cdot m - \sum_{i=1}^{m} (k - k_i)$ позиций.

Таблица 3

Грамматика	I_R^{15}	I_P^{15}	I_H^{15}
Полнота словоизменительной подсистемы	0,83	0,68	0,63
Полнота словообразовательной подсистемы	0,72	0,74	0,7I

Рассмотрим грамматики I_T^{15} и I_T^{100} некоторого текста Т. Сопоставим этим грамматикам матрицы Q_T^{15}, Q_T^{100} и множества словообразующих последовательностей V_T^{15}, V_T^{100}. Пусть z_1, z_2 - число заполненных позиций соответственно в Q_T^{15} и в Q_T^{100}; z_3, z_4 - мощности множеств V_T^{15} и V_T^{100} соответственно. Полнотой словоизменительной подсистемы I_T^{15} относительно грамматики I_T^{100} назовем отношение z_1/z_2, полнотой словообразовательной подсистемы I_T^{15} относительно I_T^{100} - отношение z_3/z_4 (табл. 3).

Пусть каждой парадигме P_i соответствует в Q_T^{15} и Q_T^{100} столбец с одним и тем же номером. Числа, записанные в таких позициях q_{ij} матрицы Q_T^{100}, которые не являются заполненными в Q_T^{15}, не превышают 7. Каждая последовательность морфем, входящая в V_T^{100}, но

не содержащаяся в V_T^{15}, является словообразующей не более чем для трех лексем. Таким образом, рассмотренные содержательные характеристики I_T^{100}, не обнаруживаемые в I_T^{15}, относятся к числу малоупотребительных.

Сопоставление I_T^{15} и I_T^{100} показывает, что I_T^{15} уже содержит наиболее употребительные в тексте Т семантические и формальные средства Г. Поэтому результирующие грамматики текстов длиной 15 тысяч символов могут служить базой не только для формальной классификации текстов (как было показано в [1]), но и для анализа содержательных характеристик языковых подсистем.

Литература

1. ТИМОФЕЕВА М.К. Индуктивная реконструкция грамматик флективных языков. -В кн.: Методы обнаружения закономерностей с помощью ЭВМ (Вычислительные системы, вып. 81). Новосибирск, 1981, с. 57-68.

2. ТИМОФЕЕВА М.К. О методике индуктивной реконструкции грамматик. -Новосибирск, 1983. - 19 с. (Препринт ИМ СО АН СССР, №40).

3. Грамматика русского языка. Том I. -М.: АН СССР, 1953. -720с.

4. СТЕПАНОВ Ю.С. Основы общего языкознания. -М.: Просвещение, 1975. -271 с.

5. КОТОВ Р.Г. Лингвистические аспекты автоматизированных систем управления. -М.: Наука, 1977. -166 с.

Поступила в ред.-изд.отд.
25 октября 1983 года