

**Н. Хомский, Дж. Миллер**

# **ВВЕДЕНИЕ В ФОРМАЛЬНЫЙ АНАЛИЗ ЕСТЕСТВЕННЫХ ЯЗЫКОВ**



**Н. Хомский, Дж. Миллер**

# **ВВЕДЕНИЕ В ФОРМАЛЬНЫЙ АНАЛИЗ ЕСТЕСТВЕННЫХ ЯЗЫКОВ**

Перевод с английского  
*Е. В. Падучевой*

**Москва • 2003**

---



**УРСС**

**Хомский Ноам, Миллер Джордж**

**Введение в формальный анализ естественных языков: Пер. с англ.**

**М.: Едиториал УРСС, 2003. — 64 с.**

**ISBN 5-354-00114-5**

Книга выдающегося американского лингвиста Н. Хомского, написанная им в соавторстве с Д. Миллером, посвящена вопросам математической лингвистики. В ней обсуждаются математические модели языка.

Рассчитана на научных работников, инженеров, аспирантов и студентов различных специальностей, занимающихся и интересующихся кибернетикой.

Издательство «Едиториал УРСС». 117312, г. Москва, пр-т 60-летия Октября, 9.


Лицензия ИД № 05175 от 25.06.2001 г. Подписано к печати 14.03.2003 г.

Формат 60×90/16. Тираж 1000 экз. Печ. л. 4. Зак. № 71.

Отпечатано в ГУП «Облиздат». 248640, г. Калуга, пл. Старый Торг, 5.

**ISBN 5-354-00114-5**

**ИЗДАТЕЛЬСТВО УРСС**  
НАУЧНОЙ И УЧЕБНОЙ ЛИТЕРАТУРЫ

 E-mail: [URSS@URSS.ru](mailto:URSS@URSS.ru)  
Каталог изданий  
в Internet: <http://URSS.ru>  
Тел./факс: 7 (095) 135-44-23  
Тел./факс: 7 (095) 135-42-46

© Едиториал УРСС, 2003

## ВВЕДЕНИЕ

Язык и коммуникация играют особую, чрезвычайно важную роль в человеческом обществе, и над языковыми проблемами думают и работают ученые самых различных специальностей. Вклад, сделанный психологами, составляет лишь относительно небольшую часть от общих усилий. Поэтому, чтобы дать перспективную картину тех проблем, с которыми сталкивается математическая психология, изучая языковое поведение, нам придется в данной главе, как и в двух следующих<sup>2)</sup>, сильно выйти за рамки психологии в традиционном смысле слова.

Существо проблемы, возникающей при описании языка и языкового поведения, состоит в следующем: носитель языка обладает способностью понимать огромное число предложений, которые он никогда ранее не слышал и не произносил, а также способностью создавать в соответствии с ситуацией новые предложения, в свою очередь понятные другим носителям языка. Возникают следующие основные вопросы:

1. Какова точная природа этой способности?
2. Каким образом эта способность реализуется?
3. Как она приобретается человеком?

В последнее время имело место несколько попыток придать вопросам такого рода точную и эксплицитную форму и построить модели, имитирующие некоторые из указанных аспектов поведения носителя языка. Если построенные модели

---

<sup>1)</sup> Chomsky N., Miller G. A., Introduction to the formal analysis of natural languages, Handbook of Mathematical Psychology, Vol. 2, New York, Wiley, 1963, 269—322.

<sup>2)</sup> Настоящая работа представляет собой главу 11 упомянутой книги. Главы 12 и 13 той же книги, на которые здесь и далее ссылаются авторы (Chomsky N., Formal properties of grammars, 323—418; Miller G., Chomsky N., Finitary models of languages users, 419—491), также посвящены математической лингвистике и тесно связаны с настоящей главой.  
— *Прим. ред.*

достаточно просты, то оказывается возможным чисто абстрактное изучение их внутреннего характера и общих свойств. Исследования такого рода находятся пока еще в младенческом состоянии; лишь очень немногие аспекты языка и коммуникации уже формализованы до такой степени, когда можно думать об абстрактном их изучении. Однако плодотворные результаты в этом направлении уже имеются, и количество их все возрастает. Мы попытаемся дать обзор этих результатов и показать, каким образом эти исследования могут углубить наше понимание природы и функционирования языка.

Первый из трех вопросов касается природы языка, а чтобы ответить на этот вопрос, необходимо выявить внутреннюю структуру, общую для всех естественных языков. В логике и лингвистике возник ряд попыток подойти к этой задаче с принципиально новой стороны; в последние годы эти попытки сконцентрировались вокруг краеугольного понятия «грамматики». Включение данной работы в книгу, посвященную математической психологии, даст возможность познакомить психологов с более реалистичным взглядом на то, что же в конечном счете приобретает человек, когда он обучается говорить и понимать естественный язык. Ассоциирование звуковых реакций с визуальными раздражителями — фактор, который больше всего привлекал внимание психологов, — является далеко не единственным существенным аспектом процесса обучения языку.

Ответом на второй вопрос должно быть формальное (т. е. основанное на построении моделей) описание носителей естественного языка. Психологи, которые, как можно было бы ожидать, должны были бы заниматься этим вопросом в общей программе изучения человеческого поведения, пока что давали лишь самые общие (и часто мало правдоподобные) ответы на этот вопрос. Некоторые ценные идеи, относящиеся к этой тематике, возникли в общей теории связи; их связь с психологией была самая непосредственная, и она довольно быстро получила признание. Однако понятия теории связи являются преимущественно статистическими и лишь незначительно соприкасаются с тем, что известно о внутренней структуре языка.

Формулируя вопросы (1) и (2) отдельно, мы хотим отчетливым образом отвергнуть то общее мнение, что язык есть не что иное, как множество словесных реакций. Если утверждается, что в некотором естественном языке действует определенное грамматическое правило, то это еще не значит, что люди, которые пользуются этим языком, могут последовательно соблюдать это правило. Дать точное описание языка — это одна задача; дать описание носителя языка — другая. Эти две задачи очевидным образом связаны друг с другом, но не тождественны.

Третий из поставленных вопросов не менее существен, чем первые два; однако здесь имеется пока гораздо меньше исследований, которые могли бы послужить основой для постановки его в абстрактной форме. Те процессы, которые вступают в действие, когда ребенок начинает говорить, находятся пока за пределами математического моделирования. Мы можем только упомянуть о том, что этот вопрос является весьма существенным, и выразить сожаление по поводу того, что ему не уделяется должного внимания в последующем изложении.

### 1. УТОЧНЕНИЕ ГРАНИЦ ИССЛЕДОВАНИЯ

Математическое изучение языка и коммуникации — это очень большая тема. Для наших целей ее необходимо сильно ограничить. Чтобы помочь читателю, мы перечислим некоторые из ограничений, наложенных на постановку задач в этой и двух последующих главах.

Первое ограничение состоит в том, что наши интересы сосредоточены в основном на естественных языках. Существуют так называемые формальные языки, которые разрабатываются логиками и математиками и изучение которых составляет основной предмет современной логики. Мы пытались, однако, ограничиться формальным изучением естественных языков, не включая в рассмотрение формальных. В некоторых случаях удобно использовать миниатюрные искусственные языки, чтобы проиллюстрировать какое-то частное свойство естественного языка в упрощенном контексте; большой интерес представляют также языки программирования, разработанные специалистами по вычислительным машинам. Однако наше главное внимание будет сосредоточено на естественных языках.

Следующее ограничение состоит в том, что исключены из рассмотрения все непрерывные системы. Звуковой сигнал, производимый говорящим, является непрерывной функцией времени и обычно представляется с помощью ряда Фурье. Разложение в ряд Фурье особенно удобно в том случае, если изучается влияние непрерывных линейных трансформаций (фильтров). К счастью, этот важный вопрос много и тщательно изучается в математике и теории связи; поэтому мы можем на нем не останавливаться без ущерба для дальнейшего изложения.

Системы связи можно рассматривать как дискретные в силу существования того, что в теории информации иногда называется *критерием точности* (Shannon, 1949). Критерий точности определяет, каким образом множество всех сигналов, которые могут быть переданы за данный промежуток времени, должно быть разбито на подмножества сигналов, эквивалентных с точки

зрения адресата. Система связи может точно передавать непрерывные сигналы, но, если адресат не может (или не хочет) обращать внимания на тонкие различия, которые система способна передавать, точность канала связи теряется впустую. Таким образом, именно адресат устанавливает степень точности передачи для всей системы. Чем выше критерий точности, тем большим должно быть число различных подмножеств сигналов, которое система связи должна различить и передать.

Адресат, который является для нас объектом изучения, — это человек-слушатель. Критерий точности определяется его способностями, навыками и интересами. На основе различий, которые имеют место в его восприятии, мы можем установить конечное множество категорий, которые будут играть роль дискретных символов. Это множество может быть алфавитом, списком слогов или словарем. Дискретные элементы этих множеств являются теми неделимыми атомами, из которых должны строиться более длинные сообщения. Восприятие этих дискретных единиц слушателем является, вообще говоря, важной психологической проблемой, но мы здесь не будем на ней останавливаться. Некоторые соображения, касающиеся именно восприятия речи, будут изложены кратко в разд. 6, где рассматривается звуковая структура языка, а также в третьей из статей цикла, где ставится вопрос о том, как знание грамматики может быть использовано для организации нашего восприятия речи. Как мы увидим, строгое описание критерия точности для речевой деятельности человека представляет весьма сложную проблему; но пока для нас существенно только то, что люди делают звуки речи на эквивалентные подмножества; это оправдывает введение дискретных обозначений.

Далее, в области дискретных систем мы ограничиваемся *системами с операцией соединения* (concatenation systems) и изучением алгебраической структуры и взаимосвязей таких систем. В частности, мы представляем речевой поток как последовательность дискретных атомов, которые непосредственно соплагаются, или сочленяются, друг с другом. Хотя это ограничение может показаться очень простым, из него вытекает несколько следствий, на которых стоит остановиться.

Пусть  $L$  — множество всех конечных цепочек (включая цепочку длины 0), которые можно построить из элементов некоторого множества  $V$ . Тогда если  $\varphi, \chi \in L$  и  $\varphi \hat{\ } \chi$  есть результат их соединения друг с другом в указанном порядке, являющийся новой цепочкой  $\psi$ , то  $\psi \in L$ . Иными словами,  $L$  замкнуто относительно операции соединения. Далее, соединение ассоциативно:

$$(\varphi \hat{\ } \chi) \hat{\ } \psi = \varphi \hat{\ } (\chi \hat{\ } \psi);$$

пустая цепочка играет роль единственного единичного символа. Множество, которое включает единицу и замкнуто относительно ассоциативной операции, называется *моноидом*. Поскольку моноиды удовлетворяют трем из четырех требований группы, они иногда называются *полугруппами*. Группой называется моноид, в котором для каждого элемента имеется обратный.

Хотя мы вынуждены строить речевые высказывания с помощью ассоциативной операции соединения, при этом следует соблюдать осторожность. Рассмотрим, например, омонимичное английское предложение *They are flying planes*, которое в действительности представляет собой два различных предложения:

$\text{They} \sim (\text{are} \sim (\text{flying} \sim \text{planes}))$ , (1a)

$\text{They} \sim ((\text{are} \sim \text{flying}) \sim \text{planes})$  (1б)

(«Они — летящие самолеты» и «Они водят самолеты»). Если имеется в виду только написание и произношение, то пример (1a) равен примеру (1б), и имеет место простое соединение. Но если мы принимаем во внимание грамматическую структуру или значение, то примеры (1a) и (1б) отчетливо различны, хотя это различие, как правило, ни фонетически, ни графически не выражается.

Лингвисты обычно решают эту проблему введением допущения о том, что естественный язык имеет несколько различных *уровней*. В этой работе мы будем рассматривать каждый уровень как отдельную систему с операцией соединения — систему со своими собственными элементами и правилами. Структура на более низком уровне определяется тем, как ее элементы соотносятся друг с другом на следующем, более высоком уровне. Поэтому, чтобы сохранить ассоциативность, мы должны ввести несколько различных систем соединения и изучать отношения между ними.

Рассмотрим две операции, выполняемые человеком над письменным текстом. Первая операция переводит последовательность букв в последовательность звуковых сигналов; мы будем называть ее *произношением*. Произношения сегментов сообщения являются (с некоторым приближением) сегментами произношения этого сообщения. Таким образом, произношение обладает своего рода линейностью (ср. разд. 6.3):

Прозн.  $(x) \sim$  Прозн.  $(y) = \text{Прозн. } (x \sim y)$ . (2)

Хотя равенство (2) не является абсолютно верным (в частности, в нем игнорируется интонация и артикуляционные переходы между последовательными сегментами), оно в гораздо

большей мере верно, чем аналогичное утверждение, касающееся второй операции.

Вторая операция переводит последовательность символов в некоторое представление ее смысла; назовем ее *пониманием*. Значения сегментов сообщения очень редко можно отождествить с сегментами значения этого сообщения. Даже если допустить, что над значениями может каким-то образом производиться операция соединения, в большинстве случаев мы скорее всего обнаружим, что при разумной интерпретации соответствующих понятий

$$\text{Поним. } (x) \sim \text{Поним. } (y) \leq \text{Поним. } (x \cup y). \quad (3)$$

Это, по-видимому, один из способов истолковать то общее положение гештальт-психологии, что значение целого больше, чем значение линейной суммы его частей. Если не следовать положениям ассоциативной психологии в духе Джемса Милля, то вообще не очень ясно, что нужно понимать под соединением двух пониманий и каким образом эта операция может быть осуществлена. По сравнению с этим операция из равенства (2) кажется уже определенной достаточно точно.

Введение в рассмотрение операции понимания ставит много сложных проблем. В недавнее время появились интересные предложения, касающиеся абстрактного изучения некоторых денотативных (Wallace, 1961) и коннотативных (Osgood, Suci, Tappenberg, 1957) аспектов значения слов естественного языка. Хотя эти вопросы весьма важны для общей теории психолингвистики, мы не будем в дальнейшем уделять им много внимания. Мы надеемся, однако, что, выяснив некоторые синтаксические вопросы, мы тем самым внесем ясность и в какие-то аспекты семантики — хотя бы потому, что сможем сказать, чем значение не является:

Наконец, как уже было сказано, в нашей работе почти ничего не говорится про обучение языку. Мы в состоянии дать формальное описание некоторых аспектов языка; кроме того, мы располагаем некоторыми математическими моделями процесса обучения. Между тем пересечение этих двух теоретических направлений пока что обескураживающе пусто.

Каким образом ребенок может, без всяких наставлений, в столь короткий срок достичь полного владения языком — это насущнейшая проблема теории обучения. Конечно, при должном старании взрослый человек с достаточными способностями может с помощью обычной грамматики и словаря достичь определенной степени владения новым языком; однако ребенок достигает абсолютного совершенства, притом с несравненно большей легкостью и без всяких явных инструкций. Подробные

инструкции и точное программирование арсенала поощрений, по-видимому, не являются обязательными. Простое наблюдение, притом в течение поразительно короткого периода времени, — это, по-видимому, все, что требуется нормальному ребенку для достижения уровня овладения языком на уровне взрослого человека.

Один из путей выяснения затронутых здесь теоретических вопросов состоит в том, чтобы построить механизм, способный дублировать поведение ребенка (Chomsky, 1962). Важной составной частью такого механизма должно быть устройство, имеющее на входе определенную выборку грамматически правильных предложений (возможно, с определенными ограничениями относительно порядка их поступления) и выдающее на выходе грамматику языка (вместе со словарем). Описание этого устройства должно основываться на определенной гипотезе относительно врожденных интеллектуальных возможностей, которые ребенок должен пускать в ход при овладении языком. Конечно, при обучении могут играть существенную роль и другие данные, поступающие на вход устройства. Так, очень важны, по-видимому, исправления со стороны окружающего языкового коллектива: исправление — это указание, что некоторое языковое выражение не является правильным предложением; таким образом, входом в устройство может быть не только множество предложений, но и множество непредложений. Далее, на входе в устройство могут, по-видимому, появляться указания о том, что один элемент является повторением другого. Какие еще входы окажутся необходимыми, покажут практические исследования.

В равной мере существенно, кроме того, определить свойства грамматики, которую наше обучающееся устройство должно в конце концов дать на выходе. Предполагается, что грамматика отражает некоторые способности владеющего языком взрослого человека. Поэтому грамматика должна, в частности, показать, каким образом человек оказывается в состоянии определить, что такое правильно построенное предложение; кроме того, она должна содержать информацию относительно объединения языковых единиц в более крупные структуры. Обучающееся устройство должно, например, понять разницу между примерами (1a) и (1б).

Описание грамматики, которая дала бы возможность перечислить в явном виде грамматически правильные предложения языка и их структурные описания, занимает центральное место в дальнейшем изложении. Наша цель состоит в том, чтобы построить формальную грамматику, которая порождает правильные структурные описания предложений, исходя из относи-

тельно малого числа общих принципов устройства предложения; эта грамматика должна быть включена в состав общей теории языковой структуры, которая давала бы возможность обосновать преимущества одной грамматики перед другими. Одна из задач профессионального лингвиста состоит в каком-то смысле в том, чтобы сделать явным тот процесс, который бессознательно осуществляет всякий нормальный ребенок.

Практически работающая модель обучения языку должна основываться на весьма сильных допущениях относительно класса потенциальных грамматик, которые может иметь естественный язык. По-видимому, обучающееся устройство должно иметь в своем распоряжении сведения об общей форме, которую может принимать грамматика, а также некоторую процедуру, позволяющую на основе заданного текста решить, является ли данная грамматика более адекватной, чем некоторая другая. Кроме того, это устройство должно располагать определенными фонетическими возможностями для распознавания и реального построения предложений; оно должно также обладать каким-то механизмом для определения, на основе заданной грамматики, грамматической структуры предложений. Все это должно быть заготовлено заранее, прежде чем начнется обучение. Думать, что адекватная грамматика может быть выбрана из бесконечного числа теоретически допустимых возможностей на основе чистой индукции по конечному числу исходных предложений, — значит совершенно недооценивать сложность проблемы.

Процесс обучения должен, таким образом, состоять в оценке различных возможных грамматик с целью найти ту, которая лучше всего согласуется с исходными данными. Обучающееся устройство должно найти такую грамматику, которая порождает бы все предложения, не захватывая ни одного непредложения, и приписывала бы порождаемым предложениям структурные описания таким образом, чтобы все предложения, которые не являются повторениями друг друга, имели описания, отличные друг от друга в соответствующих местах. Конечно, обучающееся устройство нужно будет снабдить определенными эвристическими принципами, которые позволили бы ему, при заданных входах и заданном перечне возможных грамматик, достаточно быстро выбирать несколько наиболее подходящих вариантов, которые затем могут быть подвергнуты более тщательному сравнению; эвристические процедуры нужны также для того, чтобы оценивать одни характеристики грамматик раньше, чем другие. Необходимые эвристические процедуры сильно упростятся, если заранее сделать класс потенциальных грамматик относительно узким. Практически

правильное разделение труда между эвристическими методами и ограничениями на форму грамматики должно еще уточняться в дальнейшем. Ясно, однако, что не следует особенно надеяться на возможности обнаружения правильной грамматики с помощью индукции. В конце концов, даже глупые люди обучаются говорить, а самые умные обезьяны — нет.

## 2. НЕКОТОРЫЕ АЛГЕБРАИЧЕСКИЕ АСПЕКТЫ КОДИРОВАНИЯ

Отображение одной полугруппы на другую — это операция, весьма часто встречающаяся в системах связи. Ее можно называть в целом *кодированием*, понимая под этим все разнообразные процессы кодирования, перекодирования, декодирования и передачи сигналов. Для определенности будем считать, что одна полугруппа состоит из всех цепочек букв, которые могут быть построены при помощи конечного алфавита  $A$ , а другая — из цепочек слов, составляющих конечный словарь  $V$ . Таким образом, в этом разделе мы рассмотрим некоторые абстрактные свойства систем с операцией соединения — свойства, которыми характеризуются в равной мере и искусственные, и естественные коды.

Код  $S$  есть взаимнооднозначное отображение  $\theta$  цепочек в словаре  $V$  на цепочки в алфавите  $A$ , такое, что если  $v_i, v_j$  — цепочки в  $V$ , то  $\theta(v_i \hat{v}_j) = \theta(v_i) \hat{\theta(v_j)}$ . Где  $\hat{\phantom{x}}$  — изоморфизм между цепочками в  $V$  и некоторым подмножеством всех цепочек в  $A$ ; цепочки в  $A$  определяют записи цепочек из  $V$ . В дальнейшем там, где не возникает неоднозначности, мы будем упрощать обозначение, опуская символ  $\hat{\phantom{x}}$ , обозначающий соединение, и принимая, таким образом, соглашения, обычные для орофографических систем.

Рассмотрим в качестве простого примера код  $S_1$ . Пусть  $A = \{0, 1\}$ , а  $V = \{v_1, \dots, v_4\}$ . Отображение  $\theta$  определим так:

$$\begin{aligned}\theta(v_1) &= 1, \\ \theta(v_2) &= 011, \\ \theta(v_3) &= 010, \\ \theta(v_4) &= 00.\end{aligned}$$

Это отображение может быть представлено графом вида дерева, см. рис. 1 (о формальном описании графов вида дерева см. Berge, 1958). Узлы дерева соответствуют точкам выбора; путь вниз влево от узла соответствует выбору 1; путь вниз вправо — выбору 0. Каждое слово имеет единственную запись, соответствующую цепи (последовательности ребер) кодового

дерева. Когда достигнут конец цепи и записано целое слово, система возвращается к вершине, готовая записать следующее слово:

Чтобы правильно декодировать сообщение, необходимо, конечно, обеспечить синхронизацию. Так, цепочка слов  $v_1v_1v_1v_1$  записывается как 0010011; но если первая буква записи будет утрачена, то цепочка будет декодирована как  $v_2v_2$ . Чтобы показать, что первая буква в цепочке является началом сообщения, мы будем писать в начале цепочки символ  $\#$ . В противном случае ставится многоточие.

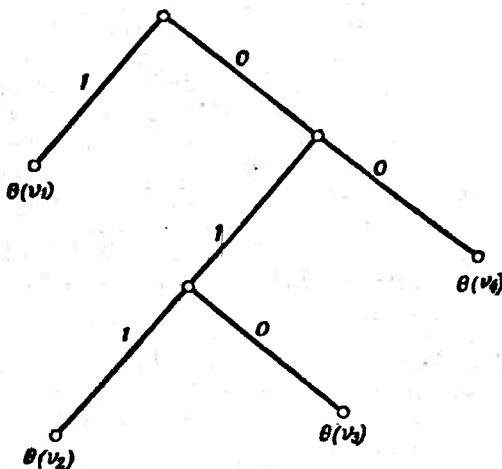


Рис. 1. Граф кодового дерева кода  $C_1$ .

В каждой данной точке последовательности букв, записывающей некоторое допустимое сообщение, имеется фиксированное множество возможных продолжений. Различные начальные цепочки могут допускать в точности одни и те же продолжения. В коде  $C_1$ , например, два сообщения, которые начинаются с  $\#000\dots$  и  $\#10\dots$ , могут быть закончены цепочками  $\dots0\#$  и  $\dots10\#$ ,  $\dots11\#$ , или же этими цепочками плюс новое слово. Введем бинарное отношение  $R$ , которое имеет место между любыми двумя начальными цепочками, допускающими одинаковое продолжение. Непосредственно видно, что  $R$  рефлексивно, симметрично и транзитивно; иначе говоря,  $R$  есть отношение эквивалентности. Две начальные цепочки, которые допускают в точности одинаковые продолжения, будут в дальнейшем называться *эквивалентными справа*. В терминах этого отношения можно определить важное понятие состояния: мно-

жество всех цепочек, эквивалентных справа, образует *состояние* кодирующей системы.

Состояние кодирующей системы — это ее память о том, что уже произошло. Каждый раз, когда к кодовой цепочке прибавляется одна буква, система может перейти в новое состояние. У кода  $C_1$  имеется три состояния: 1)  $S_0$ , когда только что было записано полностью какое-то слово; 2)  $S_1$  — после  $\#0...$  и 3)  $S_2$  — после  $\#01$ . Эти состояния соответствуют трем неким узлам дерева на рис. 1.

Следуя Шютценберже (Schützenberger, 1956), мы можем записывать переходы из одного состояния в другое в виде матриц — такая матрица существует для каждой цепочки букв. Пусть строки соответствуют состоянию после  $n$  выданных букв, а столбцы — состоянию после  $n+1$  буквы. Если переход возможен, в клетку ставится 1; в противном случае — 0. Для того чтобы задать код  $C_1$ , требуются две матрицы: одна показывает, каков результат добавления цифры 0, другая — результат добавления 1 (в общем случае это соответствует разделению кодового дерева на три подграфа, по одному на каждую букву из  $A$ ). Каждой цепочке букв  $x$  сопоставлена матрица  $M_x$ ; элементы этой матрицы  $m_{ij}$  определяют число возможных путей между состояниями  $S_i$  и  $S_j$  при появлении цепочки  $x$  в закодированном сообщении. Для кода  $C_1$  матрицы, сопоставленные элементарным цепочкам 0 и 1, таковы:

$$M_0 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Матрица для более длинной цепочки есть упорядоченное побуквенное произведение матриц. Произведение  $\mathcal{A}\mathcal{B}$  матриц  $\mathcal{A}$  и  $\mathcal{B}$  интерпретируется следующим образом: из  $S_i$  система переходит в  $S_j$  в соответствии с переходом, допустимым в  $\mathcal{A}$ , затем она переходит из  $S_j$  в  $S_k$  в соответствии с переходом, допустимым в  $\mathcal{B}$ . Число различных путей от  $S_i$  через  $S_j$  к  $S_k$  есть  $a_{ij}b_{jk}$ . Общее число путей от  $S_i$  к  $S_k$ , суммируя по всем промежуточным состояниям  $S_j$ , есть  $\sum_j a_{ij}b_{jk}$ , т. е. обычное произ-

ведение матриц «строка на столбец», которое дает элементы матрицы  $\mathcal{A}\mathcal{B}$ . В том случае, если некоторая буква вообще не может появиться в данном состоянии, строка ее матрицы, соответствующая этому состоянию, будет состоять целиком из нулей. Цепочке в  $A$ , которая не является частью записи никакой цепочки слов из  $V$ , будет соответствовать нулевая матрица. В общем случае для этих матриц не обязательно должны

существовать обратные. Матрицы не образуют группы, но изоморфно отображаются на элементы полугруппы.

Если функция, отображающая цепочки в  $V$  на цепочки в  $A$ , не является взаимнооднозначной, то в матрицах или их произведениях появятся элементы, большие единицы; это будет означать, что одна и та же цепочка из  $n$  букв является кодовой записью для более чем одной цепочки слов. Когда такая ситуация имеет место, полученное сообщение неоднозначно и не может быть понято, даже если оно было передано без помех и искажений. В этом случае нет никакого способа узнать, какая из возможных цепочек слов имелась в виду.

Поскольку между словами в нормальном потоке речи нет никаких фонетических показателей границы, неоднозначность такого рода возникает в естественных языках довольно часто. Миллер (Miller, 1958) приводит следующий пример для английского языка:

The good candy came anyway,

The good can decay many ways

(«Хорошая конфета все-таки появилась», «Хорошее может испортиться многими способами»). В этих примерах цепочку фонетических элементов можно прочесть нейтрально, так что для слушателя возникает эффект, который можно сравнить с визуальным явлением обратной перспективы. Неоднозначная сегментация гораздо более обычна во французском языке; так, следующий куплет хорошо известен как случай полной рифмы:

Gal, amant de la Reine, alla (tour magnanime),

Galamment de l'arène á la Tour Magne, á Nimes

[«Галь, любовник королевы, уехал (великодушный жест!) галантно с арены в башню Мань, в Ниме»].

Рассмотрение примеров такого рода показывает, что производство и восприятие речи есть нечто гораздо большее, чем производство и идентификация последовательных фонетических свойств, и что на разных уровнях организации происходят различные процессы переработки информации. Эти проблемы более подробно рассматриваются в применении к естественным языкам в разд. 6.

Разумеется, трудности сегментации могут быть устранены. Изучение способов их устранения приводит к простой классификации кодов (Шютценберже, в частной беседе). Общими кодами можно называть все коды, которые обеспечивают различное написание для всех различных цепочек слов. Особым видом общих кодов являются *древовидные* коды, в которых пра-

зла написания могут быть представлены графически, как на рис. 1, где каждому слову соответствует отдельная цепь в дереве. Древовидные коды могут относиться к одному из следующих двух типов: *леводревовидными* называются такие коды, в которых никакой код данного слова не является начальным (левым) сегментом кода какого-либо другого слова; *праводревовидными* называются соответственно такие коды, в которых ни одно слово не является конечным (правым) сегментом какого-либо другого слова; праводревовидный код можно получить, переписав в обратном порядке все слова, записанные в леводревовидном коде. Особую разновидность в классе древовидных кодов составляют коды, которые являются одновременно право- и леводревовидными; Шютценберже называет их *анаграмматическими*. Простейшим случаем анаграмматического кода являются *равномерные* коды, в которых все слова имеют одинаковое число букв и разделение сообщения на слова производится в воспринимающем устройстве просто на основе подсчета. Равномерные коды часто используются в инженерных приложениях теории информации; однако они не имеют большого значения для описания естественных языков.

Другим важным видом кодов являются *неравномерные* коды. Они обладают тем свойством, что если одиночная помеха или ошибка вызвала смещение границы слова, то эта ошибка не будет продолжать смещать границы слов до бесконечности; в течение конечного интервала времени ошибка будет поглощена и восстановится правильная синхронизация. Код  $C_1$  обладает этим свойством, равномерные коды им не обладают. В неравномерном коде границы слов всегда отмечаются появлением какой-либо определенной цепочки букв. Если имеется цепочка, которая заканчивает написание каждого слова, получается лево-неравномерный древовидный код. Если эта цепочка состоит из одной буквы (которая в этом случае не может встречаться ни в какой другой позиции), то получается *натуральный* код. В письменной речи синхронизация принимающего устройства осуществляется с помощью пробела между словами. В устной речи процесс сложнее; обсуждение вопроса о том, каким образом цепочки слов (морфем) отображаются в цепочки знаков фонетической транскрипции, см. в разд. 6.

Чтобы удостовериться в том, что некоторое отображение действительно является кодом, достаточно в соответствии с инженерной практикой убедиться в том, что оно является леводревовидным кодом, т. е. в том, что нет такого слова, запись которого является начальным сегментом записи некоторого другого слова. Существуют, однако, общие коды, которые не являются древовидными. Шютценберже (Schützenberger, 1956)

предлагает в качестве простейшего нетривиального примера код  $C_2$ :

$$A = \{0, 1\}, \quad V = \{v_1, \dots, v_5\},$$

$$\theta(v_1) = 00,$$

$$\theta(v_2) = 001,$$

$$\theta(v_3) = 011,$$

$$\theta(v_4) = 01,$$

$$\theta(v_5) = 11.$$

Отметим, что  $\theta(v_1)$  является начальным сегментом  $\theta(v_2)$ , так что этот код не является леводрезовидным; кроме того,  $\theta(v_4)$

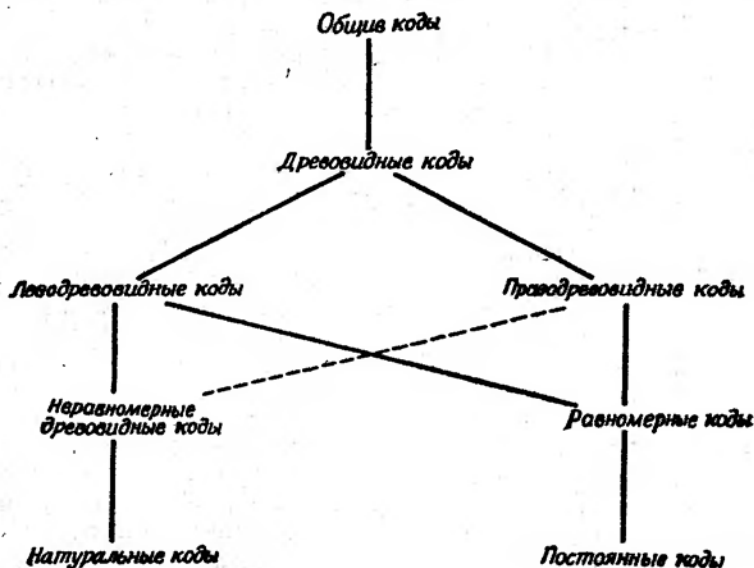


Рис. 2. Классификация кодирующих систем.

является конечным сегментом  $\theta(v_2)$ , так что он не является праводрезовидным.

При построении искусственных кодов возникает вполне понятное желание сделать длину закодированных слов как можно меньшей. В этой связи можно показать выполнение для дрезовидных кодов следующего интересного неравенства (Kraft, 1949). Пусть дан словарь  $V = \{v_1, \dots, v_n\}$ , алфавит  $A =$

$= \{a_1, \dots, a_D\}$  и отображение  $\theta$ , которое является леводревовидным кодом. Пусть  $c_i$  есть длина  $\theta(v_i)$ . Тогда

$$\sum_{i=1}^n D^{-c_i} \leq 1. \quad (4)$$

Это неравенство выводится следующим образом. Пусть  $w_j$  — число закодированных слов длины  $j$ ; тогда, поскольку  $\theta$  есть префиксный код, получаем

[illegible]

Делением последней строки на  $D^n$  получаем

$$\sum_{j=1}^n w_j D^{-j} \leq 1.$$

Но если  $n \geq c_i$  для всех  $i$ , то произошло суммирование по всем закодированным словам, откуда и получаем соотношение по формуле (4). Чем ближе для данного кода неравенство (4) подходит к равенству, тем ближе подходит данный код к минимизации средней длины своих слов. [Неравенство (4) выполняется также для недревовидных кодов (Mandelbrot, 1954; Mc Millan, 1956).]

Для всех древовидных кодов, включая, конечно, все натуральные коды, имеется некоторая функция  $w_j$ , определяющая число закодированных слов длины  $j$ . Эта функция представляет некоторый теоретический интерес, поскольку она выражает в чрезвычайно простом виде значительную информацию относительно структуры кодового дерева.

Наконец, следует отметить, что код можно рассматривать как простого вида автомат (см. след. главу), который принимает символы в одном алфавите и переводит символы в другой алфавит в соответствии с заранее составленными правилами, которые зависят только от входного сигнала и внутреннего состояния автомата. Некоторых более тонких трудностей, возникающих при построении хороших кодов, можно будет коснуться только после того, как мы припишем различным словам различные вероятности (см. гл. 13).

### 3. НЕКОТОРЫЕ ОСНОВНЫЕ ПОНЯТИЯ ЛИНГВИСТИКИ

Центральным понятием данной работы является понятие языка. Поэтому ему необходимо дать ясное определение. Мы считаем, что язык  $L$  есть множество (конечное или бесконечное) предложений, каждое из которых имеет конечную длину и построено с помощью операции соединения из конечного множества элементов. Это определение включает как естественные, так и искусственные языки логики и теории программирования.

Чтобы точно определить язык, необходимо установить некоторые общие принципы, которые отделяют последовательности атомарных элементов, являющиеся предложениями, от последовательностей, таковыми не являющихся. Это различие нельзя выразить просто списком, поскольку во всех представляющих интерес системах не накладывается ограничений на длину предложения. В таком случае определить язык можно двумя способами: можно попытаться разработать какую-либо операционную процедуру, которая будет отличать предложения от непредложений, или же можно построить рекурсивные правила, которые будут перечислять все предложения бесконечного списка. Первая из этих возможностей почти не исследовалась<sup>1)</sup> и останется за пределами нашего обзора. Вторая же составляет один из аспектов того, что естественно назвать *грамматикой* языка. Мы ограничимся здесь вторым подходом — исследованием грамматик.

При практическом исследовании естественного языка всякая операционная процедура проверки грамматической правильности, а также всякая грамматика должны удовлетворять определенным эмпирическим требованиям. До того как начата разработка анализирующей или порождающей процедуры, необходимо выделить конечный класс  $K_1$  последовательностей, которые с достаточной уверенностью могут быть включены в множество предложений, а также, по-видимому, класс  $K_2$  последовательностей, которые можно с достаточной уверенностью исключить из этого класса. Эмпирическая значимость той или иной анализирующей или порождающей процедуры в значительной мере определяется тем, сколь успешно они могут выразить различие между классами  $K_1$  и  $K_2$ . Однако вопрос эм-

<sup>1)</sup> Это утверждение едва ли верно. Алгоритмы синтаксического анализа текста, предназначенные для машинного перевода, как правило, строятся таким образом, что их без труда можно превратить в операционную процедуру для определения грамматической правильности. Ср. также одну из работ, в которых возможности построения разрешающих алгоритмов для естественных языков изучаются в непосредственном виде: И. Бар-Хилел, Разрешающие процедуры для структуры естественных языков, сб. «Математическая лингвистика», М., 1964, стр. 108—125. — *Прим. перев.*

лирической адекватности и многочисленные соображения, которые с ним связаны, выходят за рамки данного обзора.

Итак, мы займемся теперь изучением грамматик. Под грамматикой понимается набор правил, которые (в частности) рекурсивно перечисляют предложения языка. Всякое правило, которое нам может потребоваться, имеет вид

$$\Phi_1, \dots, \Phi_n \rightarrow \Phi_{n+1}, \quad (5)$$

где  $\Phi_i$  есть какая-то структура, а отношение  $\rightarrow$  интерпретируется как выражающее тот факт, что если наш процесс рекурсивного определения порождает структуры  $\Phi_1, \dots, \Phi_n$ , то он порождает также структуру  $\Phi_{n+1}$ .

Точное определение вида правил, которые могут быть разрешены в грамматике, — это одна из важных проблем математической лингвистики, и именно к ней мы сейчас обратимся. Рассмотрим следующий частный случай правил вида (5). Пусть в правиле (5)  $n=1$ , а  $\Phi_1$  и  $\Phi_2$  — цепочки символов некоторого алфавита (или словаря). И пусть имеется конечный язык, состоящий из предложений  $\sigma_1, \dots, \sigma_n$ , и абстрактный элемент  $S$  (от слова sentence — «предложение»), который мы выберем в качестве начального, данного. Тогда грамматику можно представить так:

$$S \rightarrow \sigma_1; \dots; S \rightarrow \sigma_n. \quad (6)$$

В этом тривиальном случае грамматика есть не что иное, как словарь предложений. Рассмотрим более интересный случай грамматики, состоящей из двух правил:

$$S \rightarrow aS; S \rightarrow a. \quad (7)$$

Эти два правила могут породить любое из предложений  $a, aa, aaa, aaaa, \dots$ <sup>1)</sup> (очевидно, эти предложения могут быть поставлены во взаимнооднозначное соответствие с числами натурального ряда, так что язык будет счетным). Порождение предложения, например  $aaa$ , происходит следующим образом:

- $S$  (данный, начальный символ),
- $aS$  (получается применением первого правила подстановки),
- $aaS$  (получается повторным применением первого правила подстановки),
- $aaa$  (получается применением второго правила подстановки).

<sup>1)</sup> Более точно следовало бы сказать, что мы рассматриваем грамматику с правилами вида  $\mu_1\Phi_1\mu_2 \rightarrow \mu_1\Phi_2\mu_2$ , где  $\mu_1$  и  $\mu_2$  — переменные, областью значений которых являются произвольные, быть может, пустые цепочки, а  $\Phi_1$  и  $\Phi_2$  — постоянные. Ясно, что если ограничиться только правилами этого вида, то при формулировке правил переменные могут быть опущены.

В дальнейшем мы будем изучать системы правил такого, а также более сложного видов.

Повторим еще раз, что *грамматикой* языка называется конечное множество правил, задающее этот язык (это определение будет уточнено в следующих разделах). Допустимой является та грамматика, которая задает список (в общем случае бесконечный) предложений данного языка (т. е. список цепочек символов, которые являются предложениями). Грамматика должна быть, по определению, конечной. Если допустить грамматки с неопределенным множеством правил, то сама проблема построения грамматик снимается: можно просто принять в качестве грамматики бесконечное число предложений, что будет, конечно, бессмысленно. Ясно, что грамматика должна играть роль теории повторяющихся закономерностей построения предложения, которые мы называем *синтаксической структурой* языка. В той мере, в какой грамматика формализована, она является математической теорией синтаксической структуры языка.

Очевидно, однако, что задачи грамматики не могут быть сведены к простому перечислению предложений языка (хотя практически даже эта цель никогда еще не была достигнута). Мы потребуем, чтобы грамматика приписывала каждому порождаемому предложению его *структурное описание*, которое определяло бы, из каких элементов построено предложение, каков их порядок, расположение, и т. д., — в общем задавало бы всю грамматическую информацию, необходимую для того, чтобы определить, как это предложение используется и понимается. Теория грамматики должна, таким образом, включать некоторое устройство, позволяющее для любой заданной грамматики  $G$  и предложения  $s$ , порождаемого грамматикой  $G$ , определять, какое структурное описание приписывает грамматика  $G$  предложению  $s$ . Если рассматривать грамматику как конечно характеризуемую функцию, которая перечисляет язык как свою область значений, то можно считать, что лингвистическая теория должна определить функционал, который сопоставляет каждой паре  $\langle G, s \rangle$ , где  $G$  — грамматика, а  $s$  — предложение, структурное описание  $s$  по отношению к  $G$ ; при этом одна из первоочередных задач лингвистической теории состоит, конечно, в том, чтобы дать ясное определение понятия «структурное описание».

Описанное понимание грамматики возникло недавно и может показаться непривычным. Несколько приводимых ниже примеров могут разъяснить, что здесь имеется в виду. Рассмотрим три искусственных языка, описанных Хомским (Chomsky, 1956).

Язык  $L_1$ . В  $L_1$  входят предложения  $ab$ ,  $aabb$ ,  $aaabbb$  и т. д.; иначе говоря, в  $L_1$  входят все предложения, содержащие  $n$  вхождений буквы  $a$ , за которыми следует  $n$  вхождений буквы  $b$ , и не входит никаких других предложений.

Язык  $L_2$ . В  $L_2$  входят предложения  $aa$ ,  $bb$ ,  $abba$ ,  $baab$ ,  $aabbaa$  и т. д.; иначе говоря, в  $L_2$  входят все предложения с зеркальным отображением (т. е. предложения, содержащие данную цепочку, за которой следует та же цепочка с обратным порядком символов) и не входит никаких других предложений.

Язык  $L_3$ . В  $L_3$  входят предложения  $aa$ ,  $bb$ ,  $abab$ ,  $baba$ ,  $ababab$  и т. д.; иначе говоря, в  $L_3$  входят все предложения, в которых за данной цепочкой следует та же цепочка, записанная еще раз, и не входит никаких других предложений.

Грамматика  $G_1$  языка  $L_1$  может иметь следующий вид:

Дано:  $S$ ,

$$\begin{aligned} F1: S &\rightarrow ab, \\ F2: S &\rightarrow aSb, \end{aligned} \quad (9)$$

где  $S$  играет роль аксиомы, а  $F1$  и  $F2$  — это правила образования, с помощью которых из аксиомы могут быть выведены допустимые цепочки символов. Вывод происходит так, как описано в примере (8). Вывод заканчивается, если в грамматике нет правил, позволяющих произвести замену какого-либо символа в порожденной цепочке.

Аналогичным образом грамматика  $G_2$  языка  $L_2$  может иметь следующий вид:

Дано:  $S$ ,

$$\begin{aligned} F1: S &\rightarrow aa, \\ F2: S &\rightarrow bb, \\ F3: S &\rightarrow aSa, \\ F4: S &\rightarrow bSb. \end{aligned} \quad (10)$$

Интересным и важным свойством языков  $L_1$  и  $L_2$  является то, что новые конструкции могут вставляться внутрь старых. Поэтому, например, в предложении  $aabbaa$  языка  $L_2$  имеется отношение зависимости между первым и шестым элементами; внутрь вставлено отношение зависимости между вторым и пятым элементами; и, наконец, в самой глубине расположена зависимость между третьим и четвертым элементами. В соответствии с определением грамматик  $G_1$  и  $G_2$  никаких ограничений на число таких вставленных друг в друга зависимостей в порождаемых цепочках не наложено.

Не приходится сомневаться, что естественные языки в принципе допускают «скобочные» вставки такого типа, так что

грамматики естественных языков должны быть в состоянии порождать такие последовательности. Так, английское предложение (The rat (the cat (the dog chased) killed) ate the malt) («Крыса, которую кошка, которую преследовала собака, убила, съела солод»), конечно, является запутанным и маловероятным; тем не менее оно абсолютно правильно с грамматической точки зрения и имеет ясный и однозначный смысл. Чтобы охарактеризовать более полно сложности, с которыми должна встретиться действительная грамматика естественного языка, приведем еще такой пример из английского языка:

Anyone<sub>1</sub> who feels that if<sub>2</sub> so-many<sub>3</sub> more<sub>4</sub> students<sub>5</sub> whom we<sub>6</sub> haven't<sub>6</sub> actually admitted are<sub>5</sub> sitting in on the course than<sub>4</sub> ones we have that<sub>3</sub> the room had (11) to be changed, then<sub>2</sub> probably auditors will have to be excluded, is<sub>1</sub> likely to agree that the curriculum needs revision

(«Каждый, кто понимает, что если настолько больше студентов, которых мы не пускаем, будет сидеть на лекции, чем тех, которых мы пускаем, что нужно сменить аудиторию, то слушателей придется исключить, скорее всего согласится, что необходимо изменить расписание»).

В примере (11) имеются зависимости у всех пар слов с одинаковыми индексами; в результате получается система вставленных зависимостей, как в языке  $L_2$ . Более того, картина усложняется наличием зависимостей, пересекающих указанные, например, между словами students и ones, между haven't... admitted и have на 10 слов дальше (поскольку здесь опущено подразумеваемое слово admitted). В примере (11) вставленные зависимости соответствуют нескольким конструкциям; в частном случае, когда одна и та же конструкция вставляется сама в себя, мы говорим о *самовставлении*.

Разумеется, можно с полной уверенностью утверждать, что предложение (11) никогда не будет практически использовано, разве что в качестве примера; но мы можем с той же уверенностью утверждать, что такие безусловно правильно построенные предложения, как Birds eat («Птицы едят»), Black crows are black («Черные вороны черны»), Black crows are white («Черные вороны белы»), Tuesday follows Monday («За понедельником следует вторник») и т. д., также никогда не встретятся в нормальной речи взрослого человека. Как и многие другие предложения, которые настолько очевидным образом истинны или ложны, настолько сложны, нескладны или дефектны в каком-либо ином отношении, что не имеют смысла с точки зрения повседневной жизни, эти предложения просто не используются. Тем не менее предложение (11) является абсо-

точно правильно построенным и имеет ясный и однозначный смысл; поэтому грамматика английского языка, если она претендует на психологическую адекватность, должна порождать это предложение.

Если ограничиваться правилами вида  $\varphi \rightarrow \psi$ , где  $\varphi$  и  $\psi$  — цепочки, как было предложено выше (см. об этом подробно в след. главе, разд. 3), то грамматика, порождающая язык  $L_3$ , будет заметно сложнее грамматики языка  $L_2$ . Можно, однако, построить для языка  $L_3$  совсем простую грамматику, если допустить более сильные правила. Условимся заменять через  $x$  любую цепочку, состоящую из символов  $a$  и  $b$ , и добавим граничный символ  $\#$ , отмечающий начало и конец предложения. Тогда грамматика языка  $L_3$  может иметь следующий вид:

Дано:  $\# S \#$ ,

$$\begin{aligned} F1 : S &\rightarrow aS, \\ F2 : S &\rightarrow bS, \\ F3 : \# xS \# &\rightarrow \# xx \#. \end{aligned} \quad (12)$$

Правила  $F1$  и  $F2$  порождают произвольные цепочки из букв  $a$  и  $b$ ; правило  $F3$  удваивает эти цепочки. Ясно, что порождаемый язык — это в точности язык  $L_3$  (с граничными символами в конце и в начале предложения).

Важно отметить, однако, что правило  $F3$  принципиально отличается от остальных правил, поскольку оно требует анализа цепочки, к которой применяется; такой анализ выходит за пределы возможностей, допускаемых тем ограниченным видом грамматических правил, который мы рассматривали вначале.

Допущение более разнообразных и более сильных правил, таких, как  $F3$ , приводит иногда к значительному упрощению грамматики; иначе говоря, новая форма правил дает возможность использовать закономерности структуры языка, которые в противном случае просто нельзя было бы сформулировать. Против правил такого рода едва ли можно выдвинуть какие-либо возражения, так что в разд. 5 мы рассмотрим их структуру и общие свойства более подробно. Поскольку грамматика есть теория языка, а простота и общность являются главными достоинствами теории, мы, конечно, должны попытаться сформулировать теорию языковой структуры таким образом, чтобы она допускала правила, выражающие более глубокие обобщения. Тем не менее вопрос о том, можно ли в принципе породить естественные языки с помощью правил ограниченного вида, — такого, как  $F1 - F4$  из примера (10), — представляет определенный интерес. Ответ на этот вопрос (как положительный,

так и отрицательный) выявит некоторые важные структурные характеристики естественных языков.

Система зависимостей в языке  $L_3$  совершенно иная, чем в языке  $L_2$ . Так, в цепочке *baabaa* языка  $L_3$  зависимости не вставлены друг в друга, как в цепочке *aabbaa* из  $L_2$ ; вместо этого четвертый символ зависит от первого, пятый — от второго и шестой — от третьего. Системы зависимостей такого рода также могут быть обнаружены в естественных языках (см. примеры в след. главе, разд. 4.2) и потому должны найти себе место в адекватной теории грамматики. Но в таком случае искусственные языки  $L_2$  и  $L_3$  отражают реальные свойства естественных языков, причем ниже мы увидим, что эти свойства играют принципиальную роль для определения адекватности различных моделей грамматики.

Чтобы кратко пояснить, как эти соображения применяются к естественному языку, рассмотрим следующий маленький фрагмент грамматики английского языка.

Дано:  $\# S \#$ ,

$$\begin{aligned} F_1: S &\rightarrow AB, \\ F_2: A &\rightarrow CD, \\ F_3: B &\rightarrow EA, \\ F_4: C &\rightarrow a, \text{ the, another,} \\ F_5: D &\rightarrow \text{ball, boy, girl,} \\ F_6: E &\rightarrow \text{hit, stuck.} \end{aligned} \quad (13)$$

Правила  $F_4 - F_6$  — это в действительности группы правил, поскольку они указывают несколько возможностей развертывания для каждого из символов  $C$ ,  $D$  и  $E$ . (Обычно мы интерпретируем символ  $A$  как группу существительного,  $B$  — как группу глагола и т. д.; однако эти привычные имена несущественны для формальной структуры грамматики, хотя они играют важную роль в общей лингвистической теории.) Во всякой настоящей грамматике должны быть, кроме того, фонологические правила, которые кодируют терминальные цепочки фонетическими символами. Для простоты мы отложим рассмотрение фонологического компонента грамматики до разд. 6.

С помощью грамматики (13) можно породить такие терминальные цепочки, как  $\# \text{the boy hit the girl} \#$  («Мальчик ударил девочку»). В случае этой простой грамматики все терминальные цепочки имеют одну и ту же структуру составляющих, что может быть выражено с помощью скобок и приписанных им грамматических обозначений:

$$\# (s (A (C \text{ the})_C (D \text{ boy})_D) A (B (E \text{ hit})_E (A (C \text{ the})_C (D \text{ girl})_D) A)_B) S \#$$

или, эквивалентным образом, с помощью размеченного дерева, как показано на рис. 3. Мы предполагаем, что такое дерево должно быть составной частью структурного описания для всякого предложения; в дальнейшем мы будем называть это дерево *маркером структуры составляющих (С-маркером)*. Грамматика, для того чтобы быть адекватной, должна обеспечивать

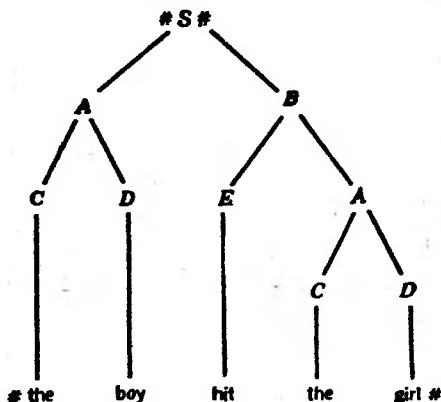


Рис. 3. Графическое представление (С-маркер) вывода грамматически правильного предложения.

С-маркером любое предложение. Каждый С-маркер содержит, в виде меток при конечных узлах, перечень словарных элементов (т. е. слов), из которых составлена данная фраза. Два С-маркера тождественны, если они имеют одинаковую структуру ветвей и одинаковые метки при соответствующих узлах. Заметим, что дерево С-маркера в отличие от кодовых деревьев в разд. 2 характеризуется определенным упорядочением ветвей слева направо — в соответствии с порядком элементов в цепочке.

Роль С-маркеров можно продемонстрировать на предложениях (1а) и (1б). Грамматика, которая только порождает заданные цепочки слов, не в состоянии охарактеризовать грамматические различия между этими двумя предложениями.

В лингвистике слова (морфемы) или последовательности, которые функционируют в качестве элементов другой конструкции, называются обычно *составляющими*. В предложении, С-маркер которого представлен на рис. 3, *girl, the girl* и *hit the girl* — это составляющие, а сочетание *hit the* — нет. Каждая составляющая возводится к некоторому узлу дерева. Если этот узел помечен символом А, то говорят, что составляющая принадлежит к типу А. Те составляющие, из которых конструкция

непосредственно образована, являются *непосредственными составляющими* этой конструкции. Например, на рис. 3 the boy и hit the girl являются непосредственными составляющими предложения; hit и the girl — это непосредственные составляющие глагольной группы *B* и т. д. Очевидно, грамматика не может считаться удовлетворительной, если она не дает порождаемым предложениям структурного описания — хотя бы в виде разложения на непосредственные составляющие.

Таким образом, грамматика должна обеспечить *C*-маркером каждое из бесконечного числа порождаемых предложений, причем каждый *C*-маркер представляется графически в виде дерева с размеченными узлами и размеченными ребрами (узлам соответствуют грамматические типы составляющих, а ребра, отходящие от одного узла, различаются своим порядковым номером). *Цепью* дерева мы будем называть последовательность ребер, каждое из которых связано с предшествующим. Так, одной из цепей дерева на рис. 3 является последовательность  $((S-B), (B-E), (E-hit))$ , другой —  $((A-D), (D-girl))$  и т. д. Поскольку дерево соответствует правильной системе скобок, его ветви не пересекаются (чтобы сделать это неформальное замечание точным в том очевидном смысле, который здесь имеется в виду, нужно было бы различать разные вхождения в дерево одного и того же символа). Символы, которыми размечаются узлы дерева, — это те самые символы, которые используются в грамматических правилах. Поскольку число правил конечно, а во всяком представляющем интерес случае число *C*-маркеров бесконечно, то должны найтись такие символы грамматического словаря, которые повторяются в *C*-маркерах сколь угодно много раз. Более того, должны найтись даже такие цепи, которые содержат некоторый символ более чем *n* раз для любого фиксированного *n*. Пусть дано множество *C*-маркеров; тогда символ словаря называется *рекурсивным элементом*, если для любого *n* найдется *C*-маркер, содержащий цепь, в которой этот символ встречается в качестве названия узла более чем *n* раз.

Мы выделим три вида рекурсивных элементов, особенно важных для последующего изложения. Рекурсивный элемент *A* называется *самостоятельным*, если он встречается в конфигурации типа той, которая изображена на рис. 4, а; *леворекурсивным*, если он встречается в конфигурации типа 4, б; *праворекурсивным*, если он встречается в конфигурации типа 4, в.

Таким образом, если *A* есть леворекурсивный элемент, то подчиненное ему дерево содержит *A* только в самой левой цепи (дерево, подчиненное *A*, — это дерево, которое может быть выведено из *A*); если *A* — праворекурсивный элемент, то подчиненное дерево содержит *A* только в крайней правой цепи; если

4 — самовставленный элемент, то подчиненное ему дерево содержит  $A$  на некоторой внутренней цепи. Эти определения нетрудно превратить в точные, однако мы будем вводить вполне точные определения только в особо интересных случаях.

Анализ рекурсивных порождающих процессов (в частности, порождающих грамматик) основан на исследованиях в области



Рис. 4. Типы рекурсивных элементов.

оснований математики и теории доказательства. Обзор работ по этому вопросу см. в книге Дэвиса (Davis, 1958); более короткое и менее формальное введение см., например, у Роджерса (Rogers, 1959) или у Трахтенброта (Трахтенброт, 1962). Более общее рассмотрение рекурсивных грамматик и их свойств см. в след. главе.

Мы уже говорили в этом разделе о некоторых основных свойствах грамматик и приводили примеры порождающих устройств, которые можно рассматривать как грамматики. Ниже мы попытаемся сформулировать характеристики грамматик более подробно. В разд. 4 мы опишем (более аккуратно) грамматики, в которых правила имеют вид (5), где  $n=1$  и всякая структура является цепочкой, как в примерах (9), (10) и (13). (В следующей главе изучаются некоторые свойства таких систем.) В разд. 5 мы обратимся к более мощному классу грамматик, родственному грамматике языка  $L_n$ , предложенной в примере (12), т. е. к грамматикам, не удовлетворяющим тому условию, что в каждом правиле типа (5) структуры являются цепочками и  $n=1$ . Наконец, в разд. 6 указываются, кратко, некоторые свойства фонологического компонента грамматики, который превращает цепочки, являющиеся выходом рекурсивного порождающего процесса, в последовательности звуков.

Мы описали порождающую грамматику  $G$  как устройство, перечисляющее определенное подмножество  $L$  множества  $\Sigma$  цепочек фиксированного словаря  $V$  и приписывающее структурные описания элементам множества  $L(G)$ , которое мы называли

языком, порожденным грамматикой  $G$ . С точки зрения того применения описываемых моделей, которое мы имеем в виду, будет более реалистично считать, что устройство  $G$  приписывает структурные описания всем цепочкам из  $\Sigma$ , причем структурное описание цепочки  $x$  показывает, в частности, в каких отношениях  $x$  отклоняется от правильной построенности, определяемой грамматикой  $G$ . Вместо того чтобы разделять  $\Sigma$  на два подмножества —  $L(G)$  (правильно построенные предложения) и  $\bar{L}(G)$  (грамматически неправильные предложения), — можно считать, что  $G$  выделяет в  $\Sigma$  класс  $L_1$  абсолютно правильно построенных предложений и частично упорядочивает все остальные цепочки из  $\Sigma$  с точки зрения *степени грамматичности*. В этом случае мы говорим, что языком, порождаемым грамматикой  $G$ , является  $L_1$ ; но одновременно мы допускаем, что цепочки не из  $L_1$  часто тоже могут быть поняты носителями языка, и даже поняты однозначно, притом на основе структурных описаний, приписанных им грамматикой. Цепочка из  $\Sigma \setminus L_1$  может оказаться понятой в результате наложения на нее интерпретации, которая основана на аналогиях и сходствах с предложениями из  $L_1$ . Степень простоты и единообразия такой интерпретации предложения можно считать степенью его грамматичности. Отклонение от грамматических закономерностей — это обычный литературный и псевдолитературный прием, и такое отклонение отнюдь не должно приводить к непонятности; действительно, как много раз отмечалось, отступления от грамматических норм могут обуславливать сжатость и глубину выражения.

Определение понятия степени грамматичности на основе интерпретации не абсолютно грамматически правильных предложений и построение грамматик, приписывающих предложениям степени грамматичности, не обязательно равные нулю или единице, — все это интересные и важные задачи. Некоторые аспекты этих задач рассматриваются в работах: Chomsky, 1955, 1961b; Ziff 1960a, 1960b, 1961; Katz, 1963. Мы вернемся к этой проблеме в гл. 13. В этой и в следующей главах мы ограничимся грамматиками, которые разделяют все цепочки ровно на две категории правильные и неправильные, и не устанавливают иерархию степеней грамматичности.

#### 4. НЕКОТОРЫЙ ПРОСТОЙ КЛАСС ПОРОЖДАЮЩИХ ГРАММАТИК

Мы рассмотрим здесь простой класс порождающих грамматик, которые могут быть названы *грамматиками непосредственных составляющих*, и введем некоторые важные обозначения,

используемые в этой и двух следующих главах. В этом разделе мы рассмотрим грамматику  $G$ :

$$G = [V, \hat{\cdot}, \rightarrow, V_T, S, \#],$$

которая представляет собой систему с операцией соединения, удовлетворяющую следующим условиям.

1.  $V$  есть конечный набор символов, называемый *словарем*. Цепочки символов этого словаря получаются с помощью операции соединения  $\hat{\cdot}$ ; это ассоциативная и некоммутативная бинарная операция над цепочками в словаре  $V$ . Если не возникает неоднозначности, символ  $\hat{\cdot}$  может опускаться.

2.  $V_T \subset V$ . Мы называем  $V_T$  *терминальным* словарем. Дополнение к  $V_T$  относительно  $V$  называется *нетерминальным*, или *вспомогательным*, словарем и обозначается  $V_N$ .

3. Отношение  $\rightarrow$  есть конечное, двуместное, иррефлексивное и асимметричное отношение, имеющее место для конечного числа пар цепочек в алфавите  $V$  и интерпретируемое как «подставляется вместо». Пары  $\langle \phi, \psi \rangle$ , такие, что  $\phi \rightarrow \psi$ , называются (*грамматическими*) *правилами* грамматики  $G$ .

4. Если  $A \in V$ , то  $A \in V_N$  тогда и только тогда, когда существуют цепочки  $\phi, \psi, \omega$ , такие, что  $\phi A \psi \rightarrow \phi \omega \psi$ . При этом  $\# \in V_T$ ;  $S \in V_N$ ;  $e \in V_T$ , где  $\#$  есть *границный* символ;  $S$  — *начальный* символ, который можно читать как «предложение»;  $e$  есть *единичный* символ, обладающий тем свойством, что для всякой цепочки  $\phi$ ,  $e\phi = \phi = \phi e$ .

Введем еще следующие дополнительные обозначения.

5. Последовательность цепочек  $D = (\phi_1, \dots, \phi_n)$  ( $n \geq 1$ ) есть  *$\phi$ -вывод* цепочки  $\psi$ , тогда и только тогда, когда выполняются следующие условия:

а)  $\phi = \phi_1$ ;  $\psi = \phi_n$ ;

б) для всякого  $i < n$  существуют цепочки  $\psi_1, \psi_2, \chi, \omega$ , такие, что  $\chi \rightarrow \omega$ ,  $\phi_i = \psi_1 \chi \psi_2$  и  $\phi_{i+1} = \psi_1 \omega \psi_2$ .

Очевидно, множество выводов полностью определяется заданием отношения  $\rightarrow$ , т. е. конечным множеством грамматических правил. Если существует  $\phi$ -вывод цепочки  $\psi$ , мы говорим, что  $\phi$  *подчиняет*  $\psi$ , и пишем  $\phi \Rightarrow \psi$ ; отношение  $\Rightarrow$  является, таким образом, рефлексивным и транзитивным.

6.  $\phi$ -вывод  $D$  цепочки  $\psi$  называется *терминальным*, если  $\psi$  есть цепочка в  $V_T$  и  $D$  не является собственным началом некоторого другого вывода (заметим, что эти условия независимы).

7.  $\psi$  есть *терминальная цепочка* грамматики  $G$ , если для цепочки  $\# S \#$  имеется терминальный  $\# S \#$ -вывод; иначе говоря, терминальная цепочка — это (с точностью до граничного

символа) последняя строка терминального вывода, начинающегося с цепочки  $\# S \#$ .

8. Терминальным языком, порождаемым грамматикой  $G$ , называется множество терминальных цепочек грамматики  $G$ .

9. Две грамматики  $G$  и  $G^*$  (слабо) эквивалентны, если они порождают один и тот же терминальный язык. Понятие сильной эквивалентности рассматривается ниже.

Естественно, чтобы граничный символ  $\#$  удовлетворял тому условию, что если  $(\varphi_1, \dots, \varphi_n)$  есть  $\# S \#$ -вывод, то для всякого  $i$ , такого, что  $\varphi_1 = \# \psi_i \#$ ,  $\psi_i$  не содержит  $\#$ . Чтобы обеспечить выполнение этого условия, мы наложим на правила грамматики (т. е. на отношение  $\rightarrow$ ) следующее дополнительное условие.

10. Если  $\alpha_1 \dots \alpha_m \rightarrow \beta_1 \dots \beta_n$  есть правило грамматики (где  $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_n \in V - \{e\}$ ), то

- а)  $\beta_i \neq \#$  для  $1 < i < n$ ;
- б)  $\beta_1 = \#$ , тогда и только тогда, когда  $\alpha_1 = \#$ ;
- в)  $\beta_n = \#$ , тогда и только тогда, когда  $\alpha_m = \#$ .

Все эти условия пока еще ничего не говорят о том, как может грамматика обеспечить  $S$ -маркер для каждой цепочки. Это требование мы будем иметь в виду при рассмотрении добавочных ограничений на грамматику.

См. работу Чулика (Culik, 1962), где содержится критика более ранних определений грамматики, изложенных в работе Хомского (Chomsky, 1959).

Возвращаясь теперь к понятию рекурсивного элемента из разд. 3, мы можем заметить, что для всякого  $A \in V_N$  имеют место следующие утверждения.

- 1. Если имеются непустые цепочки  $\varphi, \psi$ , такие, что  $A \Rightarrow \varphi A \psi$ , то  $A$  есть самовставленный элемент.
- 2. Если имеется непустой элемент  $\varphi$ , такой, что  $A \Rightarrow A \varphi$ , то  $A$  есть леворекурсивный элемент.
- 3. Если имеется непустой элемент  $\varphi$ , такой, что  $A \Rightarrow \varphi A$ ,<sup>(14)</sup> то  $A$  есть праворекурсивный элемент.
- 4. Если  $A$  есть нерекурсивный элемент, то не существует таких цепочек  $\varphi, \psi$ , что  $A \Rightarrow \varphi A \psi$ .

Обратные утверждения не обязательно верны для всех грамматик, которые мы рассматривали до сих пор, однако они будут верны для того случая, который мы будем рассматривать ниже. Предположим, что грамматика  $G$  содержит правила  $S \rightarrow VAC, VA \rightarrow BVAC$ , но не содержит правила  $A \rightarrow \chi$  ни для какого  $\chi$ . Тогда не существует цепочек  $\varphi, \psi$ , таких, что

$A \Rightarrow \varphi A \psi$ , хотя  $A$  является рекурсивным (а именно самовста-  
вленным).

При рассмотрении систем такого рода необходимо отличать терминальные цепочки от нетерминальных, а также отдельные атомарные) элементы от цепочек, образованных их соединением. Чтобы последовательно соблюдать это различие, удобно придерживаться следующих обозначений:

Тип (элемента или цепочки)	Элементы	Цепочки
Нетерминальные	$A, B, C, \dots$	$X, Y, Z, \dots$
Терминальные	$a, b, c, \dots$	$x, y, z, \dots$
Произвольные	$\alpha, \beta, \gamma, \dots$	$\varphi, \chi, \psi, \dots$

В предыдущем изложении мы уже фактически придерживались этого соглашения; оно будет использоваться далее до конца работы без дополнительных комментариев.

Теперь определим условия, при которых процесс порождения каждой терминальной цепочки будет однозначно определять ее  $S$ -маркер. Для грамматик, рассматривавшихся до сих пор, никакого естественного способа обеспечить выполнение этого требования не существует. Так, если имеется грамматика со следующими правилами:

$$S \rightarrow AB; \quad AB \rightarrow cde, \quad (15)$$

то нет никакого способа определить, является ли в цепочке  $cde$  сегмент  $cd$  составляющей типа  $A$  (подчиненной в  $S$ -маркере символу  $A$ ) и является ли сегмент  $de$  составляющей типа  $B$  (подчиненной в  $S$ -маркере элементу  $B$ ). Наиболее естественный путь достижения желаемого результата это потребовать, чтобы в каждом правиле грамматики подстановка представляла собой замену (на символ или цепочку) ровно одного символа исходной последовательности. Таким образом, грамматики могут иметь правила либо вида (16а), либо вида (16б):

$$A \rightarrow \omega \quad (16a)$$

$$\varphi A \psi \rightarrow \varphi \omega \psi \text{ (или, иначе, } A \rightarrow \omega \text{ в контексте } \varphi - \psi). \quad (16b)$$

Правила вида (16б) называются *контекстно-связанными*, а правила вида (16а) — *контекстно-свободными*. Грамматика, содержащая только правила вида (16а), называется *контекстно-свободной*. (Заметим, между прочим, что связанными контекстом или свободными являются только правила, а не элементы в терминальной цепочке.) В любом случае, если  $\varphi$  есть некоторая

строка вывода, а  $\psi$  — строка, которая непосредственно следует за  $\phi$ , то имеется единственный<sup>1)</sup> набор цепочек  $\phi_1, \phi_2, \alpha, \omega$ , таких, что  $\phi = \phi_1 \alpha \phi_2$  и  $\psi = \phi_1 \omega \phi_2$ ; тогда мы говорим, что  $\omega$  есть цепочка типа  $\alpha$  (т. е. подчинена узлу, помеченному символом  $\alpha$  в дереве, представляющем соответствующий  $S$ -маркер). В случае контекстно-свободных грамматик верны не только утверждения 1—4 из (14), но и соответствующие им обратные. Тем самым мы получаем определение различных видов рекурсивности через отношение  $\Rightarrow$ . В дальнейшем мы будем рассматривать рекурсивные элементы только в контекстно-свободных грамматиках.

Назначение правил типа (16б) состоит в том, чтобы иметь возможность выразить контекстные ограничения, наложенные на выбор элементов. Так, в число предложений типа «подлежащее + сказуемое + дополнение» входят, в частности, такие предложения, как *The fact that the case was dismissed doesn't surprise me* («Тот факт, что дело было прекращено, не удивил меня»), *Congress enacted a new law* («Конгресс ввел новый закон»), *The men consider John a dictator* («Люди считают Джона диктатором»), *John felt remorse* («Джон почувствовал угрызения совести»); однако не являются грамматически правильными предложения, которые получены из них переменной мест подлежащего и дополнения: *I don't surprise the fact that the case was dismissed* («Я не удивляю тот факт, что дело было прекращено»), *A new law enacted Congress* («Новый закон ввел Конгресс»); *John considers the men a dictator* («Джон считает людей диктатором»); *Remorse felt John* («Угрызения совести почувствовали Джона»). Носители английского языка признают, что первые четыре предложения являются абсолютно естественными, тогда как вторые четыре, если они вообще могут быть понятны, требуют специальной интерпретации, которая может быть определена по аналогии с правильно построенными предложениями. В смысле определений из разд. 3, эти предложения имеют степень грамматичности более низкую, чем предыдущие, если вообще не равную нулю. Грамматика, которая не делает различия между этими двумя видами пред-

<sup>1)</sup> В действительности, для того чтобы гарантировать единственность этих цепочек, множество правил должно удовлетворять некоторым дополнительным условиям; в частности, должна быть исключена возможность появления такой последовательности строчек вывода:  $AB, ABC$  и т. д. В будущем допустим, без дальнейших комментариев, что такие условия выполнены. На самом деле они всегда могут быть выполнены без дальнейшего сужения класса порождаемых языков, хотя, конечно, они как-то влияют на вид класса порождаемых  $S$ -маркеров. В случае грамматики с контекстно-связанными правилами эти дополнительные условия вовсе не являются лишними, как мы увидим в следующей главе, разд. 3.

дожений, безусловно не является полноценной, а чтобы естественным и экономным образом выразить это различие, достаточно ввести контекстно-связанные правила, которые накладывают ограничения на выбор подлежащего, сказуемого и дополнения.

В идеальном случае теория грамматики должна содержать определение множества допустимых грамматик, множества допустимых предложений и множества допустимых структурных описаний. Кроме того, она должна включать общий метод приписывания каждому предложению одного или нескольких структурных описаний в соответствии с грамматикой. Иначе говоря, грамматика должна быть выражена в столь явной форме, чтобы по ней можно было однозначно определить, что она утверждает относительно каждого из порожденных предложений (ср. Chomsky, 1961a). Чтобы построить теорию *грамматик непосредственных составляющих*, мы фиксируем словари  $V_N$  и  $V_T$  как заданные не пересекающиеся друг с другом конечные множества символов, удовлетворяющие описанным выше условиям. В этом случае грамматика есть просто конечное отношение, заданное в множестве цепочек словаря  $V = V_N \cup V_T$  и удовлетворяющее этим условиям. Заметим, что проблема нахождения  $V_T$  в случае естественного языка сводится, по существу, к проблеме построения универсальной фонетической теории (включая, в частности, разработку универсального фонетического алфавита и правил, определяющих общие законы распределения фонетических сегментов). Более подробно об этом см. разд. 6 и приведенные там ссылки. Кроме того, мы должны наложить ограничение на длину морфемы (чтобы словарь  $V_T$  был конечным) и установить множество «грамматических морфем» (т. е. морфем, выражающих грамматические признаки — такие, как время, вид и число). Эта последняя проблема, так же как проблема получения обоснованной интерпретации для элементов множества  $V_N$ , является классической проблемой «универсальной грамматики». Эта проблема состоит в получении независимого от языка определения множества грамматических категорий, которые могут использоваться или не использоваться в данном языке. Такое определение должно, без сомнения, опираться как на формальный анализ грамматической структуры, так и на соображения чисто семантического характера. Эти вопросы не имели популярности в последние несколько десятилетий, хотя нет никаких оснований считать их недостойными серьезного рассмотрения. Напротив, они остаются важными и первостепенными для общей лингвистики.

К изучению различных типов грамматик непосредственных составляющих мы вернемся в следующей главе.

### 5. ТРАНСФОРМАЦИОННЫЕ ГРАММАТИКИ

Как уже говорилось в разд. 3, правила грамматики всегда могут быть сведены к виду

$$\varphi_1, \dots, \varphi_n \rightarrow \psi, \quad (17)$$

где  $\varphi_1, \dots, \varphi_n$  — структуры произвольного вида, а символ  $\rightarrow$  означает, что если  $\varphi_1, \dots, \varphi_n$  были порождены в процессе вывода, то на определенном шаге может быть порождено  $\psi$ . В разд. 4 мы рассматривали простой вид грамматик, называемый *грамматиками непосредственных составляющих*; правила этих грамматик представляют собой очень специальную разновидность правила (17), а именно случай, когда  $n=1$  и каждая из структур  $\varphi_1$  и  $\psi$  является просто цепочкой символов [кроме того, мы потребовали, чтобы грамматика удовлетворяла дополнительно условию (16)]. В этом разделе мы рассмотрим класс грамматик, состоящих из двух синтаксических компонентов: из *НС-компонента* (грамматики непосредственных составляющих, удовлетворяющей условиям из разд. 4 и некоторым другим) и из *Т-компонента* (собственно трансформационной грамматики), содержащего правила вида (17), где  $n$  может быть больше 1 и структуры  $\varphi_1, \dots, \varphi_n$  и  $\psi$  представляют собой не цепочки, а *С-маркеры*. Такие грамматики мы будем называть *трансформационными*.

Целесообразность обобщения понятия грамматики, которое приводит к трансформационным грамматикам, обусловлена тем психологически очевидным фактом, что некоторые пары предложений представляются грамматически связанными друг с другом. Почему, например, предложение *John threw the ball* («Джон бросил мяч») воспринимается как связанное с предложением *The ball was thrown by John* («Мяч брошен Джоном»)? Эта связь не исчерпывается тождеством значения, поскольку аналогичного рода близость ощущается между предложениями *John threw the ball* («Джон бросил мяч») и *Who threw the ball?* («Кто бросил мяч?») или *Did John throw the ball?* («Бросил ли Джон мяч?»). Эту связь нельзя свести к какому-либо простому формальному отношению (в лингвистической терминологии — к отношению «совместной встречаемости») между наборами слов, способных заполнять определенные места в соответствующих парах предложений; действительно, такие предложения, как *The old man met the young woman* («Старый человек встретил молодую женщину») и *The old woman met the young man* («Старая женщина встретила молодого человека»), соотносятся, конечно, не таким же образом, как соответствующие друг другу предложения в активе и пассиве, хотя отличие этого соотноше-

ния от соотношения в паре актив — пассив не может быть выражено в чисто дистрибутивных терминах (см. об этом Chomsky, 1962б). В грамматике непосредственных составляющих все такие пары предложений будут порождаться более или менее независимо и, следовательно, не будут находиться в каких-либо явных отношениях друг к другу.

### 5.1. Некоторые недостатки грамматик непосредственных составляющих

Лингвистическая теория отказывается от грамматик непосредственных составляющих в пользу трансформационных грамматик, поскольку в грамматике непосредственных составляющих невозможны многие естественные обобщения и упрощения правил образования предложений. (Анализ различных соображений по этому вопросу выходит за рамки настоящей работы; см. об этом Chomsky, 1955, гл. 7—9; 1957, гл. 6, 7; 1962; Lees, 1957, 1960; Postal, 1963; ср. также след. главу, разд. 4.2.) Однако некоторые из причин формальной несостоятельности грамматик непосредственных составляющих показать довольно легко; при этом необходимо будет учесть ряд общих соображений, которые обычно упускаются из виду при оценке адекватности грамматики.

Грамматика должна породить язык, представляющий собой бесконечное множество предложений. Кроме того, она должна сопоставить каждому из этих предложений структурное описание; иными словами, она должна породить бесконечное множество структурных описаний, каждое из которых полностью определяет данное предложение (хотя обратное может быть неверно). Таким образом, при оценке порождающей способности грамматик мы можем различать два вида эквивалентности. Две грамматики называются *слабо эквивалентными*, если они порождают один и тот же язык, и *сильно эквивалентными*, если они порождают одно и то же множество структурных описаний. В этой главе, так же как в следующей, мы рассматриваем в основном слабую эквивалентность, поскольку она легче поддается описанию и потому более подробно изучена. Тем не менее сильная эквивалентность является в конечном счете гораздо более интересным понятием. Не представляют интереса такие грамматики, которые правильно порождают предложения естественного языка, но не могут приписать им правильного структурного описания.

Вопрос о том, могут ли естественные языки, рассматриваемые как множества предложений, порождаться грамматикой непосредственных составляющих того или иного типа, — это

вопрос интересный и важный. Тем не менее нет сомнений в том, что множество структурных описаний, ассоциированных с предложениями, скажем, английского языка, в принципе не может быть порождено грамматикой непосредственных составляющих, сколь бы сложны ни были ее правила. Дело в том, что грамматика непосредственных составляющих по необходимости накладывает на предложение слишком подробный анализ, что вытекает из самого способа определения *S*-маркера предложения на основе его грамматики. Суть проблемы может быть показана на таких предложениях, как

Why has John always been such an easy man to please? (18)

(«Почему Джон всегда был человеком, которому так легко угодить?»). Все это вместе — предложение; последние несколько слов образуют именную группу; каждое из слов принадлежит к определенной синтаксической категории. Однако для какого-либо описания структуры составляющих этого предложения, помимо приведенного, у нас уже нет никаких оснований. Чтобы дать такое структурное описание, грамматика непосредственных составляющих должна содержать правила вроде следующих:

$$\begin{aligned} S &\rightarrow \text{why has } NP \text{ always been } NP, \\ NP &\rightarrow \text{John}, \\ NP &\rightarrow \text{such an easy } N \text{ to } V_{\text{transitive}}, \\ N &\rightarrow \text{man}, \\ V_{\text{transitive}} &\rightarrow \text{please}. \end{aligned} \quad (19)$$

Ясно, что такой набор правил является абсурдным и оставляет все структурные закономерности этого предложения невыраженными. Более того, ассоциированный *S*-маркер этого предложения имеет тот недостаток, что он не отражает грамматической связи между словами *man* и *please* — такой же, как в предложении *it pleases the man* («Это доставляет человеку удовольствие»), так что глагольно-объектное отношение в этой паре слов остается невыраженным (ср. разд. 2.2 след. главы). Вообще грамматика непосредственных составляющих в значительной мере лишается своего смысла, если в нее входят правила, при которых некоторые конструкции состоят из шести непосредственных составляющих, как в примере (19).

Указанное серьезное затруднение перерастает в принципиальную неадекватность, когда мы обращаемся к собственно сочинительным конструкциям, таким, как в следующем примере:

The man was old, tired, tall, ..., and friendly. (20)

(«Человек был пожилым, усталым, высоким, ..., и добродушным»). Чтобы породить цепочку такого рода, грамматика непосредственных составляющих либо должна приписать ей некоторую произвольную структуру (и тогда использовать, например, право-рекурсивное правило), т. е. породить неправильное структурное описание, либо она должна содержать бесконечное число правил. Ясно, что в случае чистого сочинения в соответствии с самим значением термина никакой внутренней структуры в последовательности сочиненных элементов существовать не может.

Можно попытаться решить этот вопрос путем расширения понятия грамматики непосредственных составляющих и допущения правил типа

Предикат  $\rightarrow$  Прил.<sup>n</sup> and Прил. ( $n \geq 1$ ). (21)

При попытке сформулировать это правило таким образом, чтобы оно удовлетворяло требованиям формальной адекватности, возникнут многочисленные трудности. Кроме того, такое решение второго вопроса совершенно ничего не дает для решения тех проблем, которые возникают в примере (19). Вообще для каждого частного случая трудностей, возникающих в грамматиках непосредственных составляющих, можно придумать *ad hoc* какой-нибудь прием, который позволит эту частную трудность обойти. Гораздо предпочтительнее, однако, был бы такой пересмотр всей схемы понятий, который дал бы возможность избавляться от этих трудностей всегда одним и тем же путем. При этом грамматика непосредственных составляющих в ее прежнем виде должна описывать только тот класс предложений, для которого она адекватна и который первоначально обусловил само ее появление. Насколько нам известно, только теория трансформационных грамматик дает какие-то основания надеяться, что эта цель когда-либо будет достигнута.

Множество правил трансформационной грамматики удовлетворяет следующим условиям. Прежде всего имеется грамматика непосредственных составляющих, т. е. набор правил вида  $\Phi A \Phi \rightarrow \Phi \omega \Phi$ , где  $A$  есть единичный символ, а  $\omega$  непусто. Этот компонент грамматики порождает конечное число  $S$ -терминальных цепочек, с каждой из которых мы можем ассоциировать, как и раньше, размеченное дерево,  $S$ -маркер, представляющее ее структуру составляющих. Затем мы добавляем к грамматике множество операций, называемых *грамматическими трансформациями*, каждая из которых отображает набор  $n$   $S$ -маркеров ( $n \geq 1$ ) в новый  $S$ -маркер. Можно допустить, чтобы грамматика обладала рекурсивными свойствами за счет одних только трансформаций. Среди этих трансформаций некоторые

являются обязательными: они должны применяться к любому выводу (кроме того, некоторые трансформации обязательны по отношению к некоторым другим, т. е. должны применяться, если применены некоторые другие трансформации). Цепочка, полученная в результате применения всех обязательных и некоторых факультативных трансформаций, называется *T-терминальной цепочкой*. *T-терминальную цепочку* можно рассматривать, в сущности, как последовательность морфем. *T-терминальная цепочка*, полученная в результате применения только обязательных трансформаций, называется *ядерной цепочкой*. Если язык вообще содержит ядерные цепочки, то они будут соответствовать только наиболее простым предложениям. Идея использования грамматических трансформаций для преодоления недостатков порождающих грамматик других типов возникла из исследований З. Харриса, который предлагал использовать эти операции для «нормализации» текста (Harris, 1952a, 1952b). Описание, которое приводится ниже, основано преимущественно на работе Хомского (Chomsky, 1955); изложение такое же, как в работе Chomsky, 1961a. Ср. несколько иное развитие исходной идеи трансформации в статье Харриса (Harris, 1957).

Как уже говорилось, причины включения в грамматику правил трансформации весьма просты. Имеются предложения (простые утвердительные предложения в активной форме без сложных глагольных времен и конструкций), которые можно очень просто породить грамматикой непосредственных составляющих (если говорить точно, то это утверждение верно лишь для терминальных цепочек этих предложений). Имеются другие предложения (в пассивной или вопросительной форме, с прерванными конструкциями и сложными оборотами, получившимися в результате вставки трансформированных простых предложений, и т. д.), которые не могут быть естественным и экономным образом порождены грамматикой непосредственных составляющих, но которые тем не менее находятся в определенных закономерных отношениях с предложениями более простой структуры. Эти отношения и выражаются трансформациями. Трансформации, применяемые для порождения более сложных предложений (и их структурных описаний) из уже порожденных более простых, дают объяснение тем аспектам грамматической структуры предложения, которые не могут быть описаны грамматикой непосредственных составляющих.

Задача состоит, следовательно, в том, чтобы сформулировать общее и абстрактное понятие грамматической трансформации, которое воплотило бы в себя все эти идеи и упростило бы выражение формальных соотношений между предложениями, играющих важную роль в языке.

## 5.2. Уточнение определения грамматической трансформации

Трансформация не может быть операцией, определенной на множестве терминальных цепочек и не зависящей от структуры этих цепочек. Действительно, в этом случае трансформация пассива могла бы с одинаковым успехом применяться в предложениях (22) и (23):

The man saw the boy  $\rightarrow$  The boy was seen by the man, (22)

The man saw the boy leave  $\rightarrow$  { The boy was seen by the man leave  
The boy leave was seen by the man (23)

«Человек увидел мальчика»  $\rightarrow$  «Мальчик был увиден человеком»; «Человек увидел, как мальчик уходил»  $\rightarrow$  «\*Мальчик был увиден человеком уходить», «\*Мальчик уходить был увиден человеком»).

Чтобы применить трансформацию к некоторой цепочке, необходимо знать структуру составляющих этой цепочки. Так, при трансформации, которая превращает утвердительное предложение в вопросительное, определенный элемент глагольной группы может выноситься в начало предложения. Применение этой трансформации к предложению The man who was here was old («Человек, который был здесь, был старым») даст предложение Was the man who was here old? Для образования вопроса на первое место выносится второе was. Если вынести на первое место первое was, то получится грамматически неправильное сочетание Was the man who here was old? Поэтому мы должны каким-то образом указать, что трансформация вопроса должна применяться ко второму was, которое является элементом глагольной группы главного предложения, а не к первому. Иными словами, мы должны знать структуру составляющих исходного предложения.

Цели трансформационного анализа не были бы достигнуты также и в том случае, если бы мы стали рассматривать трансформации как правила подстановки более высокого уровня, которые применяются к цепочкам из незаконченных выводов. Так, можно показать, что трансформацию пассива нельзя рассматривать просто как правило подстановки вида

$$NP_1 + Auxiliary + Verb + NP_2 \rightarrow NP_2 + \\ + Auxiliary + be + Verb + en + by + NP_1. \quad (24)$$

Это правило относится к типу, допускаемому грамматикой непосредственных составляющих, как она была определена в разд. 4, и не удовлетворяет только условию (16) (т. е. условию

о том, чтобы подстановка производилась на место единичного символа), которое обеспечивает возможность построения соответствующего *С*-маркера. Достаточным аргументом против введения трансформации пассива с помощью правила (24). является то, что трансформации, определенные таким образом, не смогут служить способом упрощения грамматики в случае появления контекстных ограничений на выбор лексических элементов (ср. примеры в конце разд. 4). В пассивных предложениях, соответствующих этим примерам, имеют место те же самые ограничения на выбор лексических элементов, хотя здесь эти элементы встречаются в ином грамматическом оформлении. Однако если трансформация пассива будет применяться как одно из правил подстановки на той стадии вывода, когда выбор субъекта, действия и объекта еще не производился, то для соответствующего выбора агента, действия и субъекта в пассиве потребуются совершенно отдельные набор контекстно-связанных правил. Одно из достоинств трансформационной грамматики состоит в том, что она дает возможность избежать этого ненужного дублирования правил выбора и сопутствующей ему потери общности. Это преимущество будет, однако, утрачено, если трансформации будут применяться раньше, чем определен лексический состав предложения.

Представляется очевидным поэтому, что трансформационное правило должно применяться к полностью построенному *С*-маркеру, а поскольку трансформационные правила должны снова применяться к трансформам, то результатом трансформации должен быть тоже *С*-маркер — производный *С*-маркер терминальной цепочки, полученной в результате трансформации. Таким образом, трансформация превращает *С*-маркеры в новые *С*-маркеры.

Теперь мы можем определить понятие грамматической трансформации следующим образом. Пусть *Q* есть *С*-маркер терминальной цепочки *t*, причем *t* может разделяться на последовательные сегменты  $t_1, \dots, t_n$  таким образом, что всякий сегмент  $t_i$  можно возвести в *Q* к некоторому узлу, помеченному символом  $A_i$ . В таком случае мы говорим, что цепочка *t* может быть проанализирована относительно *Q* как  $(t_1, \dots, t_n; A_1, \dots, A_n)$ .

В простейшем случае трансформация *T* отчасти определяется последовательностью символов  $(A_1, \dots, A_n)$ , которая задает область ее применения по следующему правилу.

*Цепочка t с С-маркером Q входит в область применения трансформации T, если t может быть проанализирована относительно Q как  $(t_1, \dots, t_n; A_1, \dots, A_n)$ .*

В этом случае  $(t_1, \dots, t_n)$  есть собственный анализ *t* относительно  $\langle Q, T \rangle$ , а  $(A_1, \dots, A_n)$  — структурный индекс *T*.

Чтобы закончить определение трансформации  $T$ , необходимо описать те изменения, которые она производит в элементах собственного анализа цепочек, к которым применима. Например, трансформация может заключаться в перестановке или выбрасывании некоторых элементов, в замене их на некоторые другие, в постановке на заданное место некоторых постоянных цепочек и т. д. Предположим, что трансформации  $T$  сопоставлена лежащая в ее основе элементарная трансформация  $T_{el}$ , такая, что  $T_{el}(i; t_1, \dots, t_n) = \sigma_i$ , где  $(t_1, \dots, t_n)$  — собственный анализ  $t$  относительно  $\langle Q, T \rangle$ . В этом случае цепочка, получающаяся в результате применения трансформации  $T$  к цепочке  $t$  с  $C$ -маркером  $Q$ , такова:

$$T[t, Q] = \sigma_1 \dots \sigma_n$$

Ясно, что было бы нежелательно называть грамматической трансформацией любое соответствие того вида, который только что был описан. Нежелательно было бы, например, включать в грамматику трансформации, которые соотносят друг с другом такие пары, как John saw the boy («Джон увидел мальчика»)  $\rightarrow$  I'll leave tomorrow («Я уезжаю завтра»); John saw the man («Джон увидел человека»)  $\rightarrow$  Why don't you try again? («Почему бы тебе не попытаться еще раз?»); John saw the girl («Джон увидел девочку»)  $\rightarrow$  China is industrializing rapidly («Китай индустриализируется быстрыми темпами»). В грамматике должны допускаться только такие трансформации, которые выражают реальные структурные соотношения между предложениями: соотношения типа актив — пассив, утверждение — вопрос, утверждение — номинализация и т. д. Произвольного сопоставления предложений с помощью трансформаций можно избежать, если наложить дополнительное, вполне естественное ограничение на элементарную трансформацию. Это ограничение таково<sup>1)</sup>.

Если  $T_{el}$  — элементарная трансформация, то для любых чисел  $i$  и  $n$  и для любых цепочек  $x_1, \dots, x_n, y_1, \dots, y_n$  должно выполняться следующее условие:  $T_{el}(i; x_1, \dots, x_n)$  должно быть получено из  $T_{el}(i; y_1, \dots, y_n)$  в результате замены  $y_j$  на  $x_j$  для всех  $j \leq n$ .

Иными словами, результат элементарной трансформации должен быть независим от того, к какой цепочке она применяется. Тем самым исключается возможность применения трансформаций к частным цепочкам слов (или морфем). При этом условии, однако, никакая единая элементарная трансформация

<sup>1)</sup> При более точной формулировке необходимо было бы различить разные вхождения одной и той же цепочки (см. Chomsky, 1985).

не может одновременно заменить John will try («Джон попытается») на Will John try? («Попытается ли Джон?») и John tried («Джон пытался») на Did John try? («Пытался ли Джон?»), хотя это, безусловно, одна и та же простая трансформация вопроса. Элементарная трансформация, которая нужна в данном случае, должна превратить  $x_1x_2x_3$  в  $x_2x_1x_3$ . Это значит, что трансформация  $T_{el}$  должна быть определена для произвольных  $x_1, x_2, x_3$  следующим образом:

$$T_{el}(1; x_1, x_2, x_3) = x_2,$$

$$T_{el}(2; x_1, x_2, x_3) = x_1,$$

$$T_{el}(3; x_1, x_2, x_3) = x_3.$$

Чтобы такое правило дало в результате фразу Did John try?, необходимо применять трансформацию вопроса не к John tried, а, скорее, к гипотетической цепочке John + past + try (терминальной цепочке, аналогичной по структуре предложению John will try), которая лежит в основе предложения John tried. Вообще мы не можем требовать, чтобы терминальные цепочки находились в каком-то простом соотношении с реальными предложениями. Здесь приходится вводить обязательные преобразования (трансформационные и фонологические), которые определяют физический облик элементов и одновременно представляют их, выбрасывают некоторые элементы и т. д.

Чтобы понятие трансформации было эмпирически адекватным, его необходимо обобщить. Прежде всего нужно допустить применение трансформации к паре *S*-маркеров. (Трансформации, подобные рассмотренным выше, которые применялись к одному *S*-маркеру, мы будем теперь называть *сингулярными* трансформациями.) Так, терминальная цепочка, лежащая в основе предложения His owning property surprised me («То, что у него есть собственность, удивило меня»), получается из полученных ранее цепочек it surprised me («Это удивило меня») и he owns property («У него есть собственность»). Чтобы иметь возможность получать такие предложения, можно, в простейших случаях, допускать, чтобы в предшествующей грамматике непосредственных составляющих выводы предложений начинались не только с  $\# S \#$ , но и с  $\# S \# \# S \# \dots \# S \#$ . В этом случае мы введем структурный индекс ( $\#$ , *NP*, *V*, *NP*,  $\#$ ,  $\#$ , *NP*, *V*, *NP*,  $\#$ ), из которого с помощью трансформации можно получить предложение His owning property surprised me. Более подробное и аккуратное изложение этого вопроса см. (Chomsky, 1955).

Необходимо, далее, обобщить формулировку области определения трансформации и собственного анализа трансформи-

руемой цепочки. Во-первых, нет необходимости требовать, чтобы элементами структурного индекса были отдельные символы. Во-вторых, можно допустить, чтобы область определения трансформации задавалась конечным множеством структурных индексов. Более того, можно задавать область определения трансформации просто некоторым структурным условием, основанным на предикате «быть анализируемым», введенном выше. С помощью этого понятия можно определить разложение терминальных цепочек на элементы собственного анализа, после чего элементы структурного индекса могут оставаться неопределенными. С помощью этого и некоторых других обобщений можно построить эксплицитное и точное обоснование трансформационной грамматики.

### 5.3. Структура составляющих в трансформированных цепочках

Грамматическая трансформация задается структурным условием, выраженным с помощью предиката «быть анализируемым» и элементарной трансформации. Как было, однако, замечено, трансформации должны порождать не только предложения, но и их *С*-маркеры. Поэтому нужно показать, каким образом терминальной цепочке, полученной в результате трансформации, приписывается структура составляющих. Лучше всего было бы получать производные *С*-маркеры с помощью набора правил, которые составляют часть общей лингвистической теории, а не с помощью дополнительного пункта в определении каждой трансформации. Точная формулировка этих правил потребовала бы анализа исходных понятий, который выходит далеко за рамки настоящего неформального изложения (см. в этой связи Chomsky, 1955; Matthews, 1962; Postal, 1962). Тем не менее некоторые подходы к общему решению этой проблемы кажутся достаточно ясными. Прежде всего все трансформации можно разделить на небольшое число классов в соответствии с элементарной трансформацией, на которой они основаны. Для каждого класса можно задать общее правило, приписывающее трансформе производный *С*-маркер, вид которого зависит фиксированным образом от *С*-маркеров исходных терминальных цепочек. Приведем несколько примеров для иллюстрации тех общих принципов, которых необходимо при этом придерживаться.

Основным рекурсивным устройством в грамматике являются обобщенные трансформации, которые порождают новую цепочку из пары исходных. (Ясно, что число сингулярных

трансформаций, которые могут применяться к одному и тому же предложению, ограниченно.) Большинство обобщенных трансформаций основано на элементарных трансформациях, которые подставляют трансформированный вариант второго члена пары исходных терминальных цепочек вместо некоторого элемента собственного анализа первого члена пары. (В терминах,

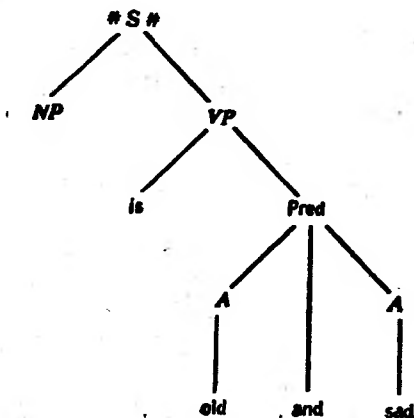


Рис. 5. С-маркер, возникающий при интерпретации предложений с союзом *and* как полученных в результате трансформации присоединения.

предложенных Лизом (Lees, 1960), это соответственно *цепочка-составляющая* и *цепочка-матрица*.) В этом случае для определения производной структуры составляющих трансформации достаточно, по-видимому, единого общего правила. Пусть трансформация заменяет символ  $\alpha$  из  $\sigma_1$  (цепочки-матрицы) на  $\sigma_2$  (цепочку-составляющую). Тогда С-маркер результирующей цепочки есть С-маркер  $\sigma_1$ , где  $\alpha$  заменен на С-маркер  $\sigma_2$ .

Все остальные обобщенные трансформации — это трансформации присоединения, которые состоят в том, что берется некоторый элемент собственного анализа  $\alpha$  вместе с элементом структурного индекса  $\beta$ , который опосредствованно подчиняет  $\alpha$  (и всеми промежуточными звеньями С-маркера, которые подчиняются элементу  $\beta$  и подчиняют  $\alpha$ ), и присоединяются (быть может, с добавлением некоторой константной цепочки) к некоторому другому элементу собственного анализа. Таким способом, например, фраза *John is old and sad* («Джон стар и печален») с С-маркером, представленным на рис. 5, получается из фраз *John is old*, *John is sad* («Джон стар», «Джон печален»)

с помощью трансформации со структурным индексом ( $NP, is, A, \# \#$ ,  $NP, is, A$ ).

Сингулярные трансформации часто являются простыми перестановками элементов собственного анализа. В частности, такая трансформация превращает рис. 6, а в рис. 6, б. Общий принцип построения производного  $S$ -маркера состоит в этом случае

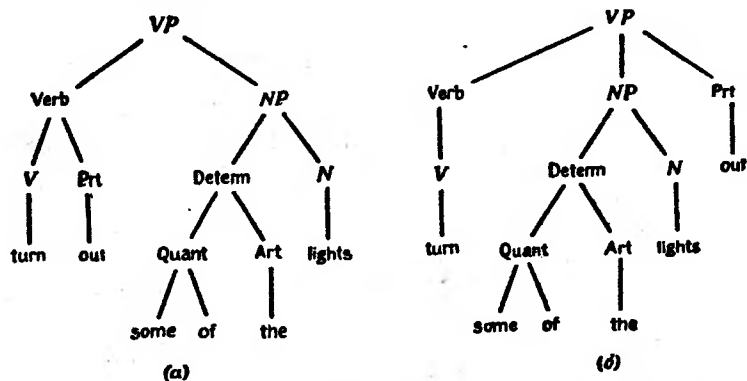


Рис. 6. Сингулярная трансформация, переводящая (а) в (б). — это перестановка; результатом ее применения является сокращение «количества структуры», приписываемой предложению.

в том, что в  $S$ -маркере исходной цепочки производятся минимальные изменения, совместимые с требованием о том, чтобы получившийся  $S$ -маркер также был деревом. Трансформация, которая дает предложение *Turn some of the lights out* («Выключи некоторые лампы»), основана на элементарной трансформации, меняющей местами второй и третий элементы трех-элементного собственного анализа; эта трансформация имеет структурный индекс ( $V, Prt, NP$ ) (разумеется, здесь приводится частный случай некоторого более общего правила).

Рис. 6 иллюстрирует весьма характерное влияние, оказываемое на структуру предложения перестановками: перестановки имеют тенденцию сокращать «количество структуры» у терминальной цепочки, к которой они применяются. Так, если на рис. 6, а представлена полностью бинарная структура, то на рис. 6, б одним бинарным разложением меньше, а вместо этого появляется одно тернарное разложение; символ  $Prt$  больше не подчиняется символу  $Verb$ . Хотя бинарные составляющие весьма характерны для простых структурных описаний, порождаемых грамматикой непосредственных составляющих, они довольно редки в  $S$ -маркерах действительных предложений. Транс-

формационный подход к синтаксическому описанию позволяет, таким образом, сохранить зерно истины, содержащееся в обычных теориях непосредственных составляющих, с их предпочтением бинарных составляющих, и в то же время освобождает от необходимости произвольным образом приписывать предложению излишнюю структурную организацию. Далее, последовательным применением трансформаций присоединения и перестановки, которые были описаны выше, можно породить разновидности *С*-маркеров, которые в принципе не могут быть порождены грамматиками непосредственных составляющих [в частности, *С*-маркеры сочиненных конструкций типа (20)]. Аналогично можно показать, каким образом трансформационный подход может помочь в решении проблем, упомянутых в связи с примером (18) и ему подобных (ср. разд. 2.2 след. главы), хотя практические трудности, которые встречаются при разработке конкретных анализов, нельзя недооценивать.

Одни сингулярные трансформации состоят в добавлении константных цепочек в заданное место собственного анализа; другие — в устранении некоторых из элементов собственного анализа. Трансформации первого типа можно рассматривать как аналогичные трансформациям присоединения. При трансформациях второго типа устраняются узлы, которые не подчиняются никакой терминальной цепочки, а в остальном *С*-маркер остается неизменным. По-видимому, эллиптические трансформации можно существенным образом ограничить. Наложение ограничений на эллиптические трансформации играет основополагающую роль при определении класса языков, который может быть порожден трансформационными грамматиками.

Итак, повторим вкратце, что трансформационная грамматика состоит из конечной последовательности контекстно-связанных правил подстановки  $\phi \rightarrow \psi$  и конечного числа трансформаций описанного типа, а также ограничений относительно порядка применения трансформаций. Структура, полученная в результате применения некоторой трансформации, в общем случае может быть подвергнута новым трансформациям, так что путем повторного применения трансформаций можно породить бесконечное число *С*-маркеров самых различных типов. На каждом этапе порождения фразы *С*-маркер представляет собой размеченное дерево, ассоциированное с порождаемой цепочкой. Полное структурное описание предложения должно состоять не только из его собственного *С*-маркера, но и из *С*-маркеров, лежащих в его основе *Т*-терминальных цепочек и их трансформационной истории. О роли такого структурного описания при определении процесса понимания предложения см. разд. 2.2 след. главы.

## 6. ЗВУКОВАЯ СТРУКТУРА ЯЗЫКА

Мы рассматриваем грамматику как состоящую из двух основных компонентов — *синтаксического компонента*, который мы уже описали, и *фонологического компонента*, к которому мы теперь переходим.

### 6.1. Роль фонологического компонента в грамматике

Синтаксический компонент грамматики содержит правила подстановки и трансформационные правила, форма и организация которых были описаны в разд. 4 и 5; в результате работы этих правил порождаются терминальные цепочки и их структурные описания. Структурное описание терминальной цепочки содержит, в частности, производный *S*-маркер, который определяет для цепочки систему размеченных скобок; в данном разделе только эта часть структурного описания и будет приниматься во внимание. Таким образом, мы обращаемся к рассмотрению следующих структур, которые мы будем считать (опуская многие детали) выходами синтаксического компонента грамматики:

$$s [_{NP} [N \# \text{Ted} \#]_N]_{NP} [_{VP} [V \# \text{see}]_V \text{past} \#]_S$$

$$[_{NP} [_{\text{Det}} \# \text{the dem pl} \#]_{\text{Det}} [_N \# \text{book}]_N \text{pl} \#]_{NP}]_{VP}]_S \quad (25)$$

(символ  $\#$  использован как показатель границ между словами). Терминальная цепочка из примера (25) является представлением предложения *Ted saw those books* («Тед видел эти книги»), которое может быть на фонетическом уровне представлено так:

$$t^1 e \cdot d + s^2 \omega + \delta^3 \dot{a} w z + b^1 u k s \quad (26)$$

(цифры обозначают силу ударения). Опять-таки здесь опущены уточнения, детали, обсуждение прочих возможностей многих фонетических признаков, которых мы в этом кратком очерке не будем касаться.

Выражение типа (26) является достаточно непосредственным представлением предложения. Можно считать, что это представление основано на универсальной фонетической системе, которая состоит из фонетического алфавита и набора общих фонетических правил. Символы фонетического алфавита определяются на основе физических (т. е. акустических и артикуляторных) понятий; общие законы универсальной фонетической системы — это законы, касающиеся способов комбинирования физических единиц, представленных этими символами, в есте-

ственном языке. Универсальная фонетическая система, так же как абстрактное определение порождающей грамматики, предложенное в разд. 4 и 5, является скорее частью общей лингвистической теории, чем специфическим разделом грамматики данного языка. Так же как и другие разделы общей теории языковой структуры, универсальная фонетическая система представляет собой гипотезу о языковых универсалиях; она может также рассматриваться как гипотеза о врожденных способностях переработки информации и образования понятий, которые пускаются в ход ребенком при обучении языку.

Роль фонологического компонента грамматики состоит в том, что он соотносит представления типа (25) и типа (26); иначе говоря, фонологический компонент включает в себя правила, определяющие фонетический облик высказывания, исходя из его морфемного состава и общей синтаксической структуры. Фонологический компонент в отличие от синтаксического не играет никакой роли в построении новых высказываний, а только сопоставляет высказываниям их фонетическую форму. Хотя по этой причине изучение фонологического компонента грамматики не относится к математическим моделям языка в узком смысле слова, процессы приписывания высказываниям их фонетического облика представляют независимый интерес. Мы укажем кратко некоторые существенные свойства этих процессов. Наше изложение основывается на работах Халле (Halle, 1959a, 1959b) и Хомского (Chomsky, 1959, 1962a).

## 6.2. Звуки и фонемы

Фонологический компонент можно представлять как устройство, которое получает на вход терминальную цепочку с различными скобками и должно перекодировать ее в фонетическую запись. Фонетическая запись есть последовательность символов фонетического алфавита, одни из которых [например, первые три символа в примере (26)] соответствуют физически определенным признакам, а другие [например, символ + в примере (26)] — символам перехода. Назовем символы первого рода *фонетическими сегментами*, символы второго рода — *фонетическими стыками*. Рассмотрим более тщательно свойства фонетических сегментов.

Каждый символ универсального фонетического алфавита является сокращенным обозначением определенного набора физических признаков; например, символ [p<sup>h</sup>] представляет губной, придыхательный, глухой, смычный звук. Эти символы не имеют независимого существования сами по себе; они служат просто сокращенными обозначениями. Следовательно, предста-

вление (26), и вообще всякую фонетическую запись, правильнее всего рассматривать как *фонетическую матрицу*: строки матрицы — это физические свойства, которые в данной лингвистической теории считаются элементарными, а столбцы — это последовательные сегменты высказывания (стыки мы пока оставляем в стороне). В этой матрице элемент (*i*, *j*) показывает, обладает ли (или в какой степени обладает) *j*-й сегмент высказывания *i*-м свойством. Фонетические сегменты соответствуют, таким образом, столбцам матрицы. Если принять за универсальную фонетическую систему систему акустических признаков, предложенную Jakobsonом, то в примере (26) символ [ə] может рассматриваться как сокращение для строчки (гласный, несогласный, низкий (grave), компактный, неокругленный, звонкий, слабый, с третьестепенным ударением). Матрицы с такими элементами и составляют выход фонологического компонента.

Что же является входом фонологического компонента? Терминальная цепочка в примере (25) состоит из *лексических морфем*, таких, как *Ted*, *book*, *грамматических морфем*, таких, как *past*, *plural*, и некоторых *показателей стыка* типа *‡*. Показатели стыка вводятся правилами синтаксического компонента, чтобы указать позиции, в которых морфологическая и синтаксическая структуры отражаются на фонетике. Для наших целей они могут рассматриваться как грамматические морфемы. Каждая грамматическая морфема, как правило, представляет собой единый терминальный символ, нерасчлененный на сегменты. Напротив, лексические морфемы обычно представлены цепочками символов, которые мы будем называть *фонемными сегментами* или просто *фонемами*<sup>1)</sup>. Тогда, если не считать размеченных скобок, входом для фонологического компонента является цепочка, состоящая из фонем и условно обозначенных грамматических морфем. Фонетическая запись в примере (25) является во всех существенных отношениях точной, если не считать того, что лексические морфемы даны в традиционной орфографии, а не в фонемной транскрипции. Так, *Ted*, *see*, *the*, *book* нужно заменить соответственно на /tɛd/, /si/, /ði/, /buk/. В примере (26) дано так мало фонетических деталей, что фонетические и фонемные сегменты почти не отличаются друг от друга.

Вскоре мы вернемся к вопросу о соотношении между фонемным и фонетическими сегментами. Пока мы только отметим,

<sup>1)</sup> Для большей точности следовало бы считать, что фонемы — это сегменты, которые появляются только на той стадии вывода, когда все грамматические морфемы уже элиминированы с помощью фонологических правил.

что не накладывается никаких требований о том, что они должны быть тесно связаны друг с другом.

Прежде чем переходить к более подробному рассмотрению фонемных сегментов, мы хотим предупредить читателя, что понимание терминов фонема, фонемная запись и т. д. в лингвистической литературе очень неустойчиво. При этом расхождения здесь не являются чисто терминологическими; они отражают глубоко идущие расхождения взглядов на сущность звуковой структуры — расхождения, не разрешенные и до настоящего времени. Здесь, разумеется, не место разбирать в подробностях эти расхождения и приводить аргументы в защиту той или иной точки зрения. (Подробно об этом см. у Халле (Halle, 19596), у Хомского (Chomsky, 1962); см. также книгу Halle, Chomsky, «Sound Pattern of English», 1968.) Данная работа основана на представлении о звуковой структуре, близком к тому, которое было у основателей современной фонологии, но сильно отклоняющемся от того, которое получило наиболее широкое распространение, особенно в Соединенных Штатах, за последние двадцать лет и часто называется *необлумфильдианским*. В частности, термин «фонема» используется здесь в смысле Сэпира (Sapir, 1933), а понятие универсальной фонетической системы имеет корни в таких классических работах, как работа Суита (Sweet, 1877) и де Соссюра (de Saussure, 1916); речь идет о Приложении к Введению «Курса общей лингвистики», которое датируется вообще 1897 г. То, что мы, следуя Сэпиру, называем *фонемной записью*, в настоящее время чаще называют *морфонемной записью*. Обычно предполагается, что существует уровень представления высказывания, промежуточный между фонетическим и морфонемным, и этот промежуточный уровень называется *фонемным*. Однако имеются, по-видимому, достаточные основания для того, чтобы отвергнуть гипотезу о существовании такого промежуточного уровня, отбросив одновременно многие допущения относительно звуковой структуры, тесно переплетающиеся с этой гипотезой в современных изложениях лингвистической теории.

Мы попытаемся обнаружить общие правила, которые применяются к большим классам элементов — таким, как согласные, смычные, звонкие сегменты и т. д., — а не к отдельным сегментам. При этом мы попытаемся заменить большое число отдельных наблюдений простыми обобщениями. Поскольку правила будут применяться к классам элементов, то сами элементы должны быть определены через их принадлежность к определенным классам. Таким образом, каждая фонема принадлежит к нескольким пересекающимся категориям, на основе которых фонему можно представить просто множеством категорий, к

которым она принадлежит; иными словами, каждую лексическую единицу можно представить *классификационной матрицей*, в которой столбцы соответствуют фонемам, а ряды — категориям и элемент  $(i, j)$  показывает, принадлежит ли фонема  $j$  к категории  $i$ . Каждая фонема может быть представлена теперь в виде последовательности категорий, которые мы называем *различительными признаками*, используя одно из имеющих хождение значений этого термина. Так же как фонетические символы, фонемы не имеют независимого существования сами по себе. Очень важным и ни в какой мере не очевидным является то обстоятельство, что различительные признаки классификационной таблицы фонем определяют категории, точно соответствующие тем категориям, которые определяются рядами фонетических матриц. Это обстоятельство было замечено Сэпиром (Sapir, 1925) и изучено в последние годы Якобсоном, Фантом и Халле (Jakobson, Fant, Halle, 1952) и Якобсоном и Халле (Jakobson, Halle, 1956); такой подход к звуковой системе имеет свои корни в классической лингвистике, процветавшей в Индии еще два тысячелетия тому назад.

### 6.3. Условие инвариантности и условие линейности

Входами для фонологического компонента грамматики являются, таким образом, матрицы различительных признаков, представляющих лексические элементы; выходами являются фонетические матрицы (и стыки). Каким должно быть отношение между категориальной матрицей различительных признаков, которая поступает на вход фонологических правил, и соответствующей фонетической матрицей, которая возникает в результате их применения? Каково, например, отношение между исходной матрицей, сокращенно записываемой  $/ted/$  (где каждый символ  $|t|$ ,  $|e|$ ,  $|d|$  заменяет собой столбец, содержащий в данном ряду плюс, если данный символ принадлежит к категории, соответствующей данному ряду, минус, если данный символ не принадлежит к этой категории, и ноль, если данный символ не определен с точки зрения принадлежности к этой категории), и результирующей матрицей, сокращенно записываемой как  $[t^h e \cdot d]$  (где каждый из символов  $[t^h]$ ,  $[e \cdot]$ ,  $[d]$  заменяет столбец, элементами которого являются фонетические свойства)?

Самое сильное требование, которое здесь может быть наложено, состоит в том, что исходная классификационная матрица должна быть собственной частью результирующей фонетической матрицы, отличаясь от этой последней отсутствием некоторых избыточных элементов. Тогда фонологические правила

будут просто заполнять пустые места классификационной матрицы, дополняя ее до соответствующей фонетической. Это сильное условие включается в определение фонологической теории, в частности, Якобсоном, а также, в неявном виде, Блохом<sup>1)</sup>. Если это требование соблюдено, то фонемная запись удовлетворяет условию, которое мы называем *условием линейности*.

Под *условием линейности* мы понимаем требование о том, что каждая фонема должна быть связана с определенным звуковым отрезком высказывания, причем если фонема *A* находится слева от фонемы *B* в фонемной записи, то отрезок высказывания, сопоставленный фонеме *A*, должен предшествовать отрезку, сопоставленному фонеме *B* в фонетической записи. (Мы ограничиваемся здесь фонемами, которые называют *сегментными*, поскольку так называемые суперсегментные фонемы мы рассматриваем как признаки сегментных фонем.)

*Условие инвариантности* состоит в том, что каждой фонеме *A* должно быть сопоставлено некоторое определяющее множество  $\Sigma(A)$  физических фонетических свойств, таких, что каждый вариант (аллофон) фонемы *A* обладает свойствами из  $\Sigma(A)$  и никакой фонетической сегмент, не являющийся вариантом (аллофоном) фонемы *A*, не имеет всех свойств из  $\Sigma(A)$ .

Если бы звуковые системы удовлетворяли условиям линейности и инвариантности, то задача создания машин, способных распознавать фонемы в нормальной человеческой речи, сильно упростилась бы. Соблюдение этих условий дало бы также возможность построить модель восприятия, основанную на сегментации и классификации, что явилось бы аргументом в пользу той точки зрения, что методы анализа, необходимые в лингвистике, могут быть сведены к сегментации и классификации. Однако выполнение этих требований зависит от реальных фактов, а не от принятия какого-либо решения, и, по нашему мнению, есть много оснований сомневаться, что в естественных языках эти условия выполняются. Поэтому мы не будем считать, что для каждой фонемы имеется набор фонетических свойств, которые однозначно характеризуют все ее варианты, и что эти наборы действительно располагаются во временной последовательности в точном соответствии с линейным порядком фонем.

<sup>1)</sup> Эти авторы не стали бы считать то, что они называют *фонемными записями*, входами в фонологический компонент нашей грамматики. Однако, как было сказано выше, мы не видим оснований для допущения о существовании промежуточной записи того типа, который называется *этими* и другими фонологами *фонемными*.

У нас нет возможности останавливаться на этом вопросе подробно, но одного примера будет достаточно, чтобы продемонстрировать те трудности, которые заставляют отказаться от условий линейности и инвариантности. Английские слова *write* («писать») и *ride* («ехать верхом») в любой разумной фонемной записи должны быть представлены соответственно как /*raɪt*/ и /*raɪd*/; иным словами, их фонемный состав различается звонкостью конечной согласной. В фонетическом отношении они различаются также свойствами гласных. Есть, например, такой диалект, где *write* должно быть представлено фонетически как [raɪt], а *ride* — как [ra·yd] с характерным автоматическим удлинением гласной перед звонкой согласной. Чтобы вывести фонетическую запись из фонемной, мы применяем следующее правило:

*Гласные перед звонкими согласными удлиняются.* (27)

Это правило имеет достаточно общий вид и вполне может быть включено в рассматриваемую нами систему. Рассмотрим теперь слова *writer* («писатель») и *rider* («всадник») в том же диалекте. В соответствии с синтаксическим компонентом грамматики *writer* есть *write*+деятель, а *rider* есть *ride*+деятель, причем лексические единицы имеют в точности тот вид, в котором они приведены; а именно фонемная запись для *writer* и *rider* будет соответственно /*raɪt+r*/ и /*raɪd+r*/ . Имеется, однако, общее правило, что фонемы /*t*/ и /*d*/ в некоторых контекстах сливаются в альвеолярный дрожащий звук [D], в частности после главного ударения, как в словах *writer* и *rider*. Таким образом, грамматика для этого диалекта может содержать следующее фонетическое правило:

[*t*, *d*] → D после главного ударения. (28)

Применяя к фонемным записям /*raɪt+r*/ и /*raɪd+r*/ сначала правило (27), а потом правило (28), мы получим сначала [raɪt+r], [ra·yd+r] в соответствии с правилом (27), а затем [raɪDr], [ra·yDr] в соответствии с правилом (28); эти выражения и будут фонетическими представлениями слов *writer*, *rider*. Теперь мы видим, что фонемные представления этих слов различаются только четвертым сегментом (звонкая/глухая согласная), тогда как фонетические представления различаются только вторым сегментом (более долгая/более краткая гласная). Следовательно, мы не можем утверждать, что последовательность фонем  $A_1 \dots A_m$  связана с последовательностью сегментов  $a_1 \dots a_m$  таким образом, что  $a_i$  содержит, помимо избыточных

признаков, множество признаков, которые однозначно определяют  $A_i$ . Этот типичный пример показывает несостоятельность условий линейности и инвариантности в применении к фонологическим системам. Отсюда следует, что фонемы не могут быть получены из фонетических записей в результате простых операций сегментации и классификации по выделенным признакам; во всяком случае, обычно описываемые операции являются заведомо непригодными.

Заметим, между прочим, что мы нигде не говорили о том, что фонетические признаки, образующие универсальную систему, должны быть определены в абсолютных терминах. Так, один из универсальных признаков может быть признаком «передний/задний» или «краткий/долгий». Если фонетический сегмент  $A$  отличается от фонетического сегмента  $B$  только тем, что  $A$  имеет признак «краткий», а  $B$  — «долгий», то это означает, что в любом контексте  $X—Y$  более долгий элемент отождествляется с  $B$ , а более краткий — с  $A$ . При этом может получиться так, что  $A$  в одном контексте окажется в действительности длиннее, чем  $B$  в другом контексте. Многие лингвисты выдвигали требование, что фонетические признаки должны определяться в абсолютных терминах. Вместо признака «краткий/долгий» они требовали установления (с какой-то степенью приближения) абсолютной длины каждого сегмента. Если добавить это требование к требованию инвариантности, то исключается даже частичное пересечение фонем, т. е. приписывание звука  $a$  к фонеме  $B$  в одном контексте и к фонеме  $C$  — в другом. Такова, по-видимому, точка зрения Блоха (Bloch, 1948, 1950). Это ограничение слишком сильное, и оно опровергается не только примером, который был приведен здесь, но и большим числом других (см. Bloch, 1940). Действительно, в работах по акустической фонетике (Liberman, Delattre, Cooper, 1952; Schatz, 1954) показано, что если требовать соблюдения этого условия, то при артикуляторном и акустическом подходе к признакам (как предлагается в работе Bloch, 1950) нельзя произвести даже анализ смычных /p, t, k/, поскольку они частично пересекаются. Такое следствие можно считать доказательством через доведение до абсурда.

Требования относительной и абсолютной инвариантности позволяют построить определенные модели восприятия речи; однако трудность (или невозможность) соблюдения этих требований показывает, что эти модели неверны. Возникает необходимость новых моделей, которые мы еще будем рассматривать ниже.

Теперь же мы вернемся к нашей основной теме.

#### 6.4. Некоторые фонологические правила

Как было сказано, входами для фонологического компонента грамматики являются терминальные цепочки, состоящие из лексических и грамматических морфем и стыков, с указанием структуры составляющих. Выходом фонологического компонента является фонетическая матрица, в столбцах которой записаны последовательные сегменты высказывания, а в строках — фонетические признаки. Разумеется, желательно, чтобы правила фонологического компонента были по возможности общими и немногочисленными. В частности, предпочтительны такие правила, которые применяются к крупным и естественно выделенным классам элементов и основаны на простых и кратких характеристиках данного контекста. Предпочтительны, далее, такие правила, в которых один и те же классы элементов используются несколько раз. Чтобы проверить выполнение этих и других требований, можно определить сложность фонологического компонента через число признаков, упомянутых в правилах; при этом форма правил должна быть задана таким образом, чтобы она допускала правильные обобщения [см. Halle, 1961)]. Тогда мы можем отдать предпочтение более простым (более общим) грамматикам перед более сложными и использующими большее число признаков (т. е. частных случаев).

Задача фонологического анализа состоит в том, чтобы сопоставить каждому высказыванию фонемное представление, состоящее из матриц, в которых столбцами являются фонемы, а строками — различительные (классификационные) признаки, а также вывести самый простой набор правил (простота должна быть при этом точно определенным понятием), которые определяют фонетические матрицы, соответствующие данным фонемным представлениям. На фонемное представление не накладывается требований линейности и инвариантности. Интересное и важное наблюдение состоит в том, что эти требования в сущности, как правило, выполняются и имеется лишь один важный класс исключений.

Чтобы определить фонетическую форму высказывания, фонологические правила должны использовать не только фонемное представление, но и некоторую другую информацию; в частности, они должны использовать информацию о структуре составляющих. Отсюда следует, что вообще невозможно (идет ли речь о лингвисте или о ребенке, обучающемся языку) построить правильное фонемное представление высказывания без существенного использования синтаксической информации. Аналогично, естественно ожидать, что и при восприятии речи

установление фонемного представления высказывания опирается на синтаксические ключевые показатели. Иначе говоря, фонетическое отождествление высказывания основано в какой-то мере на частичном его понимании, и этот вывод вовсе не является парадоксальным.

Фонологический компонент состоит: 1) из последовательности правил подстановки, включающей, в частности, набор правил, определяющих структуру морфемы; 2) из последовательности трансформационных правил и 3) из последовательности правил подстановки, которые можно назвать *фонетическими правилами*. Эти группы правил применяются к терминальной цепочке в указанном порядке.

Правила, определяющие структуру морфемы, позволяют упростить матрицы, определяющие индивидуальные лексические морфемы, за счет использования общих свойств всего множества матриц. В английском языке, например, если ни один из трех начальных сегментов лексической единицы не является гласной, то первый из них должен быть согласной /s/, второй — смычным согласным, а третий — плавным, или глайдом. Таким образом, в матрицах представляющих такие морфемы, как *string* или *square*, эту информацию можно уже не задавать. Аналогичным образом, глайд, который является последним элементом начального сочетания согласных, не требует дальнейшего определения, поскольку он полностью определяется последующей гласной: при всех согласных, кроме /s/, перед /u/ должен стоять /y/, а во всех остальных случаях — /w/. Так, имеются сочетания *cure* /kyūr/ (лечить) и *queer* /kwīr/ (странный), но не /kwūr/ или /kyūr/. Имеется и много других правил такого рода. Они позволяют сократить число признаков, упоминаемых в грамматике, поскольку одно и то же правило, определяющее структуру морфемы, может применяться ко многим матрицам, что способствует увеличению простоты описания (в том смысле, который был указан выше). Заметим, что правила, определяющие структуру морфемы, позволяют провести различие между обоснованным и необоснованным выделением допустимых и недопустимых бессмысленных слогов.

Трансформационные фонетические правила определяют фонетические процессы, обусловленные структурой составляющих (вспомним то основное свойство трансформационных правил, как они были определены выше, что они применяются только к цепочке с заданной структурой составляющих). В английском языке имеется такое переплетение правил, определяющих ударение и редукцию гласных, что фонетическая система имеет несколько степеней ударения и сложное распределение редуци-

рованных и нередуцированных гласных [Хомский, Халле и Лукков (Chomsky, Halle, Lukoff, 1956)]; Халле и Холмский (Halle, Chomsky, 1968). Эти правила существенным образом используют структуру составляющих как на морфологическом, так и на синтаксическом уровне; следовательно, они должны рассматриваться скорее как трансформационные правила, чем как правила подстановки. Эти правила упорядочены и применяются циклами — сначала к минимальным составляющим (т. е. лексическим морфемам), затем к более крупным и т. д., пока не будет достигнута максимальная из сфер действия фонетических процессов. Поразительный факт, по крайней мере для английского языка, состоит в том, что вне и внутри слова действуют, по существу, одни и те же правила. Таким образом, имеется один цикл трансформационных правил, повторное применение которых определяет фонетическую форму как отдельных слов, так и сложных сочетаний. Циклическое упорядочение правил позволяет определять фонетическую структуру более сложных сочетаний, как морфологического, так и синтаксического характера, через фонетическую структуру исходных элементов.

Основными правилами трансформационного цикла в английском языке являются правила приписывания ударения и правила редукиции гласных. Место основного ударения определяется типом составляющей и конечным аффиксом. После того как место главного ударения определено, все остальные ударения автоматически ослабляются. Последовательное применение этого правила к последовательно увеличивающимся составляющим цепочки, не имеющей начальной расстановки ударений, дает в результате систему с несколькими степенями ударения. Гласная редуцируется в [i] в некоторых фонологических условиях, если она ни на одной из предшествующих стадий вывода не получила главного ударения или если последующие циклы ослабили ее первоначальное главное ударение до третьестепенного (или, в некоторых позициях, до вторичного). Правило редукиции гласных применяется в трансформационном цикле один раз, а именно тогда, когда мы достигаем уровня слова.

Подробное описание этих правил в объеме данной главы невозможно, поэтому мы ограничимся лишь несколькими замечаниями о том, как они работают. Рассмотрим, в частности, следующие четыре правила<sup>1)</sup>, которые применяются в том порядке, в каком они написаны:

<sup>1)</sup> Эти правила несколько отличаются от правил, описанных в более подробной и общей грамматике (см. Халле и Хомский, в печати).

Правило *существительного*, которое в существительных (и их основах) ставит ударение на первом слоге при довольно общих условиях. (29а)

Правило *ядерного ударения*, которое превращает последнее главное ударение в доминирующее, ослабляя тем самым все остальные ударения в конструкции. (29б)

Правило *редукции гласных*. (29в)

Правило *приспособления ударения*, которое ослабляет на единицу все неглавные ударения в слове. (29г)

В трансформационном цикле, который описан ниже, мы производим существительные *pérmit* (разрешение), *tórmént* (мучение) от глаголов *permit* (разрешать), *tórmént* (мучить) с помощью правила существительного; при этом главное ударение на втором слоге автоматически заменяется на вторичное. Правило приспособления ударения определяет окончательную последовательность ударений: главное — вторичное. Второй слог не редуцируется в [t], поскольку на той стадии, когда применяется правило редукции гласных, он защищен вторичным ударением.

Таким образом, *pérmit*, *tórmént* имеют следующий вывод:

1.  $[_N [{}_V \text{per} + \text{mit}]_V]_N$   $[_N [{}_V \text{t} \text{órmént}]_V]_N$
2.  $[_N [{}_V \text{per} + \text{mit}]_V]_N$   $[_N [{}_V \text{t} \text{órmént}]_V]_N$
3.  $[_N \text{per} + \text{mit}]_N$   $[_N \text{t} \text{órmént}]_N$
4.  $[_N \text{per} + \text{mit}]_N$   $[_N \text{t} \text{órmént}]_N$
5.  $\text{per} + \text{mit}$   $\text{t} \text{órmént}$
6.  $\text{per} + \text{mit}$   $\text{t} \text{órmént}$
7.  $p^h \text{írmít}$   $t^h \text{órmént}$

Строка 1 фонологическая; строка 7 — фонетическая (детали опущены). Строка 2 получается применением общего правила (которое мы не приводим) для *tórmént* и с помощью правила (29) для *pérmit* (поскольку главное ударение в этом случае отсутствует). Строка 3 заканчивает первый трансформационный цикл, стирая самые внутренние скобки. Строка 4 получается с помощью правила (29а). Строка 5 заканчивает второй трансформационный цикл, стирая самые внутренние скобки.

Строка 6 получается с помощью правила (29б) [при этом правило (29в) оказывается неприменимым в силу наличия вторичного ударения на второй гласной]. Строка 7 получается применением других фонетических правил.

Рассмотрим, напротив, слово *torrent* (поток). В нем, как и в *torrent*, второй гласной фонемой является /e/, ср. *torrential* (поточный). Однако это слово не образовано, как *tórmént*, от глагола *torrént*. Следовательно, вторая гласная не получает главного ударения в первом цикле; поэтому правило (29в) редуцирует ее в [i]. Таким образом, противопоставление редуцированной и нередуцированной гласных в *tórmént* — *tórgént* является результатом различий в синтаксическом анализе. Начальное ударение в *tórgént* опять-таки является результатом применения правила (29а).

То же самое правило, которое образует *pérmit* и *tórmént* из *permit* и *torment*, изменяет вторичное ударение конечного слога в глаголе *advocate* (защищать) ['ædvəkɛɪt] не третьестепенное, в результате чего правило редукции гласных (29а) редуцирует эту гласную в [i]. Таким образом мы получаем противопоставленную друг другу редуцированную и нередуцированную гласные в существительном *advocate* (защитник) ['ædvəkɪt] и в глаголе *advocate*, так же как во всех других словах с суффиксом -ate. В точности те же самые правила определяют различие между редуцированными и нередуцированными гласными в существительном *compliment*, [...mɪnt] (комплимент, привет) и в глаголе *compliment* [...mɛnt] (приветствовать) и в других аналогичных случаях.

Рассмотрим теперь слово *condensation* (конденсация). На одном из первых циклов мы приписываем главное ударение второму слогу глагола *condense* (сгущать). На следующем цикле правила применяются к форме *condensation* как к целому, поскольку это следующая по величине составляющая. Суффикс -ion всегда дает главное ударение на непосредственно предшествующем слоге, в данном случае на -ate. Применение этого правила ослабляет ударение на слоге *dens* до вторичного. Правило редукции гласных не применяется к этой гласной, поскольку она защищена вторичным ударением. Следующее правило, имеющее достаточную степень общности, заменяет последовательность ударений XX1 на 231, а правило приспособления ударения дает окончательную последовательность 3414. Таким образом, результирующая форма имеет нередуцированный гласный во втором слоге с ударением четвертой степени. Для сравнения рассмотрим слово *compensation* (компенсация). Второй гласной в этом слове является также фонема /e/ (ср. *com-*

pensatory, заменительный); однако она не получает ударения ни на одном из предшествующих циклов (поскольку compensation не является производным от compense, как condensation от condense). Поэтому она редуцируется в [i]. Таким образом, мы получаем противопоставление редуцированной и нередуцированной гласных под слабым ударением в паре compensation — condensation как автоматическое, хотя и не непосредственное следствие различий в структуре составляющих.

В качестве последнего примера рассмотрим переплетение правил (29a) и (29б) по мере усложнения синтаксических моделей предложения. Ср. сочетания John's blackboard eraser (Джонова тряпка для стирания с доски), small boys' school (в смысле small school for boys, маленькая школа для мальчиков) и small boys school (в смысле school for small boys, школа для маленьких мальчиков). Выводы для этих трех сочетаний (после того как начальные циклы припишут главное ударение словам) таковы:

- I. 1. [<sub>NP</sub><sup>1</sup>John's [<sub>N</sub><sup>1</sup>black board<sup>1</sup>]<sub>N</sub> eraser<sup>1</sup>]<sub>NP</sub>.
2. [<sub>NP</sub><sup>1</sup>John's [<sub>N</sub><sup>1</sup>black board<sup>2</sup> eraser<sup>1</sup>]<sub>N</sub>]<sub>NP</sub>.  
(Применяем правило (29a) к самой внутренней составляющей и стираем скобки.)
3. [<sub>NP</sub><sup>1</sup>John's black board<sup>3</sup> eraser<sup>2</sup>]<sub>NP</sub>.  
(Применяем правило (29a) к самой внутренней составляющей и стираем скобки.)
4. John's black board<sup>3</sup> eraser<sup>3</sup>.  
(Применяем правило (29б) и стираем скобки.)
- II. 1. [<sub>NP</sub><sup>2</sup>small [<sub>N</sub><sup>1</sup>boys' school<sup>1</sup>]<sub>N</sub>]<sub>NP</sub>.
2. [<sub>NP</sub><sup>1</sup>small boys' school<sup>1</sup>]<sub>N</sub>]<sub>NP</sub>.  
(Применяем правило (29a) к самой внутренней составляющей и стираем скобки.)
3. small boys' school.  
(Применяем правило (29б) и стираем скобки.)
- III. 1. [<sub>N</sub><sup>1</sup>[<sub>NP</sub><sup>1</sup>small boys]<sub>NP</sub> school<sup>1</sup>]<sub>N</sub>.
2. [<sub>N</sub><sup>2</sup>small boys' school<sup>1</sup>]<sub>N</sub>.  
(Применяем правило (29б) к самой внутренней составляющей и стираем скобки.)

3. <sup>3</sup>small <sup>1</sup>boys <sup>2</sup>school.

(Применяем правило (29а) и стираем скобки.)

4. <sup>3</sup>small <sup>1</sup>boys <sup>3</sup>school.

(На основе широко действующего правила, которое мы здесь не приводили.)

Короче говоря, фонетическая запись высказываний, которая на первый взгляд кажется весьма сложной и беспорядочной, может быть порождена систематическим циклическим применением простых трансформационных правил, причем порядок применения этих правил определяется синтаксической структурой высказывания. Поэтому разумно предположить, что именно эти правила лежат в основе порождения и восприятия действительной речи. При таком допущении мы получаем вполне правдоподобное объяснение того, почему носители языка единообразно и последовательно порождают и воспринимают новые высказывания со всеми этими сложными характеристиками (не осознавая, конечно, всех происходящих при этом процессов и их фонетических последствий). Это дает возможность построить новую теорию восприятия речи, по которой установление того, что наблюдаемое акустическое явление представляет собой звуковую последовательность с такими-то свойствами, в какой-то мере требует определения ее синтаксической структуры (т. е. частичного ее понимания). Более обычная точка зрения состоит в том, что мы определяем фонетическое и фонологическое строение высказывания, распознавая в звуковых волнах последовательность физических свойств, каждое из которых является определяющим свойством некоторой фонемы. Как мы уже пытались показать, эта точка зрения, основанная на требованиях линейности и инвариантности, является несостоятельной.

Можно представить себе устройство, распознающее предложения (т. е. модель восприятия), которое включает в себя грамматические правила, а также эвристический компонент, который должен опробовать входные предложения, извлечь из него некоторые ключевые признаки, соотносящие его с правилами, необходимыми для его порождения, и выбрать из альтернативных вариантов нужный путем последовательного приближения. При таком подходе нет оснований предполагать, что каждый сегмент высказывания имеет определенное характеризующее его свойство или вообще что сегменты речи действительно образуют последовательность. Далее, эта модель устраняет маловероятное допущение о том, что имеется одна грамматика для говорящего и другая для слушающего.

Такой подход к процессам восприятия неоднократно предлагался в различных работах последнего времени (MacKay, 1951; Bruner, 1958; Halle и Stevens, 1959, 1962; Stevens, 1960); что касается восприятия речи, то такая точка зрения в совершенно отчетливой форме была предложена Вильгельмом фон Гумбольдтом (Humboldt, 1836). Ясно, что для развития этих идей необходимо изучать распознавание и понимание речи. На основе изучения ключевых признаков можно построить достаточно обоснованную гипотезу относительно того, что является входом для устной речи человека; опираясь на эту гипотезу можно, далее, говорить о том, какую форму принимают высказывания в мозгу человека. Все эти гипотезы могут быть проверены в результате сравнения входов с представлениями, порождаемыми в мозгу; на основе таких сравнений гипотеза может быть принята или может быть отвергнута (ср. с гл. 13, разд. 2). Хотя восприятие речи есть чрезвычайно сложный процесс, для нормального взрослого человека он является абсолютно естественным; при этом среди других сложных процессов восприятия он занимает особое место, поскольку в данном случае мы находимся уже у истоков правдоподобной и точной порождающей теории, которая даст возможность понять организуемые принципы, лежащие в его основе.

## ЛИТЕРАТУРА

- Berge C., *Théorie des graphes et ses applications*, Paris, Dunod, 1958. (Русский перевод: К. Берг, Теория графов и ее применение, ИЛ, М., 1962.)
- Bloch B., A set of postulates for phonemic analysis, *Language*, 24 (1948), 3—46.
- Bloch B., Phonemic overlapping, *Am. Speech*, 16 (1940), 278—284; Joos M. (Ed.), *Readings in linguistics*, Washington, 1957, 93—96.
- Bloch B., Studies in colloquial Japanese IV: Phonemics, *Language*, 26 (1950), 86—125; Joos M. (Ed.) *Readings in linguistics*, Washington, 1957, p. 329—348.
- Bruner J. S., *Neural mechanisms in perception*, Solomon H. C., Cobb S., Penfield W. (Ed.), *The brain and human behavior*, Baltimore, 1958, 118—143.
- Chomsky N., *Logical structure of linguistic theory*, Microfilm, Mass. Inst. Tech. Library, 1955.
- Chomsky N., Three models for the description of language, *IRE Trans. on Inform. Theory*, IT-2 (1956), 113—124. (Русский перевод: Хомский Н., Три модели описания языка, Кибернетический сб. № 2, М., 1961, стр. 237—266.)
- Chomsky N., *Syntactic structures*, The Hague, 1957. (Русский перевод: Хомский Н., Синтаксические структуры, сб. «Новое в лингвистике», вып. 2, 1962, стр. 412—527.)
- Chomsky N., *The transformational basis of syntax*, Hill A. A. (Ed.), IV-th Univer. of Texas Symp. on English and syntax, 1959 (unpublished).
- Chomsky N., On the notion of rule of grammar, Jakobson R. (Ed.) *Structure of language and its mathematical aspects*, Proc. 12 Symp. in App. Math., Providence, 1961, 6—24. (a); Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.

- Chomsky N., Some methodological remarks on generative grammar, *Word*, 17 (1961), 219—239. (Русский перевод: Хомский Н., Несколько методологических замечаний о порождающей грамматике, сб. «Вопросы языкознания», 1962, № 4.)
- Chomsky N., Explanatory models in linguistics, Nagel E., Suppes P., Tarsky A. (Eds.), *Logic, Methodology and Philosophy of Science. Proceedings of the 1960 international congress*, Stanford, 1962, 528—550 (a). Русский перевод: Объяснительные модели в лингвистике, сб. Математическая логика и ее применения, М., «Мир», 1965 г. 245.)
- Chomsky N., The Logical basis of linguistic theory, *Proc. of the 9-th Int. Congr. of linguists*, 1962, Preprints, Cambridge, Mass., 1962, 509—574 (6); Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Chomsky N., Halle M., Lukoff F., On accent and juncture in English, For Roman Jakobson, The Hague, 1956.
- Culik K., On some axiomatic systems for formal grammars and languages, Mimeographed, 1962.
- Davis M., Computability and unsolvability, New York, 1958.
- Halle M., Sound pattern of Russian, The Hague, 1959 (a).
- Halle M., Questions of linguistics, *Nuovo Cimento*, 13 (1959), 494—517 (6); Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Halle M., On the role of simplicity in linguistic descriptions, Jakobson R. (Ed.), *Structure of language and its mathematical aspects. Proc. of the 12 Symp. in App. Math.* Providence, 1961, 89—94.
- Halle M., Chomsky N., Sound pattern of English, New York, 1968.
- Halle M., Stevens K. N., Analysis by synthesis, *Proc. of the seminar on speech com. and production*, AFCRC-TR-59—198, 1959; Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Halle M., Stevens K. N., Speech recognition: a model and a program for research, *IRE trans. on inform. theory*, IT-8 (1962), 155—159.
- Harris Z. S., Discourse analysis, *Language*, 28 (1952), 1—30 (a); Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Harris Z. S., Discourse analysis: a sample text, *Language*, 28 (1952), 474—494 (6).
- Harris Z. S., Co-occurrence and transformation in linguistic structure, *Language*, 33 (1957), 283—340; Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Humboldt W. von, Über die Verschiedenheit des menschlichen Sprachbaues, Berlin, 1836, Bonn, 1960.
- Jakobson R., Fant C. G. M., Halle M., Preliminaries to speech analysis, Tech. Rep. 13. Acoustics Laboratory, Mass. Inst. Tech., Cambridge, Mass., 1952. (Русский перевод: Якобсон Р., Фант Г. М., Халле М., Введение в анализ речи, сб. «Новое в лингвистике», вып. 2, М., 1962.)
- Jakobson R., Halle M., *Fundamentals of language*, The Hague, 1956. (Частичный русский перевод: Якобсон Р., Халле М., Фонология и ее отношение к фонетике, сб. «Новое в лингвистике» вып. 2, М., 1962.)
- Katz J., Semi-sentences, Katz J., Fodor J. (Eds.), *Readings in philosophy of language*, New York, 1963.
- Kraft L. G., A device for quantifying, grouping and coding amplitude modulated pulses, MS thesis, Dept. Elec. Eng., Mass. Inst. Techn., 1949.
- Lees R. B., Review of Chomsky, «Syntactic structures», *Language*, 33 (1957), 375—408.
- Lees R. B., A grammar of English nominalizations, Supplement to «International J. Am. linguistics», 1960.
- Liberman A. M., Delattre P., Cooper F. S., The role of selected stimulus variables in the perception of unvoiced stop consonants, *Amer. J. Psychol.*, 65 (1952), 497—516.
- MacKay D. M., Mindlike behavior in artefacts, *Brit. J. Philos. Science*, 2 (1951), 105—121.
- Mandelbrot B., On recurrent noise limiting coding, *Proc. Symposium on information networks*, Polytechnic Institute of Brooklyn, 1954, 205—221.
- Matthews G. H., *Hidatsa syntax*, Mimeographed, Mass. Inst. Tech., 1962.
- McMillan B., Two inequalities implied by unique decipherability, *IRE trans. on inform. theory*, IT-2 (1956), 115—116.
- Miller G. A., Speech and communication, *J. of acous. Soc. Amer.*, 30 (1958), 397—398.

- Osgood C. E., Suci G. J., Tannenbaum P. H., The measurement of meaning, Urbana, 1957.
- Postal P., Some syntactic rules in Mohawk, Ph. D., Dept. of Anthropology, Yale University, 1962.
- Postal P., Constituent analysis, Supplement to «International J. Amer. linguistics», 1964.
- Rogers H., The present theory of Turing machine computability, *J. soc. indust. appl. math.*, 7 (1959), 114—130.
- Sapir E., Sound patterns in language, *Language*, 1 (1925), 37—51; Mandelbaum D. G. (Ed.), Selected writings of Edward Sapir, Berkeley, 1949, 33—45.
- Sapir E., La réalité psychologique des phonèmes, *J. de psychologie normale et pathologique* (1933), 247—265; Mandelbaum D. G. (Ed.), Selected writings of Edward Sapir, Berkeley, 1949, 46—60.
- Saussure F. de, Cours de linguistique générale, Paris, 1916.
- Schatz C. D., The role of context in the perception of stops, *Language*, 30 (1954), 47.
- Schützenberger M. P., On an application of semi-group methods to some problems in coding, *IRE trans. on inform theory*, IT-2 (1956), 47—60.
- Shannon C. E., Communication in the presence of noise, *Proc. IRE*, 37 (1949), 10—21. (Русский перевод: сб. Шеннон К., Работы по теории информации и кибернетике, Связь при наличии шума, ИЛ, М., 1963, стр. 433.)
- Stevens K. N., Toward a model for speech recognition, *J. acoust. Soc. Amer.*, 34 (1960), 47—55.
- Sweet H., A handbook of phonetics, Oxford, 1877.
- Трахтенброт Б. А., Алгоритмы и машинное решение задач, изд. 2, М., 1960.
- Wallace A. F. C., On being just complicated enough, *Proc. Nat. Acad. Sci.*, 47 (1961), 458—464.
- Ziff P., Semantic Analysis, Ithaca, New York, 1960 (a).
- Ziff P., On understanding «Understanding utterances», Mimeographed, 1960 (b); Katz J., Fodor J. (Eds.), Readings in philosophy of language, New York, 1963.
- Ziff P., About ungrammaticalness, Mimeographed, University of Pennsylvania, 1961.

## СОДЕРЖАНИЕ

Введение .....	3
1. Уточнение границ исследования .....	5
2. Некоторые алгебраические аспекты кодирования .....	11
3. Некоторые основные понятия лингвистики .....	18
4. Некоторый простой класс порождающих грамматик .....	28
5. Трансформационные грамматики .....	34
5.1. Некоторые недостатки грамматик непосредственных составляющих .....	35
5.2. Уточнение определения грамматической трансформации .....	39
5.3. Структура составляющих в трансформированных цепочках .....	43
6. Звуковая структура языка .....	47
6.1. Роль фонологического компонента в грамматике .....	47
6.2. Звуки и фонемы .....	48
6.3. Условие инвариантности и условие линейности .....	51
6.4. Некоторые фонологические правила .....	55
Литература .....	62

## Издательство УРСС

специализируется на выпуске учебной и научной литературы, в том числе монографий, журналов, трудов ученых Российской Академии наук, научно-исследовательских институтов и учебных заведений.



### Уважаемые читатели! Уважаемые авторы!

Основываясь на широком и плодотворном сотрудничестве с Российским фондом фундаментальных исследований и Российским гуманитарным научным фондом, мы предлагаем авторам свои услуги на выгодных экономических условиях. При этом мы берем на себя всю работу по подготовке издания — от набора, редактирования и верстки до тиражирования и распространения.

---

Среди вышедших и готовящихся к изданию книг мы предлагаем Вам следующие:

#### Серия «Женевская лингвистическая школа»

*Балли Ш. Жизнь и язык.*

*Сеше А. Очерк логической структуры предложения.*

*Сеше А. Программа и методы теоретической лингвистики.*

#### Серия «Новый лингвистический учебник»

*Кобозева И. М. Лингвистическая семантика.*

*Баранов А. Н. Введение в прикладную лингвистику.*

*Плунган В. А. Общая морфология: Введение в проблематику.*

*Бурлак С. А., Старостин С. А. Введение в лингвистическую компаративистику.*

#### Серия «Лингвистическое наследие XX века»

*Кибрик А. Е. Очерки по общим и прикладным вопросам языкознания.*

*Комлев Н. Г. Компоненты содержательной структуры слова.*

*Кацнельсон С. Д. Типология языка и речевое мышление.*

*Блумфилд Л. Язык.*

*Семереньи О. Введение в сравнительное языкознание.*

*Бенвенист Э. Общая лингвистика.*

*Балли Ш. Французская стилистика.*

*Балли Ш. Упражнения по французской стилистике.*

*Балли Ш. Общая лингвистика и вопросы французского языка.*

---

*Ажеж К. Человек говорящий: Вклад лингвистики в гуманитарные науки.*

*Шухардт Г. Избранные статьи по языкознанию.*

*Дельбрюк Б. Введение в изучение языка.*

*Романенко А. П. Советская словесная культура: образ риторика.*

*Лаптева О. А. Русский разговорный синтаксис.*

*Лаптева О. А. Живая русская речь с телеэкрана.*

*Новиков Л. А. Художественный текст и его анализ.*

*Комлев Н. Г. Слово в речи. Денотативные аспекты.*

*Кузнецов В. Г. Женевская лингвистическая школа: от Соссюра к функционализму.*

---

1536 ID 6024



9 785354 001149 >

E-mail: [URSS@URSS.ru](mailto:URSS@URSS.ru)

Каталог изданий

в Internet: <http://URSS.ru>

Тел./факс: 7 (095) 135-44-23

Тел./факс: 7 (095) 135-42-46