АКАДЕМИЯ НАУК СССР
ИНСТИТУТ ЯЗЫКОЗНАНИЯ
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ

Р. Г. ПИОТРОВСКИЙ

Текст, машина, человек

ИЗДАТЕЛЬСТВО «НАУКА»
Ленинградское отделение
Ленинград 1975

В работе описывается отечественный опыт реализации машинного перевода и автоматического аннотирования текстов на иностранных языках, а также обобщаются теоретические исследования в области вероятностной переработки текста в системе «человек—машина—человек».

Особое внимание уделяется проблеме машинного распознавания смыслового образа текста в свете антиномий человека и робота, языка и речи, синхронии и диахронии. В заключение рассматриваются вопросы лингвистического обеспечения АСНТИ.

Работа адресована специалистам в области информатики, программирования, а также языковедам, интересующимся вопросами математической лингвистики. Книга может быть использована в качестве учебного пособия на отделениях прикладной математики, а также структурной и математической лингвистики.

П $\frac{20205\text{-}580}{042\,(02)\text{-}75}$ 344-75

ВВЕДЕНИЕ

В эпоху научно-технической революции язык значительно расширил свои коммуникативно-общественные функции. Продолжая оставаться важнейшим орудием общения между людьми, естественный язык становится той основой, на которой организуется диалог между человеком и электронновычислительной машиной. Опыт построения и работы современных автоматизированных систем управления, а также систем сбора, хранения и переработки научно-технической информации говорит о том, что в ближайшем будущем естественный язык и его разновидности станут основным средством передачи информации в коммуникационных системах «человек—машина—человек».

Функционируя в машине, естественный язык переносится с привычной для него ассоциативно-эвристической почвы человеческого мышления на чуждый ему последовательно-алгоритмический субстрат компьютера. Эта трансплантация языка и возникающая в связи с ней инженерно-лингвистическая проблематика имеют для языкознания ряд методологических последствий.

Развитие современного языкознания характеризуется стремлением ввести в лингвистические исследования такие технологические ограничения и условия, которые давали бы, с одной стороны, логическую строгость теории, а с другой — обеспечивали бы объективность и полноту описания лингвистического материала. Стремление к логической строгости теории и объективности описания характеризует общую типологию, социолингвистику, квантитативное и структурное языкознание. Эти технологические требования выполняются и в инженерной лингвистике, логическую непротиворечивость и полноту построений которой проверяет и бескомпромиссно оценивает кибернетический автомат.

Вместе с тем инженерная лингвистика выполняет в системе лингвистических дисциплин особую методологическую функцию. Чтобы охарактеризовать эту функцию, вспомним о том, как наука проходит «диалектический путь познания истины, познания объективной реальности» (Ленин, 1973, с. 152—153). «Истина, — говорит В. И. Ленин, — есть процесс. От субъективной идеи человек идет к объективной идее через „практику“ (и технику)» (Ленин, 1973. с. 183).

Основная задача инженерной лингвистики состоит в моделировании на ЭВМ лингвистических знаков, систем и функций. Для искусственного воспроизведения объектов нужна не только техническая база, но и в первую очередь достаточно сильная и объективная теория. Поэтому практические результаты инженерной лингвистики выступают в качестве эффективного критерия для оценки тех или иных лингвистических теорий. Такова методологическая функция, которую выполняет инженерная лингвистика относительно общей теории языка.

Решая технологические и некоторые теоретические задачи, инженерная лингвистика является также источником новых идей и методов не только для смежных с языкознанием дисциплин, например для методики преподавания языков (Алексеев и др., 1974) или информатики (Salton, 1968/1973, с. 22 и сл.), но и для таких, казалось бы, далеких от инженерной лингвистики отраслей знаний, как структурное литературоведение (Григорьев, 1971; Aitken, Bailey, Hamilton-Smith, 1973), психиатрия (Пашковский, Сребрянская, 1971), техническая кибернетика (Михлин, Пиотровский, Фрумкин, 1974), теория программирования (Yngve, 1958/1963; Pratt, 1973; Попескул*, 1975; Скитневский, 1974), теория нечетких множеств и алгоритмов (ср. гл. 8).

Каковы же источники и основные направления в распространении инженерно-лингвистических идей?

Первым источником является стержневая проблема инженерной лингвистики — проблема автоматического распознавания смыслового образа. Если в какой-либо области знания появляется необходимость произвести машинное или просто достаточно строгое разбиение семантического пространства с последующим иерархическим упорядочением выделенных областей (ср. задачи по построению отраслевых или универсальных тезаурусов), то исследователи все чаще обращаются в этом случае к опыту инженерной лингвистики. Еще более полезной оказывается теория и методика инженерной лингвистики при построении программ машинного отнесения неизвестного объекта к той или иной области искусственного пространства (ср. в этом плане работы в области опознания неизвестных метео- и радиолокационных объектов — СтРААТ, 1974, с. 18).

Вторым источником распространения новых идей и методов является семиотическая конфронтация естественного языка и искусственных языков (в первую очередь языка математики). Эта конфронтация (ср. **11**, **15**, **50—52**) обнаруживает те принципиальные различия, которые существуют между знаковой системой языка и конвенциональными знаковыми системами, — системами, выступающими по отношению к естественному языку в качестве вторичного искусственного кода (Ярцева и др., 1974, с. 9). Исследование отношений, в которые вступает открытый естественный язык с закрытыми конвенциональными семиотическими системами, является одной из центральных методологических и технологических задач инженерной лингвистики. Правильное решение этих задач является одним из условий научного прогресса в области семиотики, информатики, программирования и теории математических машин, моделирования мыслительных процессов.

Представителей смежных гуманитарных дисциплин инженерная лингвистика привлекает также конструктивностью своей научной стратегии и методики. Этот конструктивизм состоит в том, что жизненность каждой идеи и гипотезы проверяется здесь, как уже было сказано, путем воспроизведения на ЭВМ тех объектов, с которыми имеет дело эта теория. То положение, которое заняла инженерная лингвистика среди других языковедческих дисциплин, явилось результатом усилий, которые прилагали к решению лингвистических задач на ЭВМ сотни языковедов и математиков-программистов. Путь, который прошла за четверть века эта наука, не был простым и легким ни для лингвистов, ни для программистов. На этом пути, пожалуй, было больше разочарований и неудач, чем эффектных результатов. Тем не менее за этот период накоплен значительный опыт. Поэтому настало время оглянуться на путь, пройденный инженерной лингвистикой, и, оценив полученные результаты, рассмотреть основные теоретические проблемы и стоящие перед ней технологические задачи, а затем наметить ближайшие лингвистические, информационно-семиотические, программистские и технические перспективы их решения.

В настоящей книге делается попытка ретроспективного и проспективного рассмотрения вопросов инженерной лингвистики, опирающегося прежде всего на опыт группы «Статистика речи», а также на некоторые результаты, полученные другими отечественными и зарубежными коллективами. Основные идеи этой книги излагались в лекционных курсах по общему языкознанию, инженерному языкознанию и лингвистическому обеспечению АСНТИ. Эти лекционные курсы были прочитаны автором в Воронежском, Горьковском, Дальневосточном, Латвийском, Ленинградском, Московском, Самаркандском, Тартуском, Тбилисском университетах, Кишиневском и Ленинградском политехнических институтах, ЛГПИ им. А. И. Герцена, Минском педагогическом институте иностранных языков, а также в других педвузах. Обсуждение содержания лекций способствовало уточнению некоторых вопросов, рассматриваемых в книге.

Рукопись книги была прочитана П. М. Алексеевым, Г. М. Перцовой, которым автор выражает свою признательность за деловую критику. Автор благодарит также Л. Н. Беляеву, В. Н. Билан, И. С. Кравцову, Е. М. Найвельт, П. В. Садчикову и других своих учеников из группы «Статистика речи», принимавших участие в подготовке рукописи к печати.

Часть первая

ПЕРЕРАБОТКА ТЕКСТА ЧЕЛОВЕКОМ И МАШИНОЙ

Глава 1

ИНФОРМАЦИЯ И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

1. Энергетические и информационные процессы

Между энергетическими и информационными явлениями обычно трудно провести четкую границу. Действительно, такие, казалось бы, сугубо энергетические воздействия одного объекта на другой, как удар молнии в дерево или толчок старшины в спину парашютиста, замешкавшегося перед открытым люком, можно рассматривать и как акты передачи некоторой информации от первого объекта ко второму. Вместе с тем такие чисто информационные процессы, как подача старшиной десантного подразделения команды «прыгай» или чтение лекции, невозможны без некоторого, хотя и очень слабого, энергетического воздействия голоса источника сообщения (старшины или лектора) на органы слуха приемника сообщения (парашютиста или учащегося).

При изучении функционирования человеко-машинных комплексов оказывается принципиально важным определение того, какой аспект рассматриваемого процесса является здесь релевантным: энергетический или информационный. Чтобы дать ответ на этот вопрос, необходимо определить существо каждого из указанных аспектов.

Энергию принято определять как общую меру различных форм материального движения (БСЭ-II, статья «Энергия»); менее ясные определения приводятся для понятия «информация», которая иногда рассматривается как функция отражения новизны, иногда как мера организации системы (БСЭ-III, статья «Информация»; Кондаков, 1971, статья «Информация»).

Рассмотрение вопросов, связанных с функционированием языка в человеко-машинных системах, требует более точного описания таких понятий, как «информация» и ее виды — «информационный процесс», «знак». Обратимся поэтому к их определению.

2. Семиозис, знак и информация

Передача информации осуществляется в коммуникативной системе, включающей в идеале следующие компоненты:
1) источник (отправитель) сообщения;

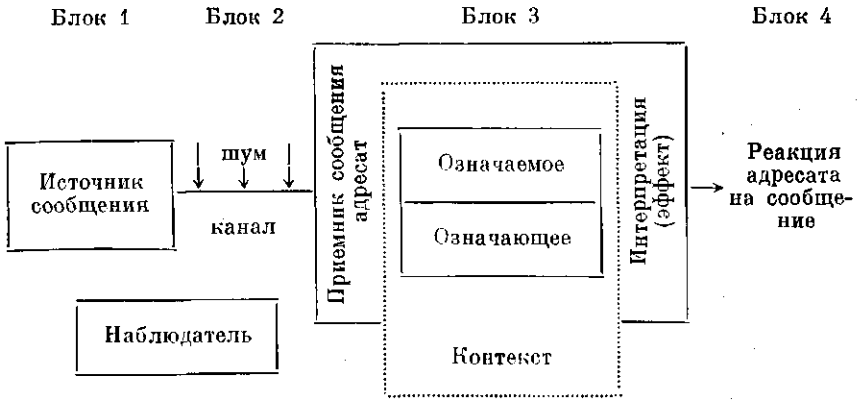


Рис. 1. Схема информационного процесса.
Прерывистой линией обозначен семиозис.

2) канал с шумом, по которому передается сообщение;
3) знак,[1] компонентами которого являются:
   а) означающее (имя, носитель информации),
   б) означаемое, которое может включать денотат (т. е. отражение предмета из внешнего мира), десигнат (т. е. понятие о предмете, смысл имени, концепт денотата), коннотат (т. е. комплекс чувственно-оценочных, в том числе эстетических оттенков значения); [2]
4) приемник сообщения;
5) интерпретация сообщения в приемнике, т. е. готовность последнего реагировать на означаемое знака и выдавать ответный эффект;
6) контекст, т. е. окружение, в котором осуществляется интерпретация сообщения;
7) реализация реакции адресата на распознанное и проинтерпретированное в ходе семиозиса сообщение;
8) наблюдатель извне — метанаблюдатель (рис. 1).

Присутствие компонентов 1—5 является обязательным условием для осуществления информационного процесса. Легко заметить, например, что при отсутствии знака или интерпретационной расположенности приемника воздействие источника на приемник может иметь только энергетический характер; отсутствие канала связи вообще снимает вопрос о взаимодействии источника и приемника. Что же касается контекста и внешнего наблюдателя, то их присутствие вообще не является обязательным при реализации

[1] Мы будем определять знак как материальный объект, который условно представляет некоторый предмет, явление, свойство, связь или отношение предметов, явлений, свойств, а также чувственное состояние источника сообщения (ср.: Кондаков, 1971, статья «Знак»).

[2] Некоторые авторы, например: Lewis, 1952, с. 58 и сл.; Langer, 1958, с. 64, дают иные определения терминам «коннотат», «коннотация».

информационного процесса. В тех случаях, когда адресат реагирует на сообщение, информационный процесс может рассматриваться в игровом аспекте (Wittgenstein, 1933/1958). Однако эта реакция может быть также равна нулю. Подробнее об информационном процессе см.: Morris, 1955, Glossary; Nauta, 1972, с. 27—30; ср.: Ветров, 1968; Степанов, 1971, с. 159; Урсул, 1971; Hayakawa, 1964; Schaff, 1960/1963; Kirschenmann, 1969.

Импульсы, идущие от отправителя по каналу связи, являются функцией внутреннего состояния, или, иными словами, структурного разнообразия источника сообщения. Однако эти импульсы становятся реальными носителями информации и формируют сообщение только в том случае, когда приемник обладает состоянием готовности осуществить отражение и интерпретацию той части внутреннего состояния источника, которая воплотилась в переданной совокупности импульсов. Без этого отражения импульсы могут осуществлять лишь энергетическое воздействие на приемник. Поэтому основным звеном информационного процесса является не столько передача, сколько прием сообщения. Этот прием осуществляется с помощью семиозиса, представляющего собой пятичленное отношение: означающее (носитель информации)—приемник сообщения — означаемое [1]—контекст (если таковой существует) — интерпретация (Morris, 1938; Nauta, 1972, с. 35—36). Семиозис, опирающийся на механизмы распознавания и сведения, которые заранее заложены в приемнике (см. 6—7), обеспечивает отражение и интерпретацию приемником структурного разнообразия источника сообщения. Общая схема информационного процесса и положение в ней семиозиса показаны на рис. 1.

Если согласиться с только что приведенной схемой информационного процесса, то саму информацию можно рассматривать как величину (или меру), характеризующую различные формы отражения в приемнике структурного разнообразия взаимодействующего с ним источника сообщения.

Прежде чем обратиться к определению разных форм информации, рассмотрим подробнее различные виды знака.

Наиболее примитивной формой знака является с и г н а л, образующийся в результате чувственного воздействия импульса на ощущения («нервную систему») адресата. Сигнал, образующийся из соединения импульса (означающего) с чувственным коннотатом (означаемым) (рис. 2, *а*), является основным звеном семиозиса у растений и низших животных, использующих врожденные программы поведения (Thorpe, 1956; Tinbergen, 1958, с. 38—44; Лурия, 1973, с. 12). В естественном языке к сигналам можно условно отнести междометия и другие формы выражения эмоций.

Более сложной формой знака является Д - з н а к. Его означаемое включает денотат, рядом с которым могут присутствовать

[1] Означаемое знака необязательно должно включать денотат, десигнат и коннотат; достаточно присутствия одного из этих трех аспектов (подробнее см. ниже, стр. 51).

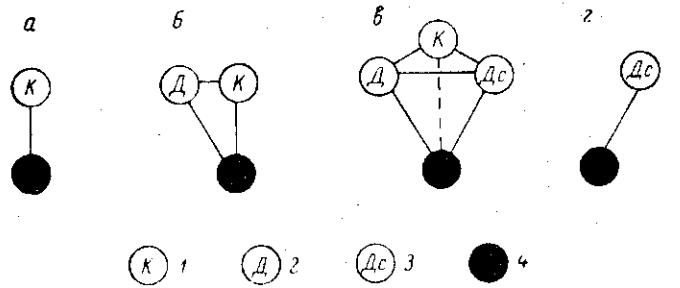


Рис. 2. Различные виды знака.

а — сигнал; *б* — Д-знак; *в* — символ; *г* — искусственный знак (ИЗ). У с л о в н ы е о б о з н а ч е н и я: *1* — коннотат; *2* — десигнат; *3* — денотат; *4* — носитель информации (имя).

и коннотативные значения (рис. 2, *б*). Д-знаки используются наряду с сигналами в коммуникативных системах высших животных, способных к индивидуально изменчивому поведению и особенно к тому уровню поведения, который был назван Н. А. Бернштейном «уровнем синэргий» (Бернштейн, 1947; ср. также: Uexkuell, Kriszat, 1956; Жинкин, 1960). В отличие от сигнала Д-знак задается не генетически, но вырабатывается в ходе обучения или, точнее, адаптивного взаимодействия организма с внешней средой. Примером Д-знака в языке человека является имя собственное (ср.: Степанов, 1971, с. 90).

Наиболее сложной формой знака является с и м в о л, служащий основным звеном семиозиса во всех аспектах сознательной деятельности человека и в различных формах его социального общения, в том числе в естественном языке (Степанов, 1971, с. 32—74; Nauta, 1972, с. 29; Лурия, 1973, с. 294—296). Наиболее типичными примерами символов являются значимые единицы естественного языка — слова и их формы, словосочетания, предложения.[1] Лингвистический символ включает четыре компонента: имя (материальный носитель информации), затем денотат (предмет, явление действительности, обозначаемое именем) и десигнат, или концепт (т. е. смысл, понятие о предмете или явлении), а также коннотат, охватывающий дополнительные экспрессивно-оценочные, а также эстетические значения [2] (рис. 2, *в*).

В ходе своей истории человечество вынуждено искусственно воспроизводить все большее количество энергетических явлений

[1] В разряд символов не входят лингвистические «полузнаки» — так называемые знаки-дистинкторы и знаки-делимитаторы (Маслов, 1967, с. 112); их мы будет относить, по традиции, к классу «фигур» (Hjelmslev, 1953/1960, с. 305).

[2] Ч. Моррис наряду с коннотативным, т. е. чувственно-оценочным (apraisive), компонентом выделял в означаемом символе прескриптивный (prescriptive) аспект, связанный с моторными центрами коры больших полушарий (Morris, 1964).

и их компонентов. Аналогичным образом развитие цивилизации требует построения искусственных информационно-семиотических процессов. Производящими системами в этих процессах являются искусственные языки (в частности, язык математики), использующие искусственный знак (ИЗ). ИЗ представляет собой соединение имени (означающего) с его концептом (десигнатом) (рис. 2, *г*). Подробнее об ИЗ см.: Church, 1956/1960, с. 17—36; Beth, 1963, с. 1—16; Пиотровская, Пиотровский, 1974, с. 361—363.

При обсуждении проблем, связанных с переработкой текста в ЭВМ, особый интерес представляет соотношение лингвистического символа и искусственного знака. Этот вопрос мы подробно рассмотрим в **15**.

3. Виды информации

В зависимости от того, к каким аспектам отражения и к каким компонентам знака прилагается информационная мера, различаются следующие виды информации.

1. Прагматическая информация, характеризующая ценность сообщения с точки зрения той цели, которую преследует в данный момент приемник сообщения, и с точки зрения выбора того решения, которое является наилучшим для достижения этой цели (Ackoff, 1962, с. 160 и сл.; Wójcik, 1969; Nauta, 1972, с. 220—225, 267—271).

2. Семантическая информация, которая оценивает отношения, возникающие в ходе семиозиса, между означающим и десигнатом знака (Bar-Hillel, 1964, с. 299; Hintikka, 1968, с. 312; Bogodist et al., 1974, с. 19—25).

3. Сигматическая информация, характеризующая отношение означающего и денотата (Klaus, 1965/1967, с. 15—16).

4. Аффективная информация, включающая эстетическое воздействие. Эта информация характеризует отношение означающего и экспрессивно-оценочное и эстетическое восприятие знака, обобщенное в его коннотате (Tillich, 1957; Moles, 1958/1966; Nauta, 1972, с. 60).

5. Синтактическая информация, представляющая собой оценку приемником тех ограничений, которые накладываются на комбинаторику и частоту употребления знаков (Пиотровский, 1968, с. 10 и сл.; Georgiev, Piotrowski, 1974, с. 125—127).[1]

Д. Наута убедительно показал, что отраженную в приемнике синтактическую информацию не следует смешивать с энтропией источника информации (Д. Наута обозначает это понятие терминами «преформация», «трансмиссионная информация», «потенциальная информация» — ср.: Nauta, 1972, с. 39, 156, 176—195, 203—204). Энтропия источника оценивает комбинаторно-статистические ограничения, заложенные в том языке, которым пользуется источник, — ограничения, которые лимитируют выбор импульсов при порождении сообщения. Энтропия не отражается в приемнике сообщения (приемник может воспринимать сообщение как не имеющую смысла последовательность импульсов, более того, приемника может и не быть вовсе), она фиксируется и измеряется наблюдателем извне (метанаблюдателем). Короче говоря, энтропия может рассматриваться в качестве меры структурной упорядоченности источника сообщения.[1]

[1] Мы не рассматриваем такие не имеющие прямого отношения к содержанию книги информационные концепции, как репрезентативную информацию Мак-Кея, научную информацию Фишера и, разумеется, донаучное житейское представление об информации как о сведениях, которыми обмениваются люди в ходе общения (Урсул, 1971, с. 276; Nauta, 1972, с. 205—214).

4. Три уровня интерпретации сообщения

Рассматривая биологические системы коммуникации, современная кибернетика выделяет три уровня готовности (интерпретационной способности) приемника сообщений.

Первый уровень характеризует интерпретационную способность растений и низших животных, которые, получив сигнал или цепочку сигналов, реализующих синтактическую и аффективную информацию, отвечают на них простейшими реакциями (ср. явление тропизма, этограммы и т. п. — Степанов, 1971, с. 30—31, 82—83; Thorpe, 1956; Tinbergen, 1958).

Второй уровень — уровень адаптивной обучаемости, характеризует информационную деятельность высших животных. Эти последние способны принимать не только синтактическую и аффективную информацию, но могут также перерабатывать сигматическую информацию, передаваемую с помощью Д-знаков. Интерпретационная способность высших животных характеризуется такими высокими формами адаптации и взаимодействия со средой, как способность к обучению (Ramsay, 1966; Nauta, 1972, с. 144—147).

Третий — децизивный уровень готовности характеризует информационное (в том числе и речевое) поведение человека, опирающееся на прагматическую информацию, которая обобщает в себе все остальные виды информации. Это поведение предусматривает не только возможность простейших реакций и адаптивного обучения, но отличается от других интерпретационных уровней способностью к принятию осмысленных решений (Лурия, 1974, с. 24; Nauta, 1972, с. 147—150).

[1] Может возникнуть вопрос, не следует ли рассматривать наблюдателя извне также в качестве приемника сообщения. В этом случае потенциальная информация представляла бы не что иное, как синтактическую информацию, отраженную приемником. Однако следует иметь в виду, что мера упорядоченности источника оценивается обычно по-разному наблюдателем и адресатом (ср. 40). Эти расхождения являются технологическим доводом в пользу целесообразности различения энтропии источника сообщения и синтактической информации.

С точки зрения наших задач целесообразно выделить еще один — нулевой или досемиотический уровень готовности адресата, связанный с передачей по каналу потенциальной информации, оцениваемой через величину энтропии. Адресат, имеющий нулевую интерпретационную способность, воспринимает последовательность импульсов, образующих сообщение, только как энергетическое воздействие или просто как информационный шум. В то же время метанаблюдатель может распознать в этой последовательности потенциальную информацию (энтропию), характеризующую структурную упорядоченность источника сообщения (ср. выше, стр. 10).

При моделировании информационных процессов применительно к человеко-машинным системам в ЭВМ могут возникать ситуации, аналогичные всем только что описанным уровням интерпретационной готовности.

Глава 2

ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ В МОЗГУ ЧЕЛОВЕКА И В МАШИНЕ

5. Технические возможности человека и ЭВМ при переработке текста

Классическая лингвистика отвлекается обычно от той материальной среды, в которой осуществляются изучаемые ею информационные процессы. Этот подход неприемлем для инженерной лингвистики, стратегия и технология которой зависят от особенностей функционирования естественного языка в системе «человек—машина—человек». Поэтому, прежде чем обратиться к рассмотрению лингвистических вопросов переработки текста в ЭВМ, необходимо сравнить структурно-функциональные характеристики человеческого мозга и памяти компьютера, а также определить особенности машинной переработки текстового сообщения. Это тем более необходимо и потому, что, хотя возможности машины и человека неоднократно обсуждались в научной литературе, среди языковедов, философов, кибернетиков и математиков нет единого мнения по поводу перспектив использования вычислительной техники в лингвистике. Одни, вспоминая об ЭВМ первого и второго поколения, имевших небольшой объем памяти, низкую скорость работы (см. табл. 1), нестандартные накопители на МЛ, а также характеризовавшихся программной несовместимостью, — полны скепсиса по поводу любой автоматизации в лингвистике. Другие собираются ставить на ЭВМ сложнейшие семантико-синтаксические и даже стилистические задачи, которые по-настоящему смогут быть решены только на ЭВМ пятого поколения.

Между тем в настоящее время осуществляется полное обновление технической базы наших информационных органов, которые повсеместно переходят на пользование ЭВМ третьего поколения (ЭВМ-III), в первую очередь на ЭВМ Единой системы — ЕС-1020 (Ряд-20), ЕС-1022 (Ряд-22), ЕС-1030 (Ряд-30), ЕС-1040 (Ряд-40) и проектируемых ЕС-1035, ЕС-1050, ЕС-1060.

Характерными признаками машин третьего поколения являются, во-первых, их микроэлектронная компонентная основа (интегральные схемы), а во-вторых, исключительно широкое использование операционных систем, значительно расширяющих возможности взаимодействия человека и компьютера. Эти признаки отсутствовали или были очень слабо развиты у машин первого поколения (ламповые ЭВМ) и второго поколения (ЭВМ на дискретных полупроводниковых устройствах).

Что же касается перечисленных выше ЕС-ЭВМ, то они представляют собой семейство программно-совместимых машин, предназначенных для решения очень широкого класса вычислительных и логических задач, а также для использования в различных системах обработки технических, экономических, информационных, лингвистических и других данных. Эта универсальность ЕС-ЭВМ достигается за счет:

— универсального набора команд и их программной совместимости, опирающейся на байтовый слог,

— модульного принципа в построении связей и процедур управления,

— расширенного набора внешних устройств, особенно необходимых при решении лингвистических задач,

— мощной системы математического обеспечения.

С точки зрения лингвистов-пользователей, важной особенностью ЕС-ЭВМ является их программная совместимость, которая позволяет решать однажды запрограммированную лингвистическую задачу на любой модели ЕС-ЭВМ без перепрограммирования и повторной трансляции.

Технические средства, которые входят в состав каждой модели ЕС-ЭВМ, можно разделить на четыре уровня (см. рис. 3). Первый уровень образует процессор, объединяющий оперативную память, центральное устройство управления, арифметическо-логическое устройство. В процессоре объединяется основная часть функциональных средств ЕС-ЭВМ. Здесь сосредоточены средства, позволяющие выполнять арифметические и логические операции, средства обращения к памяти, средства, управляющие выполнением заданной последовательностью команд и организующие обмен между оперативной памятью и системой ввода—вывода (ср.: Шелихов, Селиванов, 1973, с. 6).

Обмен данными между процессором и внешними устройствами (четвертый уровень) осуществляется через каналы (второй уровень) и средства управления внешними устройствами (третий уровень).

Рис. 3. Структурная схема ЕС-ЭВМ.

ЕС-ЭВМ располагает внешними устройствами, которые можно разбить на 7 групп:

1) накопители информации на магнитных барабанах (МБ). дисках (МД) и картах (МК);

2) накопители на магнитной ленте (МЛ);

3) печатающие устройства (ПУ);

4) перфокарточные и перфоленточные устройства ввода;

5) дисплеи — устройства связи оператора с ЭВМ на электронно-лучевых трубках;

6) устройство связи оператора с ЭВМ с помощью электрической пишущей машинки;

7) терминалы — устройства обработки информации через абонентские пункты.

С точки зрения лингвиста-потребителя, особого внимания заслуживает использование в ЭВМ-III дисплея и накопителей

информации на МБ, МД и МК. Последние представляют собой запоминающие устройства с прямым доступом. В отличие от накопителей на МЛ, в которых поиск информации осуществляется по зонам, в накопителях с прямым доступом каждая запись имеет свой адрес, по которому обеспечивается прямое обращение к этой записи. Поскольку при решении лингвистических задач большая часть необходимой лингвистической информации хранится во внешних ЗУ, использование ЗУ с прямым доступом значительно убыстряет процесс автоматической переработки текста.

Что же касается дисплея, то с его помощью можно осуществлять редактирование машинных результатов еще до вывода их на ПУ. Кроме того, дисплей позволяет легко пополнять и корректировать автоматические словари, а также грамматические алгоритмы, использующиеся для переработки текста.

После этой краткой характеристики тех машин, на которых в настоящее время решаются лингвистические задачи, обратимся к сопоставлению возможностей человека и ЭВМ-III при переработке больших массивов текста. Скорость переработки текстовой информации в процессоре и оперативной памяти ЭВМ-III в несколько сотен и тысяч раз превосходит скорость преобразования этой информации в мозгу человека (ср. табл. 1). Хотя эти скорости несколько обесцениваются более медленным функционированием внешних накопителей на магнитных лентах, картах, барабанах и дисках, а также вводящих и выводящих устройствах, общая скорость ЭВМ при переработке текста намного превосходит возможности человека.

Текст с перфокарт или с перфоленты вводится в ЭВМ со скоростью до 1500 символов/сек. (Шелихов, Селиванов, 1973, с. 9). Современные оптические читающие автоматы считывают печатный текст со скоростью от 25 до 70 символов/сек. (Beauclair, 1973, с. 269—270; Olf, 1973; Crosfield 7000, 1973, с. 14). К концу 80-х годов ожидается появление оптических читающих устройств, работающих со скоростью 10—15 тыс. символов/сек., использующих знаковый ансамбль 500—1000 знаков и обладающих надежностью не более одной ошибки на 10^6 знаков с возможностью быстрой замены алфавитов и шифров.

В то же время скорость чтения у человека не превышает 25 символов/сек. Стандартное печатающее АЦПУ-128 выводит лингвистическую информацию из ЭВМ со скоростью до 900 символов/сек., новые печатающие устройства ЕС-7030 и ЕС-7032 смогут выводить машинные результаты со скоростью до 1800 символов/сек. (Шелихов, Селиванов, 1973, с. 9). Скорость выдачи информации человеком через устную речь или скоропись не превышает 10 символов (букв, фонем)/сек.

Машина обладает большей по сравнению с человеком помехоустойчивостью и надежностью. Если функционирование мозга при переработке текста (например, при его реферировании или переводе) зависит от физиологического и психического состояния человека, а также быстро нарушается в результате утомляемости, то

работа ЭВМ характеризуется большой устойчивостью и продолжительностью (например, среднее время безотказной работы ЭВМ «Минск-32» находится в пределах 50 часов).

Для машины характерно устойчивое и практически ничем не ограниченное во времени хранение информации. Напротив, человек при отсутствии тренировки последовательно забывает усвоенный материал (Florès, 1973, с. 317 и сл.).

Вместе с тем суммарные объемы памяти современных ЭВМ пока еще на несколько порядков меньше объема памяти человека (табл. 1).

Все эти характеристики электронного «мозга» (да простит нам читатель применение этой антропоморфической метафоры: здесь и дальше такие метафоры являются не более чем стилистико-дидактическим приемом) в некоторой степени влияют на выбор инженерно-лингвистической технологии. Однако в своих основных чертах стратегия инженерной лингвистики диктуется своеобразием структуры и функционирования компьютера, принципиально отличающим его от механизмов высшей нервной деятельности человека.

6. Схема высшей нервной деятельности человека

Схему высшей нервной деятельности человека упрощенно можно представить в виде системы, состоящей из трех блоков: рецепторного блока, воспринимающего информацию, которая поступает из внешнего мира, блока переработки и хранения информации (нервная сеть)[1] и эффекторного блока, служащего средством вывода информации из нервной сети (Arbib, 1964/1968, с. 15) (рис. 4).

Информационные процессы, связанные с переработкой лингвистической информации, осуществляются в основном в нервной сети, поэтому мы сосредоточимся на описании ее структуры и функционирования, оставляя в стороне вопросы, связанные с работой рецепторов и эффекторов.

Нервная сеть представляет собой совокупность большого числа двух видов клеток — нейронов и глиальных клеток. Информационные процессы определяются в основном свойствами нейронов, их функционированием и связями. Целесообразно рассматривать нейрон в качестве основного элемента нервной сети.

Поскольку нейроны весьма разнообразны по своему строению и выполняемым ими функциям, мы ограничимся рассмотрением некоторой их упрощенной схемы, ориентируясь при этом на ассоциативные нейроны, т. е. такие нейроны, которые анализируют и синтезируют информацию, полученную от рецепторных нейронов, и принимают решение об определенной форме активности.

[1] С этим блоком связаны два подблока: подблок, регулирующий тонус и бодрствование, и подблок программирования, регуляции и контроля психической деятельности (Лурия, 1973, с. 84).

Таблица 1

Характеристики современных ЭВМ, использовавшихся и предназначенных для решения лингвистических задач, в сравнении с информационными возможностями человека (см.: Грубов, Кардан, 1969, с. 40; Ларионов. Левин, Пржиялковский, Фатеев, 1973, с. 3; Küpfmüller, 1959, с. 68—74; George, 1961/1963, с. 438)

Параметры	Кибернетические автоматы								Человек[2]
	ЭВМ первого поколения	ЭВМ второго поколения		ЭВМ третьего поколения					
	«Урал-4»	М-222	«Минск-32»	ЕС-1020	ЕС-1022	ЕС-1030	ЕС-1040	ЕС-1050	
Скорость переработки информации (для ЭВМ — сложение—вычитание с плавающей запятой), в дв. ед./сек.	$0.5 \cdot 10^4$	$3.5 \cdot 10^4$	до $2.5 \cdot 10^4$	до $2 \cdot 10^4$	$8 \cdot 10^4$	до 10^5	до $3 \cdot 10^5$	до $7.50 \cdot 10^5$	15—45
Суммарная емкость памяти, в дв. ед., в том числе основной памяти (ОЗУ).	$6 \cdot 10^3$ 2048 48-разрядных слов $\simeq 10^5$ дв. ед.	$6 \cdot 10^6$ до 32768 50-разрядных слов $\simeq 1.7 \cdot 10^6$ дв. ед.	10^9 32768 37-разрядных слов $\simeq 1.2 \cdot 10^6$ дв. ед.	$9.1 \cdot 10^8$ 64÷256 Кбайт $\simeq 5.7 \cdot 10^6 \div$ $\div 2.3 \cdot 10^7$ дв. ед.	$9.1 \cdot 10^8$ 128÷512 Кбайт $\simeq 1.1 \cdot 10^7 \div$ $\div 4.6 \cdot 10^7$ дв. ед.	$9.1 \cdot 10^8$ 128÷512 Кбайт $\simeq 1.1 \cdot$ $10^7 \div 4.6 \cdot$ 10^7 дв. ед.	$2 \cdot 10^9$ 128÷512 Кбайт $\simeq 1.1 \cdot 10^7 \div$ $\div 4.6 \cdot 10^7$ дв. ед.	$2 \cdot 10^9$ 128÷1024 Кбайт $\simeq 1.1 \cdot 10^7 \div$ $\div 9.2 \cdot 10^7$ дв. ед.	10^{10}—10^{15}

Примечания.
[1] Теоретически объемы внешней памяти современных ЭВМ могут быть увеличены до любых размеров: все зависит от наличия сменных пакетов дисков и МД.
[2] В мозгу человека насчитывается около 10 млрд нейронов. Диапазон оценок объема памяти человека зависит от того, какое количество возможных состояний приписывается нейрону. Так, например, Д. Вулридж считает, что нейрон соответствует 4—5 двухпозиционным переключателям (Wooldridge, 1963/1965, с. 270). В этом случае объем памяти составлял бы $(4 \div 5) \cdot 10^{10}$ дв. ед.

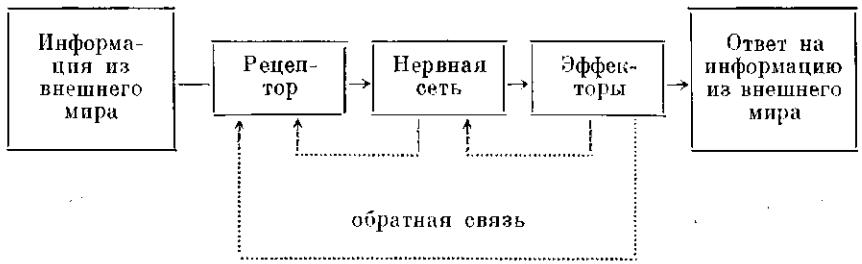


Рис. 4. Схема высшей нервной деятельности.

Схематично нейрон можно представить в виде системы, состоящей из следующих частей:

1) тела клетки (сомы), содержащей ядро и клеточную протоплазму;

2) дендритов — ветвящихся отростков, играющих роль входов в нейрон;

3) аксона — длинного, ветвящегося на конце нервного волокна, выполняющего функцию выходного канала, по которому передаются импульсы от сомы к другим клеткам;

4) синапсов — маленьких пластинок (бляшек), которыми заканчиваются ветвления аксонов и которые передают выходной сигнал с аксона данного нейрона на дендриты или тело других нейронов (рис. 5).

Аксон, его ветвления, синапсы, контактирующие с каждым из них ветки дендритов и, наконец, сами дендриты образуют прямые пути передачи информации от одного нейрона к множеству других нейронов. Имея друг с другом прямые каналы связи, мозговые нейроны оказываются замкнутыми на себя через другие нейроны, в результате чего формируются нейронные кольца и системы колец, образующие основную структуру нервной сети (Chang, 1950).

Работу нейрона и нейронной сети можно представить следующим образом.

Входные сигналы, передающиеся или путем химической диффузии, или в виде электрических импульсов, поступают через множество дендритов в нейрон. Под воздействием возбуждающих импульсов в соме происходит плавное накопление внутреннего потенциала нейрона. Как только этот потенциал достигнет некоторого критического порога, нейрон выдает импульс, передающийся по его единственному аксону на синапсы его разветвлений, а оттуда на дендриты других нейронов. При этом возбуждение одного нейрона в кольце может активизировать не только данное кольцо, но одновременно и другие нейронные кольца и их системы. Если же сумма весов входных синапсов, на которые поступают возбуждающие импульсы, не достигнет порога нейрона, нейрон импульса не выдает и передача сигнала по нервной сети прекращается.

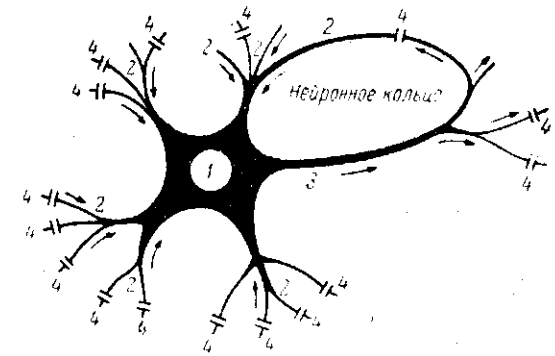


Рис. 5. Схематическое изображение нейрона.
1 — тело клетки (сома); *2* — дендриты; *3* — аксон; *4* — синапсы.

Из всего сказанного видно, что нейрон выступает в качестве порогового преобразователя информации с хорошо калиброванными входами (синапсами дендритов) и калиброванными выходами (синапсами аксонных разветвлений). Этот преобразователь обладает острой настройкой на временное расположение импульсов на входном сигнале.

Когда речь идет о преобразовании информации в нейроне, то имеется в виду тот простой факт, что входные сигналы могут либо возбудить нейрон, либо оказаться недостаточными для его возбуждения. В первом случае сигнал будет передан от нашего нейрона к другим нейронам, во втором — будет заторможен и прекратит свое существование.[1]

Если говорить о содержательной речевой информации, то нам пока мало известно о том, как записывается эта информация от-

[1] В кратком обзоре, разумеется, нельзя описать все многообразие гипотез, с помощью которых делаются попытки описать структуру функционирования мозга (о теории нейронных ансамблей см.: Hebb, 1949; Reitman, 1965/1968; о молярной теории резонансных аналогий и т. д. см.: Arbib, 1964/1968, с. 88—134, 142—149; George, 1961/1963, с. 158—202, 305—309). Не рассматриваются и такие важные вопросы нейрофизиологии, как механизмы торможения, в том числе пресинаптическое торможение, самогенерация нервного импульса (фоновая активность), адаптация и рефрактерность нейронов, наконец, проблема оценки количества возможных состояний одного нейрона, взаимоотношение нейрона и глии и т. п. Рассмотрение этих вопросов увело бы нас в сторону от проблемы соотношения мозга и компьютера при решении лингвистических задач. Изложение основных положений современной нейрофизиологии и нейропсихологии читатель найдет в следующих книгах: Анохин, 1973; Лурия, 1973; Прогресс биологической и медицинской кибернетики, 1974; Ashby, 1960/1964; Ranson, Clarc, 1959; Rosenblatt, 1962/1965; Wooldridge, 1963/1965; Pribram, 1974/1975; и сборниках: Biology of Memory, 1970; Models of Human-memory, 1970; Perception Mechanisms and Models, 1972/1974.

Все эти работы были использованы нами при описании структуры и функционирования нейрона и нейронных сетей.

носительно колец и микроансамблей нейронов, а также о том, как происходит ее считывание и переработка в мозгу.[1] Однако способность нейрона передавать определенный сигнал множеству других нейронов, затормаживая другие сигналы, а также наличие нейронных колец, способных в течение некоторого времени сохранять реверберирующую информацию, составляют те нейрофизиологические условия, при которых реализуется замечательная способность нашей памяти, с одной стороны, к ассоциативной записи информации, а с другой — к эвристическому, т. е. целенаправленному и избирательному ее поиску. Когда речь идет об эвристическом поиске, осуществляемом в мозгу человека, то чаще всего используется гипотеза, согласно которой вся зона человеческого поиска состоит из иерархически организованных областей и подобластей, связи между которыми и проявляются в виде ассоциаций (Miller, Galanter, Pribram, 1960/1965).

Ход ассоциативного поиска при восприятии текста представляется в этом случае как последовательное соотнесение поступающих в мозг человека цепочек графем или фонем при помощи морфем с областями смыслового пространства (семантического уровня). Однако если осуществлять это связывание путем последовательных операций, то время, необходимое для обработки одного предложения, было бы непомерно большим. Поэтому все цепочки обрабатываются одновременно.

Каждая совокупность нейронов, в которой записана морфема, может иметь два состояния: возбужденное и латентное. Каждая принятая мозгом фонемно-графемная цепочка возбуждает соответствующие морфемы, комбинации которых образуют на синтагматической оси множество морфемных цепочек. Система и норма языка отбирает из этого множества разрешенные комбинации, затормаживая остальные цепочки.

Морфемные цепочки возбуждают в свою очередь те области нейронной сети, в которых записаны соответствующие этим цепочкам семантические единицы. Последние формируются в последовательности семантем, среди которых отбирается лишь та семантическая цепочка, которая соответствует общей смысловой стратегии сообщения.

[1] Впрочем, кое-какие результаты в этом направлении имеются. Так, например, экспериментальные исследования с помощью биоэлектронной системы, обеспечивающей селективную связь альтерирующего стимула с характеристиками биопотенциалов коры головного мозга, показывают зависимость процессов функциональной реорганизации мультиклеточной активности нейронных популяций мозга человека в зависимости от вида словесных сигналов, используемых при тестировании кратковременной вербальной памяти (Бехтерева, Бундзен, Матвеев, Каплуновский, 1971, с. 1745—1760). Об исследованиях речевой организации памяти с помощью сравнительноанатомических наблюдений, физиологического метода раздражения отдельных участков мозга и клинических наблюдений над поведением больных с локальными поражениями мозга см.: Лурия, 1973; Лурия, 1974.

Порождение текста осуществляется, по всей видимости, в обратном порядке. Отправитель информации определяет сначала общую стратегию высказывания, возбуждая соответствующую область семантического пространства. Затем выбирается цепочка нужных семантем, которая реализуется в соответствующих последовательностях морфем, а каждая морфема возбуждает необходимую цепочку графем или фонем.

Следует иметь также в виду, что в ходе эвристического поиска человек формирует, очевидно, такие программы, которые непосредственно не приводят к цели, но создают предпосылки, обеспечивающие получение желаемого результата. Важное значение имеет здесь формирование новых «подкрепляющих раздражителей», использующихся для накопления полезной информации, которое в итоге может привести к инсайту.[1] Но сам инсайт может породить новые задачи и формировать новые программы, помогающие решать эти задачи, опираясь на прежний опыт.

С только что описанным нейролингвистическим и психолингвистическими механизмами работы мозга хорошо согласуются данные семиотической лингвистики. Известно, что каждая лингвистическая единица представляет собой многомерную систему, входя одновременно в несколько парадигм или ассоциативных рядов (Saussure, 1922/1933, с. 123—124; ср.: Михайлов, Черный, Гиляревский, 1968, с. 437—454; Miller, 1969; Aitken, Bailey, Hamilton-Smith, 1973, с. 103 и сл.; Клименко, 1974; Лурия, 1974, с. 14—15).Примером такого ассоциативного пучка может служить словоформа *шепчут*, стоящая на пересечении многих парадигм (рис. 6). Эти парадигмы могут быть образованы не только на основе морфологических словообразовательных, лексических и лексико-морфологических связей данного слова (ср. ряды *1—4* на рис. 6), они могут отражать тонкие семантические отношения (ср. синонимические — ряд *5* — или антонимические — ряд *6* — группы), наконец, они могут быть построены на основе общности акустических образцов (ср. аллитерационные и рифмообразующие группы слов — ряды *1*, *7*, *8*) и т. д. В результате в слове, как и в любом другом знаке естественного языка, оказывается заложенным не только основное понятийное значение с указанием на реализующие его денотаты, а также грамматические характеристики: одновременно в слове «упаковано» большое число дополнительных коннотативных значений. Если бы словесно-логическая (вербальная) память не располагала бы механизмами обобщения и сжатого хранения в языковом знаке информации, получаемой от рецепторов и от более низких уровней памяти, то для хране-

[1] Инсайт — «внезапное» творческое решение задачи, «озарение, приходящее в результате мобилизации прошлого опыта и полезной информации. Такое понимание инсайта не совпадает с его трактовкой в гештальт-психологии, согласно которой инсайт возникает независимо от прошлого опыта системы.

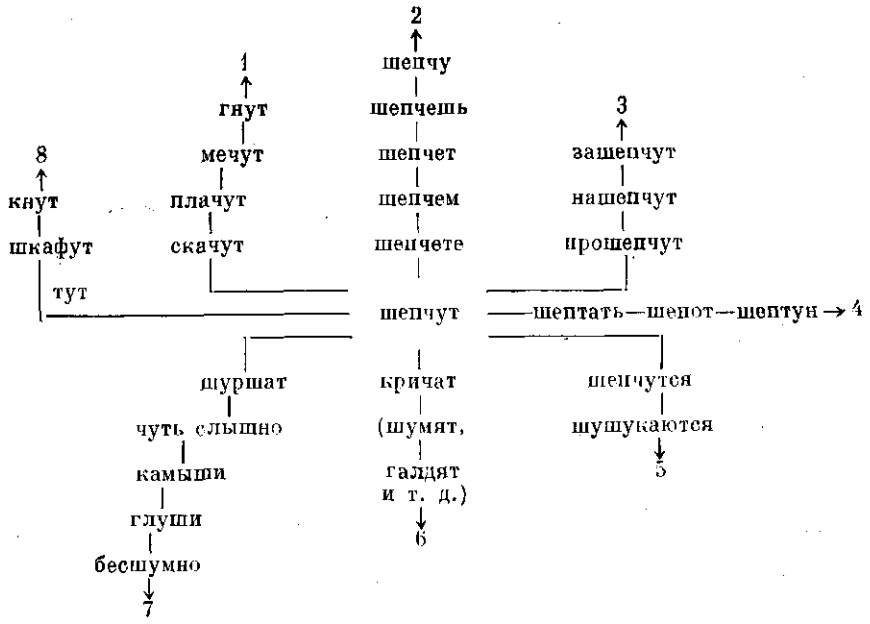


Рис. 6. Ассоциативные ряды словоформы *шепчут*.

ния лингвистической информации потребовалось бы астрономическое число элементов, на несколько порядков превышающее число нейронов, имеющихся в мозгу человека (ср.: George, 1961/1963, с. 435).[1]

[1] Последовательное насыщение каналов приема информации и ее последовательное укрупнение (chunk) в обобщенные единицы, которые записываются в лингвистической памяти с помощью некоторого сжатого кода, можно проиллюстрировать следующим примером.

Предположим, что каждая буква русского текста изображается набором из 200 черных точек. Тогда светочувствительные рецепторы нашего глаза при чтении шестибуквенного слова принимают 6·200=1200 дв. ед. информации (Steinbuch, 1965/1967, с. 140 и 299). Для кодирования одной буквы русского 32-х буквенного алфавита требуется всего лишь 5 дв. ед. информации (30). Поэтому при побуквенной записи информация, получаемая от рецепторов, сильно компрессируется: эта запись обходится нервной сети всего лишь в 5·6=30 дв. ед. Если же учесть, что в словарном (внетекстовом) русском слове на одну букву приходится около 2 дв. ед. информации (см.: Пиотровский, 1968, табл. 20, стлб. 27—28), то станет ясным, что для структурно-семантического кодирования шестибуквенного слова мозг использует не более 12 дв. ед.

Аналогичным образом для структурного кодирования китайского иероглифа требуется всего лишь 65 дв. ед., в то время как при его чтении глаз воспринимает, очевидно, многие тысячи дв. ед. информации (ср.: Chang, 1973, с. 257—265).

Из сказанного не следует, что мозг представляет собой систему, состоящую из фиксированного числа элементов. Еще в 30-х годах была сформулирована теория роста нервных элементов в связи с появлением новых нервных связей (Holt, 1931; Kappers, Huber, Crosby, 1936).

Чтобы понять смысл речевого сообщения, человек должен переходить от потенциальных (языковых, «словарных») значений знаков к их актуальным (текстовым) значениям. Если бы этот переход осуществлялся путем сканирования (т. е. последовательного перебора) всех парадигм, в которые входит каждый знак, то, учитывая небольшую скорость распространения нервного импульса (не более 120 м/сек.) и ограниченную скорость переработки информации мозгом (не более 45 дв. ед./сек.), декодирование самых простых текстов занимало бы несколько минут. Что же касается распознавания более сложных сообщений, то здесь расход времени на сканирующую дешифровку текста был бы гораздо большим. Так, например, распознание художественного эффекта словоформы *шепчут* в известном стихотворении Бальмонта «Камыши» (1909):

Полночной порою в
болотной глуши
Чуть слышно, бесшумно,
шуршат камыши.
О чем они шепчут?
О чем говорят?
Зачем огоньки между ними
горят? —

в котором эта словоформа взаимодействует с другими членами аллитерационной парадигмы *глуши—шуршат—шепчут*, потребовало бы у русского читателя несколько часов напряженных раздумий.[1]

Только ассоциативная организация нашей памяти, использующей многоканальный и многоуровневый эвристический поиск, в ходе которого нужные с точки зрения заданной стратегии сведения подкрепляются дополнительным возбуждением, а ненужные погашаются торможением, дает возможность человеку почти мгновенно распознавать смысл получаемых им речевых сообщений.

7. Переработка текста естественного языка на ЭВМ

Структура и функционирование памяти современной ЭВМ принципиально отличается от построения и работы человеческого мозга: информация, записанная в структурной единице (байте, слове) машины, не может быть передана одновременно тысячам

[1] В нашей лингвистической памяти живут самые неожиданные ассоциативные ряды слов. Например, Л. Н. Толстой писал в «Юности»: «. . .у нас с Володей установились . . . следующие слова с соответствующими понятиями: *изюм* означало тщеславное желание показать, что у меня есть деньги, *шишка* . . . означало что-то свежее, здоровое, изящное, но не щегольское. . .» (Толстой, 1966, с. 299).

В. Маяковский признавался, что уменьшительное *Сережа* (речь шла о Сергее Есенине) вело у него «за собой массу других. . . словечек: *ты*, *милый*, *брат* и т. д.» (Маяковский, 1952, с. 26).

других ячеек, как это имеет место в нервной сети. Машинная ячейка может одновременно принимать сообщение только от одной такой же ячейки, передавать также только одной ячейке. В связи с этим на современных универсальных ЭВМ не могут быть осуществлены ассоциативная запись и поиск информации. Эти машины решают предлагаемые им задачи, в том числе и лингвистические, на основе правил одноканальности передачи сообщения, его поэлементной запоминаемости и строгой последовательности выполняемых операций.

Рассмотрим указанные особенности переработки текста в машине на примере информационного процесса, заключающегося в том, что компьютеру передается для перевода на русский язык английское словосочетание (the) 20-th CENTURY FOX (XX CENTURY FOX) '20-й век Фокс (XX век Фокс)'.

Исходя из схемы ИП, показанной на рис. 1, можно разбить этот информационный процесс на четыре блока.

Первый блок состоит в порождении некоторым источником сообщения (автором) сегмента XX CENTURY FOX.

Второй блок, предусматривающий подготовку к вводу и ввод в ЭВМ указанного сегмента, представляет собой последовательность следующих операций.

1. Перенос текста с помощью специальных печатающих устройств на перфоленту, перфокарты, т. е. на такие носители информации, с помощью которых она может быть введена в машину.[1] Поскольку на перфоносителе, например на перфоленте, буквы текста кодируются с помощью сочетаний пробитых и непробитых позиций (ср. рис. 7), то в ходе рассматриваемой операции графическая субстанция, передающая содержание текста, заменяется «дырочной» субстанцией.

2. Перфоносители подаются в устройство ввода, где они пропускаются между источником света и набором фотоэлементов, расположенных таким образом, чтобы каждая позиция перфоленты или перфокарты была соотнесена с экраном одного определенного фотоэлемента. Если в данной позиции перфоноситель имеет пробивку, то луч света, пройдя через нее, падает на экран соответствующего фотоэлемента. Если пробивки нет, фотоэлемент не получает светового сигнала. Таким образом, при осуществлении этой операции «дырочная» субстанция заменяется световой: каждая буква текста теперь кодируется определенной комбинацией световых сигналов (рис. 8).

3. Фотоэлемент, на экран которого упал луч света, отправляет в оперативную память машины (рис. 8) электрический импульс. Фотоэлемент, не получивший светового сигнала, электрического импульса не выдает. Наш текст опять перекодируется в новую субстанцию: каждая буква оказывается записанной теперь с помощью особой комбинации электрических импульсов.

[1] Мы не касаемся здесь возможности записи текста прямо на МЛ, или ввода текста в ЭВМ с дисплея, или, наконец, через оптические читающие устройства.

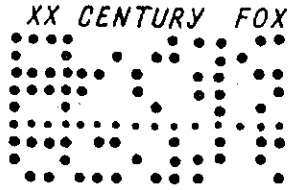
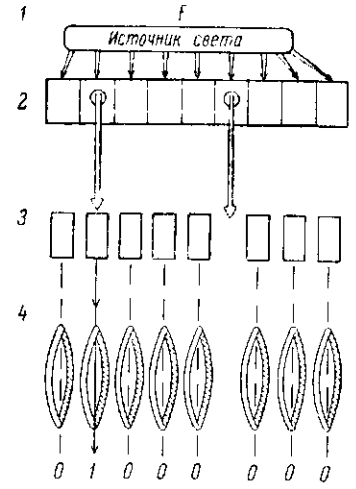


Рис. 7. Запись английского сегмента XX CENTURY FOX на восьмипозиционной перфоленте.



Написание (графика).

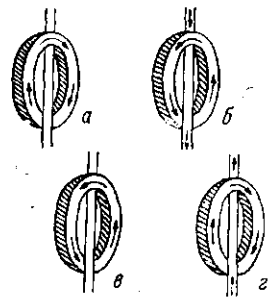
Фотоэлементы.

Машинный код буквы F, записанный в один байт оперативной памяти ЕС-ЭВМ на ферритовых кольцах.

Рис. 8. Последовательные преобразования буквы F из графической субстанции (*1*) в перфорационную (*2*), световую (*3*) и электромагнитную (*4*) субстанции.

Условные обозначения: ⟹ — пучок света; ⟶ — импульс тока от фотоэлемента на обмотку ферритового кольца; — — — — — — отсутствие импульса тока; 1 — ферритовое кольцо намагничено; 0 — ферритовое кольцо не намагничено.

4. Электрический импульс от фотоэлемента передается на обмотку определенного ферритового кольца (множество этих колец и образует оперативную память компьютера). Каждое ферритовое кольцо имеет два устойчивых магнитных состояния: одно соответствует нулю, другое — единице. Каждое состояние соответствует отсутствию или наличию тока в обмотке кольца, а также зависит от направления этого тока. Предположим, например, что кольца, предназначенные для записи буквы F, находятся в нулевом состоянии (рис. 9, *а*). Поступающий от второго фотоэлемента импульс тока перебрасывает второе кольцо в противоположное состояние намагниченности (рис. 8 и 9, *б*), условно соответствующей единице, остальные же семь колец, не получившие импульса, сохраняют нулевое состояние (рис. 9, *г*).



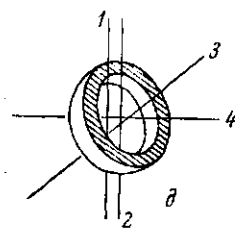


Рис. 9. Перемагничивание ферритового кольца.

а — кольцо намагничено по часовой стрелке (нулевое состояние), тока нет; *б* — подан ток в направлении сверху вниз, кольцо намагничивается против часовой стрелки (состояние «единица»); *в* — ток сброшен, кольцо находится в состоянии намагниченности «единица»; *г* — подан ток в направлении снизу вверх, кольцо перемагничивается в нулевое состояние; *д* — ферритовое кольцо, нанизанное на четыре провода в плате памяти.

Таким образом, наш текст оказывается выраженным в электромагнитной субстанции: каждая его буква воплощается в комбинации колец, находящихся в нулевом состоянии намагничивания и в состоянии «единица» (рис. 8).

Третий блок описываемого ИП представляет собой реализацию машинного семиозиса, включающего распознавание и интерпретацию смысла принятого машиной сообщения.

Машинный семиозис строится в виде последовательности следующих операций.

1. Записанный на ферритовых кольцах текст с помощью разного рода последовательных промежуточных электромагнитных преобразований [1] поэлементно сравнивается с уже записанными в памяти ЭВМ лексическими единицами, с тем чтобы получить необходимые для дальнейшей работы сведения. Так, например,

[1] Дальнейшие смены состояний намагничивания кольца осуществляются следующим образом.

Если ток выключается, то кольцо остается в прежнем намагниченном состоянии (рис. 9, *а*, *в*). Последующий ток, имеющий силу не менее 0.5 ампера, либо вовсе не влияет на кольцо, если он пропущен в том же направлении, что и первый импульс (рис. 9, *б*), либо перебрасывает феррит в противоположное состояние намагниченности, если он пропущен в обратном направлении (рис. 9, *г*).

Через каждое кольцо пропущено четыре провода (рис. 9, *д*). Первые два устанавливают ферриты в состоянии 0 или 1. Необходимо также уметь определять, в каком состоянии находилось раньше кольцо. Для определения предыдущего состояния феррита на его обмотку подается ток для перевода кольца в состояние 0. Если оно уже находилось в состоянии 0, изменений не произойдет. Если же феррит был в состоянии 1, его переброска в состояние 0 вызовет возникновение небольшого тока, который можно обнаружить в проводе считывания (провод *3* на рис. 9, *д*).

Четвертый провод, пропущенный через кольцо, называется проводом запрета или проводом записи. Этот провод упрощает занесение информации на кольца (Germain, 1967/1971, с. 57—59).

словоформа fox должна быть сопоставлена сначала со списком ядерных словоформ, образующих непереводимые пословно словосочетания (обороты). Последовательно сканируя этот список, ЭВМ в конце концов обнаруживает, что электромагнитные коды текстовой буквенной последовательности fox полностью совпадают с кодами имеющейся в списке словоформы fox. Такое совпадение свидетельствует о том, что анализируемая единица текста опознана машиной и соотнесена с заложенным в ней описанием языка.

2. Извлекается информация, стоящая при словарной лексеме fox. Согласно этой информации, fox может образовывать обороты двух типов: двухсловный оборот, в котором fox стоит на втором месте (ср.: old fox 'хитрец', slow fox 'медленный фокстрот'), и трехсловный оборот, в котором fox занимает первое или последнее место (ср.: fox and geese 'волки и овцы', XX century fox 'XX век, Фокс', play the fox 'хитрить').

3. С указанными оборотами посимвольно сравнивается текстовой сегмент XX century fox. Сравнение показывает полное совпадение текстового сегмента с предпоследним оборотом.

Четвертый блок представляет собой реакцию компьютера на принятое им сообщение. Здесь имеют место две операции.

1. Выбор русского эквивалента для оборота 20 th Century Fox и передача его на печатающее устройство (или дисплей, или абонентский пункт).

2. Перевод русского эквивалента из электромагнитной субстанции в графическую запись, которая и выдается через АЦПУ читателю (рис. 10).[1]

В тех случаях, когда автомат переводит не идиому, но свободное словосочетание, например the fox has caught a cock 'лисица поймала петуха', проверка на наличие этого сегмента в словаре оборотов дает отрицательный ответ. Тогда машина, пользуясь двуязычным автоматическим словарем, осуществляет пословный перевод. Если словарь дает для отдельных слов несколько эквивалентов (ср.: fox 'лиса', 'лисица', 'первокурсник', 'хитрить', 'обманывать' и т. д.; cock 'петух', 'кран', 'флюгер', 'курок', 'кубрик', 'вожак' и т. д.), то вслед за программой пословного

[1] Многократное перекодирование текста естественного языка при его обработке автоматом еще раз говорит о том, что лингвистические структуры могут воплощаться в различных субстанциях. Действительно, в какой бы субстанции — фонетической, графической, световой или электромагнитной — ни был воплощен английский или русский текст, он всегда остается английским или, соответственно, русским текстом: английский текст должен анализироваться английской, а русский — русской грамматикой, также независимо от того, воплощены ли эти грамматики в звуковой, графической, «дырочной», электромагнитной или химической субстанции.

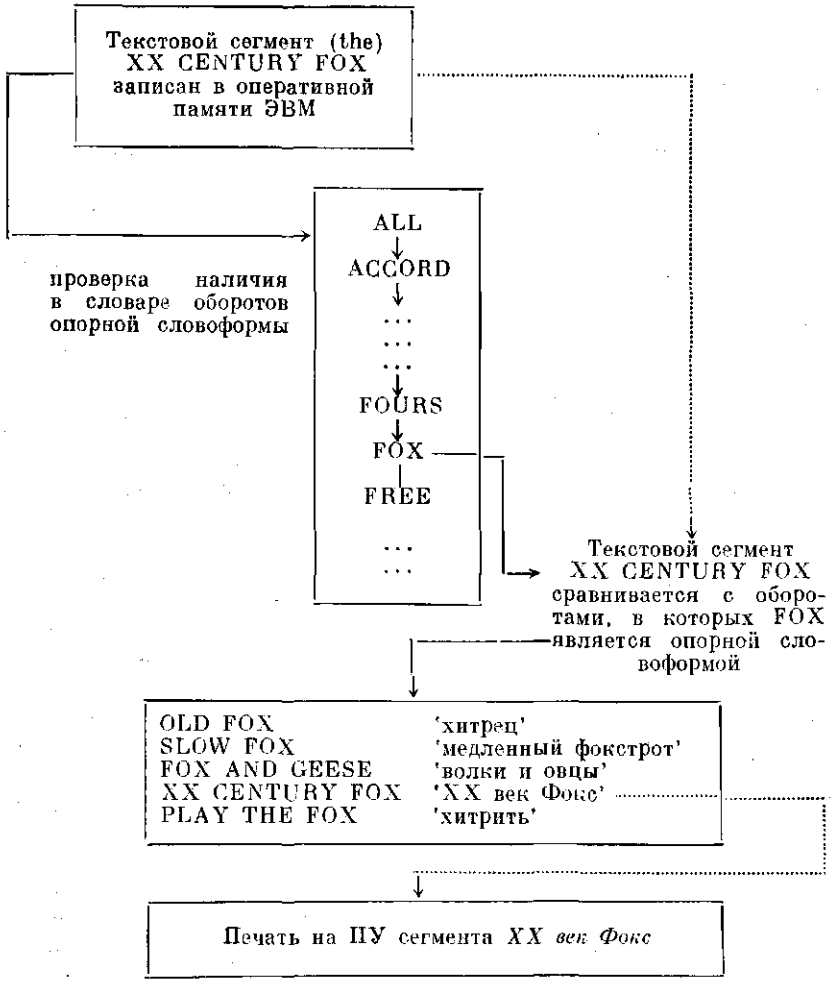


Рис. 10. Схема МП английского текстового сегмента XX CENTURY FOX.

перевода включаются программы устранения многозначности, опирающиеся на исследование комбинаторики словоформ (см. **21—23, 62**). После этого работают морфологические и синтаксические программы, поочередно осуществляющие грамматический анализ английского текста. Наконец, включается программа русского грамматического синтеза, которая, используя информацию, полученную в результате грамматического анализа, произведет морфолого-синтаксическое упорядочение русских лексем, выбранных из двуязычного словаря.

Предположим теперь, что компьютер должен не только провести лексико-грамматический разбор уже цитировавшегося стихотворения Бальмонта, но и осуществить его стилистический анализ. В этом случае после программ фразеологического, лексического, семантического, морфологического и синтаксического анализа будут введены в действие стилистические программы, которые будут последовательно просматривать текст, используя при этом каждый раз один определенный стилистический критерий. Так, например, чтобы обнаружить в тексте нашего стихотворения скопление шипящих, которое должно, по замыслу автора, передавать коннотативный эффект ночного шелеста камышей, автомат должен установить сначала частоту каждой буквы и ее распределение по словам текста. Если буква *Ш* или какая-либо другая буква покажет заметное отклонение этих характеристик от типичного для нее эталона, заложенного в памяти ЭВМ, это будет означать, что данная буква или группа букв выполняют в анализируемом тексте особую стилистическую функцию (ср. **29**). Автомат сможет ответить и на вопрос о том, какую именно стилевую функцию выполняет скопление указанных букв. Для этого он должен будет после выполнения предшествующей программы сканировать заранее введенную в него парадигму стилевых функций, каждая из которых связана со скоплением определенных букв или их комбинаций.

Чем тоньше и сложнее лингвистическая задача, тем большее число языковых парадигм и текстовых дистрибуций вовлекается в поиск. Если эта сложная задача решается автоматом, то ассоциативный пучок информации, характеризующий каждую лингвистическую единицу, расчленяется и преобразуется в цепочку парадигм и дистрибуций, последовательно записанных в памяти ЭВМ.

Сегодняшний компьютер лишен интегративных функций и способностей работать с проблемной ситуацией, характеризующих мозг человека (Лурия, 1974; Miller, Galanter, Pribram, 1960/1965, с. 155). В частности, при переработке текста одновременный, подчиненный его смысловой стратегии поиск заменяется в машине последовательным перебором, имеющим очень слабую целенаправленность.[1]

При этом быстродействие ЭВМ, являющееся ее основным преимуществом перед человеком, растрачивается на сканирование материала. Медленно работающий мозг благодаря использованию целенаправленного ассоциативного поиска избегает в этих случаях непроизводительных затрат и решает сложную семантико-синтаксическую задачу скорее, чем это делает современная ЭВМ.

[1] Целенаправленность поиска лингвистической информации в современных серийных ЭВМ ограничивается обычно тем, что просматривается не весь массив сведений, заложенный в том или ином эталоне, а лишь часть его, отмеченная некоторым формальным признаком. Например, при отождествлении словоформы fox просматривается не весь английский словник, но лишь та его часть, которая охватывает слова, начинающиеся на fo (foal, foam, fob, focal и т. д.), или только массив трехбуквенных слов (Вертель, Вертель, Крисевич, Пиотровский, Трибис, 1971, с. 95 и сл.). О других приемах целенаправленного поиска см.: Крисевич, 1970, с. 321—330.

Напротив, когда речь идет о «рутинных» лингвистических задачах, т. е. о таких задачах, в ходе решения которых не возникает проблемных ситуаций и которые предусматривают поиск и упорядочение лингвистических объектов по одной, максимум двум-трем парадигмам, машинное алгоритмическое решение дает более точный и быстрый результат, чем ассоциативный поиск, осуществляемый человеком.

8. Уровни восприятия текста

Чтобы прогнозировать развитие инженерной лингвистики и определить в связи с этим стратегию исследований на ближайшее десятилетие, а также выбрать оптимальную технологию при составлении конкретных алгоритмов и программ, нельзя ориентироваться только на современное состояние кибернетической техники и замыкаться в узкие рамки конкретных лингвистических и нейрофизиологических знаний. Здесь необходим более широкий подход, с одной стороны, учитывающий перспективы и реальные направления развития вычислительной техники, а с другой — ориентирующийся на методологические вопросы кибернетики, языкознания и нейрофизиологии.

Приступая к обсуждению этих вопросов, начнем с проблемы порождения текста и различных уровней его восприятия.

Текст формируется в результате взаимодействия порождающей системы языка, его нормы и узуса, с одной стороны, и независимой от языка внешней ситуации — с другой. Эта внешняя ситуация включает с о б ы т и е, описываемое текстом, и о б щ и й с и т у а т и в н ы й к о н т е к с т (ср.: Germain, 1972, с. 117и сл.), который может смыкаться с узусом употребления языка (рис. 11).

При распознавании смысла сообщения его живой получатель также опирается на знание системы языка, его нормы и узуса, а также на знакомство с ситуацией.

Если полное владение системой, нормой и узусом языка сочетается у получателя сообщения с хорошим знанием и целенаправленной прагматической оценкой ситуации, в которой родилось высказывание, получатель может осуществлять полное или г л о б а л ь н о е распознавание смысла текста.

В том случае, когда получатель сообщения в ситуации не ориентируется, он способен осуществить лишь л и н г в и с т и ч е с к о е распознавание его смысла. Вспомним в этой связи шуточный вариант отрывка из «Полтавы», якобы полученный однажды путем переложения на русский язык немецкого перевода пушкинской поэмы:

Был Кочубей богат и горд.
Его поля обширны были,
И очень много конских морд,
Мехов, сатина первый сорт
Его потребностям служили.
(Чуковский, 1968, с. 57)

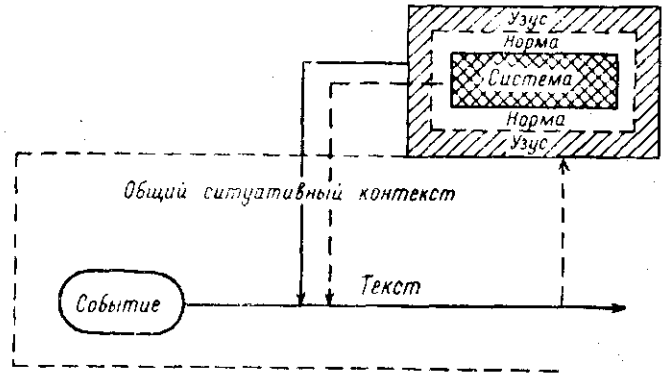


Рис. 11. Схема текстообразования.

Незадачливый переводчик (если он только действительно существовал) воспринял немецкий текст вне связи с исторической ситуацией и литературной традицией.

Оценка общего ситуативного контекста имеет решающее значение при распознавании смысла таких текстов, в которых отправитель сообщения говорит совсем не то, что он действительно думает. Чисто лингвистическое восприятие такого сообщения — например, прозы Достоевского, Хемингуэя или Джойса — выполняет лишь роль стимула для ситуативного распознавания подлинного смысла художественного произведения.

Степень глобального понимания сообщения часто определяется коллективной или индивидуальной прагматикой его потребителя. Так, например, текст газетной рекламы сигарет «Уинстон»: bad grammar, good taste 'плохая грамматика, хороший вкус' едва ли будет понятен некурящему американцу, который не следит за рекламами табачных изделий и не помнит предыдущего рекламного текста: Winston tastes good, like a cigarette should 'у сигареты «Уинстон» такой приятный вкус, какой должна иметь сигарета', вызвавшего нарекания туристов из-за нарушения грамматических норм, которое состоит в употреблении like вместо as (Швейцер, 1973, с. 240).

Теперь рассмотрим уровни лингвистического восприятия сообщения.

Низшими ступенями лингвистического восприятия текста являются:

а) грамматический уровень, на котором распознается только грамматическая схема предложения (ср. щербовскую *глокую куздру* или псевдоамериканский текст . . . *Ди вемп оф совена Джи-эй*. . , у Маяковского (1952, с. 12));

б) словарный уровень, на котором воспринимаются только лексические значения словоформ (ср.: *председатель*, *комитет*, *я*, *этот*, *записка*, *писать* — Пешковский, 1956, с. 53).

Более высокой ступенью словарного распознавания является фразеологический уровень восприятия, на котором узнаются значения целых словосочетаний, но не учитываются грамматические связи между этими словосочетаниями (ср. рис. 58).

Следующей ступенью является лексико-грамматическое распознавание текста, представляющее собой обобщение первых трех уровней и воплощающее ту лингвистическую способность человека, принимающего сообщение, которую Н. Хомский называет компетенцией (Chomsky, 1968/1972, с. 15, 34). Однако это распознавание не дает еще полного лингвистического восприятия текста.

Дело в том, что, распознав значение отдельных слов и словосочетаний, а также выявив грамматическое построение текста, человек может не уловить дальних семантико-синтаксических (в том числе и стилистических) связей. Примером такого неполного лингвистического распознавания смысла текста может служить русское предложение: *В этих случаях машина может оказаться не более беспомощной, чем оказываюсь я, человек, пытаясь понять, скажем, женскую ревность* (РО, с. 231), представляющее собой перевод английской фразы: In such cases the machine is no more handicapped than I am, being a man, in trying to understand, say, female jealousy (Recognizing Patterns. Studies in Living and Automatic Systems. The MIT Press, Cambridge Mass. and London, 1968, с. 183).

Переводчик, увлеченный изложением проблемы человек-машина, не заметил, что семантическая связь machine-man перекрывается более сильной семантической антитезой man-female, оттесняющей в данном тексте основное значение слова man 'человек' и актуализирующее его второе значение 'мужчина'. Не уловив этих семантико-стилистических отношений, переводчик вместо полной лингвистической интерпретации английского оригинала: «. . . я, *мужчина, пытающийся* понять. . . женскую ревность. . .» дает неполное, хотя и правильное с лексико-грамматической точки зрения его понимание — «. . . я, *человек, пытающийся понять женскую ревность. . .*».

Учет дальних смысловых и синтаксических связей позволяет нам подняться на семантико-синтаксический уровень распознавания содержания сообщения, — уровень, соответствующий, очевидно, тому знанию языка, в основе которого лежит полная языковая компетенция (ср.: Chomsky, 1968/1972, с. 72).

Однако следует помнить, что сообщение может быть неточно сформулировано или неверно понято в связи с тем, что получателю остались неизвестными некоторые нормативно-узуальные аспекты употребления отдельных лексико-грамматических единиц.

Например, москвич, едущий ленинградским трамваем в направлении Парка Победы, вероятно, не придает особого значения сообщению вагоновожатого о том, что *вагон идет в парк*. Между тем слово *парк* употреблено здесь в чисто «ленинградском» узуальном значении — 'трамвайное депо'. Аналогичным образом автомобилист англичанин, попадая в американский город, воспринимает надпись: cars must be kept on the pavement 'стоянка разрешена только на проезжей части' как требование парковать автомобиль на тротуаре, поскольку в Англии pavement имеет значение не 'проезжая часть', но 'тротуар' (Cherry, 1965/1972, с. 94—95).

Итак, настоящее лингвистическое восприятие сообщения достигается тогда, когда семантико-синтаксический уровень распознавания сочетается с узуально-нормативным аспектом его восприятия.

Переход от лингвистического к глобальному уровню понимания требует, как уже говорилось, учета всей широкой ситуации, в рамках которой осуществляется передача сообщения. Эта ситуация включает наряду с широким контекстом, в котором происходит описываемое событие, оценку личности собеседника и структуры его мнений (Abelson, Carroll, 1965, с. 24—30; Weizenbaum, 1968/1970, с. 223—230; ср.: Schank, 1972a), а также такие нелингвистические каналы коммуникации, как ритм речи, интонация, выражение лица и телодвижения говорящего. Именно на этом уровне обнаруживается «безэнтропийный» характер языка, отличающий его от систем физического мира (Кобозев, 1971).

В настоящее время созданы работающие алгоритмы, которые позволяют ЭВМ распознавать текстовое сообщение на грамматическом, словарном, фразеологическом, а отчасти на лексико-грамматическом и даже семантическом уровнях (см. гл. 9).

Возникает вопрос, какой уровень понимания текста следует считать будущим пределом возможностей автомата: лексико-грамматический, лингвистический или ЭВМ сможет достичь глобального «человеческого» уровня восприятия текста? Применительно к машинному переводу этот вопрос формулируется следующим образом: сможет ли машина переводить иностранный текст так, как это делает квалифицированный переводчик, обладающий достаточными специальными познаниями в той области, к которой относится переводимый текст, или выдаваемый компьютером перевод никогда не будет высококачественным и будет всегда нуждаться в дополнительной интерпретации (постредактировании) со стороны читателя?

9. Проблемы искусственного интеллекта

Вопрос о глобальном распознавании автоматом смысла сообщения, — распознавании, предусматривающем целенаправленную оценку ситуации и принятие решения, тесно связан с проблемой искусственного интеллекта и моделирования психических процессов. Вокруг проблемы машинного мышления шли и продолжают идти оживленные споры, полярными пунктами которых служат две точки зрения.

Первая точка зрения, опирающаяся на неокартезианский тезис о том, что человек — это кибернетическая машина [1] высокой организации с заложенным в ней генетическим алгоритмом самообучения, исходит из предпосылки, что в основе мышления (а вместе с ним и языка) лежит жесткая детерминированная структура. Поэтому, как утверждают сторонники этой точки зрения, пределов формализации и машинного моделирования человеческого мышления не существует (Ивахненко, 1968, с. 31—34; Mac Culloch, 1956/1960, с. 61—67; Chomsky, 1968/1972, с. 17—34; Steinbuch, 1965/1967, с. 427—445; Saterland, 1968/1971, с. 33 и др.).

Вторая точка зрения реализует виталистическое представление о том, что мыслительные (в том числе и ментально-речевые) процессы и социальный интеллект принципиально не алгоритмизуемы и поэтому не могут быть перенесены в автомат (Гансовский, 1968, с. 101—105; Krämer, 1958, с. 511—522; Taube, 1961/1964, с. 58 и др.).

Не останавливаясь на деталях этих споров (см.: Полетаев, 1971), можно выделить два результирующих положения этих дискуссий.

С одной стороны, виталистам не удалось привести сколько-нибудь серьезных методологических аргументов, отрицающих алгоритмические возможности человеческого мышления и опровергающих идею создания машинных программ, имитирующих и моделирующих некоторые мыслительные процессы. Выяснилось также, что утверждение о принципиальной невозможности воспроизведения на ЭВМ мыслительных процессов не вытекает из приводимых обычно виталистами философских доводов (Баженов, 1964, с. 338; Братко, 1969, с. 38, 42—51).

С другой стороны, оказалось, что с моделированием психики и построением «машинного интеллекта» связаны парадоксы, являющиеся следствием известной теоремы Гёделя о неполноте. Некоторые из этих парадоксов воплощаются в тех технологических трудностях, которые возникают на пути создания алгоритмов, моделирующих отдельные рече-мыслительные процессы.

Виталистические опровержения возможности построения машинного интеллекта мало что дают для определения стратегий инженерной лингвистики. Зато вытекающие из теоремы Гёделя парадоксы, которые, как выясняется, связаны с некоторыми лингвистическими антиномиями, представляют для нас первостепенный интерес (Piotrowski, 1972, с. 155—156).

Начнем с рассмотрения математических парадоксов.

Вторая теорема Гёделя (теорема о неполноте) формулируется обычно так: если L является ω-непротиворечивой и адекватной арифметической логикой, то L неполна (Gödel, 1931, 173—198).

[1] Под термином «машина» здесь имеется в виду не только автомат, выполняющий заранее предусмотренные человеком операции, но и такие автоматические устройства, которые способны ставить перед собой задачу и находить необходимые способы ее решения.

Более обобщенно это означает, что если имеется достаточно богатая формальная и непротиворечивая система L, то не существует доказательства ее непротиворечивости, осуществленного средствами, формализуемыми самой системой L. При этом в системе L всегда существуют неразрешимые предложения F, т. е. такие предложения, справедливость или ложность которых нельзя алгоритмически доказать средствами системы L (доказательство этих положений см.: Kleene, 1952/1957; Tarski, 1930, с. 210—211; ср.: Hartmanis, Hopcroft, 1970/1971, с. 376). Противоречивость или непротиворечивость этих предложений можно доказать, построив другую, более мощную систему L'. Однако в ней появятся новые неразрешимые предложения F', которые снова нужно анализировать с помощью более сильной системы L'', и т. д. Таким образом, становится ясным, что далеко не для каждого класса задач можно построить в рамках данной логической системы разрешающий машинный алгоритм (Трахтенброт, 1974, с. 60—65).

Все эти соображения были истолкованы некоторыми кибернетиками как доказательство ограниченности математических способностей машины сравнительно с математическими возможностями человека. Так, например, Э. Нагель и Дж. Ньюмен указывали, что современные ЭВМ содержат конечное число заложенных в них команд, которые соответствуют фиксированным правилам вывода формализованной аксиоматической процедуры. Эти машины дают ответы на вопросы, функционируя согласно заложенным в них программам. Однако, как показал Гёдель в своей теореме о неполноте, существует бесконечное число проблем элементарной теории чисел, решение которых невозможно никаким данным аксиоматическим методом и на которые машины не смогут дать ответа, как бы совершенны ни были их конструкции и с какой бы скоростью они ни выполняли операции. Человеческий мозг, возможно, так же ограничен, но теорема Гёделя указывает, что структура и деятельность человеческого ума являются более сложными и тонкими, чем строение и функционирование любой машины, какую мы можем себе вообразить в настоящее время (Nagel, Newman, 1958).

Однако такое толкование Гёделя исходит из спорной посылки, заключающейся в том, что те математические операции, которые не может выполнить компьютер, может произвести человек, оперирующий теми же понятиями истины, что и машина. В действительности как для человеческого мозга, решающего формальные задачи, так и для автомата присущи одни и те же математические ограничения. Эти ограничения «присущи любой системе, поведение которой упорядочено и подчинено определенным законам. Система, которая могла бы обойти эти ограничения, обладала бы волшебными свойствами» (Эшби, 1968, с. 34—35).

Итак, если отказаться от агностических представлений о том, что психика и поведение человека управляются непознаваемой энтелехией, и твердо встать на материалистические позиции по-

знаваемости процессов мышления (Анохин, 1973, с. 83), то придется признать, что не существует методологических барьеров, принципиально разделяющих «естественный» и «искусственный» интеллекты, при решении математических задач. «Трагедия» проблемы создания «искусственного разума» состоит в том, что мышление человека, хронологически находясь впереди создаваемого им математического аппарата, всегда оказывается в данный момент богаче тех дедуктивных форм, которые должны быть заложены в автомат, реализующий этот искусственный разум. Следствием того, что человеческий интеллект опережает возможности его формализации, и является неспособность ЭВМ к самостоятельной инициации творческого акта (ср.: Смолян, 1973; Тихомиров, 1975).

Теперь обратимся к вопросу о возможности глобального понимания текста со стороны кибернетического автомата. Этот вопрос мы будем рассматривать с точки зрения некоторых антиномий и парадоксов современного языкознания.

10. Индивидуальный и социальный аспекты лингвистического знака

Объекты внешней среды, являясь каждый источником громадного количества информации, образуют в совокупности непрерывный континуум. Искусственные и естественные кибернетические системы при восприятии и описании этого континуума разбивают его на дискретные области (ср. 55).[1] При этом каждая область описывается не всем множеством характеризующей ее информации, но некоторыми существенными, инвариантными признаками, способными выполнять мнемоническую функцию. Этот принцип лежит в основе отражения внешнего мира не только у человека, но также и в искусственных кибернетических системах. На этом же принципе «квантования» непрерывной объективной действительности («субстанции содержания») построен семиозис во всех знаковых системах передачи информации, в том числе в естественном языке (Hjelmslev, 1953/1960, с. 307 и сл.).

Квантование всегда связано с необратимой потерей более или менее существенной информации, содержащейся в описываемом объекте. В современных дискретных кибернетических системах процесс распознавания вызывает разрушение большей части информации, содержащейся в объекте: автомат удерживает в памяти лишь те признаки объекта, которые оказываются существенными с точки зрения выполняемой машиной задачи. Человек,

[1] Эта разбивка непрерывного потока информации на дискретные кванты определяется, в частности, апперцепционным свойством нашего сознания, состоящим в том, «что отдельные содержания сознания сливаются, если разность времен их поступления бывает ниже определенной минимальной необходимой величины. Этот временный квант составляет (у здорового человека) примерно 1/16 сек. или около 65 мсек. . .» (Steinbuch, 1965/1967, с. 301).

как и другие организмы, являясь многоцелевой системой, в которой очередность появления той или иной задачи заранее не определена, стремится удержать максимум сведений об объекте. В частности, выясняется, что на сенсорном уровне приема, т. е. на уровне ультракороткой памяти (ср.: Wickelgren, 1971, с. 67—69; Лурия, 1974, с. 11), дискретная запись и переработка сведений сочетаются с работой аналоговой (непрерывной) системы, обеспечивающей максимальную сохранность информации (Siebert, 1968/1970, с. 134 и сл.).

Продвижение на более высокие — синтетические уровни распознавания и запечатления материала (Norman, 1969; Лурия, 1974, с. 11—12), в ходе которого познавательные механизмы высшей нервной деятельности осуществляют избирательную фильтрацию, квантование и обобщение получаемых от рецепторов сведений в укрупненные информационные единицы (Nauta, 1972, с. 140—156), связано с потерей большого объема сведений, полученных от объекта на сенсорном уровне (George, 1961/1963, с. 417 и сл.; Steinbuch, 1965/1967, с. 299—301, ср. также с. 22, примеч. 1).

Наиболее интенсивно процесс активной интеграции и квантования сведений, сопровождающийся отбрасыванием побочного информационного шума, происходит на словесно-логическом уровне распознавания, — уровне, на котором осуществляется формирование индивидуальных «биолингвистических» знаков. Однако, несмотря на многократное уплотнение и селекцию сведений, предшествующую формированию языкового знака, в последнем заключено много сведений об объекте: ассоциативная форма их записи (см. 6) позволяет включить в содержание знака не только основные сигнификативные и денотативные значения, но также сведения о дополнительных ассоциативных рядах, на пересечении которых находится данный знак, — сведения, которые обобщаются в его коннотативной информации (ср. рис. 2 и 6).

Процесс интеграции сведений не останавливается на уровне словесно-логической памяти отдельного индивидуума: если взглянуть на проблему не с бионической, но с более широкой, семиотической позиции, нетрудно заметить, что следующий этап обобщения и отбора информации переносится с индивидуального психо- и физиолого-лингвистического уровня на социолингвистический уровень.

В социолингвистическом знаке удерживается та денотативная, сигнификативная и коннотативная информация, которая содержится во всех обобщаемых им лингвистических знаках психофизиологического уровня.

Из сказанного следует, что, говоря о глобальном понимании текста, надо последовательно различать, с одной стороны, его распознавание отдельным человеком, опирающееся на индивидуальную интерпретацию системы языка и субъективно смещенную выборку из совокупности знаний и идей социального кол-

лектива (Moles, 1967/1973, с. 53), а с другой — коллективное распознавание, опирающееся на полную систему знаний, идей и социолингвистических знаков, — язык в полном смысле этого слова.[1]

11. «Идеальный» лингвистический автомат и его парадоксы

Кибернетический автомат, способный к глобальному, «человеческому» распознаванию текста, не может, естественно, имитировать лингвистическое поведение только одного конкретного человека, познавательный и языковой опыт которого четко ограничен его индивидуальностью и верхней временной границей его жизни. Лингвистический компьютер будущего должен вобрать в себя интеллектуальный, культурно-социальный и лингвистический опыт некоторого коллектива — народа, социально-профессиональной группы и т. п. Короче говоря, наш компьютер должен выступать в роли некоторого homo sapiens universalis et immortalis, владеющего языком на социальном уровне. Однако построение такого автомата связано с рядом антиномий. Наиболее важными из них являются парадокс дискретности и непрерывности в речевой деятельности человека, а также так называемый парадокс Ахиллеса и черепахи, в котором преломляется соссюровская антиномия синхронии и диахронии.

Рассмотрим сначала парадокс дискретности и непрерывности. Коллективная речевая деятельность определяется не только схемой: «язык — норма—узус—событие—общий ситуативный контекст—текст» (ср. 8, рис. 11), но также временными (диахрония), территориальными (диалектология), социальными (стилевая вариативность) характеристиками.

Принято считать, что схема: «язык—ситуация—текст» имеет в основном дискретную природу. Вместе с тем следует иметь в виду, что при обобщении индивидуальных лингвистических знаков в социолингвистический знак у последнего не формируется четких семантических границ. Семантические поля одной лингвистической единицы обычно наплывают на поля соседних единиц (ср.: Zadeh, 1973/1974, с. 14; Бектаев, Пиотровский, 1973, с. 122—131).

Что же касается остальных лингвистических векторов, то к ним приложима не столько дискретная, сколько непрерывная мера. В частности, выясняется, что даже диахронические скачки — быстротекущие лингвистические процессы — описываются непрерывно возрастающими или убывающими монотонными кривыми вида:

$$p = \frac{A}{\pi} \operatorname{arctg}\left(\frac{t-\tau}{100}\right) + B \qquad (1)$$

и

$$1 - p = \frac{A}{\pi} \operatorname{arcctg}\left(\frac{t-\tau}{100}\right) + B, \qquad (2)$$

где p — вероятность употребления развивающейся лингвистической категории в момент t; $1-p$ — вероятность отступающей категории: A, B, τ — параметры формулы. С помощью выражения (2) описывается, например, характерное для конца XIX и начала XX в. отступление форм родительного падежа множественного числа русских существительных *вольтов*, *радианов*, *рентгенов*, а формула (1) моделирует распространение современных форм родительного падежа множественного числа типа *вольт*, *радиан*, *рентген* (Граудина, 1964, с. 215; Пиотровская, Пиотровский, 1974, с. 366—374). С помощью непрерывных функций должны описываться также диалектные и стилистические явления в языке и речи (Будагов, 1974; Пиотровский, 1960, с. 8—15; Piotrowski, 1964, с. 106; Piotrowski, 1973; Филин, 1966).

Таким образом, в глобальном распознавании и порождении текста участвуют не только дискретные механизмы, опирающиеся на двухпозиционную логику (альтернатива «истинно—ложно»). С помощью такой логики можно анализировать и синтезировать только лексико-грамматическую структуру текста, однако она не пригодна для описания семантических и диахронических процессов, а также диалектных и стилистических коннотатов. Для описания этих аспектов языка и речи должен быть применен аппарат вероятностной логики, в которой основными элементами являются не категорические ответы «истинно» или «ложно», а оценки вероятности того или другого из этих ответов (ср.: Zadeh, 1973/1974). Отсюда следует, что полное глобальное распознавание и порождение текста не может быть осуществлено на ЭВМ-III, использующих аппарат двухпозиционной логики: оно должно выполняться на «гибридных» дискретно-аналоговых автоматах большой мощности (ср.: Edmundson, 1971, с. 110). Хотя в настоящее время и ведется лабораторная разработка разных видов искусственных нейронов с суммирующими устройствами и изменяемым порогом, до создания промышленных образцов пока еще очень далеко.

Теперь рассмотрим второй парадокс — так называемый парадокс Ахиллеса и черепахи. В этом парадоксе, как уже говорилось, обобщается соссюровская антиномия синхрония—диахрония и противоречие между открытым характером естественного

[1] Напомним в этой связи следующее утверждение Ф. де Соссюра: «. . . язык — это . . . система, потенциально существующая в каждом мозгу или, лучше сказать, в мозгах целой совокупности индивидов, ибо язык не существует полностью ни в одном из них, он существует в полной мере лишь в массе» (Saussure, 1922/1933, с. 38). О соссюровской интерпретации соотношения психофизиологического и социального аспектов языка см.: Körner, 1973, с. 140 и сл.

языка,[1] опирающимся на способность человеческого сознания к неограниченному отбору и целенаправленному ассоциированию получаемой информации вместе с эвристическим и закрытым построением лингвистического машинного алгоритма. Действительно, закрытое описание языка для ЭВМ всегда строится на основе того синхронного среза, который совпадает с началом разработки этой формализации. На алгоритмизацию и программирование этого описания уходит обычно несколько лет, а то и десятилетий. За это время подверженная действию новых ситуаций и внутренних диахронических процессов открытая система языка успевает, подобно зеноновской черепахе, уйти вперед от исходного синхронного среза. В результате заложенный в ЭВМ алгоритм оказывается несколько устаревшим и не способным осуществить стопроцентную переработку тех текстов, которые порождаются новыми ситуациями и вновь образованными единицами системы и нормы.

Преодолеть парадокс Ахиллеса и черепахи и обеспечить глобальное распознавание и порождение текста можно лишь в том случае, если идеальный лингвистический автомат сможет не только отвечать простейшими реакциями на принимаемое им сообщение, но будет обладать вторым (адаптивным) и третьим (децизивным) уровнями интерпретационной готовности (3). Для этого наш автомат должен отвечать следующим условиям.

1. Этот автомат должен образовывать вместе с внешней лингвистической средой (текстами) замкнутую систему, обладающую обратной связью (ср.: Амосов, Касаткин, Касаткина, Талаев, 1973, с. 61—63). Эта связь должна давать возможность автомату, с одной стороны, приспосабливаться путем самообучения к изменениям внешней среды, а с другой — противодействовать тем воздействиям среды, которые грозят расстройством или даже уничтожением этой системы и входящих в нее лингвистических алгоритмов (ср.: Lofgren, 1962/1966).

2. Автомат должен обладать способностью к гомеостатическому саморегулированию и самоорганизации, самовосстановлению и самовоспроизводству, обеспечивающим «бессмертие» искусственного лингвистического интеллекта. В частности, идеальный лингвистический робот должен быть снабжен алгоритмами, прогнозирующими и реализующими внутреннее развитие системы и нормы языка, а во-вторых, должен располагать программами обучения, в частности программами автоматического пополнения словаря и грамматики новыми единицами, возникающими в связи с появлением во внешней среде новых ситуаций и объектов.

[1] Свойство открытости языка позволяет ему описывать изменения своей собственной структуры, а затем приспосабливать эти изменения для лучшего описания вновь возникающих и предварительно не предусмотренных ситуаций (ср.: Higman, 1969/1974, с. 180—181).

3. Автомат должен уметь анализировать, формализовать и обобщать принимаемую из внешней среды информацию и использовать эти обобщения для принятия лингвистических решений и постановки собственных целей. Указанная способность позволит автомату решать экстраполяционные задачи, без которых невозможна любая рассудочная деятельность, в том числе глобальное распознавание и порождение текста. Для реализации этих условий необходимо вычислительное устройство, работающее уже не по схеме последовательной очередности (принцип машины Тьюринга), но опирающееся на «нейроноподобную» ассоциативную память (принцип автоматов Неймана), имеющую подобно живому мозгу многоуровневую архитектуру. Может показаться, что рассуждения по поводу идеального лингвистического автомата более уместны в научно-фантастической повести, чем в исследовании по инженерной лингвистике.

Однако не следует забывать о быстрых темпах развития кибернетической техники. Еще на рубеже 50-х и 60-х годов большинство физиологов, математиков, философов, лингвистов и даже программистов, ориентировавшихся на ЭВМ первого и второго поколений, уверенно заявляли, что машина никогда не сможет выйти из рамок предопределения, она не овладеет способностью к классификации и обобщению стимулов внешней среды, самообучению, адаптации, целенаправленному поведению, самовоспроизведению и т. д. Утверждалось также, что функционирование ЭВМ будет всегда осуществляться по последовательно-алгоритмическому принципу, а ассоциативную машинную память, моделирующую память человека, построить никогда не удастся.

Однако прошло всего лишь пятнадцать-двадцать лет, и появились созданные А. Ньюэллом, Дж. Шоу и Г. Саймоном машинные алгоритмы «Логик-теоретик» и «Шахматист НШС», робот Винограда (АИС «Вопрос — ответ» SHRDLU), имитирующие творческую, заранее не предопределенную деятельность человеческого мозга, предусматривающую альтернативное принятие решений (см.: Feigenbaum and Feldman, 1963/1967, с. 31—32, 113—144, 239—254), созданы кибернетические модели «Мадалина» и «Минос» (Nilsson, 1971/1973), реализована на ЭВМ-III семантико-синтаксическая программа «Элиза», с помощью которой осуществляется диалог «живой пациент—автоматический психиатр» (Weizenbaum, 1968/1970, с. 215—244). Созданы читающие автоматы, которые используют алгоритмы распознавания образов, построенные на машинной классификации объектов внешнего мира, обобщении результатов этой классификации и самообучении (Вапник, Червоненкис, 1974; Загоруйко, 1972). Построены и опробованы на ЭВМ второго поколения М-220 и БЭСМ 6 алгоритмы децизивного (3) целенаправленного адаптивного поведения, — так называемые М-автоматы РЭМ и МОД (Амосов, Касаткин, Касаткина, Талаев, 1973, с. 52 и сл.). Де-

лаются попытки сконструировать самовоспроизводящиеся автоматы (Jacobson, 1958/1963, с. 223—258). Что же касается разработки ассоциативных «нейроноподобных» запоминающих устройств, то за последние десять лет созданы опытные образцы таких устройств (ср.: Findler, 1968, с. 229—254; Радченко, 1968; Digby, 1973, с. 768—772; Thurber, Patton, 1973, с. 1140—1143).

Разумеется, все эти экспериментальные разработки имитируют лишь отдельные черты естественного разума. Ни один из имеющихся алгоритмов не может рассматриваться в качестве модели интеллекта. Это и понятно. В настоящее время не выработано общепринятой формулировки для самого понятия интеллекта, не выяснен его состав и не определены решающие механизмы отдельных интеллектуальных операций. Не изучены такие узловые блоки интеллекта, как принятие решения, цель, предсказание, не вскрыто их операциональное взаимодействие в ходе осуществления интеллектуальных актов (Анохин, 1973, с. 87).

Однако широкий фронт и глубина исследовательских работ в области загадочных проблем интеллекта позволяет надеяться, что в ближайшие десятилетия будут построены такие модели, которые смогут быть использованы при построении алгоритмов искусственного разума, — алгоритмов, которые в свою очередь будут введены в математическое обеспечение ЭВМ пятого и, возможно, четвертого поколений. Учитывая эту перспективу, мы не можем ограничиваться изготовлением только лингвистических программ, но должны также думать о создании эвристических машинных алгоритмов переработки текста, включающих такие основные блоки интеллекта, как принятие решения, цель, предсказание.

На первых порах эти эвристические алгоритмы придется, разумеется, реализовать на ЭВМ последовательно-очередного действия. Дело в том, что в настоящее время развитие современной вычислительной техники идет по линии совершенствования внешних устройств ЭВМ, миниатюризации памяти и расширения ее мощности, а также увеличения быстродействия машины (подробнее см. стр. 17). Что же касается основных принципов построения и функционирования компьютера, то они остаются пока без изменения: выпускаемые промышленностью индустриальных стран ЭВМ третьего поколения, на которые в течение ближайших десяти-пятнадцати лет будут ориентированы алгоритмы лингвистического обслуживания АСУ, ИПС и МП, а также разрабатываемые в настоящее время машины четвертого поколения сохраняют последовательно очередной принцип машины Тьюринга. Использование «нейроноподобной» памяти типа автоматов Неймана можно ожидать у ЭВМ пятого поколения, которые появятся только в конце века.

12. Глобальное распознавание текста на ЭВМ последовательно-очередного действия

Чтобы реализовать на ЭВМ последовательно-очередного действия алгоритм глобального распознавания текста, потребуется расчленить ассоциативный пучок, характеризующий каждую лингвистическую единицу, на отдельные грамматические, словообразовательные, семантические и прочие парадигмы. Адреса этих парадигм вместе с валентностями должны быть последовательной цепочкой записаны при каждой лингвистической единице языка.

Если предположить, что один развитой язык имеет словарь, включающий не менее 1 млн лексических единиц (общеупотребительных слов и устойчивых словосочетаний, простых и сложных терминов),[1] то для записи такого словаря понадобится память объемом не менее 300 Мбайт $\simeq 2.75 \cdot 10^9$ дв. ед. (64). Это заметно выше суммарных объемов памяти самых мощных отечественных и зарубежных ЭВМ.

Морфологические и синтаксические алгоритмы сравнительно невелики, и ими в наших расчетах можно пренебречь,[2] зато объем

[1] При оценке объема лексики и фразеологии языка мы исходим из следующих количественных характеристик общих и терминологических словарей русского языка.

1. «Обратный словарь русского языка» (М., 1974), обобщивший словники четырех словарей русского литературного языка, включает около 125 тыс. слов, в число которых вошло сравнительно небольшое число специальных терминов.

2. «Фразеологический словарь русского языка» под ред. А. И. Молоткова (М., 1967) и русско-английский словарь идиом (V i t e k A. Russian-English Idiom Dictionary. Detroit, 1973) содержат в сумме около 10 тыс. общеупотребительных фразеологизмов.

3. Терминология отдельных отраслей знаний содержит каждая многие тысячи лексических единиц. Так, например, «Словарь дескрипторов по химической промышленности» (М., 1966) содержит 20 тыс. дескрипторов и до 8 тыс. синонимов, жаргонов и других терминов, извлеченных из корпуса химических и нефтехимических терминов, общим объемом 180 тыс. простых и сложных терминов, многие из которых не засвидетельствованы в словарях литературного языка. Аналогичную картину показывают «Большая медицинская энциклопедия» (М., 1956—1971), «Физический энциклопедический словарь» (М., 1960—1966), словарь «Автоматизация производства и промышленная электроника» (М., 1962/1964) и другие словари.

К той же количественной оценке «генерального» корпуса лексических единиц можно прийти на основании данных словарей английского языка. Словарь Вебстера (Webster's Third New International Dictionary of the English Language. Springfield Mass., 1961) дает около 450 тыс. словарных статей. Современные тезаурусы по различным областям знаний включают от 3 до 14 тыс. дескрипторов и ключевых слов каждый, причем эти последние извлекаются из списков, содержащих многие десятки простых и сложных терминов (ср.: Михайлов, Черный, Гиляревский, 1968, с. 469—503), часть из которых отсутствует в словаре Вебстера.

[2] Программа порождения лично-временных форм русского глагола насчитывает около 2.5 тыс. 37-ми разрядных ячеек или приблизительно 0.01 Мбайт (Пиотровская, 1973, с. 260—277). Для описания морфологии и синтаксиса русского языка понадобилось бы примерно 20 таких программ общим объемом в 0.2 Мбайт.

управляющих программ составит не менее 30% машинной записи словаря и других программ, т. е. займет в памяти ЭВМ около 100 Мбайт = $0.9 \cdot 10^9$ дв. ед.

Таким образом, общий объем лексико-грамматического и семантического алгоритма, необходимого для осуществления глобального распознавания текста в условиях одноязычной ситуации составит не менее 400 Мбайт = $3.8 \cdot 10^9$ дв. ед.

Эта информация также заметно превышает объем памяти таких мощных ЭВМ, как ЕС-1050 и IBM-370 (Крайзмер, Матюхин, Майоркин, 1971; Piotrowski, Georgiev, 1974, с. 75; ср. табл. 1; Katzan, 1971/1974).[1]

13. Глобальное распознавание текста на ЭВМ и лингвистическая теория

В предыдущих разделах мы познакомились с рядом антиномий и парадоксов, стоящих на пути осуществления автоматического распознавания текста. Эти антиномии, как и большинство информационно-лингвистических и семиотических парадоксов, с которыми нам еще придется встретиться, являются принципиально неразрешимыми. Преодоление или ослабление эффекта этих парадоксов следует искать в конструктивных решениях, состоящих в том, что мы без сожаления должны жертвовать несущественными деталями ради того, чтобы решить в целом поставленную задачу.

Недостаточный объем памяти современных ЭВМ не является основным барьером при осуществлении глобального распознавания текста: память машины можно всегда расширить путем использования сменных пакетов дисков, магнитных лент и карт. Принципиальные трудности, лежащие на пути создания глобального алгоритма, заключены в лингвистическом существе проблемы.

Первое препятствие связано с тем колоссальным объемом лингвистической и технической работы, который необходим для построения такого алгоритма.

Предположим, что для разработки одной словарной статьи, содержащей полную семантическую и морфолого-синтаксическую информацию (ср.: Мельчук, 1974, с. 113—131), а также перфорации этой статьи нужен день работы одного исполнителя.[2] Тогда

[1] В будущем (80—90-е годы) можно ожидать появление систем ЭВМ коллективного пользования через сеть связи с программной коммутацией абонентов, — систем с памятью до 10^{15} дв. ед. Наличие таких систем полностью сняло бы вопрос об ограничениях, накладываемых емкостью машинной памяти на алгоритмы переработки текста.

[2] При разработке таких статей обычно используются словари, грамматики и интроспекция составителя. Если же, помня о правиле остранения (Шкловский, 1919, с. 106—112; Chomsky, 1968/1972, с. 35—36), разрабатывать эти статьи с помощью объективных данных текстового и психолингвистического эксперимента, позволяющего выявить лексико-грамматические

общий объем времени, необходимого для лингвистической разработки словаря и его перфорации, составит 1 млн человеко-дней. Если отвлечься от вопросов существования разрешающего алгоритма для машинной реализации такого словаря (ср.: Трахтенброт, 1974, с. 60—65) и времени построения такого алгоритма, то лингвистическая разработка и перфорация глобального словаря потребовала бы 30 лет работы коллектива, состоящего из 100 лингвистов и операторов-перфораторщиков. Но такого срока вполне достаточно для того, чтобы результаты этой работы подверглись действию антиномии Ахиллеса и черепахи: заложенная в макрословарь лингвистическая информация окажется устаревшей к моменту его промышленной реализации.

Второе и, по существу, основное препятствие заключается в недостаточной разработанности лингвистической теории. Строить алгоритм глобальной переработки текста на ЭВМ можно исходя из двух идеальных подходов, один из которых опирается на вероятностную, а другой на детерминистскую концепции языка.

Существо первого подхода — будем называть его иконическим — состоит в том, что язык рассматривается как популяция всех реализованных когда-либо и потенциально возможных высказываний-текстов, предложений или их сегментов (ср.: Herdan, 1966, с. 28). Если ввести эти сегменты и тексты в память ЭВМ и приписать каждому из них эквивалент на информационном (машинном) или выходном естественном языке, то машина будет способна осуществлять глобальное распознавание смысла текста. Однако иконический подход может быть реализован лишь относительно ограниченного числа текстовых сегментов: память компьютера вряд ли когда-либо сможет вместить все реализованные тексты и их сегменты. Основной же методологический барьер на пути иконического подхода состоит в том, что функционирование языка носит новаторский характер. Это значит, что в результате взаимодействия открытой системы языка и его нормы, с одной стороны, и постоянно меняющейся ситуации — с другой (см. 8), порождается большое число совершенно новых текстовых цепочек, которые не являются повторениями уже реализованных предложений и сегментов и которые не могут быть выведены с помощью простой аналогии из уже имеющихся в памяти ЭВМ моделей. Таким образом, глобальное машинное распознавание текста на основе иконического подхода оказывается невозможным.[1]

и семантические характеристики слова не с точки зрения лингвистического поведения составителя, но исходя из лингвистического опыта коллектива (см. 10), то на составление каждой статьи уйдут уже не дни, а месяцы.

[1] На иконическом подходе строятся только системы переработки таких текстов, которые порождаются подъязыками, представляющими собой закрытую и строго фиксированную совокупность лингвистических единиц (словоупотреблений, словосочетаний, предложений). Примером могут служить английский и русский подъязыки переговоров «земля—воздух» («аэро-

Более перспективным является как будто второй — детерминистский подход, отражающий концепцию жесткой структуры языка. Его последовательным выражением являются теории порождающих грамматик. Здесь в качестве языка выступают словарь, который включает исходные лингвистические элементы (фонемы, морфемы, грамматические классы слов, ядерные предложения), а также циклические алгоритмы порождения (Chomsky, Miller, 1963/1965, с. 231—246).

Высказывается предположение, что полная порождающая грамматика конкретного языка вырастает из некоторой врожденной универсальной грамматики или умственной структуры (ср.: Lashley, 1951, с. 112—136; Lenneberg, 1967, с. 106—107, 294 и сл., 331—334), которая с помощью ограниченного числа рудиментарных рекурсивных грамматических функций способна порождать бесконечный класс глубинных структур. В свою очередь полная порождающая грамматика включает фиксированную систему «порождающих принципов, которые характеризуют и связывают глубинные и поверхностные структуры некоторым определенным образом. . .» (Chomsky, 1968/1972, с. 30; ср. также с. 38 и сл., 105 и сл.).

Несмотря на всю ее глубину и заманчивую эвристичность, теория порождающих грамматик не может быть в настоящем ее виде использована в качестве основы глобального распознавания и порождения текста на ЭВМ-III. Для этого имеется несколько причин теоретического и технологического характера.

1. Чтобы стать инструментом глобального распознавания, порождающая грамматика должна уметь формально эксплицировать языковую компетенцию коллектива или, иными словами, должна включать полное описание механизмов создания текстов. Между тем выясняется, что эти описания строятся в настоящее время по типу так называемых грамматик контекстно-свободных языков, использующих такие процедуры, которые позволяют осуществлять преобразования над элементами словаря без учета контекстных ограничений (ср.: Ginsburg, 1966/1970). Контекстно-свободные грамматики являются очень удобным инструментом при исследовании искусственных языков программирования. Однако их построение находится в явном несоответствии с существом информационного процесса и механизмами порождения текста естественного языка (ср. 8 и рис. 11). В результате применения контекстно-свободных грамматик, с одной стороны, порождается много цепочек, не являющихся реальными предложениями данного языка, с другой — обнаруживаются реальные предложения и сегменты, построения которых не соответствуют правилам порождающей грамматики данного языка

дром—пилот»), каждый из которых состоит из фиксированного числа фраз-штампов (в английском подъязыке их 755, в русском — 820) типа: англ. mai I take off, русск. «разрешите взлет». Аналогичным образом строятся коммуникативные системы «земля—вода», «вода—земля».

(Chafe, 1968, с. 109—127; Danielsen, 1972, с. 260—285; ср.: Ту Рак, 1974).

Естественные языки должны описываться контекстно-зависимыми грамматиками, создание которых предусматривает выявление всех контекстных ограничений, накладывающихся на порождение текста. Формальной дедуктивной процедуры, выявляющей эти ограничения, пока не существует, и ее, по всей вероятности, создать нельзя, учитывая тот факт, что текст порождается не только системой языка, но также «несистемной» статистической нормой (Пиотровский, Турыгина, 1971) и узусом (Степанов Г. В., 1966) и, наконец, независимой от языка ситуацией.

Контекстные ограничения можно, разумеется, выявить путем индуктивного описания текста, но такой подход выпадает из компетенции порождающих грамматик, оказываясь в сфере лингвистики текста.

2. Теории порождающих грамматик не создали однозначной детализованной процедуры для обнаружения глубинных структур, — процедуры, которая так необходима при осуществлении глобального семантического машинного перевода.

В подавляющем большинстве случаев глубинные структуры постулируются путем обращения языковеда к собственной интроспекции и здравому смыслу. Этот подход, пренебрегающий правилом «остранения», немедленно приводит нас к ситуации, описываемой парадоксом актера и зрителя, при котором языковед, выступающий одновременно и наблюдателем и объектом самонаблюдений, полностью попадает во власть собственных субъективных оценок.

В результате выделение глубинных структур с помощью обращения к здравому смыслу оправдывает себя лишь в элементарных очевидных случаях типа: a wise man is honest (поверхностная структура) $\rightarrow {}_{S}[{}_{NP}$[a man$_{S}[{}_{NP}$[man]$_{NP}$ $_{VP}$[is wise]$_{VP}]_{S}]_{NP}$ $_{VP}$[is honest]$_{VP}]_{S}$ (глубинная структура) (см.: Chomsky, 1968/1972, с. 40—41).

Однако при массовой переработке текста постоянно приходится иметь дело с менее очевидными ситуациями. Действительно, следует ли считать, что русские фразы: *Иван — сын Петра*, *Иван — сын Петру*, *Иван — Петров сын* реализуют одну глубинную структуру или этих структур здесь три? Или, опираясь на их французские эквиваленты: Jean est le fils de Pierre, Jean est un fils de Pierre, можно полагать, что эти предложения отражают шесть глубинных структур?

Исследования по межъязыковой идиоматичности (Добрускина*, 1973; Каменева*, 1973) особенно отчетливо показывают всю зыбкость и субъективность принципов, по которым осуществляется в настоящее время выделение глубинных структур.

3. Теория порождающих грамматик рассматривает язык и текст с точки зрения говорящего. Между тем машинное описание языка

должно строиться в первую очередь как распознающее устройство, ориентированное на «грамматику для слушающего» (ср.: Hockett, 1961/1965, с. 139—166).

4. В теории порождающих грамматик пока не выработано эффективных приемов остранения (ср. выше, с. 44, примеч. 2), т. е. таких приемов, которые превратили бы исследователя-лингвиста из невольного участника исследуемого информационного процесса в его стороннего метанаблюдателя.

Недостаточная разработанность теоретических основ и технологии трансформационных грамматик является причиной того, что эта теория не смогла стать фундаментом инженерной лингвистики. И поэтому не случайно, что построенная на этих принципах переработка текста ограничивается пока несколькими более или менее удачными экспериментами (Klein, Kuppin, 1970, с. 144—162; Walker, 1973; O'Malley, 1973).

Глава 3

КОММУНИКАТИВНАЯ СИСТЕМА «ЧЕЛОВЕК—МАШИНА—ЧЕЛОВЕК»

14. Переработка сообщения в системе «человек—машина—человек»

В предыдущей главе были подробно описаны те технические, программистские и лингвистические препятствия, которые стоят на пути осуществления глобального распознавания и порождения текста на ЭВМ-III.

Возникает вопрос, существуют ли прагматические пути и приемы, с помощью которых можно было бы обойти перечисленные барьеры или, по крайней мере, ослабить действие рассмотренных выше парадоксов и антиномий? Может быть, все эти препятствия непреодолимы и в связи с этим промышленная переработка текста на ЭВМ-III принципиально не может быть реализована?

На эти вопросы можно ответить следующим образом. ЭВМ-III, как, впрочем, и машины последующих поколений, нельзя рассматривать как автономно функционирующее средство переработки текстового сообщения. Полная переработка текста, предусматривающая его глобальное восприятие и порождение, может быть осуществлена только в коммуникативной системе «человек—машина—человек» ($Ч_1МЧ_2$),[1] в которой человек выступает в качестве источника ($Ч_1$) и конечного приемника ($Ч_2$) сообщения, а ЭВМ служит средством преобразования этого сообщения, значительно ускоряющим его восприятие человеком — адресатом сообщения.

[1] Под термином *машина* понимается сочетание ЭВМ и программ, позволяющих осуществлять автоматическую переработку текста.

Рис. 12. Передача сообщения в информационной схеме: человек—машина—человек.

Передачу сообщения в схеме $Ч_1МЧ_2$ (см. рис. 12) можно рассматривать в виде комбинации двух ИП. Первый процесс предусматривает передачу текстового сообщения от человека к машине ($Ч_1$—М). Второй заключается в передаче переработанного в ЭВМ текста от машины к человеку (М—$Ч_2$).

В соответствии со схемой, представленной на рис. 12, каждый из рассматриваемых нами ИП включает по четыре блока.

Первый ИП ($Ч_1$—М) включает следующие блоки:

1.1) порождение текста отправителем сообщения ($Ч_1$);

1.2) передача в ЭВМ цепочки импульсов, кодирующих этот текст;

1.3) машинный семиозис принятой совокупности импульсов, в результате осуществляется распознавание и формализация смысла текста;

1.4) преобразование внутри ЭВМ распознанного текста (например, перевод его на другой язык или компрессия в целях его индексирования, аннотирования или реферирования), а затем выдача машинных результатов на естественном языке.

Второй ИП (М—$Ч_2$) состоит из следующих блоков:

2.1) выдача компьютером машинных результатов (этот блок совпадает с блоком 1.4;

2.2) чтение машинных результатов адресатом $Ч_2$;

2.3) семиозис машинного сообщения;

2.4) реакция $Ч_2$ на полученное сообщение.

«Горячими точками» при построении системы $Ч_1МЧ_2$ являются, с одной стороны, формирование блока 1.3, где происходит машинный семиозис, а с другой — организация блока 2.3, в котором осуществляется распознавание, интерпретация человеком результатов, выдаваемых ЭВМ. В обоих случаях успех семиозиса зависит от того, насколько удачно решается вопрос о переходе от естественного языка к языку машин, который по своему существу является искусственным математическим языком.

Третьей «горячей точкой» инженерной лингвистики является задача обучения компьютера, т. е. составление лингвистических алгоритмов и их программирование. Эта задача также не может быть решена без правильного понимания соотношения естественного и искусственного языков.

Поскольку и естественный и математический (а следовательно, и машинный) языки являются семиотическими системами [1] — язы-

[1] Эта точка зрения, правда, не является общепринятой. Так, например, такие философы и логики, как Д. Гильберт и Р. Карнап, считали матема-

ком в широком смысле этого слова, рассмотрим их основные свойства.

Будем называть системой языка совокупность его единиц и отношения между этими единицами. В качестве единиц естественного языка выступают звуки и фонемы, буквоварианты и буквы, морфы и морфемы, словоформы, слова и структурные схемы их построения, словосочетания и предложения и их скелетные схемы. В машинном языке такими единицами будут буквы, цифры и другие символы (точнее, их коды), машинные морфемы и слова, в качестве которых выступают особым образом ограниченные слева и справа цепочки букв, формально выделенные словосочетания и предложения, а также их скелетные схемы.

В качестве структуры этой системы рассматриваются отношения между единицами языка: их парадигматическое соположение и противопоставление, синтагматическое размещение (дистрибуция) этих единиц относительно друг друга в тексте, вхождение одной единицы в другую, представление одной единицы через другую.

В систему естественного языка заложены три функции: а) номинативная, которая формирует лингвистический символ и благодаря которой становится возможным передать в сообщении семантическую и сигматическую информацию.

б) предикативная, которая связывает между собой знаки и которая характеризуется синтактической информацией текста;

в) прагматическая, соотносящая в ходе передачи сообщения первые две функции с личностью и интересами отправителя или получателя — в том числе коллективного — сообщения (10).

Чтобы выяснить, как обстоит дело с указанными тремя функциями в машинном языке, рассмотрим с этой точки зрения математический язык, являющийся основой всякого информационного языка, применяемого в компьютере.

Язык математики представляет собой исчисление, т. е. некоторую систему правил работы с математическими знаками. Исчисление строится следующим образом: строится алфавит первичных знаков, задаются начальные термины («слова») исчисления, указываются правила вывода новых терминов («слов») из начальных терминов. Математические структуры образуются путем композиций всех этих элементов (Налимов, 1974, с. 155). Таким образом, предикативная функция оказывается представленной в языке математики и всех других его ответвлениях, в том числе и в машинном языке.

Хотя математические структуры строятся путем задания отношений между составляющими их элементами, физическая

тику (точнее, математическую логику) семиотикой, а по мнению Л. Ельмслева, математика (соответственно, математическая логика) может рассматриваться лишь как план выражения какого-либо языка (подробнее см.: Степанов, 1971, с. 157).

природа этих последних остается неизвестной. Отсюда следует, что отношение имя—денотат (т. е. отражение конкретного референта из объективной действительности) в каждом математическом языке остается невыраженным. Номинативная функция в математическом языке ограничивается значимостью (десигнатом) знака, которая определяется правилами его построения и использования.

Что же касается прагматической функции, то она отсутствует в языке математики, как, впрочем, и в большинстве других семиотических систем.

Различия между естественным и искусственным языками особенно отчетливо прослеживается в структуре лингвистического символа и математического (машинного) искусственного знака.

15. Лингвистический символ и искусственный машинный знак

Как уже говорилось, символ включает четыре компонента: имя, денотат и десигнат, в которых отражается специфика номинации, а также коннотат. В искусственных языках эта четырехкомпонентная структура символа свертывается в характеризующее ИЗ двухчленное построение имя — десигнат (см. 2). Роль имени (означающего) здесь выполняет материальный носитель информации (цифра, буква или любой другой символ), а в качестве десигната (означаемого) выступает значимость этой буквы, цифры или символа в данной математической системе или какой-либо другой искусственной семиотике. В машинном знаке роль означающего выполняет закодированная в электромагнитной субстанции цепочка букв (или других единиц алфавита языка), а функцию означаемого (десигната) исполняет лексико-грамматическая информация, записанная в сопровождающем цепочку машинном документе. Этот десигнат, записываемый обычно в МД вручную или программно человеком-алгоритмизатором или вырабатываемый самой машиной, соотносит машинное имя (означаемое) с той искусственной семиотической системой лингвистических знаний, которая была введена в компьютер в ходе его обучения. Таким образом, номинативная функция в искусственном машинном языке ограничивается десигнативными отношениями между машинными означаемым и означающим.[1]

[1] Некоторые авторы пытаются представить машинный знак в виде триады единиц: входного текста—МД—выдаваемая ЭВМ выходная единица, соответствующая обобщенному путем вращения треугольнику Фреге. При этом выходная лингвистическая единица рассматривается в качестве референта. Легко заметить, что термин «референт» употреблен здесь не в смысле «денотат». Референт — денотат машинного знака обнаруживается лишь при его соотнесении человеком — приемником сообщения со знаком естественного языка. Знак может восприниматься и интерпретироваться по-разному в зависимости от индивидуальных особенностей носителя языка, — находя-

Поливалентность языкового символа проявляется особенно ярко тогда, когда однозначный и четко определенный математический знак помещается в лингвистический контекст. Превращаясь в символ естественного языка, математический знак немедленно включается в различные ассоциативные ряды и, обрастая дополнительными коннотативными оттенками, теряет свою строгость и однозначность. Вот как, например, интерпретируется в различных лингвистических текстах математическое понятие мнимого числа ($\sqrt{-1}$).

«. . . Квадратный корень из минус единицы, — писал Ф. Энгельс, — есть не просто противоречие, а даже прямо абсурдное противоречие, действительная бессмыслица. И все же $\sqrt{-1}$ является во многих случаях необходимым результатом правильных математических операций; более того, что было бы с математикой, как низшей, так и высшей, если бы ей запрещено было оперировать с $\sqrt{-1}$?» (Маркс, Энгельс, т. 20, с. 125).

По-иному метафорически интерпретировалось понятие мнимого числа у Г. В. Лейбница: «Мнимые числа — это поразительный полет духа божьего, это почти амфибии, пребывающие где-то между бытием и небытием» (Лейбниц, 1908, с. 126).

У В. Хлебникова мнимые числа вызывали совсем неожиданные ассоциации: «Полюбив выражения вида $\sqrt{-1}$, которые отвергали прошлое, мы обретаем свободу от вещей» («Курган Святогора»). Или: «Мы взяли $\sqrt{-1}$ и сели в нем за стол. Наш Ходнырлет был глыбой стекла, мысли и железа» (наброски неопубликованного рассказа; цит. по: Лейтес, 1973).

Итак, лингвистический символ, являющийся продуктом ассоциативно-эвристического мышления человека, оказывается семантически богаче математического знака. Благодаря этому с помощью естественного языка мы можем ставить и решать, пусть субъективно и неполно, интуитивно сформулированные задачи, — задачи, для которых строгое и полное решение либо неизвестно, либо вообще не существует. Поскольку такие задачи не могут решаться на языке математики, естественный язык оказывается мощнее и богаче любого формализованного искусственного языка.

Вместе с тем моносемия математического знака, его строго определенная значимость прозрачной анатомии формализованного языка позволяют, во-первых, недвусмысленно и исчерпывающе формулировать задачу, а во-вторых, решать ее алгоритмическим путем с помощью автомата.

Определив специфику естественного и машинного языка, а также образующих их знаков и отношений, обратимся к семиотическому анализу поэтапной (см. выше, с. 49) передачи сообщения в системе $Ч_1МЧ_2$.

Рассмотрим поведение лингвистических символов и ИЗ внутри каждого из описанных выше ИП (14).

В блоке 1.1 человек — отправитель информации порождает некоторые лингвистические символы, каждый из которых включает все четыре компонента — имя, денотат, десигнат и коннотат. В качестве примера можно взять рассматривавшееся в 7 словосочетание (the) XX CENTURY FOX или образующие его словоформы с их семантическими полями и оттеночными значениями.

В блоке 1.2 происходит разрушение каждого из этих символов: путем многократного перекодирования (7, рис. 7) в память ЭВМ вводится лишь цепочка импульсов, соответствующая означающему (имени) текстового символа, что же касается денотата, коннотата и десигната этого символа, то они, естественно, в компьютер не попадают.

В блоке 1.3 осуществляется машинный семиозис, в результате которого происходит как бы восстановление разрушенного знака. Принятая машиной цепочка импульсов последовательно сравнивается с записанными в машинных словарях цепочками до тех пор, пока не произойдет полное совпадение текстовой цепочки с некоторой словарной цепочкой (примером может служить описанный в 7 поиск по машинному словарю слова FOX). В случае такого совпадения лексико-грамматическая информация, заключенная в МД словарной цепочки, передается текстовой цепочке, которая тем самым получает свой десигнат и превращается в машинный ИЗ. Следует, однако, иметь в виду, что восстановленный путем этого машинного семиозиса искусственный знак является обедненной моделью исходного речевого символа. В машинный документ вкладывается лишь основная лексико-грамматическая информация, он не может вместить все дополнительные лексико-грамматические значения и коннотативные оттенки, присущие данной лингвистической единице в естественном языке (ср. 2).

В блоках 1.4 (2.1) и 2.2 происходит новое разрушение знаков текста, которые превращаются в цепочку импульсов.

Эта цепочка поступает в блок 2.3, в ходе работы которого человек — приемник сообщения, опираясь на свои знания предмета и лингвистические навыки, восстанавливает у полученных от ЭВМ цепочек их десигнативные и денотативные значения. Оживающие знаки обрастают также коннотативными оттенками (ср. восприятие русским читателем перевода входного английского сегмента «XX век Фокс» или метафорическое употребление ИЗ $\sqrt{-1}$ у Хлебникова и Лейбница).

Итак, в ходе переработки текста в системе $Ч_1МЧ_2$ дважды происходит разрушение и восстановление знака: первый раз при передаче сообщения от человека к машине и второй раз при передаче машинных результатов от компьютера к человеку. Эффект работы

щиеся вне прагматики языка математический и машинные знаки расшифровываются одинаково всеми искусственными приемниками сообщения (ЭВМ) либо вообще не воспринимаются, если машина не имеет необходимых для такой расшифровки знаний.

системы $Ч_1МЧ_2$ зависит, во-первых, от объема содержательной информации исходного сообщения, который удается восстановить живому приемнику $Ч_2$, а во-вторых, от степени готовности массового читателя $Ч_2$ пользоваться услугами системы. Вопросы быстродействия системы и ее экономичности имеют пока второстепенное значение.

Достаточно полное восстановление исходной информации на заключительном этапе работы системы $Ч_1МЧ_2$ может быть достигнуто при соблюдении следующих условий.

1. В системе $Ч_1МЧ_2$ перерабатываются сообщения, основное содержание которых передается не коннотативными средствами или экстралингвистическими приемами (мимика, жест и т. д.), но оказывается заложенным в десигнатах лингвистических единиц или входящих в текст этого сообщения единиц искусственных языков.

2. Перерабатываемые в системе $Ч_1МЧ_2$ тексты ориентированы на одинаковое описание внешнего мира («тезаурус») у отправителя и получателя сообщения (ср.: Neubert, 1970). Это условие обычно выполняется относительно научно-технических и деловых документов такой тематики, которая входит в сферу профессиональной компетенции как источника ($Ч_1$), так и адресата ($Ч_2$) сообщения.

3. Порождение текста автором $Ч_1$ и семиозис машинных результатов конечным адресатом $Ч_2$ осуществляются только на базе тождественных тезаурусов, а также на основе одинаковых (если речь идет об одноязычной переработке текста) или близких (если речь идет о машинном переводе) лингвистических навыков.

4. Обучение компьютера ориентировано, с одной стороны, на ту тематику текстов, которые предполагается перерабатывать в системе $Ч_1МЧ_2$, а с другой — на тот тезаурус и те лингвистические навыки, которыми должны обладать эвентуальный отправитель и приемник сообщения.

5. Вводимое в машину описание языка должно содержать также лексико-грамматические сведения, использование которых позволило бы свести до минимума потерю информации при переходе от символов входного сообщения к текстовой последовательности машинных знаков, при преобразовании машинного текста внутри ЭВМ и, наконец, при расшифровке приемником $Ч_2$ машинных результатов. Минимизации информационных потерь можно добиться лишь тогда, когда в десигнаты машинных знаков вкладываются те лексические и грамматические значения, которые с достаточно большой вероятностью могут считаться существенно необходимыми для правильной формализации и переработки входного сообщения.[1]

[1] Легко заметить, что условия 1, 2, 3, 5 не выполняются относительно художественных и разговорно-бытовых текстов, которые, по всей вероятности, никогда не будут перерабатываться в системе $Ч_1МЧ_2$.

Что касается готовности массового потребителя пользоваться услугами системы $Ч_1МЧ_2$, то, как показывает опыт эксплуатации указанных систем (Гончаренко*, 1972а, б; Добрускина*, 1973), эта готовность возникает лишь тогда, когда ЭВМ выдает текстовые сегменты, которые без особого напряжения восстанавливаются читателем в символы естественного языка, образуя при этом текст, не противоречащий системе, норме и узусу данной разновидности языка.

16. Инженерное языкознание и лингвистика текста

Выше было показано, что правильный отбор и детальная экспликация материала, предназначенного для обучения ЭВМ, работающей в системе $Ч_1МЧ_2$, имеет первостепенное значение. Этот отбор и экспликация могут быть осуществлены либо путем использования дедуктивных процедур в духе порождающих грамматик, либо с помощью индуктивных приемов лингвистики текста.

Как уже говорилось (**13**), теория порождающих грамматик не выработала пока детализованной методики однозначного выделения глубинных структур, из которых можно было бы построить машинный язык. Кроме того, правило остранения (**13**) требует, чтобы все дедуктивные лингвистические схемы проверялись с помощью массового психолингвистического эксперимента на носителях языка. Если такой эксперимент можно организовать для оценки русского материала, то его постановка относительно иностранных языков связана с большими организационными и технологическими трудностями.

Более надежным и реалистичным является построение машинных описаний языка, опирающееся на индуктивную лингвистику текста.

Подобно каждой индуктивной теории, ЛТ опирается на простейшие законы, согласующиеся с опытом, при этом иногда используются скорее психологическое, чем логическое основание (ср.: Wittgenstein, 1933/1958, парадоксы 6.363, 6.3631). Тем не менее конечная цель ЛТ совпадает с задачами дедуктивной трансформационной грамматики: эта цель состоит в том, чтобы установить инвариантные схемы построения текста или, иными словами, сформулировать правила текстообразования. Однако в отличие от дедуктивно-интроспективного подхода порождающей грамматики ЛТ стремится решить эту задачу индуктивным путем, рассматривая большие массивы текста.

Реализация этой идеи связана с преодолением антиномии языка, выступающего в роли социального кодекса лингвистического поведения (ср. **10**) и индивидуального текста, являющегося актом воли и понимания конкретного отправителя сообщения и конкретного адресата (ср.: Saussure, 1922/1933, с. 38—39). Выйти из круга этой антиномии можно двумя путями: либо следует отказаться от всяких попыток типизации текстов и сосредоточить свое

внимание на рассмотрении его индивидуальности, либо необходимо сосредоточить все внимание на выявлении в конкретных текстах некоторых релевантных признаков и схем, т. е. правил речи, отражающих «дух языка» (Гак, 1966, с. 11; Ольшанский, 1974). По первому пути идет стилистика индивидуальной — в первую очередь, индивидуально-художественной речи. Второй путь — это путь лингвистики текста, которая не рассматривает текст как индивидуальный и единичный сегмент речи, но принимает его за единицу языка (ср.: Bense, 1962; Prütze, 1970; Wittmers, 1970; Sgall, 1973). При таком подходе задача исследователя состоит в том, чтобы извлечь из исследуемой выборки то общее, что лежит в основе отдельных образующих ее текстов, и таким образом выявить лексико-фразеологические, морфологические, синтаксические и даже ситуативные «формулы» построения некоторого усредненного типового текста.

При решении этой задачи следует исходить из следующей вероятностной схемы порождения и распознавания текста. Предположим, что система языка включает некоторое число элементов. Норма и узус задают каждому из этих элементов априорную вероятность, которая для элемента λ будет равна $P(\lambda)$. Как уже говорилось, в порождении текста участвуют не только система, норма и узус языка, но и ситуация.

Предположим теперь, что имеется множество ситуаций H, включающее отдельные несовместимые ситуации $h_1, h_2, \ldots, h_k, \ldots, h_n$, каждая из которых имеет свою вероятность.[1] Тогда вероятность появления лингвистического элемента λ в тексте T_k, описывающем ситуацию h_k, можно представить в виде условной вероятности

$$P(\lambda/h_k) = \frac{P(h_k \lambda)}{P(h_k)}.$$

Если речь идет об употреблении элемента λ в некоторой определенной позиции n нашего текста T_k, то возникает новая условная вероятность появления λ, которая определяется выражением

$$P(\lambda/h_k,\ b_i^{n-1}) = \frac{P(h_k b_i^{n-1}\lambda)}{P(h_k)\,P(b_i^{n-1}/h_k)},$$

где b_i^{n-1} — цепочка лингвистических элементов, предшествующих λ.

Дальнейшее уточнение условий употребления λ потребует введения новых условных вероятностей для появления λ в тексте (подробнее см.: Бектаев, Пиотровский, 1973, с. 70—76).

Опираясь на только что описанную схему, ЛТ получает возможность использовать наряду с чисто лингвистическими приемами различные статистико-вероятностные и теоретико-информационные модели. Так, например, соотношение задаваемых нормой языка априорных вероятностей $P(\lambda)$ и текстовых апостериорных вероятностей $P(\lambda/h_k,\ b_i^{n-1})$ может быть проанализировано как в русле Бейесовых стратегий (Бектаев, Пиотровский, 1973, с. 82—86; Налимов, 1974, с. 94—98), так и в плане теории распределений (Бектаев, Пиотровский, 1973; 1974; Налимов, 1974, с. 99—100; Fucks, 1955; Herdan, 1966). Дистрибуции и валентности отдельных лингвистических элементов могут быть описаны с помощью аппарата марковских связей (СтР, 1968, с. 4), а при количественных оценках семантики и прагматики текста используется теоретико-информационная методика (Пиотровский, 1968; Endres, 1973; Bogodist et al., 1974).

Использование всех этих приемов, с одной стороны, приводит к разрушению индивидуальности и целостности конкретного текста, с другой стороны, эта методика дает возможность выделить в чистом виде системно-языковые, нормативные и узуальные явления, которые характеризуют данную текстовую выборку и репрезентируемую ею разновидность языка. Эти лингвистические сведения как раз и необходимы для обучения машины: именно на их основе строятся надежно работающие алгоритмы анализа и синтеза текста.

Однако лингвистика текста не может ограничиться статистико-дистрибутивным и информационным анализом самого текста. Она должна располагать описанием производящей этот текст системы языка вместе с ограничивающими ее нормой и узусом. Вопрос о том, на основе какого математического аппарата должно строиться это описание, будет затронут в 8-й главе.

Что касается чисто технологической стороны дела, то лингвистика текста характеризуется следующими чертами.

1. Первичная обработка лингвистического материала осуществляется либо на машине, либо с помощью формализованной ручной процедуры.

2. В ходе исследования текста, а затем и построения алгоритмов применяются методы проверки лингвистических гипотез (Бектаев, Пиотровский, 1974, с. 190—289), а также статистические оценки достоверности получаемых результатов.

3. Наряду с его практическим и чисто народнохозяйственным значением машинный эксперимент рассматривается как эффективное средство проверки формальной строгости, корректности, а также теоретической ценности лингвистического построения, лежащего в основе той или иной системы автоматической переработки текста.

[1] Для каждого конкретного носителя языка это будет некоторая персональная или субъективная вероятность.

Часть вторая

ЛИНГВИСТИКА ТЕКСТА

Глава 4

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЕКСТА

17. Базовый язык и статистическое моделирование текста

В первой части настоящей работы было показано, что порождение текста является результатом взаимодействия многих переменных, одни из которых относятся к системе, норме и узусу языка, другие имеют ситуативную природу, а третьи отражают индивидуальность источника сообщения (10). Выявить все эти переменные и представить их в формализованном виде в машинном описании языка практически невозможно: построение такого алгоритма потребовало бы многих десятков лет, а реализующая его машинная программа оказалась бы слишком громоздкой и недостаточно эффективной по временным характеристикам даже относительно современных ЭВМ большой мощности и высокого быстродействия.

Отсюда следует, что общение человека с ЭВМ — в том числе и с совершенными машинами большой мощности — может осуществляться лишь при условии, что в память компьютера будет введена некоторая модель, представляющая собой сокращенное описание естественного языка. Эту модель, включающую наиболее существенные с точки зрения текстообразования его элементы и связи, мы будем называть машинным базовым языком (МБЯ).[1]

Основная задача, возникающая в ходе построения МБЯ, заключается в том, чтобы найти такую процедуру, которая позволила бы достаточно быстро и эффективно выделять среди всего многообразия и многоплановости языка и речи те элементы и связи, композиция которых образует такую модель языка, которая, функционируя в памяти ЭВМ, могла бы обеспечить эффективную переработку текста.

Препятствия, связанные со сложностью, многоплановостью и неопределенностью изучаемого объекта, преодолеваются обычно путем изучения его с помощью приемов теории вероятностей и

[1] Если указанная модель языка используется не для нужд машинной переработки текста, а для других прикладных целей, например в интересах оптимизации преподавания языка, то мы будем называть эту модель просто базовым языком (БЯ).

статистики. Этим способом удается выделить основные, релевантные характеристики объекта и отделить их от второстепенных деталей. Этот подход вполне может быть применен в инженерно-языковедческих построениях, опирающихся на лингвистику текста: построению машинного БЯ предшествует вероятностно-статистическое описание речи, выявляющее основные элементы и текстообразующие свойства конкретного языка. Это описание осуществляется с помощью последовательного использования различных вероятностно-статистических и информационных моделей все возрастающей сложности.

Основная идея вероятностно-статистического моделирования состоит в том, что текст рассматривается как случайный процесс, а единицы текста (буквы, буквосочетания, морфемы, словоупотребления, словосочетания, грамматические схемы) — как случайные события.[1]

Каждый случайный процесс является результатом действия некоторой производящей системы. Для текста такой производящей системой является система, норма и узус языка или подъязыка. Появление лингвистической единицы, являющейся случайным событием, предполагает некоторый комплекс условий, при выполнении которых осуществляется или не осуществляется данное лингвистическое событие. В этот комплекс входит описываемая ситуация, жанровые особенности текста и личность автора, окружающий контекст и т. д. При моделировании текста комплекс условий определяется в зависимости от поставленной задачи. Это значит, что мы заранее отказываемся от описаний менее важных для нас сторон текста с тем, чтобы получить возможность моделировать те его свойства, которые особенно важны с точки зрения поставленной задачи.

Наиболее простой вероятностно-статистической моделью текста, использующейся при обучении ЭВМ, является частотный список (словарь) словоформ и слов (ср. табл. 2, 3).

18. Частотный словарь

Частотные словари (ЧС) слов и словоформ строятся исходя из следующих упрощающих допущений.

1. Моделирующий текст (опытная выборка) и текст, выступающий в роли натурного объекта (генеральная совокупность текстов, относящаяся к данному подъязыку в фиксированный период времени), рассматриваются как реализация стационарного случай-

[1] Подробное определение таких базовых понятий квантитативного языкознания и лингвостатистики, как модель, натурный объект (оригинал), коэффициенты подобия и интерпретация модели, подъязык, случайный процесс и случайное событие, выборка, частота и математическое ожидание, частость (относительная частота) и вероятность, относительная ошибка, частотный словарь и т. д., можно найти в работах: Штофф, 1966; Пиотровский, 1966, с. 15—25; Бектаев, Пиотровский, 1973, 1974).

Таблица 2

Фрагмент ЧС словоформ немецких газетных текстов ГДР и ФРГ (по данным Ротарь*, 1971).

i	Словоформы	Тексты ГДР F	Тексты ФРГ F	Суммарные данные по газетным текстам обоих немецких государств						
				F	F^*	f	f^*	$\check{H}_i = -f_i \log_2 f_i$	$\check{H}^* = -\sum_{h=1}^{i} f_h \log_2 f_h$	φ%
1	der (*art.*)	4598	4478	9076	9076	0.0454	0.0454	0.2024764	0.202476	1.87
2	die (*art.*)	3859	3906	7765	16840	0.0388	0.0842	0.1819677	0.384444	3.55
3	und	2880	2827	5707	22548	0.0285	0.1127	0.1464166	0.530861	4.90
4	in (*prp.*)	2282	2154	4436	26984	0.0222	0.1349	0.1218701	0.652731	6.04
5	den (*art.*)	1485	1476	2961	29945	0.0148	0.1497	0.0899814	0.742512	6.87
6	des	1223	1208	2431	32376	0.0127	0.1619	0.0773338	0.820046	7.59
7	mit (*prp.*)	983	1023	2006	34382	0.0100	0.1719	0.0665945	0.886641	8.21
8	von	955	913	1908	36290	0.0095	0.1814	0.0640305	0.950671	8.80
9	das (*art.*)	911	968	1879	38169	0.0094	0.1908	0.0632649	1.010139	9.35
10	für (*prp.*)	875	851	1726	39894	0.0086	0.1995	0.0591709	1.073107	9.93
. . .	. . . . .	. . .	. . .	. . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
95	ihr	118	83	201	90754	0.0010	0.4538	0.0100083	3.131188	28.99
96	Jahre	75	116	191	90945	0.0010	0.4548	0.009580	3.140768	29.07
97	um (*prtc.*)	123	68	191	91136	0.0010	0.4558	0.009581	3.150349	29.16
98	sei	67	123	190	91329	0.0010	0.4568	0.009538	3.169425	29.25
99	sehr	94	96	190	91516	0.0010	0.4578	0.009538	3.169425	29.35
100	jetzt	103	84	187	91703	0.0009	0.4587	0.009409	3.178834	29.43
101	Jahr	87	97	184	91887	0.0009	0.4596	0.009279	3.188113	29.51
. . .	. . . . .	. . .	. . .	. . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
1000	Sprache	18	2	20	135032	0.0001	0.6752	0.0013288	5.722406	52.98
. . .	. . . . .	. . .	. . .	. . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
11622	-			2	183175	0.00001	0.9159	0.0001660	9.347291	86.54
. . .	. . . . .	. . .	. . .	. . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .

Примечания. Расшифровку индексов, использующихся для различения конверсионных омонимов см. в списке сокращений. N (ГДР) = 100 000; N (ФРГ)=100 000; N(общенем.) = 200 000.

Таблица 3

Фрагмент ЧС слов русского подъязыка электроники (см.: Калинина, 1968а, 1968б)

i	Слово	F	f
1	в (во)	6686	0.0332639
2	и	5742	0.0285671
3	при	3924	0.0195224
4	на	3914	0.0195223
5	с (со)	2228	0.0110845
. . .	. . . .	. . . .	. . . . .
21	к (ко)	1052	0.0052338
22	время (времени, временем, ...)	942	0.0046865
. . .	. . . . .	. . . .	. . . . .

Примечание. N = 20094.

ного процесса. Иными словами, предполагается, что производящие текст система, норма и узус, а также сопутствующий им комплекс условий остаются неизменными, в связи с чем распределение вероятностей в тексте для всех рассматриваемых лексических единиц также остается неизменным. В действительности каждый реальный текст является нестационарным процессом. Распределение вероятностей употребления многих слов и словоформ зависит от темы текста, времени его написания (эта зависимость имеет место даже в том случае, когда речь идет о коротком хронологическом периоде), от возраста, образования и стилевых вкусов автора.

2. Моделирующий текст рассматривается как последовательность независимых испытаний, не учитывающих вероятностные связи (лексико-грамматические и стилистические валентности), существующие между словоупотреблениями текста (ср.: Калинина, 1968а, с. 66; Mandelbrot, 1961, с. 191—193).

Хотя оба упрощающих предположения делают текстовую модель мало похожей на реальный текст и приводят к заметным потерям смысловой и синтактической информации (ср. 3), они дают возможность применить для исследования построенной модели математический аппарат, а также организовать на основе этих исследований объективный отбор лексики и морфологии, необходимой для построения тех или иных программ автоматической переработки текста.

19. Машинное построение частотного словаря

Методика подготовки текстового материала и ручного составления ЧС была подробно описана во многих работах (ср.: Алексеев, 1971б; Ешан*, 1966; Лукьяненков*, 1969; Новак*, 1962).

Поэтому этих вопросов мы касаться здесь не будем, а остановимся на машинных алгоритмах составления ЧС.

Машинные алгоритмы ЧС словоформ и словосочетаний, ориентированные на ЭВМ *второго поколения*, предусматривали последовательное выполнение нескольких самостоятельных задач. При этом текст вводился порциями, каждая из которых включала от одной до пяти тысяч словоупотреблений (ср.: Зубов, Лукьяненков, Пиотровский, Хотяшов, 1968, с. 108—119; Вертель, Вертель, 1970, с. 290—311). Такой подход значительно увеличивал время получения ЧС, а ограниченный объем памяти ЭВМ второго поколения не позволял перерабатывать за один выход большие массивы текста.

Эти недостатки удается преодолеть при использовании машин с байтовой структурой (ЕС-ЭВМ).

Теперь разбивка всей задачи на отдельные блоки (см. рис. 13) является чисто условной, поскольку процесс получения ЧС происходит непрерывно.

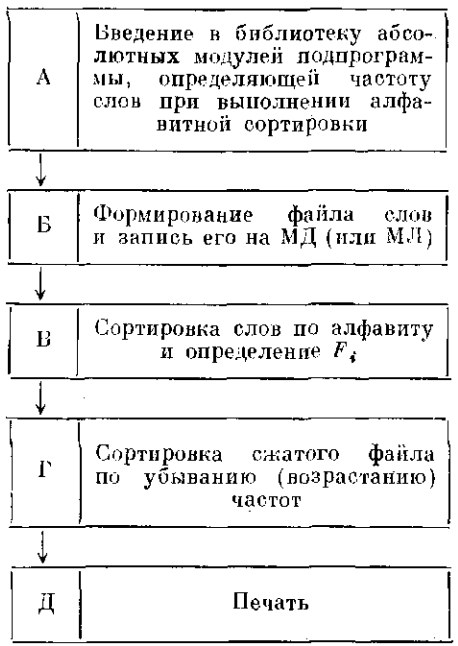


Рис. 13. Принципиальная блок-схема алгоритма получения ЧС на ЕС-ЭВМ (Карпиловч, 1973, с. 454).

Программирование алгоритма осуществлено на Ассемблере. Для записи словоупотребления отводится поле в 28 байт. При этом машинные документы (коды словоформы и его частоты) не накапливаются в оперативной памяти (см. рис. 14), но переносятся сразу

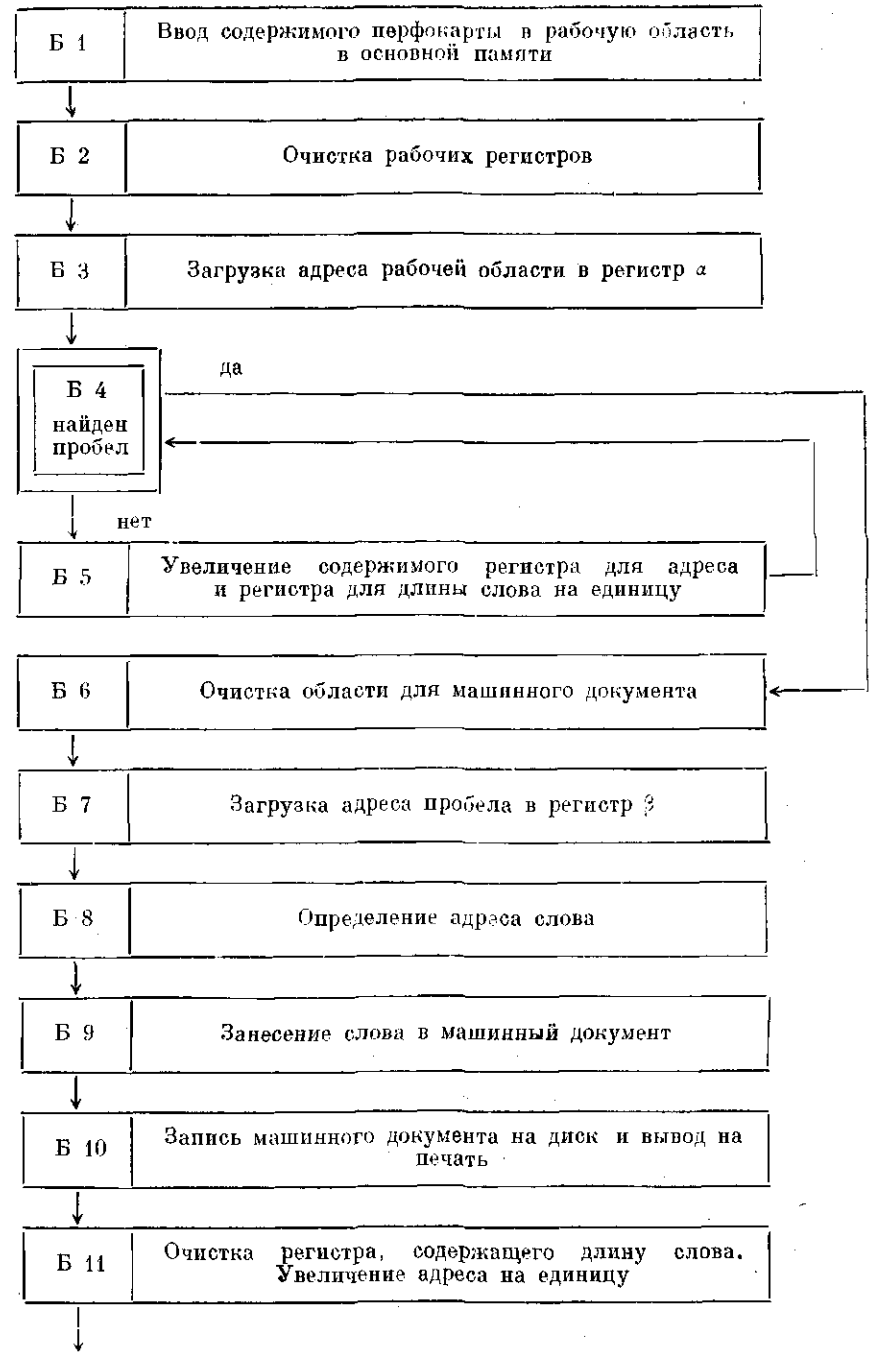


Рис. 14. Расшифровка блока Б в алгоритме получения ЧС на ЕС-ЭВМ (Карпилович, 1973, с. 455—456).

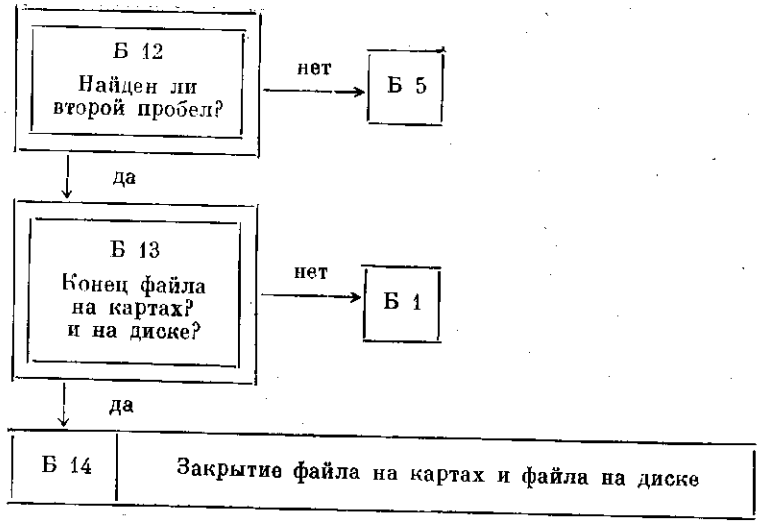


Рис. 14 (*продолжение*).

на МД. При этом отпадает необходимость вводить в ЭВМ текст малыми порциями, что значительно ускоряет его переработку. За один выход удается перерабатывать тексты общим объемом в 500 тыс. словоупотреблений. Сортировки на МД и МЛ осуществляются с помощью универсальной программы, имеющейся в библиотеке программ ЕС-ЭВМ (подробнее см.: Карпилович, 1973, с. 450—459).

20. Отбор лексики из частотного словаря

Как известно, частотный словарь представляет собой список лексических единиц, расположенных в порядке убывания их частот употребления в исследуемой текстовой выборке длиной в N словоупотреблений. Всем лексическим единицам ЧС присвоены номера (i), которые возрастают в соответствии с убыванием частот F_i. Одновременно все выявленные в тексте лексические единицы могут упорядочиваться по алфавиту, образуя алфавитно-частотные списки (рис. 15). Если при каждой словоформе списка указывается ее частота в каждой из внутренних выборок, а также приводятся общие характеристики этих эмпирических распределений, то мы имеем дело с так называемыми распределительными словарями (Копылова и др., 1974, с. 194—196) (см. рис. 16, 17).

Словоформы могут объединяться в так называемые обратные словари, в которых лексические единицы располагаются по алфавиту последних букв (рис. 18). Алфавитно-частотные списки используются для приведения в АС найденных словоформ к одной

	i		F
	129	КВТ	37
	130	КГ4	2
○	131	КГ4АС	6
	132	КГ	5
	133	КГМ	1
○	134	КГСЕК	2
	135	КГСМ	13
	136	КГССМ	6
○	137	КЕРАМИ4ЕСКИЕ	1
	138	КЕРАМИ4ЕСКУЮ	1
	139	КИЛОВАТТ-4АСА	1
	140	КИЛОВАТТА	2
	141	КИЛОГРАММЕ	1
	142	КИНЕМАТИ4ЕСКУЮ	1
	143	КИНЕТИ4ЕСКОЙ	3
	144	КИНЕТИ4ЕСКУЮ	1
	145	КИНЕТИКИ	1
○	146	КИНЕТИКОЙ	1
	147	КИНОКАДРЫ	1
	148	КИПЕНИЕ	11
○	149	КИПЕНИИ	5
	150	КИПЕНИЯ	17

Рис. 15. Фрагмент алфавитно-частотного словаря русских научно-технических текстов, полученного на ЭВМ ЕС-1020.

лексеме (основе, исходной форме слова и т. п.). Обратные словари применяются при выявлении истинных и ложных флексий, использующихся в исследуемом подъязыке, а также для построения статистической модели его морфологии (Пиотровская, 1973).

Кроме значений i и F_I, в частотном, а иногда также в обратном и алфавитно-частотном списках указываются следующие величины: накопленная частота F_i^*, представляющая сумму данной частоты F_i и всех предшествующих ей частот; относительная частота (частость) $f_i = \frac{F_i}{N}$; накопленная относительная частота (накопленная частость) f_i^* (ср. табл. 2); удельная энтропия

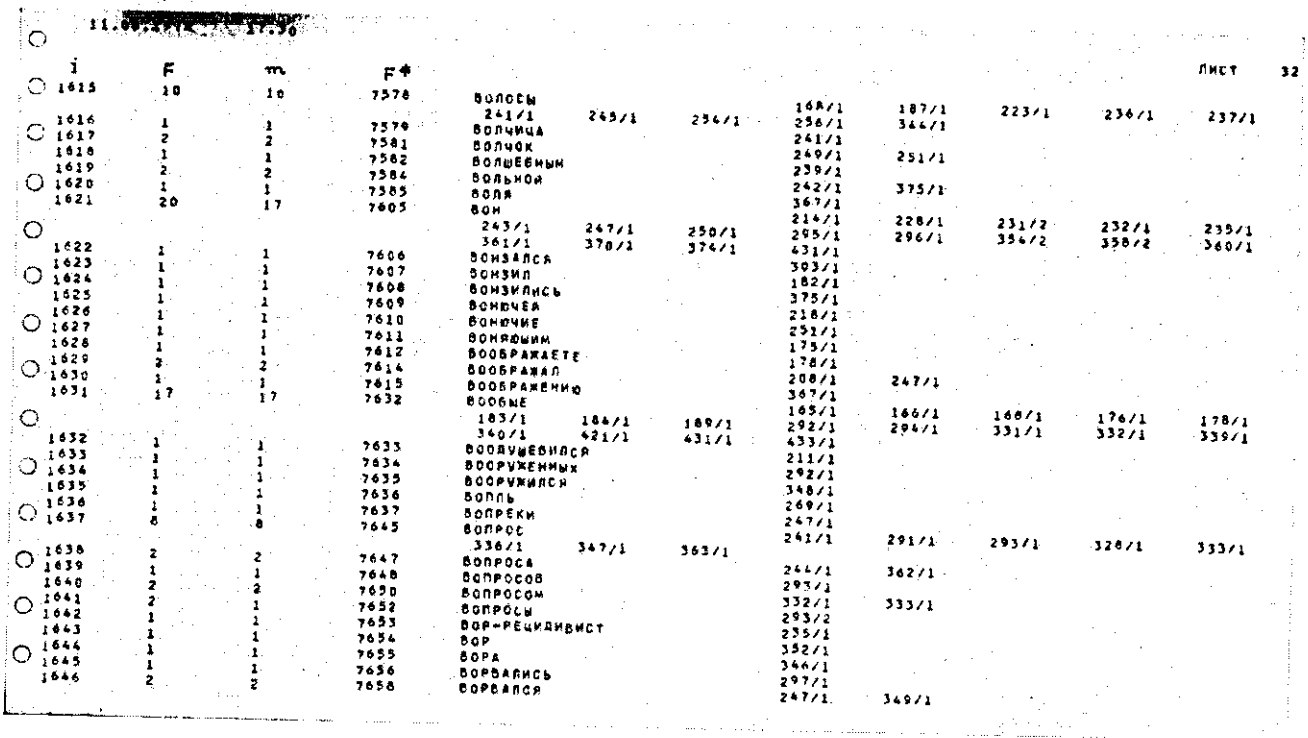
$$\bar{H}_i = -f_i \log_2 f_i; \tag{1}$$

накопленная удельная энтропия

$$\bar{H}_i^* = -\sum_{h=1}^{i} f_h \log_2 f_h. \tag{2}$$

Кроме того, для каждого значения f_i задается максимальная относительная ошибка наблюдения δ_i, вычисляемая с помощью упрощенного выражения

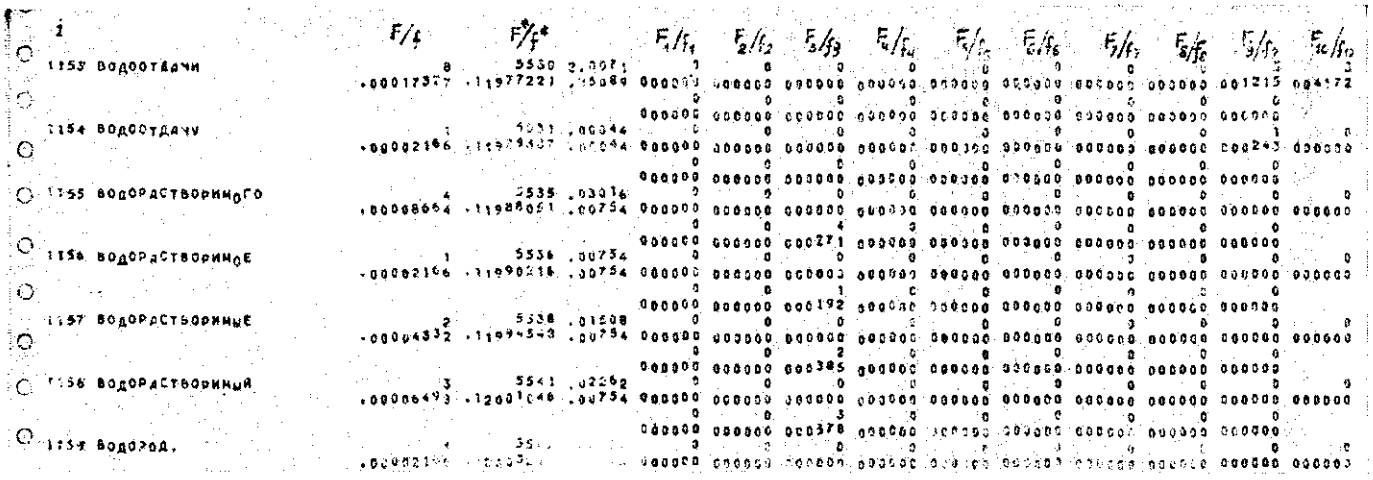
$$\delta_i = \frac{Z_p}{\sqrt{Nfi}} = \frac{Z_p}{\sqrt{Fi}}, \tag{3}$$



i	F	m	F*												Лист 32
1615	10	10	7578	ВОЛОСЫ											
				241/1	265/1	256/1	168/1	187/1	223/1	236/1	237/1				
1616	1	1	7579	ВОЛЧИЦА			256/1	344/1							
1617	2	2	7581	ВОЛЧОК			241/1								
1618	1	1	7582	ВОЛШЕБНЫМ			249/1	251/1							
1619	2	2	7584	ВОЛЬНОЙ			239/1								
1620	1	1	7585	ВОЛЯ			242/1	375/1							
1621	20	17	7605	ВОН			367/1								
				243/1	247/1	250/1	214/1	228/1	231/2	232/1	235/1				
				361/1	370/1	374/1	295/1	296/1	356/2	358/2	360/1				
1622	1	1	7606	ВОНЗАЛСЯ			431/1								
1623	1	1	7607	ВОНЗИЛ			303/1								
1624	1	1	7608	ВОНЗИЛИСЬ			182/1								
1625	1	1	7609	ВОНЮЧЕЙ			375/1								
1626	1	1	7610	ВОНЮЧИЕ			218/1								
1627	1	1	7611	ВОНЯЮЩИМ			251/1								
1628	1	1	7612	ВООБРАЖАЕТЕ			178/1								
1629	2	2	7614	ВООБРАЗИЛ			175/1								
1630	1	1	7615	ВООБРАЖЕНИЮ			178/1								
1631	17	17	7632	ВООБЩЕ			208/1	247/1							
				183/1	184/1	189/1	367/1								
				340/1	421/1	431/1	165/1	166/1	168/1	176/1	178/1				
1632	1	1	7633	ВООБЩЕВИЛСЯ			292/1	294/1	331/1	332/1	339/1				
1633	1	1	7634	ВООРУЖЕННЫХ			433/1								
1634	1	1	7635	ВООРУЖИЛСЯ			211/1								
1635	1	1	7636	ВОПЛЬ			292/1								
1636	1	1	7637	ВОПРЕКИ			348/1								
1637	8	8	7645	ВОПРОС			269/1								
				336/1	347/1	363/1	247/1								
1638	2	2	7647	ВОПРОСА			241/1	291/1	293/1	328/1	333/1				
1639	1	1	7648	ВОПРОСОВ			244/1	362/1							
1640	2	2	7650	ВОПРОСОМ			293/1								
1641	2	1	7652	ВОПРОСЫ			332/1	333/1							
1642	1	1	7653	ВОР-РЕЦИДИВИСТ			293/2								
1643	1	1	7654	ВОР			235/1								
1644	1	1	7655	ВОРА			352/1								
1645	1	1	7656	ВОРВАЛИСЬ			346/1								
1646	2	2	7658	ВОРВАЛСЯ			297/1								
							247/1	349/1							

Рис. 16. Фрагмент машинного алфавитно-частотного распределительного словаря русских беллетристических и поэтических текстов (автор алгоритма и программы для ЭВМ «Минск-32» — Е. В. Вертель).

Справа от словоформ указаны: в числителе — номер серии (объем серии 1000 словоупотреблений), *в знаменателе* — количество употреблений словоформы в данной серии.



i	F/f	F*/f*	F_1/f_1	F_2/f_2	F_3/f_3	F_4/f_4	F_5/f_5	F_6/f_6	F_7/f_7	F_8/f_8	F_9/f_9	F_{10}/f_{10}
1153 ВОДООТДАЧИ	8	5530 2.00715	1	0	0	0	0	0	0	0	1	1
	.00017327	.11977221 .05089	000000	000000	000000	000000	000000	000000	000000	000000	001215	004172
			0	0	0	0	0	0	0	0	0	
1154 ВОДООТДАЧУ	1	5531 .00044	000000	000000	000000	000000	000000	000000	000000	000000	000000	
	.00002166	.11979387 .00754	0	0	0	0	0	0	0	0	1	0
			000000	000000	000000	000000	000000	000000	000000	000000	000243	000000
1155 ВОДОРАСТВОРИМОГО	4	5535 .00316	0	0	0	0	0	0	0	0	0	0
	.00008664	.11988051 .00754	000000	000000	000000	000000	000000	000000	000000	000000	000000	000000
			0	0	4	0	0	0	0	0	0	
1156 ВОДОРАСТВОРИМОЕ	1	5536 .00734	000000	000000	000271	000000	000000	000000	000000	000000	000000	
	.00002166	.11990218 .00754	0	0	0	0	0	0	0	0	0	0
			000000	000000	000003	000000	000000	000000	000000	000000	000000	000000
1157 ВОДОРАСТВОРИМЫЕ	2	5538 .01508	0	0	1	0	0	0	0	0	0	
	.00004332	.11994543 .00754	000000	000000	000192	000000	000000	000000	000000	000000	000000	
			0	0	0	0	0	0	0	0	0	0
1158 ВОДОРАСТВОРИМЫЙ	3	5541 .02262	000000	000000	000385	000000	000000	000000	000000	000000	000000	000000
	.00006493	.12001040 .00754	0	0	0	0	0	0	0	0	0	0
			000000	000000	000578	000000	000000	000000	000000	000000	000000	000000
1159 ВОДОРОД,	4	55..	0	0	0	0	1	1	0	1	0	0
	.00008...	.....	000000	000000	000000	000000	000...	000...	000000	000...	000000	000000

Рис. 17. Фрагмент машинного алфавитно-частотного распределительного словаря русских научно-технических текстов (авторы алгоритма и программы для ЭВМ «Минск-22» — И. С. Кравцова и М. В. Кулешова).

В 5-м столбце указаны: значения коэффициентов употребительности — $K_у$ (*числитель*) и $K_р$ (*знаменатель*), определенные по формулам:

$$K_р = \frac{H(Я|w_i)}{H(Я)},$$

где $H(Я)$ — энтропия вероятностных весов подъязыков, а $H(Я/w_i)$ — энтропия этих весов относительно словоформы w_i;

$$K_у = P(w_i)\, K_р,$$

где $p(w_i)$ — вероятность словоформы w_i (см.: Копылова, Кравцова, Кулешова, 1974, с. 194—196).

В *6—15-м столбцах показаны*: значения F и f словоформы w_i для каждой из девятнадцати частных выборок (серий).

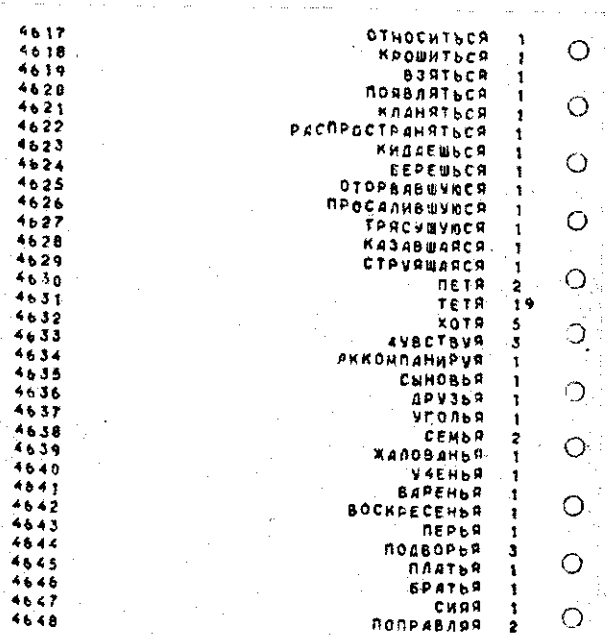


Рис. 18. Фрагмент обратного частотного словаря русских беллетристических текстов, полученного на ЭВМ «Минск-32» (автор алгоритма и программы — Е. В. Вертель).

где $\mathfrak{p}$ — надежность (доверительная вероятность наших утверждений), а $Z_{\mathfrak{p}}$ — коэффициент, связанный с $\mathfrak{p}$ табулированной зависимостью, показанной в табл. 4. Иногда указываются также верхняя ($p_i^в$) и нижняя ($p_i^н$) границы того интервала, в котором лежит истинная вероятность i-той лексической единицы. При этом

$$p_i^в = \frac{Nf_i + \frac{1}{2}Z_{\mathfrak{p}}^2 + Z_{\mathfrak{p}}\sqrt{Nf_i(1-f_i) + \frac{1}{4}Z_{\mathfrak{p}}^2}}{N + Z_{\mathfrak{p}}^2};$$

$$p_i^н = \frac{Nf_i + \frac{1}{2}Z_{\mathfrak{p}}^2 - Z_{\mathfrak{p}}\sqrt{Nf_i(1-f_i) + \frac{1}{4}Z_{\mathfrak{p}}^2}}{N + Z_{\mathfrak{p}}^2}. \qquad (4)$$

Подробнее о выводе зависимостей (3), (4) см.: Калинина, 1968а, с. 68; Бектаев, Пиотровский, 1973, с. 249—252; 1974, с. 128—134, 156—161, 295—296.

При решении прикладных задач отбор лексики осуществляется обычно исходя из заданной величины накопленной частости,

которая вместе с относительной ошибкой наблюдения δ_i позволяет оценить то количество наиболее частых словоформ или слов, которое следует отобрать в учебный или автоматический словарь для того, чтобы обеспечить заданное покрытие наугад взятого текста.

Рассмотрим эту задачу на примере отбора лексики в словарь-минимум БЯ, предназначенный для оптимизации преподавания иностранных языков.

Исследования Л. П. Герман-Прозоровой показали, что владение элементами грамматики иностранного языка, сочетающееся со знанием не менее 70% словоупотреблений специального иностранного текста, позволяет читателю, знакомому с его тематикой, понять смысл этого текста (Алексеев, Герман-Прозорова, Пиотровский, Щепетова, 1974, с. 220). Это значит, что входной словник учебного словаря-минимума, призванного обеспечить общее понимание иностранного текста, должен включать начальный массив ЧС, ограниченный снизу словоформой, имеющей накопленную частоту $f^* \simeq 0.7$.

Таблица 4

Зависимость коэффициента $Z_{\mathfrak{p}}$ от величины надежности $\mathfrak{p}$

$\mathfrak{p}$	$Z_{\mathfrak{p}}$
0.9000	1.64
0.9500	1.96
0.9544	2.00
0.9750	2.24
0.9800	2.33
0.9900	2.58

В английском подъязыке электроники 70-процентную покрываемость текста обеспечивают около 500 наиболее частых словоформ (Алексеев, 1971б, с. 11). Поэтому в базовый учебный словарь по английской электронике Л. П. Герман-Прозоровой (1973) было включено, за вычетом интернационализмов, 463 высокочастотные словоформы из ЧС словаря, составленного П. М. Алексеевым (1968, с. 151—161), что соответствует 270 словам. Соответственно во французский базовый словарь по электронике следовало бы взять около 500 наиболее частых словоформ (или 200 слов) — см.: Кочеткова, Скрелина, 1968, с. 167—170; в немецкий — около 1000 словоформ (ок. 400 слов) — Зореф *, 1972, с. 7.

В русский же словарь-минимум по электронике для обеспечения покрываемости в 70% следовало бы взять около 1800 словоформ или соответственно 400 слов (Калинина, 1968а, с. 72—73).

Само собой разумеется, что человеку удается извлечь необходимую информацию из текста с помощью этого ограниченного словаря только потому, что он опирается на обширные знания тематики текста и на богатые эвристические способности своего мозга. Этими возможностями современный компьютер не обладает. Поэтому, как показывает опыт, для того чтобы обеспечить распознавание общего смысла перерабатываемого текста, машине должны быть заданы массивы лексики, обеспечивающие гораздо более высокую покрываемость.

Таблица 5

Основные характеристики частотных списков словоформ для различных языков и стилей

Строй языка	Язык	Функциональные стили											
		разговорная речь				научно-техническая и деловая речь				публицистика			
		тематика, авторы ЧС	N тыс.	L тыс.	покрываемость контрольного текста, %	тематика, авторы ЧС	N тыс.	L тыс.	покрываемость контрольного текста, %	тематика, источники, авторы ЧС	N тыс.	L тыс.	покрываемость контрольного текста, %
Флективно-аналитический с остаточным синтетизмом	Английский	Интервью — Schonell, Middelton, Shaw, 1956.	500	13.0	—	Судовые механизмы — Лукьяненков*, 1969.	400	13.0	99.0	Газеты Великобритании и США, политика и экономика — Алексеев, Турыгина, 1974, с.4.	200	23.6	0.98
		Устная речь — Howes, 1966.	250	9.7	—	Полупроводниковая техника,—Гончаренко*, 1972; Тарасова*, 1974.	300	12.1	98.5				
						Электроника — Алексеев, 1971.	200	10.6	97.3				
	Французский					Тракторное и с/х машиностроение — Рахубо*, 1974.	300	13.0	—				
						Электроника—Кочеткова*, 1969.	200	10.3	97.9				
	Шведский												
Флективно-аналитический с элементами синтетизма	Румынский и молдавский					Румынская электроника — Ешан*, 1966.	200	14.3	96.0	Газеты МССР—Маткаш*, 1967.	200	25.8	93.4
	Испанский					Радиотехника—Михайлова*, 1972.	212	13.5	97.0				
Флективный с преобладанием синтетизма	Немецкий					Экономическая литература ГДР—Вертель*, 1972, с. 212—213.	400	24.2	—	Газеты ГДР и ФРГ—Ротарь*, 1971.	200	28.4	91.4
						Электроника—Зореф*, 1972.	209	20.4	94.8				
	Русский					Радиотехника—Междумова, 1973, с. 7, 15.	400	28.7	98.0				
						Электроника — Калинина, 1968б, с. 87.	201	21.5	—				
Агглютинирующий	Казахский												

Строй языка	Язык	Функциональные стили											
		беллетристика				смешанные выборки							
		тематика, источники, авторы ЧС	N тыс.	L тыс.	покрываемость контрольного текста, %	авторы ЧС	N тыс.	L тыс.	покрываемость контрольного текста, %				
Флективно-аналитический с остаточным синтетизмом	Английский	Роман Дж. Джойса «Улисс» — Hanley, 1965.	260	29.9	—	Thorndike, Lorge, 1944.	1800~30.0		—				
						Kučera, Francis, 1967.	1000~50.4		99.2				
						Carroll, 1971.	5089~86.7		—				
	Французский												
	Шведский					Allén, 1970.	1000~112.4						
Флективно-аналитический с элементами синтетизма	Румынский и молдавский												
	Испанский												
Флективный с преобладанием синтетизма	Немецкий					Meier, H, 1964, с. 114.	10911	258.2	—				
								25.6	94.4				
								41.1	96.3				
	Русский												
Агглютинирующий	Казахский	Роман М. Ауэзова «Абай жолы» — Джубанов, 1973.	466	61.4	—								

Современный опыт построения и эксплуатации автоматических систем реферирования, аннотирования и перевода текста (Вертель и др., 1971, с. 89, 98; Тарасова*, 1974; Рахубо*, 1974) показывает, что более или менее удовлетворительное распознавание смысла текста достигается тогда, когда в память машины вводится лексика, извлеченная из частотных списков, обеспечивающих покрываемость контрольного текста[1] на 97—99%. Объем выборки N, необходимой для вывода ЧС заданной покрываемости, равно как и объем L получаемого словника, зависят, с одной стороны, от строя языка, а с другой — от стилевой характеристики текста.

Для аналитических фузионных языков (английский, французский) 97—99-процентную покрываемость относительно научно-технических текстов определенной тематики показывают словники, которые извлекаются из выборок, состоящих не менее чем из 200 тыс. словоупотреблений. По мере нарастания синтетичности языка объем списка L словоформ, извлекаемого из выборки фиксированной длины, естественным образом увеличивается. В то же время покрываемость контрольного текста при этом заметно падает (см. табл. 5). Таким образом, при построении входных автоматических словарей научно-технической тематики для таких синтетических языков, как русский и немецкий, необходимо пользоваться статистикой, опирающейся на выборки не менее чем в 300—400 тыс. словоупотреблений.

Лингвостатистические исследования последних десятилетий показывают также, что богатство словника возрастает, а покрываемость контрольных текстов падает при переходе от научно-технических текстов к публицистическим, беллетристическим текстам и особенно к выборкам смешанного характера (ср. табл. 5). Все это, разумеется, указывает на необходимость увеличения объемов исходных выборок при построении ЧС, которые затем должны быть использованы при формировании словников автоматических словарей, предназначенных для перевода, аннотирования и реферирования общественно-политических и беллетристических текстов.

21. Статистическая модель текста на уровне словосочетаний. Иконический выбор сегментов

Как уже говорилось, ЧС слов и словоформ строится на допущении, что моделирующий текст выступает в качестве последовательности независимых испытаний, пренебрегающих лексико-грамматическими связями между словоупотреблениями текста. Это упрощение приводит к тому, что построенный на основе ЧС автоматический словарь перерабатывает лексические единицы текста только на уровне их виртуальных (словарных) значений и не учитывает актуальных значений у словоупотреблений текста.

[1] Контрольным текстом считается текст, принадлежащий данному подъязыку, но не входящий в экспериментальную выборку.

Эксплуатация экспериментальных систем МП, использующих эти словари, показала их слабую эффективность (Кочеткова*, 1969, с. 11; Вертель и др., 1971, с. 122—127; Linguistic and Engineering Studies . . . , 1960, с. 183—348; Masterman, 1967). Выдаваемый ЭВМ выходной текст содержит при этом много лишних эквивалентов, которые создают сильный информационный шум, часто мешающий потребителю понять содержание этого текста.

Автоматическая переработка текста, учитывающая актуальные значения составляющих его словоформ, становится возможной при условии, если используемая в ЭВМ лингвистическая модель включает списки типовых контекстов и валентностей.

Выявление типовых контекстов и дистрибуций также осуществляется с помощью статистических моделей текста, в роли которых выступают ЧС словосочетаний.

Словосочетания, реализующие типовые контексты и дистрибуции, чаще всего представляют собой трехсловные сочетания (триады), либо двухсловные (диады), четырехсловные (тетрады), пятисловные (пентады) и т. д. цепочки знаменательных и служебных словоформ (ср.: Kaplan, 1955). Выборка этих словосочетаний в ЧС может производиться вручную или на ЭВМ с помощью иконической или эвристико-алгоритмической процедур.

Иконическая процедура предусматривает выбор из текста сегментов фиксированной длины, например триад или тетрад. Из текста могут выбираться либо все сегменты подряд, либо может быть применен целенаправленный отбор.

При сплошном выборе триады и тетрады извлекаются путем последовательного продвижения по тексту со сдвигом на одно словоупотребление вправо. Например, имеется следующий русский текст:

Совместимость с машинами «Минск-2, 22, 22М». Выполнение программ машин «Минск-2, 22, 22М» на машине «Минск-32» обеспечивается аппаратурными и программными средствами (Кушнерев, Неменман, Цагельский, 1973, с. 7).

Из этого текста указанным формальным путем можно выбрать следующие контактно связанные триады:

1) Δ *Совместимость с*
2) *Совместимость с машинами*
3) *с машинами «Минск-2, 22, 22М»*
4) *машинами «Минск-2, 22, 22М»* Δ
5) *«Минск-2, 22, 22М»* Δ *Выполнение*
6) Δ *Выполнение программ*
7) *Выполнение программ машин*
8) *программ машин «Минск-2, 22, 22М»*
9) *машин «Минск-2, 22, 22М» на*
10) *«Минск-2, 22, 22М» на машине*
11) *на машине «Минск-32»*
12) *машине «Минск-32» обеспечивается*
13) *«Минск-32» обеспечивается аппаратурными*

14) *обеспечивается аппаратурными и*
15) *аппаратурными и программными*
16) *и программными средствами*
17) *программными средствами* Δ

Среди перечисленных триад есть вполне осмысленные и грамматически оформленные словосочетания (триады 2, 3, 8, 11), вместе с тем имеется немало бессмысленных цепочек (триады 5, 9, 10, 12), а также не полностью оформленных с грамматической и семантической точки зрения трехсловных сочетаний (триады 1, 4, 6—8, 13—17).

Хотя описанная процедура дает возможность извлекать из текста все не только контактно связанные, но и оформленные дистантными связями словосочетания заданной длины, применить ее при статистическом описании больших массивов текста не удается. Во-первых, число различных триад или тетрад в несколько раз больше, чем количество разных словоформ, образующих словник ЧС, одновременно частоты каждого из этих словосочетаний очень малы. Поэтому для получения достоверных статистических данных даже о наиболее частых триадах и тетрадах пришлось бы использовать выборки текста, во много раз превосходящие выборки, применяющиеся при построении ЧС словоформ и слов. Во-вторых, получаемый список содержит десятки и даже сотни тысяч случайных бессмысленных комбинаций, а также не полностью оформленных цепочек, которые не всегда удается однозначно отделить от семантически и грамматически оформленных словосочетаний.

В этих условиях ручной поиск в полученном списке оформленных цепочек связан с большими временными затратами и, что самое главное, осуществляется исходя из лингвистической интроспекции исследователя. В итоге вся процедура выделения типовых контекстов теряет здесь свой формальный и объективный характер.

Более экономным является целенаправленный выбор из текста триад и тетрад, опирающихся на заранее заданные опорные слова-ядра, которые находятся в строго фиксированной позиции относительно остальных элементов словосочетания.

Так, например, в рассмотренном выше тексте можно выделить все предложно-именные конструкции, задав следующую процедуру выборки: в качестве ядра принимается каждый встретившийся в тексте предлог, от которого берется два словоупотребления вправо. Сформированные таким образом триады:

с машинами «Минск-2, 22, 22М»
на машине «Минск-32»

выбираются в ЧС предложно-именных микроконтекстов. Используя эту процедуру, Т. Р. Зуева составила на выборке в 2 млн словоупотреблений (русские тексты по электронике) ЧС четырехсловных предложно-именных словосочетаний объемом в 100 тыс. тетрад (см.: Зуева, 1970).

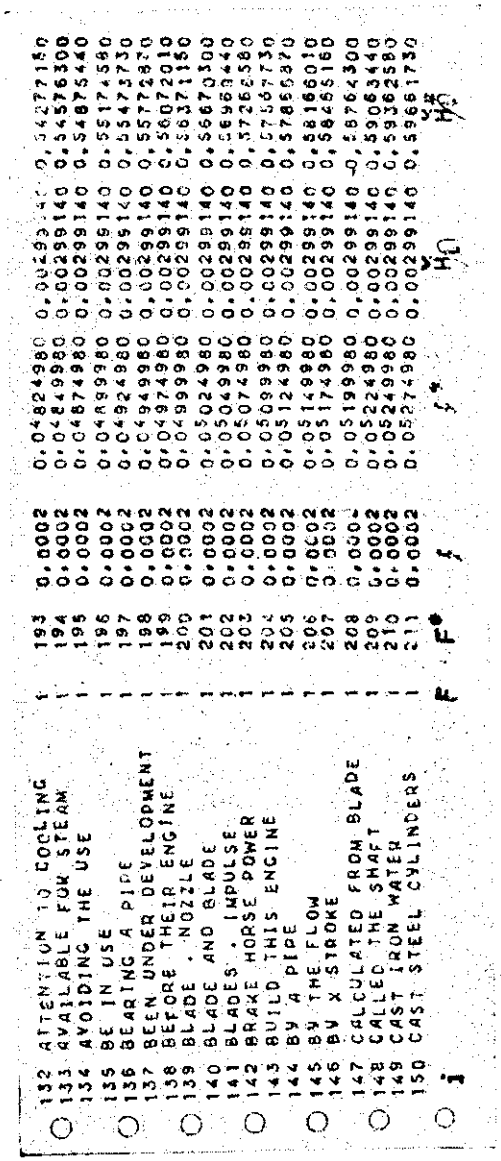
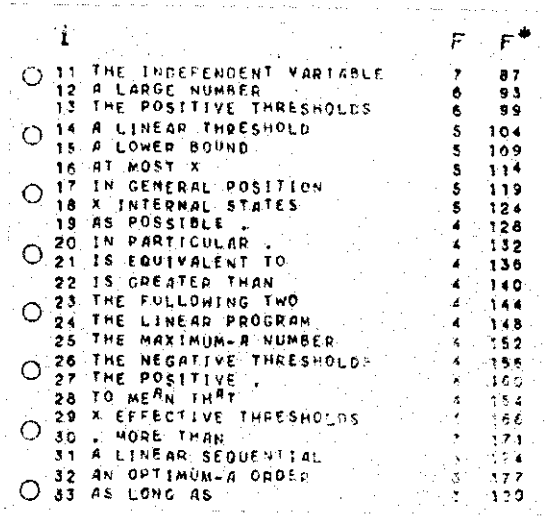


Рис. 19. Фрагмент частотного списка английских трехсловных сочетаний, извлеченных из текстов по судовым механизмам (авторы алгоритма и программы — А. В. Зубов и К. Ф. Лукьяненков).



i		F	F*
○	11 THE INDEPENDENT VARIABLE	7	87
	12 A LARGE NUMBER	6	93
	13 THE POSITIVE THRESHOLDS	6	99
○	14 A LINEAR THRESHOLD	5	104
	15 A LOWER BOUND	5	109
	16 AT MOST X	5	114
○	17 IN GENERAL POSITION	5	119
	18 X INTERNAL STATES	5	124
	19 AS POSSIBLE .	4	128
○	20 IN PARTICULAR .	4	132
○	21 IS EQUIVALENT TO	4	136
	22 IS GREATER THAN	4	140
○	23 THE FOLLOWING TWO	4	144
○	24 THE LINEAR PROGRAM	4	148
	25 THE MAXIMUM-A NUMBER	4	152
○	26 THE NEGATIVE THRESHOLDS	4	156
○	27 THE POSITIVE .	4	160
	28 TO MEAN THAT	4	164
○	29 X EFFECTIVE THRESHOLDS	4	168
○	30 . MORE THAN	3	171
	31 A LINEAR SEQUENTIAL	3	174
	32 AN OPTIMUM-A ORDER	3	177
○	33 AS LONG AS	3	180

Рис. 20. Фрагмент частотного списка английских трехсловных сочетаний (авторы алгоритма и программы — В. С. Крисевич и В. А. Ноздрина).

С помощью аналогичных процедур на больших текстовых выборках были получены частотные и алфавитно-частотные списки именных триад для русского подъязыка электроники (Белоцерковская, 1973, с. 3) и русских публицистических текстов (Никитина, 1971, с. 262—285), для английских текстов по виноделию и полупроводниковой технике (Тарасова*, 1974, с. 10), геологии нефти и газа (Колесникова*, 1974), радиоэлектронике (Соркина*, 1969, с. 473—505), сельхозмашиностроению (Добрускина*, 1973), строительству (Борисевич*, 1969, с. 506—514), судовым механизмам (Лукьяненков*, 1969; Каменева*, 1973) — ср. рис. 19, французского публицистического (Орлова, Буравцева, 1973, с. 103—137) и сельскохозяйственного (Рахубо*, 1974, с. 9) текста, а также для немецкого подъязыка публицистики (Чижаковский*, 1971). Получены также списки адъективных триад для английского подъязыка электроники (Ноздрина, 1969) — см. рис. 20, а также ЧС глагольных триад и тетрад для английского подъязыка электроники (Данейко, 1969, с. 452—472), для французских (Ионицэ*, 1971, с. 5—8) и немецких (Мануков*, 1972, с. 4—5) публицистических текстов (в последних двух случаях выбирались текстовые сегменты не только с контактной, но и дистантной связью).[1]

[1] Кроме того, для разных европейских языков построено большое количество ЧС триад, для которых ядрами послужили наиболее частые единицы соответствующих ЧС словоформ (ср.: Боркун, 1969, с. 441—451; Булашева, 1969, с. 376—387; Машкина*, 1968; Нехай*, 1968; Чапля*, 1967, с. 4—7; Шаранда*, 1968).

22. От фиксированных сегментов к устойчивым словосочетаниям и типовым контекстам

Хотя только что рассмотренная целенаправленная выборка дает гораздо меньшее число бессмысленных и не до конца оформленных цепочек, совокупность всех выбранных с ее помощью триад (диад, тетрад и т. д.) нужно рассматривать в качестве лингвистического полупродукта. Обычно все эти сегменты подвергаются дальнейшей обработке с целью выделения достаточно часто употребляющихся грамматически и семантически оформленных словосочетаний разной длины — назовем их условно устойчивыми словосочетаниями (УС). Эти словосочетания в дальнейшем либо помещаются в словарь машинных оборотов, либо анализируются с точки зрения выделения типовых дистрибуций и контекстных схем. При этом имеют место следующие ситуации.

1. Триада (диада, тетрада и т. д.) является в полном смысле слова свободным словосочетанием, образованным простым контактным примыканием словоформ без какого бы то ни было согласования, управления, а также без лексической идиоматичности — например, англ., фр. cf. fig. 1; ср. рис. 1. Такие триады в списки машинных оборотов и типовых контекстов не включаются.

2. Триада включает УС, состоящее из двух словоформ (соответственно в тетраду могут входить УС, состоящие из двух или трех словоформ и т. д.). Например, в триады at most X, as possible Δ, in particular Δ, Δ more than, приведенные в машинном ЧС английских адъективных триад (рис. 20), содержат двусловные УС at most 'самое большее', as possible 'как можно', in particular 'в частности', more than 'более чем'. Все эти двусловные УС легко выделяются путем отсечения лишних элементов X и Δ, а затем переносятся в соответствующие списки.

3. Триада полностью совпадает с УС. Ср., например, такие триады из ЧС подъязыка судовых механизмов (рис. 19), как be in use 'быть в употреблении, использоваться' или as long as 'пока'. Такие триады целиком переносятся в списки машинных оборотов или типовых контекстов.

4. Триада (диада, тетрада и т. д.) представляет собой сегмент, являющийся частью УС, которое может быть «достроено» путем добавления к триаде недостающих словоформ. Рассмотрим подробнее методику этого «достраивания».

Недостающие в триаде или тетраде элементы УС можно иногда восстановить путем обращения к интуиции или фразеологическому словарю. Например, по сегменту . . .*Национальностей Верховного Совета* (Никитина, 1971, с. 266) восстанавливается с вероятностью, близкой к единице, устойчивое словосочетание *Совет* (*а*, *у*, *ом*, *е*) *Национальностей Верховного Совета*; очень частотная английская триада at the same (Боркун, 1969, с. 443) скорее всего входит во фразеологизм at the same time 'вместе с тем, одновременно', и т. д. (Кунин, 1967, с. 933). Однако, как показывает

опыт, обращение к словарю или интуиции не всегда дает возможность правильно достроить оборот. Так, например, согласно словарю В. К. Мюллера (Англо-русский словарь, изд. 7, М., 1960, с. 1170), триада of the Common, обнаруженная в газетных текстах, должна восстанавливаться до УС of the Common Council 'муниципалитета'. В действительности эта триада обычно является частью УС of the Common market 'общего рынка' — оборота, не засвидетельствованного многими англо-русскими словарями (исключение составляет словарь-минимум, составленный П. М. Алексеевым и Л. А. Турыгиной — 1974, с. 119).

Более эффективным приемом выявления типовых контекстов и УС является их «сборка» из не полностью оформленных триад (диад, тетрад и т. д.). Метод сборки строится исходя из следующих соображений. Каждый текст (предложение, словосочетание) можно рассматривать как результат некоторого развертывания, стоящего в начале текста словоупотребления *или словосочетания путем* так называемых $(k+1)$ — вероятностных приближений (Chomsky, Miller, 1963/1967, с. 148—152). Так, например, предложение *Выполнение программы машин «Минск-2, 22, 22М» на машине «Минск-32» обеспечивается аппаратурными и программными средствами* (см. в **21**) может быть представлено как развертывание начальной триады:

Путем ее дополнения диадами и триадами Δ *Выполнение программ машин «Минск-2, 22, 22М» | на машине «Минск-32»|обеспечивается аппаратурными|и программными| средствами* Δ

Выполнение программ машин «Минск-2,22,22М» на машине «Минск-32» обеспечивается аппаратурными и программными средствами Δ

Δ *Выполнение программ машин*
Выполнение программ машин «Минск-2,22,22М»
программ машин «Минск-2,22,22М» на
машин «Минск-2,22,22М» на машине
«Минск-2,22,22М» на машине «Минск-32»
на машине «Минск-32» обеспечивается
машине «Минск-32» обеспечивается аппаратурными и
обеспечивается аппаратурными и программными
аппаратурными и программными средствами
и программными средствами
программными средствами Δ

Этот же текст можно представить как теоретико-множественную сумму триад, образуемых путем сдвига вправо на одну словоформу (см. с. 73).

Только что описанную методику развертывания можно применить для вероятностной сборки УС из ранее выделенных триад и тетрад. Необходимость в такой «сборке» возникает в связи с тем, что из-за циклопических объемов необходимых выборок не удается получить достоверную статистику словосочетаний, состоящих более чем из трех-четырех словоупотреблений. Между тем значительное количество устойчивых и употребительных словосочетаний, в том числе сложных терминов, состоит из пяти и более словоупотреблений. Чтобы получить достаточно большое число типичных для данного подъязыка оборотов длиной в пять и более словоупотреблений, целесообразно собирать их путем напластования частых «заготовок» триад и тетрад.[1] Нижний порог употребительности этих заготовок принимается обычно равным $f=10^{-5}$ (такой порог соответствует нижнему порогу частости слов и словоформ, включенных обычно в списки УС и типовых контекстов).

Если сборка оборота осуществлена при условии, что каждая предыдущая триада (тетрада) имеет по крайней мере две общих словоформы с последующей триадой (тетрадой), а получаемый в результате теоретико-множественного сложения оборот является вполне осмысленным, можно предполагать, что такой оборот реально встречается в текстах и его можно условно включить в список УС или типовых контекстов.

Воспользовавшись данными Н. С. Булашевой (1969, с. 380) и Н. И. Исабековой (1969, с. 384—387), возьмем четыре триады:

1) *в зависимости от* $(3.3\cdot10^{-4})$
2) *в зависимости от напряжения* $(1.6\cdot10^{-5})$
3) *от напряжения на* $(2.7\cdot10^{-5})$
4) *напряжения на выходе* $(2.6\cdot10^{-4})$,

теоретико-множественное сложение которых дает вполне осмысленный, с точки зрения радиоэлектроники, сегмент: *в зависимости от напряжения на аноде*. Верхний порог потенциальной частости этого сегмента, определяемый всегда по относительной частоте наиболее редкой триады, составляет $1.6\cdot10^{-5}$.

С помощью указанных приемов формируются списки именных, глагольных, наречных УС (ср. табл. 6—12), построенные по дан-

[1] Во избежание недоразумений следует подчеркнуть, что развертывание начального словосочетания с помощью k-ограниченного марковского источника не может быть использовано для формального выделения машинных оборотов. Увеличение k не ведет к выделению множества грамматически правильных сегментов. Хотя число высоковероятных грамматически правильных сегментов растет, число маловероятных грамматически правильных последовательностей растет еще быстрее (Chomsky, Miller, 1963/1967, с. 151—152). Кроме того, при любом конечном k возникнут такие неправильные с точки зрения нормы языка сегменты, состоящие более чем из k символов, которые ЭВМ не сможет опознать как неправильные.

Таблица 6

Начало списка русских именных УС

№	Устойчивые словосочетания	$f \cdot 10^6$
1	*анодный ток в лампе*	1.1
2	*анодный ток в цепи*	2.7
3	*анодный ток в цепи анода*	1.0
4	*анодный ток в цепи коллектора*	3.0
5	*анодный ток в цепи эмиттера*	2.7
6	*анодного тока в цепи анода*	1.0
7	*анодного тока в цепи коллектора*	1.0
8	*анодного тока в цепи сетки*	1.0
9	*анодного тока в цепи эмиттера*	1.0
10	*анодного тока лампы*	1.0
11	*в зависимости от величины*	1.6

Таблица 7

Начало списка русских глагольных УС

№	Устойчивые словосочетания	$f \cdot 10^5$
1	*будем считать, что*	2.3
2	*будет иметь вид*	4.3
3	*будет равно*	1.0
4	*будет установлено, что*	4.5
5	*в качестве примера рассмотрим*	1.7
6	*говорит о том, что*	1.1
7	*дело в том, что если*	1.0
8	*для этого нужно*	1.2
9	*должен быть равен*	1.0
10	*должна быть равна*	1.4

Таблица 8

Начало списка наречных, предложных и союзных устойчивых словосочетаний

№	Устойчивые словосочетания	$f \cdot 10^5$
1	*а вместе с тем*	1.6
2	*в зависимости от того*	2.9
3	*в зависимости от того, где*	2.7
4	*в зависимости от того, как*	2.6
5	*в несколько раз больше, чем в*	1.9
6	*в результате чего*	5.5
7	*в результате этого*	3.0
8	*в соответствии с этим*	2.0
9	*в то время как*	7.0
10	*в то же время*	7.5

ным работы: Бектаев, Белоцерковская, Борисевич и др., 1973, с. 111—114). Эти списки используются затем при составлении автоматических словарей оборотов, которые применяются в системах МП, а также автоматического аннотирования и реферирования, построенных по принципу информационного бланка (см. 62).[1]

Прием напластования триад может быть применен и при построении программ предсказуемостного синтеза (Rhodes, Alt, 1962/1971, с. 85).

Например, имеем триады:

1) Δ *в зависимости* $(1.3 \cdot 10^{-4})$
2) *в зависимости от* $(3.3 \cdot 10^{-4})$
3) *зависимости от величины* $(3 \cdot 10^{-5})$
4) *от величины поля* $(7 \cdot 10^{-6})$
5) *от величины сопротивления* $(7 \cdot 10^{-6})$
6) *от величины тока* $(9 \cdot 10^{-6})$

Напластование триад 1—3 дает достаточно частый ($f \leqslant 3 \cdot 10^5$) сегмент: *в зависимости от величины*. Однако дальнейшее его развертывание на лексическом уровне нецелесообразно, поскольку находящиеся справа от него словоупотребления имеют слишком

Таблица 9

Начало списка именных типовых контекстов с правой грамматической валентностью

№	Типовые контексты	$f \cdot 10^5$
1	*анодный ток лампы в* + сущ. предл. пад.	2.7
2	*анодный ток лампы будет* + инф.	1.0
3	*анодный ток при* + сущ. ед. ч. предл. пад.	1.0
4	*в зависимости от величины* + сущ. ед. ч. род. пад.	4.0
5	*в зависимости от напряжения на* + сущ. ед. ч. предл. пад.	2.5
6	*в качестве примера* + спрягаемая форма глагола	1.2
7	*в результате этого* + сущ. ед. ч. род. пад.	3.0
8	*величина которого зависит от* + им. ф. род. пад.	1.0
9	*во время работы* + сущ. ед. ч. род. пад.	2.0
10	*как и в области* + сущ. ед. ч. род. пад.	1.5

[1] Путем напластования триад можно построить и целое предложение, которое по своему синтаксису не отличается от реальных предложений, взятых из соответствующих текстов. Например, теоретико-множественное сложение трех триад:

1) *анодный ток возрастает* $(1.1 \cdot 10^{-5})$
2) *возрастает с увеличением* $(1.4 \cdot 10^{-5})$
3) *с увеличением температуры* (10^{-5}),

дает осмысленное, нормативно построенное квазипредложение: *Анодный ток возрастает с увеличением температуры.* Предложения этого типа также могут быть использованы для выделения типовых контекстов.

Таблица 10

Начало списка глагольных типовых контекстов с правой грамматической валентностью

№	Типовые контексты	$f \cdot 10^5$
1	*будет зависеть от* + им. ф. род. пад.	1.0
2	*будут зависеть от* + им. ф. род. пад.	1.0
3	*выполняется в виде* + сущ. ед. ч. род. пад.	1.0
4	*для этого необходимо* + инф.	1.0
5	*зависит от величины* + сущ. ед. ч. род. пад.	1.0
6	*зависит от концентрации* + сущ. ед. ч. род. пад.	1.0
7	*зависит от напряжения на* + сущ. ед. ч. предл. пад.	1.0
8	*зависит от напряженности* + им. ф. род. пад.	1.0
9	*зависит от температуры* + сущ. ед. ч. род. пад.	1.0
10	*зависит от частоты* + сущ. ед. ч. род. пад.	1.0

Таблица 11

Список наречных, предложных и союзных типовых контекстов с правой грамматической валентностью

№	Типовые контексты	$f \cdot 10^5$
1	*даже в случае* + сущ. ед. ч. род. пад.	1.4
2	*до того, как* + спрягаемая форма глагола	1.1
3	*как в этом* + сущ. ед. ч. предл. пад.	1.0
4	*по отношению к* + сущ. ед. ч. дат. пад.	1.5
5	*по сравнению с* + им. ф. твор. пад.	1.0
6	*точно так же, как и в* + им. ф. предл. пад.	2.0
7	*точно так же, как и для* + им. ф. род. пад.	1.0
8	*точно так же, как и при* + им. ф. предл. пад.	1.0
9	*точно так же, как и у* + им. ф. род. пад.	1.0
10	*это следует из* + им. ф. род. пад.	1.0

Таблица 12

Начало списка типовых контекстов с левой и правой грамматической валентностью

№	Типовые контексты	$f \cdot 10^5$
1	глагол + *в виде* + сущ. род. пад.	1.0
2	глагол + *в зависимости от* + сущ. ед. ч. род. пад.	2.6
3	глагол + *в зависимости от типа* + им. ф. род. пад.	1.8
4	глагол + *в качестве* + им. ф. род. пад.	5.0
5	глагол + *в результате диффузии*	1.2
6	глагол + *в результате ионизации*	1.0
7	глагол + *в результате* + им. ф. ед. ч. род. пад.	8.0
8	глагол + *в течение* + сущ. ед. ч. род. пад.	2.0
9	глагол + *в течение некоторого времени*	1.4
10	глагол + *в течение нескольких* + сущ. мн. ч. род. пад.	4.5

малую вероятность. Поэтому в этих продолжениях выделяется грамматический инвариант — существительное ед. ч. род. пад. В итоге получаем следующий типовой контекст:

в зависимости от величины + сущ. ед. ч. род. пад.

Типовые контексты также объединяются в списки (ср. табл. 9—12, заимствованные из работы: Бектаев, Белоцерковская и др., 1973, с. 115—117). Эти списки используются при составлении морфологических и особенно синтаксических алгоритмов (см. 23, 69).

23. Алгоритмический выбор устойчивых словосочетаний

Иконический выбор сегментов фиксированной длины с последующим ручным преобразованием их в УС представляет собой трудоемкую процедуру, не всегда дающую достаточно точные лингвистические результаты. Более эффективным является непосредственный выбор из текста грамматически и семантически оформленных словосочетаний переменной длины. Извлечение таких сегментов можно проводить либо исходя из интуиции носителя языка, либо путем использования формального алгоритма, моделирующего лингвистическую интуицию человека.

При обследовании больших массивов текста ручное извлечение сегментов себя не оправдывает. Дело здесь не столько в большом объеме работы, требующей длительного срока, в течение которого проявляется действие парадокса Ахиллеса и черепахи, сколько в том, что при длительной обработке текста выборщик в результате самообучения начинает незаметно для себя отклоняться от заданных критериев извлечения сегментов.

Более эффективным является машинный выбор УС, который позволяет в сравнительно короткий срок обследовать на основе строгой формальной процедуры большие массивы текста и получить достаточно надежную статистику УС.

Автоматический выбор УС переменной длины успешно применяется относительно русского (Исаков, Лоскутов, 1971, с. 435—443), английского (Приймов, Соркина, 1970, с. 189—214; Данейко, Петровская, 1970, с. 215—290; Волосевич*, 1971; ср.: Kupo, 1966, с. 83—106; Foster, 1968/1974, с. 40—44), немецкого (Мануков, Зубов, 1972, с. 311—331), французского (Коверин, Скитневский, 1973, с. 13—48) и испанского (Михайлова*, 1972) текстов. Каждый из алгоритмов извлечения УС опирается на сегментацию текста, осуществляющуюся путем автоматического распознавания формальных границ, отделяющих одно грамматически оформленное словосочетание от другого.

В качестве примера рассмотрим алгоритм сегментации и извлечения словосочетаний, построенный И. И. Волосевичем (1973, с. 298—312). Этот алгоритм рассчитан на формальное извлечение из текста таких сегментов, как

1) именные группы,
2) глагольные группы,
3) начала придаточных предложений,
4) причастные и герундиальные обороты,
5) инфинитивы,
6) устойчивые словосочетания типа to and fro, as well as, as nearly as, а также вводных слов и групп типа first, in general.[1]

В качестве именной группы рассматриваются:

а) имя существительное нарицательное или собственное,

б) сочетание существительного с препозитивными определениями,

в) сочетание существительного с определяющими его предложными определительными конструкциями, выполняющими атрибутивную функцию,

г) сочетание существительного с постпозитивными определениями.

В число глагольных групп включаются:

а) простые личные формы глагола,

б) аналитические формы глагола, а также глагольные словосочетания, в состав которых входят личные, указательные и другие местоимения (эти местоимения, таким образом, рассматриваются как своеобразные препозитивные аналитические форманты глагольных словоформ).

Построение алгоритма, а затем и его работа начинается с определения формальных признаков именных сегментов.

Проанализировав 4 тыс. именных групп, взятых из научно-технических текстов, автор выяснил, что английская именная группа обычно представляет собой формально выделяемую цепочку, левой границей которой являются детерминативы (определенный и неопределенный артикли и притяжательные местоимения), глагольные формы, если именная группа является дополнением к глаголу, отдельные морфемы и т. д. Ее правой границей, не входящей в состав группы, являются обычно личные местоимения, глаголы, предлоги и т. д. (ср. табл. 13). В качестве ядра именной группы можно принять крайнее правое существительное, т. е. существительное, соседствующее с правым граничным сигналом.

Граничные сигналы именной группы подразделяются на безусловные и промежуточные.

Безусловной формальной границей считается такой структурный сигнал, использование которого позволяет машине однозначно определить начало или конец какой-либо группы. Так, например, правой безусловной границей именной группы являются детерминативы (определенный и неопределенный артикли, притяжательные местоимения), предлоги, личные формы глагола и т. д.

[1] Алгоритм И. И. Волосевича предусматривает морфологическую индексацию каждой словоформы, входящей в сегмент.

Таблица 13

Левые и правые границы именной группы в английском тексте и их частости

Левые границы	*f*	Правые эксклюзивные границы	*f*
1) Детерминативы (инкл.)	0.46	1) Глагольные словоформы	0.32
2) Предлоги (инкл.)	0.16	2) . (точка)	0.20
3) Местоименные слова (экскл.)	0.14	3) Предлоги	0.19
4) Глагольные словоформы (экскл.)	0.12	4) Отдельные морфемы	0.07
5) Сочинительные союзы (экскл.)	0.04	5) Личные местоимения	0.04
6) Подчинительные союзы (экскл.)	0.03	6) Детерминативы	0.04
7) Отдельные словоформы согласно списков 1,4, приведенных в Приложении (экскл.)	0.03	7) Отдельные словоформы (см. списки 3,10 в Приложении)	0.04
8) Словоформы, оканчивающиеся на -ing (экскл.)	0.02	8) Подчинительные союзы	0.04
		9) Местоименные слова	0.03
		10) Сочинительные союзы	0.03

Безусловные формальные границы именной группы, в свою очередь, подразделяются на инклюзивные, т. е. входящие в группу, и эксклюзивные, не входящие в группу. Разделение по этому принципу можно рассмотреть на примере следующего предложения: If there is a voltage, the oscillator is working, although not necessarily at the right frequency. Здесь неопределенный артикль, вводящий именную группу a voltage, рассматривается как инклюзивный компонент именной группы, во второй же группе — the oscillator — безусловный граничный сигнал the рассматривается двояко. С одной стороны, не включаясь в предыдущую именную группу a voltage, он эксклюзивно указывает на ее конец, а с другой стороны, the является левым инклюзивным сигналом группы the oscillator.

Безусловные инклюзивные формальные границы именной группы являются одновременно эксклюзивными границами глагольной группы. Так, в нашем примере неопределенный артикль является безусловной эксклюзивной границей глагольной группы there is. Другой эксклюзивной границей глагольной группы is working, although not necessarily служит предлог at, рассматриваемый одновременно и как инклюзивный граничный сигнал именной группы at the right frequency.

Безусловными инклюзивными сигналами глагольной группы и одновременно безусловными эксклюзивными сигналами именной группы являются также личные местоимения в общем падеже, указательные, неопределенные и другие местоимения, вводящие

глагольную группу (ср. список 6 из Приложения). Примером такого двойного граничного сигнала является местоимение we в следующем предложении: Last month we saw how some common accepted thoughts on transistors are not true. К промежуточным формальным границам относятся такие структурные сигналы, опознание которых не позволяет ЭВМ окончательно и однозначно определить границу группы и установить принадлежность самого сигнала к какой-либо группе. Рассмотрим в этой связи следующие английские предложения: Technicians working with transistors will frequently encounter circuits which have no counterpart in vacuum-tube work (1). The second interesting feature is the method used to sypply DC operating currents (2). There are several reasons for using such an arrangement (3). Base bias voltage remains essentially unchanged whether or not the oscillator is functioning (4).

В каждом из приведенных предложений имеются словоформы, оканчивающиеся на буквокомбинацию -ing, которую мы можем принять в качестве граничного сигнала. Однако с помощью этого сигнала нельзя однозначно установить, какую функцию выполняет в сегменте словоформа, оканчивающаяся на -ing (ср. Приложение). Действительно, такая словоформа может быть:

— ядром именной группы: for using — ср. предложение (3);

— компонентом именной группы: the second interesting feature; DC operating current — ср. (2);

— компонентом глагольной группы: is functioning — ср. (4);

— границей (дее)причастного оборота: working with transistors — ср. (1).

Выявление промежуточного граничного сигнала является указанием на то, что необходим дальнейший анализ, в ходе которого, опираясь на левую и правую контактную дистрибуцию, можно определить природу опознанного граничного сигнала.

Так, например, анализируя в предложении (1) правую и левую дистрибуцию, можно установить, что словоформа с морфемой -ing — безусловная формальная граница именной группы technicians, с одной стороны, а с другой — что эта словоформа указывает на начало (дее)причастного оборота.

Опознание граничного сигнала -ing в предложениях (2), (3) и (4) и проверка правой дистрибуции обнаруживает, что этот сигнал указывает на принадлежность анализируемой словоформы в предложениях (2) и (3) к именной группе, а в предложении (4) — к глагольной группе.

Аналогичным образом выделяются и другие типы словосочетаний (см. выше).

Алгоритм сегментации текста и выбора словосочетаний включает следующие семь блоков:

Б1 — ввод исходной синтаксической информации и текста
Б2 — поиск именной группы
Б3 — поиск глагольной группы

Б4 — поиск (дее)причастного оборота
Б5 — поиск начала придаточного предложения
Б6 — поиск инфинитивов, устойчивых словосочетаний и т. д.
Б7 — вывод на печать найденных словосочетаний.

Проиллюстрируем работу алгоритма на примере следующего английского предложения: The schematic for the audio amplifier used in the Westinghouse H-602P7 receiver is shown in Fig. 1.

Читая текст слева направо, машина ищет левую границу первой группы, последовательно сравнивая прочитанную словоформу со списками формальных границ. Опознав по списку 1 (см. Приложение) граничный сигнал the, ЭВМ выполнит следующие операции:

1) проверку наличия информации в именной буферной области (буф. 1),

2) проверку присутствия информации в глагольной буферной области (буф. 2),

3) проверку наличия информации в промежуточной буферной области (буф. 3).[1]

Поскольку в каждой из трех областей информации нет, цепочка the, опознанная по списку 1 как левая граница именной группы, заносится в буф. 1. Сюда же переносится и словоформа schematic.

Следующей словоформой текста оказывается предлог for, принадлежащий к списку 4. Анализируя эту словоформу, машина оказывается перед альтернативой, сущность которой состоит в том, что словоформа for является, с одной стороны, безусловным правым эксклюзитивным граничным сигналом накопленной группы the schematic, а с другой стороны — for может оказаться левой инклюзивной границей второй именной группы, если за for следует словоформа из списка 2 или из списка 7а. Прежде чем нам принять однозначное решение, словоформа for временно засылается в буф. 3. Затем читается следующее словоупотребление, которое необходимо для определения функции словоформы for. Этим словоупотреблением оказывается вторая цепочка the. Цепочка the последовательно сравнивается со всеми словоформами списков 2 и 7. Тождества the со словоформами в указанных списках не обнаруживается. В связи с этим делается вывод, что словоформа for есть безусловный эксклюзивный граничный сигнал именной группы the schematic.

Прежде чем осуществить печать именной группы, проводится ее анализ, т. е. устанавливается ядро группы и определяющие его компоненты. Суть этого анализа состоит в следующем. Хотя ЭВМ и информирована о том, что последняя словоформа накопленной именной группы является ядром этой группы, на это условие накладываются два ограничения. Последняя словоформа группы

[1] Под буферной областью понимается поле в оперативной памяти ЭВМ, где накапливается получаемый в ходе анализа сегмент текста и откуда этот сегмент поступает на печатающее устройство. На иллюстративной блок-схеме (рис. 22) буферы 1, 2, 3 объединены в блоке 7.

при сравнении не должна отождествляться ни с запятой [,], ни со словоформой из списка 7в.

Осуществив указанную проверку, ЭВМ определяет, что ядром группы является слово schematic, которому она сообщает индекс «n». Установив, что компонентом группы является в данном случае определенный артикль, который не индексируется, машина печатает всю группу, сообщив ей общегрупповой индекс «именная группа»:

THE SCHEMATIC$_{n}$ ИМЕННАЯ ГРУППА

Выше было установлено, что словоформа for, находящаяся в буфере промежуточной информации, является безусловным граничным сигналом предшествующей именной группы.

Теперь машина выполняет следующие операции:

а) пересылает словоформу for из буф. 3 в буф. 1,

б) засылает словоформу the, прочитанную ранее для определения природы граничного сигнала for, в буф. 1,

в) засылает сюда же следующие за the словоформы audio и amplifier.

Засыл словоформ audio и amplifier в буф. 1 объясняется тем, что их сравнение со словоформами из списков граничных сигналов дает отрицательный ответ.

Анализ следующей словоформы used обнаруживает в ней морфему -ed, которая может быть граничным сигналом (ср. Приложение). Однако пока неизвестно, является ли словоформа с морфемой -ed компонентом именной группы, как в предложении: A figure X shows a forced draft blower limiting governer (9), или анализируемая словоформа used служит безусловной эксклюзивной границей именной группы for the audio amplifier.

Чтобы решить эту задачу, машина обращается к контексту. Поместив слово used в буф. 3, машина читает следующее слово in, которое помогает ЭВМ однозначно установить, что used является безусловной эксклюзивной границей второй именной группы.

Прежде чем напечатать группу и сообщить ей индекс, машина проводит внутригрупповой анализ и индексацию компонентов группы. Иными словами, машина устанавливает ядро группы amplifier$_{n}$ и определяющий его компонент (ему приписывается индекс «a») — audio$_{a}$, а также фиксируется, что группа введена предлогом for$_{pr}$. В итоге на печать выдается следующая информация:

FOR$_{pr}$ THE AUDIO$_{a}$ AMPLIFIER$_{n}$ ИМЕННАЯ ГРУППА

Отпечатав эту группу, ЭВМ возвращается к анализу словоформы used. Здесь необходимо выяснить, вводит ли словоформа used глагольную группу, как в предложении: The original plan suggested a rather conservative indirect cycle reactor system, либо эта словоформа вводит причастный оборот. Решение в пользу причастного оборота выносится после того, как вправо от used обнаруживается словоформа is, являющаяся безусловной левой инклюзивной границей глагольной группы (список 3). Правая граница начала причастного оборота определяется с помощью словоформы in (список 4). В итоге словоформа used пересылается из буф. 2 (буфер глагольной группы) и печатается с общим индексом «начало (дее)причастного оборота»:

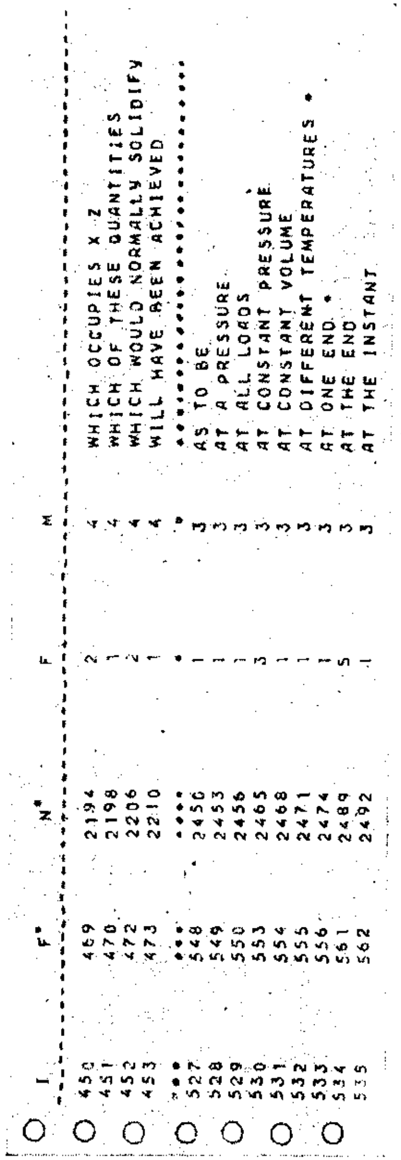
USED НАЧАЛО (ДЕЕ)ПРИЧАСТНОГО ОБОРОТА

Обработка сегмента in the Westinghouse H-602P7 receiver осуществляется тем же путем, что и переработка именных групп the schematic и for the audio amplifier. Выдача на печать группы вместе с морфологическими и общегрупповым индексом —

IN$_{pr}$ THE WESTINGHOUSE$_{n}$ H-602P7 RECEIVER$_{n}$ ИМЕННАЯ ГРУППА

осуществляется по сигналу словоформы is, выступающей в роли правой эксклюзивной границы именной группы (список 3). Выдав на печать именную группу, машина заносит словоформу is в буф. 2 и читает словоформу shown. Отождествив ее со словоформой списка 9, ЭВМ присоединяет ее к словоформе is в буф. 2. Внутригрупповой анализ с присоединением «v», печать и общегрупповая индексация осуществляются по сигналу следующей за shown словоформы in:

IS SHOWN$_{v}$ ГЛАГОЛЬНАЯ ГРУППА



I	F*	N*	F	M	
450	469	2194	2	4	WHICH OCCUPIES X Z
451	470	2198	1	4	WHICH OF THESE QUANTITIES
452	472	2206	2	4	WHICH WOULD NORMALLY SOLIDIFY
453	473	2210	1	4	WILL HAVE BEEN ACHIEVED
•	•	•	•	•	••••••••••••
527	548	2456	1	3	AS TO BE
528	549	2453	1	3	AT A PRESSURE
529	550	2456	1	3	AT ALL LOADS
530	553	2465	3	3	AT CONSTANT PRESSURE
531	554	2468	1	3	AT CONSTANT VOLUME
532	555	2471	1	3	AT DIFFERENT TEMPERATURES
533	556	2474	1	3	AT ONE END
534	561	2489	5	3	AT THE END
535	562	2492	1	3	AT THE INSTANT

Рис. 21. Фрагмент машинного ЧС грамматически и семантически оформленных сегментов нефиксированной длины (автор алгоритма и программы В. М. Петровская).

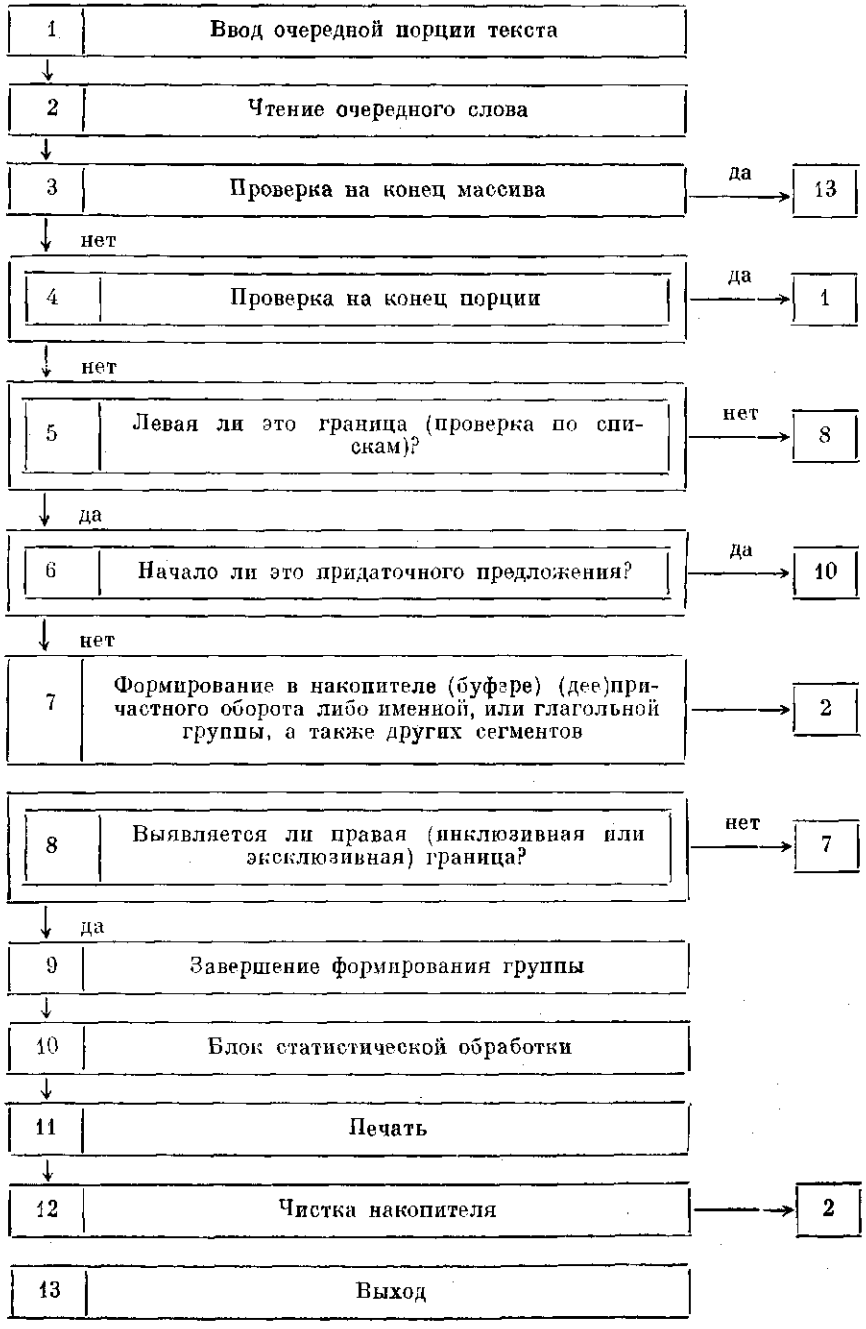


Рис. 22. Иллюстративная блок-схема сегментации текста и составления ЧС грамматически и семантически оформленных словосочетаний.

Обработка сегмента in Fig. 1 происходит по той же схеме, что и выделение групп the schematic, for the audio amplifier, in the Westinghouse H-602P7 receiver. Внутригрупповой анализ, общегрупповой индекс и печать осуществляется по сигналу ⊡, который воспринимается ЭВМ как отдельная словоформа, включающаяся в сегмент:

IN_{pr} FIG. 1_n ИМЕННАЯ ГРУППА

Сигналом конца разбивки предложения на грамматически и семантически оформленные сегменты является «словоформа» ⊡ (точка).

Аналогичные принципы сегментации текста по формальным граничным сигналам, задаваемым в списках, использованы и в других программах выбора оформленных словосочетаний (ср.: Данейко, Петровская, 1970; Коверин*, 1972; Приймов, Соркина, 1970, и др. — см. выше, с. 83). Некоторые из этих программ осуществляют также статистическое упорядочение полученных сегментов (ср. рис. 21). Типовой алгоритм сегментации текста со статистическим упорядочением словосочетаний показан на рис. 22.

24. Грамматическая статистика

Статистическое моделирование текста предусматривает также состояние частотных списков отдельных грамматических форм (Алексеев, Навальна, 1971, с. 78—81; Гурова, 1974), схем построения отдельных членов предложения (Сафонова, 1973, с. 49—57) и самих предложений (Копачева, 1966, с. 249—265). Эти частотные списки составляются чаще всего вручную. Пример ЧС грамматических схем построения сказуемого в английском научно-техническом тексте приведен в табл. 14.

25. Парадокс вероятности и информации

Частотные словари, составленные на основе малых выборок (200—500 тыс. словоупотреблений), являются, как уже было сказано, упрощенными и грубыми приближениями к информационно-статистической структуре лексики текста. Эти приближения оправдывали себя при построении МБЯ с малыми АС. Эти МБЯ, предназначавшиеся для использования в машинах второго поколения, с ограниченной памятью, могли обеспечить лишь общее распознавание смыслового образа — распознавание на уровне темы текста. Что же касается ремы, т. е. новизны текста, то ее распознавание с помощью малых АС в достаточной степени затруднительно. Дело здесь не только в том, что ЧС слов и словоформ не дают информации о контекстных связях этих лексических единиц. Эти ЧС, как это было показано в 21—23, могут быть дополнены вероятностно-статистическими моделями, учитываю-

Таблица 14

Частотный список наиболее употребительных грамматических схем построения сказуемого в английских научно-технических текстах (N=4185)
По данным И. Н. Сафоновой

i	Грамматические схемы сказуемого	F	f	f^*
1	V_{cs}	742	0.178	0.178
2	V_{cp}	360	0.086	0.264
3	*is*+Pr-n	355	0.085	0.349
4	*are*+Pr-n	252	0.060	0.409
5	*is*+U	234	0.056	0.465
6	Vh, -ed	206	0.049	0.514
7	*was*+Pr-n	153	0.039	0.553
8	*were*+Pr-n	138	0.033	0.586
9	*is*+N	119	0.028	0.614
10	*can be*+Pr-n	94	0.022	0.636
11	*will*+Inf	89	0.021	0.657
12	*are*+U	82	0.020	0.677
13	*have*+Pr-n	61	0.015	0.692
14	*may be*+Pr-n	60	0.014	0.706
15	*is*+число	54	0.013	0.719
16	Vh	51	0.012	0.731
17	*may*+Inf	51	0.012	0.743
18	*can*+Inf	49	0.012	0.755
19	*have been*+Pr-n	48	0.011	0.766
20	*will be*+Pr-n	47	0.011	0.777
21	*has been*+Pr-n	44	0.010	0.787
22	*are*+N	42	0.010	0.797
23	*has*	42	0.010	0.807
24	*is*	42	0.010	0.817

Примечание. В таблице используются следующие сокращения: Inf — инфинитив, N — существительное, Pr-n — причастие (Participle I), U — прилагательное, Vh — глагол прошедшего времени (Past Indefinite для неправильных глаголов), Vh, -ed — глагол прошедшего времени (Past Indefinite), V_{cp} — личная форма глагола мн. ч. настоящего времени (Present Indefinite), V_{cs} — личная форма глагола 3-го лица ед. ч. настоящего времени (Present Indefinite). Курсивом отмечены вспомогательные и модальные глагольные формы.

щими эти связи. Основное затруднение заключается здесь в том, что частотный подход к отбору лексики и фразеологии в автоматический, как, впрочем, и в любой традиционный словарь, имплицитно содержит в себе парадокс вероятности и информации. Этот парадокс вытекает из статистического определения информации, согласно которому неопределенность (энтропия) лингвистического опыта i, которая количественно равна потенциальной информации, получаемой метанаблюдателем при осуществлении этого опыта, измеряется как минус логарифм вероятности этого опыта:

$$h_i = -\log p_i.$$

Это значит, что редкие лингвистические единицы несут больше потенциальной информации, чем единицы, часто употребляю-

щиеся в тексте. Хотя строгой зависимости между величинами потенциальной и семантической информации, содержащейся в лексических единицах, пока не установлено; полученные из эксперимента данные говорят о том, что редкие слова и словоформы также несут заметно больше смысловой информации, чем частые лексические единицы (Bogodist, Guéorguiev, Pestunova, Piotrowski, Raitar, 1974, с. 40).

Это соотношение между вероятностью и информацией имеет два следствия.

Во-первых, частотная покрываемость текста отдельными массивами частотного словаря не совпадает с их информационным покрытием, определяемым из равенства

$$\varphi = \frac{\check{H}_i^*}{H} 100\% = \frac{-\sum_{h=1}^{i} f_h \log_2 f_h}{-\sum_{i=1}^{n} f_i \log_2 f_i} 100\%. \quad (5)$$

Хотя ввиду малой величины $p_i \simeq f_i$ удельная потенциальная энтропия

$$p_i h_i \simeq -f_i \log_2 f_i = \check{H}_i \quad (1)$$

у редких лексических единиц ниже удельной энтропии частых единиц, рост информационной покрываемости почти на всем протяжении ЧС значительно отстает от роста частотного покрытия (ср. рис. 23).

Во-вторых, лишь в самом конце частотного списка находящиеся там редкие лексические единицы дают резкое возрастание информационной покрываемости текста. Именно поэтому при формировании словников автоматических словарей в них приходится обычно включать все лексические единицы, зафиксированные в ЧС, в том числе и такие, которые имеют F=1.

Частотные словари составляются на основе экспериментальных выборок, каждая из которых отражает лишь малую часть текстов, входящих в генеральную совокупность. Поэтому в частотные словари не попадает большое число редких слов или словоформ, образующих массив резервных лексических единиц, каждая из которых имеет в данном частотном словаре F=0.

Не попадая в составляемый на основе ЧС автоматический словарь, эти нуль-частотные слова и словоформы образуют находящийся за пределами машинного БЯ лексический массив, являющийся источником той частотной непокрываемости текста в 1—3%, о которой говорилось выше. Эти неопознаваемые автоматическим словарем нуль-частотные единицы являются обычно теми высокоинформативными словоформами, которые необходимы для формального распознавания ремы текста. Нуль-частотные слова и словоформы входят также в сложные лексические

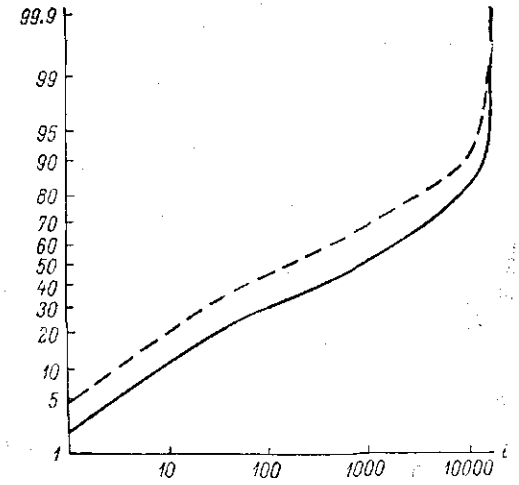


Рис. 23. Рост частотной (*прерывистая кривая*) и информационной (*сплошная кривая*) покрываемости немецкого газетного текста, по данным Ротарь*, 1971 (ср. также: Ешан*, 1966, с. 9).

единицы, которые также служат средством выражения новизны текста.

Чему же равно общее количество нуль-частотных словоформ, остающихся за пределами наших отраслевых ЧС и АС?

Чтобы ответить на этот вопрос, необходимо иметь сведения о количестве лексических единиц, использующихся в отдельных подъязыках.

К сожалению, точных данных об общем объеме лексики, использующейся в той или иной области знаний, у нас нет. По некоторым косвенным данным каждая область знаний использует несколько десятков, а иногда и сотен тысяч простых и сложных терминов. Так, например, в русском языке при обработке «исчерпывающей представительной коллекции информационных документов по химии, химической и нефтехимической промышленности» выделено до 180 тыс. сложных и простых терминов (см.: Словарь дескрипторов по химии и химической промышленности. М., 1966, с. 3). Обычно считается, что биология и медицина используют каждая до миллиона сложных и простых терминов, включая греко-латинские образования.

Оценивая объем терминологической лексики в отдельном подъязыке, можно воспользоваться опытом построения автоматических двуязычных словарей. Так, например, исследование вопросов машинного перевода немецких экономических текстов, показало, что входной словник полного немецко-русского словаря по этой тематике должен включать около 80 тыс. лексических единиц (Вертель*, 1974, с. 11). Одновременно оказалось, что для перевода 6.6 тыс. английских словоформ, обозначающих специальные понятия подъязыка полупроводниковой техники, используется сокращенный выходной русский словник, охватывающий 142 тыс. словоформ (полный словник должен был бы включать более 260 тыс. словоформ — см.: Гончаренко*, 1972б).

Таблица 1

Прирост новых слов в зависимости от увеличения объема выборки

Объем выборки	Русские тексты по электронике (Калинина, 1968а, с. 89)		Английские тексты по судовым механизмам (Лукьяненков*, 1969)		Смешанная выборка (Kučera, Francis, 1967, с. 417)	
	всего словоформ в выборке	количество новых словоформ по сравнению с предыдущей выборкой	всего словоформ в выборке	количество новых словоформ по сравнению с предыдущей выборкой	всего словоформ в выборке	количество новых словоформ по сравнению с предыдущей выборкой
50 000	9 464	9 464	—	—	—	—
100 000	14 062	4 598	6 731	6 731	13 706	13 706
150 000	17 263	3 201	—	—	—	—
200 000	21 468	4 205	9 181	2 450	—	—
250 000	—	—	—	—	23 655	9 949
300 000	—	—	11 314	2 133	—	—
400 000	—	—	12 971	1 657	—	—
1 000 000	—	—	—	—	50 406	26 751

Все сказанное выше свидетельствует о том, что полные словники отраслевых словарей в аналитических языках включают десятки тысяч, а в синтетических языках, вероятно, сотни тысяч словоформ.[1] Отсюда следует, что нуль-частотные лексические единицы составляют в каждом из подъязыков значительный массив словоформ. Хотя эти словоформы и покрывают всего 2—3% контрольного текста, они, с одной стороны, сами выражают новые научно-технические понятия, а с другой — участвуют в формировании сложных терминов, также использующихся для воплощения ремы текста.

Тогда возникает вопрос, можно ли с помощью уже оправдавшей себя на небольших массивах текста статистической методики выявить весь массив нуль-частотных лексических единиц и каковы ориентировочные объемы необходимых для выполнения этой задачи выборок?

Чтобы ответить на этот вопрос, К. Ф. Лукьяненковым * (1969) и автором этих строк проведена грубая геометрическая экстраполяция объемов выборок научно-технических текстов, необ-

[1] Эти оценки в целом соответствуют темпу прироста новых словоформ при увеличении объема выборок на 50 тыс. (см. табл. 15, стб. 3) и на 100 тыс. (см. табл. 15, стб. 5) словоупотреблений в пределах тех частных выборок в 200—400 тыс. словоупотреблений, которые используются для построения ЧС научно-технических подъязыков. Темп прироста здесь достаточно велик (ср.: табл. 15; рис. 24). Что же касается смешанных выборок, отражающих лексику литературного языка в целом (ср. стб. 6—7 в табл. 15), то здесь прирост новых словоформ еще выше, чем в научно-технических подъязыках.

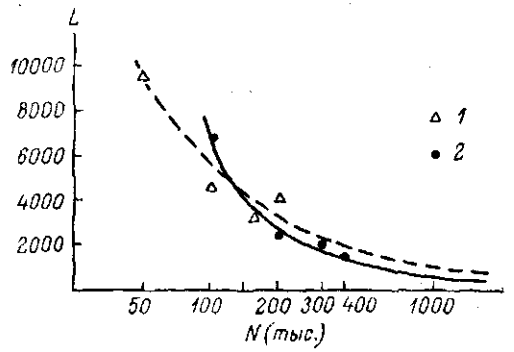


Рис. 24. Зависимость числа новых слов на порцию текста от длины выборки.

1 — русские тексты по электронике (Калинина, 1968а, с. 89); *2* — английские тексты по судовым механизмам (Лукьяненков*, 1969).

ходимых для построения ЧС, словники которых включали бы несколько десятков тысяч словоупотреблений (ср. табл. 16, стб. 3—6). Эти данные сопоставляются с экстраполяцией, полученной на смешанной выборке английских текстов Дж. Кэрроллом с помощью логнормальной модели распределения частот (Carroll, 1968; Kučera, Francis, 1967, с. 411—412, 417 — см. табл. 16, стб. 2. В этом же плане ср.: Тулдава, 1974, с. 248—249).

Приведенные в табл. 16 данные показывают, что для выделения словников в 60—90 тыс. словоформ [1] понадобится обработка текста в 10^{10} словоупотреблений (около 30 млн страниц). Обработать этот текст вручную, естественно, невозможно. Использовать для решения этой задачи ЭВМ можно было бы при наличии читающих устройств, надежно работающих относительно разных типов шрифтов. Поскольку таких устройств пока нет, указанный текстовой массив перед его обработкой на ЭВМ следует перфорировать. Эта работа потребовала бы 3 млн человеко-дней на один массив (т. е. работы тысячи перфораторщиков в течение десяти лет). Легко понять, что необходимые для такой перфорации материальные затраты себя не оправдывают, а срока выполнения этой работы вполне достаточно, чтобы антиномия Ахиллеса и черепахи свела бы на нет результаты всей работы (ср. 11).

Итак, на данном этапе развития вычислительной техники статистический путь выявления всего массива редких высокоинформативных лексических единиц оказывается нереальным. Поэтому приходится использовать компромиссный подход, сущность которого состоит в следующем. Статистическим путем выявляются те лексические единицы, располагая которыми можно обеспечить автоматическую переработку текста на уровне темы.

[1] По всей вероятности, именно таковы объемы АС, необходимые для лингвистического обеспечения банков данных.

Таблица 16

Наблюдаемый и ожидаемый рост словника частотного словаря L в зависимости от увеличения объема выборки N

N (тыс. словоупотреблений)	L (словоформ)								
	английский язык				румынские тексты по электронике	русские тексты по электронике	Наблюдаемый объем словника		
	смешанная выборка (Kučera, Francis, 1967, с. 417)	электроника	судовые механизмы						
50	—	5 399	4 857		6 785	9 464			
100	13 706 (15 865) *	7 853	6 731		10 284	14 062			
150	—	9 361	—		12 477	17 263			
200	—	10 582	9 364		14 292	21 468			
250	23 655 (26 218)	—	—		—	—			
300	—	—	11 314		—	—			
400	—	(14 000) **	12 971		(19 000)	(29 000)	Ожидаемый объем словника		
1 000	50 406 (50 963)	(18 000) **	(17 000)		(23 000)	(33 000)			
2 000	—	(21 000) ***	(20 000) ***		(27 000)	(38 000)			
5 000	—	(26 000) ***	(24 000) ***		(32 000)	(45 000)			
10 000	(118 624)	(29 000)	(27 000)		(35 000)	(50 000)			
30 000	—	(35 000)	(32 000)		(41 000)	(60 000)			
50 000	—	(38 000)	(35 000)		(44 000)	(65 000)			
100 000	(206 309)	(41 000)	(38 000)		(48 000)	(70 000)			
1 000 000	—	(52 000)	(49 000)		(61 000)	(74 000)			
3 000 000	—	(55 000)	(52 000)		(65 000)	(79 000)			
10 000 000	—	(63 000)	(60 000)		(75 000)	(91 000)			

* Цифры, приведенные без скобок, указывают на наблюдаемые объемы словников, в скобках указываются ожидаемые объемы словников ЧС.

** По экстраполяционным оценкам П. М. Алексеева (1969, с. 37), выборка в 1 млн словоупотреблений должна дать в английском подъязыке электронных словник объемом в 29 тыс. словоформ (27 тыс. словоформ без учета омографии).

*** ЧС Ринсленда, построенный на выборке около 6 млн словоупотреблений, дал словник в 25 32 словоформы (Rinsland, 1945).

Из этих единиц формируется исходный отраслевой автоматический словарь, который в ходе его эксплуатации последовательно пополняется теми незарегистрированными в нем словоформами и словосочетаниями, которые будут обнаружены в перерабатываемых текстах (ср.: Вертель и др., 1971, с. 50—51).[1] Такая организация обучения компьютера имитирует усвоение человеком специальной терминологии. Известно, что основами своей терминологии специалист овладевает в учебном заведении, а затем в процессе работы последовательно пополняет свой терминологический запас.

26. Вероятностные модели текста. Закон Ципфа—Мандельброта

Наряду с формированием грубых статистических моделей — частотных списков — в лингвистике текста делались попытки построить два типа вероятностных моделей текста.

Авторы моделей первого типа строили их исходя из следующих рассуждений.

Пусть имеется конечное или бесконечное множество дискретных объектов w, каждый из которых отмечен определенной меткой, взятой из некоторого дискретного множества меток. Обозначим число различных меток, появляющихся ровно F раз в выборке из N объектов через $m(F, N)$. При этом выясняется, что для больших N

$$m(F, N) = G(N) F^{-(1+\alpha)}, \tag{6}$$

где $\alpha > 0$, а $G(N)$ — коэффициент, зависящий от объема выборки N. Соотношением (6) описывается ряд ситуаций, встречающихся в разных отраслях знаний. Этой закономерности подчиняется отношение между числом людей, получающих определенный доход, и величиной этого дохода (так называемый закон Парето); она описывает отношение между числом родов $m(F, N)$, содержащих ровно F видов, и самой величиной F (так называемый закон Уиллиса—Юла). Применительно к лингвистике эта закономерность связывает число словоформ (или слов) m, встретившихся ровно F раз в выборке N с самой частотой F.

Расположим теперь все слова в порядке убывания их частот и определим их ранги i как число словоформ (или слов), встретившихся F и более раз:

[1] В отраслевые АС включается также и некоторое количество лексических единиц, взятых из специальных двуязычных, а также терминологических и энциклопедических одноязычных словарей. Однако целиком положиться на эти словари нельзя, поскольку они обычно с опозданием фиксируют изменения в лексике подъязыков.

$$i = \sum_{j=F}^{\infty} m_j. \tag{7}$$

Объединив (6) и (7) и аппроксимировав сумму интегралом, имеем для больших F_i:

$$i \approx \frac{G(N)}{\alpha F_i^{\alpha}} \tag{8}$$

и

$$F_i \approx \sqrt[\alpha]{\frac{G(N)}{\alpha i}} = \left(\frac{G(N)}{\alpha i}\right)^{\frac{1}{\alpha}} = kNi^{-\gamma} \tag{9}$$

при

$$\frac{1}{\alpha} = \gamma \quad \text{и} \quad \left(\frac{G(N)}{\alpha}\right)^{\frac{1}{\alpha}} = kN.$$

Поделив обе части равенства (9) на N и принимая $\frac{F_i}{N} \approx p_i$, приходим к выражению

$$p_i = ki^{-\gamma}, \tag{10}$$

при логарифмировании обеих частей которого имеем:

$$\log p_i = \log k - \gamma \log i. \tag{11}$$

Эти две зависимости известны под названием закона Ципфа [1] (см.: Zipf, 1949; Chomsky, Miller, 1963/1967, с. 180—181).

Используя формулу (10) и определив экспериментальным путем значения ее коэффициентов относительно русских текстов по электронике, Е. А. Калинина сделала попытку вывести зависимость для определения теоретической покрываемости p_i^* при заданном i (Калинина, 1968а, с. 72—73).

При этом было выяснено, что закон Ципфа хорошо выполняется в интервале $300 \leqslant i \leqslant 4200$, причем $k=0.145$, $\gamma=1$, а среднеквадратическая ошибка в определении k не превышает 4.2%. Эмпирическим путем было установлено также, что $p^*_{300} = 0.447$.

[1] Ципф считал, что для естественных языков $k=0.1$; $\gamma=1$, в связи с чем формула (10) принимает вид

$$p_i = \frac{k}{i} = \frac{0.1}{i}.$$

Исходя из этих данных, имеем

$$p_i^* = \sum_{h=1}^{i} p_h = \sum_{h=1}^{300} p_h + k \sum_{h=301}^{i} \frac{1}{h} = p_{300}^* + k\left[\left(1 + \frac{1}{2} + \ldots + \frac{1}{i}\right) - \right.$$

$$\left. - \left(1 + \frac{1}{2} + \ldots + \frac{1}{300}\right)\right] = p_{300}^* + k\,[(0.577 + \ln i) - (0.577 + \ln 300)] =$$

$$= 0.447 - 0.145 \ln 300 + 0.145 \ln i. \qquad (12)\ ^1$$

Переходя от натуральных логарифмов к десятичным и решая равенство (12), получаем

$$p_i^* = 0.334 \lg i - 0.38. \qquad (13)$$

По этой формуле были определены теоретические значения накопленной вероятности для русских текстов по электронике. Эти значения показали хорошее схождение с опытными накопленными частостями f_i^* в интервале $300 \leqslant i \leqslant 5000$.

Вторая серия вероятностных моделей текстообразования, наиболее полно изложенная в работах Б. Мандельброта (ср.: Mandelbrot, 1961, с. 190—219), строилась исходя из более сильных теоретических посылок, учитывавших, с одной стороны, стохастические схемы Маркова, а с другой — идеи шенноновской теории информации.

С одной стороны, при построении своих моделей Б. Мандельброт и его последователи рассматривают текст в качестве конечной марковской цепи, т. е. такой последовательности случайных событий, в которой влияние лингвистического события A на следующие за ним события быстро затухает. В простейших случаях текст рассматривается как последовательность независимых испытаний (ср. 18).

С другой стороны, главным инструментом теоретических построений Б. Мандельброта является введение понятия «стоимости» (cost) носителя информации (в нашем случае некоторой текстовой единицы), — стоимости, определяемой из отношения

$$h_i = -\log_2 p_i, \qquad (14)$$

где p_i — вероятность i-того носителя информации, а h_i — «стоимость» его передачи в оптимальном двоичном коде. Нетрудно заметить, что по величине и определению «стоимость» h_i совпадает с потенциальной информацией («преформацией») носителя i (ср. также: Орлов, 1970, с. 11—16). «Стоимость» кодирования текстовой единицы интерпретируется Б. Мандельбротом как длительность (длина) текстовой единицы в фонемах и буквах.

Далее предполагается, что наш лингвистический код практически согласован с вероятностью текстовых единиц. Это значит, например, что редкие словоформы или слова имеют длинное и поэтому дорогостоящее кодовое обозначение, а самые употребительные и семантически не нагруженные лексические единицы получают короткие, «дешевые» коды.

Записав (14) в виде

$$\beta_1 h_i = -\ln p_i, \qquad (15)$$

где β_1 — коэффициент перехода к натуральным логарифмам, приходим к равенству

$$p_i = \exp(-\beta_i h_i), \qquad (16)$$

характеризующему обратную связь вероятности лингвистической единицы с ее «стоимостью».

Для того чтобы связать «стоимость» с ранговым номером лингвистической единицы в вероятностном списке, используются следующие рассуждения.

Пусть $S_1, S_2, \ldots, S_j, \ldots, S_n$ — символы кода (словоформы или слова), а $h_1, h_2, \ldots, h_j, \ldots, h_n$ — их «стоимости». Опуская некоторые сложные рассуждения, можно утверждать, что число символов ценой меньше h_i или равной h_i составляет для устойчивого кода

$$i = \rho_1 e^{\beta h_i} - \rho, \qquad (17)$$

где ρ_1 и ρ — некоторые постоянные (подробнее см.: Brillouin, 1956/1960, с. 54, 60, 73).

Расположим теперь все символы нашего вероятностного словаря в порядке возрастания «стоимости» h. Тогда i будет указывать порядковый номер слова или словоупотребления в этом списке. Соотношение между h_i и i будет выражаться равенством:

$$h_i = -\frac{\ln \rho_1}{\beta} + \frac{\ln (i + \rho)}{\beta} = h_0 + \frac{1}{\beta} \ln (i + \rho). \qquad (18)$$

Чтобы исключить величину стоимости, сопоставим это выражение с формулой (15), характеризующей связь вероятности символа S_i с его стоимостью. Этим путем удается связать вероятность символа S_i непосредственно с его номером. Действительно, поскольку

$$\beta_1 h_i = \beta_1 h_0 + \frac{\beta_1}{\beta} \ln (i + \rho), \qquad (19)$$

а

$$\beta_1 h_i = -\ln p_i,$$

[1] При выводе использовалось вспомогательное равенство

$$1 + \frac{1}{2} + \ldots + \frac{1}{i} \approx \ln i + 0.577,$$

то

$$\ln p_i = -\beta_1 h_0 - \frac{\beta_1}{\beta} \ln (i + \rho). \qquad (20)$$

Приняв $\exp(-\beta_1 h_0) = k$, а $\frac{\beta_1}{\beta} = \gamma$ и потенцируя (20), приходим к формуле Мандельброта

$$p_i = k (i + \rho)^{-\gamma}, \qquad (21)$$

выражающей соотношение вероятности и ранга S_i в вероятностном словаре.

В тех случаях, когда $k=0.1$, $\rho=0$, $\gamma=1$, формула (21) включает в виде частного случая закон Ципфа (см. выражение (10) и примеч. на с. 99).

В тех случаях, когда $\gamma > 1$, а $1 \leqslant i \leqslant \infty$, Б. Мандельброт предлагает модифицировать выражение (21) в виде:

$$p_i = (\gamma - 1)\, \rho^{\gamma-1} (i + \rho)^{-\gamma},$$

где коэффициент k оказывается определенным через параметры γ и ρ (подробнее см.: Mandelbrot, 1961, с. 195—197).

27. Зависимости, связанные с законом Ципфа—Мандельброта

Опираясь на полученные выше формулы, описывающие отношение между вероятностью p_i и рангом i при условии, что $\gamma > 1$ и $1 \leqslant i \leqslant \infty$, Б. Мандельброт попытался определить ряд важных для лингвистики текста зависимостей (Mandelbrot, 1961, с. 214—218).

Первой из этих зависимостей является отношение между числом слов m, имеющих частоту F в выборке N, и величиной F.

Вероятность появления словоформы (или слова) w_i, обладающей вероятностью p_i, ровно F раз в выборке длиной в N словоупотреблений, составляет, согласно формуле Бернулли,

$$P_N (w_i, F) = C_N^F p_i^F (1 - p_i)^{N-F}.$$

Одновременно количество слов m, которое употребляется F раз в выборке N, составляет

$$m (F, N) = \sum_i U_i (i, N, F),$$

при условии, что $U=1$, когда w_i появляется; и $U=0$, когда w_i не появляется.

При этих условиях математическое ожидание m будет равно:

$$M [m (F, N)] = \sum_i M [U_i (N, F)] = C_N^F \sum_i p_i^F (1 - p_i)^{N-F}.$$

Применив к этому выражению интегрирование и используя гамма-функцию, Мандельброт приходит к приближенному равенству:

$$M [m (F, N)] \simeq \frac{\sqrt[\gamma]{kN}}{\gamma} \frac{\Gamma\left(F - \frac{1}{\gamma}\right)}{\Gamma (F + 1)} \qquad (22)$$

для случаев, когда N велико.

В тех случаях, когда F также велико, имеем

$$M [m (F, N)] \simeq \frac{\sqrt[\gamma]{kN}}{\gamma F^{\left(\frac{\gamma+1}{\gamma}\right)}}. \qquad (23)$$

При малых F предлагается вводить ряд коррекций, чтобы получить теоретические величины, близкие к опытным данным (Mandelbrot, 1961, с. 215).

Второй интересующей нас зависимостью является уже упоминавшееся выше (20) отношение между объемом выборки (N) и объемом получаемого из нее словника (L_N).

Это отношение можно записать в виде:

$$L_N = \sum_i L^0 (i, N),$$

где $L^0 (i, N)=1$, если лексическая единица с рангом i появляется хотя бы раз в выборке, и $L^0 (i, N)=0$, если она не появляется ни разу. Хотя переменная $L^0 (i, N)$ не является независимой, математическое ожидание суммы равно здесь сумме средних

$$M (L_N) = \sum_i M [L^0 (i, N)].$$

Исходя из введенного выше допущения, согласно которому словоупотребления в тексте считаются взаимно независимыми (18), имеем

$$M (L_N) = \sum_i [1 - (1 - p_i)^N].$$

(ср.: Бектаев, Пиотровский, 1973, с. 175).

Используя промежуточное равенство

$$M (L_{N+1}) - M (L_N) = \sum_{i=1}^{\infty} (1 - p_i)^N p_i$$

и интеграл

$$\int_0^1 (1 - x)^{N-F} x^F di (x) = \frac{1}{\gamma} \sqrt[\gamma]{k} \int_0^1 (1 - x)^{N-F} x^{F-1-\frac{1}{\gamma}} dx$$

где $x=p_i$, а $i=\sqrt[\gamma]{\frac{k}{p_i}}$, Б. Мандельброт (1961, с. 217) приходит в итоге к приближенному равенству:

$$M(L_N) \simeq \sqrt[\gamma]{kN}\,\Gamma\frac{\gamma-1}{\gamma}, \qquad (24)$$

описывающему отношение объема словника и величины выборки (мы не рассматриваем здесь другие модели этого типа — см.: Carroll, 1968, с. 213—230).

28. О распределении лингвистических единиц в генеральной совокупности текстов

Сопоставление теоретических данных, получаемых с помощью формул (10) и (21), с результатами статистического эксперимента показало, что закон Ципфа—Мандельброта удовлетворительно описывает поведение высокочастотных и среднечастотных словоформ и слов в интервале $0 \leqslant i \leqslant 2000$ (Бектаев и др., 1966, с. 189—198; Калинин*, 1964б, с. 9; Кочеткова*, 1969, с. 5; Маткаш*, 1967, с. 11). В том случае, когда ЧС построены на достаточно репрезентативных и сопоставимых выборках, получаемым в указанном интервале значениям γ и ρ (см. табл. 17) [1] может быть дана лингвистическая интерпретация. Так, например, увеличение величины коэффициента γ связано обычно с уменьшением богатства словаря: наибольшего значения коэффициент γ достигает в детской речи и в текстах, записанных от больных шизофренией, находящихся на такой стадии заболевания, когда навязчивое состояние больного приводит к заметному сужению его словаря.

Величина поправочного коэффициента ρ характеризует, очевидно, степень синтетичности языка (ср. данные по английскому и казахскому языкам — см.: Бектаев, 1975).

Что касается низкочастотной зоны ЧС, то она плохо описывается законом Ципфа—Мандельброта.

С одной стороны, начиная примерно с $i=1000$, все заметнее увеличиваются группы словоформ, имеющих одинаковую частоту. Эти словоформы образуют на графике «ступеньки» (рис. 25), которые не могут быть целиком аппроксимированы теоретической кривой.[2]

С другой стороны, при распространении зависимостей (10) и (21) на низкочастотные ступенчатые участки ЧС обнаружива-

[1] Судя по данным Г. Кучеры и У. Френсиса, значение коэффициента γ несколько увеличивается в зависимости от роста объема выборки, достигая более или менее стабильного уровня при выборке в 100—250 тыс. словоупотреблений (Kučera, Francis, 1967, с. 357).

[2] Теоретическая прямая Ципфа чаще всего проходит через левые концы ступенек [Калинин, 1964а, с. 125].

Таблица 17

Значения коэффициентов γ, ρ, k для разных языков, стилей и подъязыков (см.: Джубанов, 1973, с. 20; Зореф *, 1972, с. 8; Кочеткова*, 1969; Новак*, 1962, с. 8; Ротарь*, 1971, с. 6; Brillouin, 1956/1960, с. 74)

Тип языка	Язык и его разновидность	N (тыс. словоупотреблений)	γ	ρ	k
Фузионно-аналитический с остаточным синтетизмом	Афганская проза	205	1.01	0.000	0.100
	Английские тексты, записанные от шизофреника	Нет данных	1.40	Нет данных	
	Английская детская речь	Нет данных	1.60	Нет данных	
	Английская художественная проза	260	1.04	Нет данных	0.013
	Английские тексты по электронике	200	0.99	Нет данных	0.096
	Французские тексты по электронике	200	0.98	Нет данных	0.119
Фузионно-аналитический с элементами синтетизма	Румынская и молдавская проза	300	1.04	1.350	0.123
	Румынские тексты по электронике	200	0.99	0.990	0.090
	Испанские тексты по радиотехнике	200	1.07	0.050	0.160
Фузионный с преобладанием синтетизма	Немецкая публицистика	200	1.00	2.354	0.104
	Немецкие тексты по электронике	200	0.99	1.590	0.100
	Русские тексты по электронике	200	0.84	1.900	0.140
	Чешские литературные тексты	1500	1.04	1.350	0.132
Аглютинирующий	Казахская проза	466	1.02	7.835	0.146

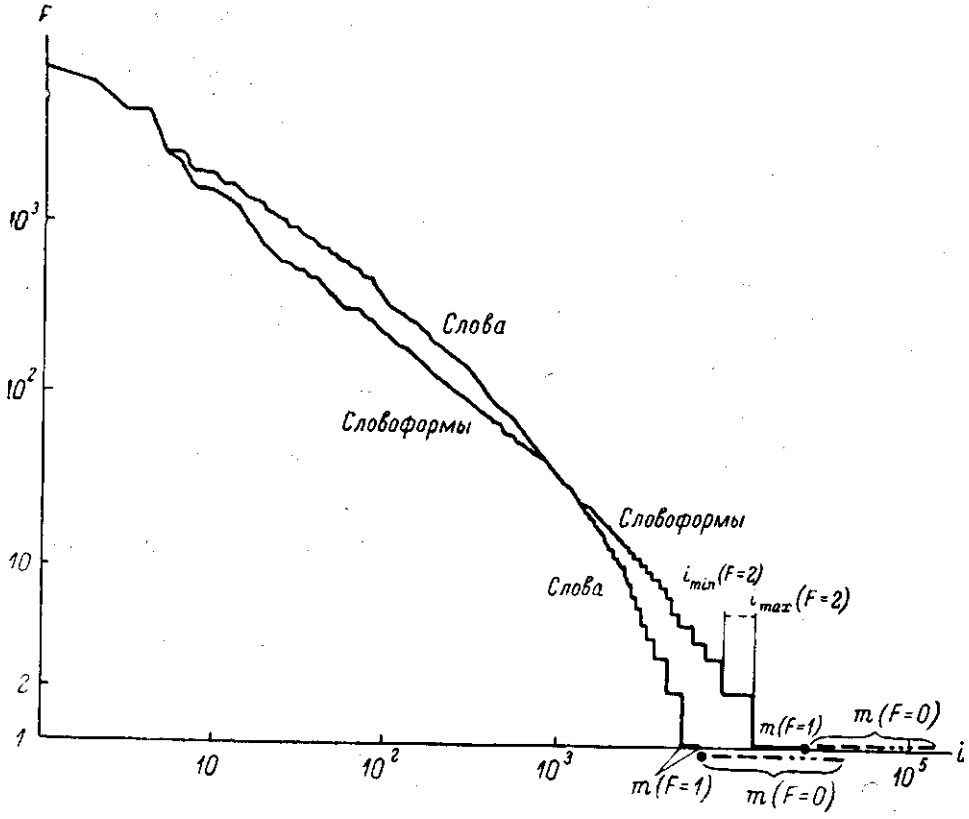


Рис. 25. Кривые зависимости ранг—частота для слов и словоформ в русских текстах по электронике (*Калинина, 1968а, с. 89—91*).

ется изменение значений коэффициентов γ и k в зависимости от роста i (ср. табл. 18). При этом В. М. Калининым (1964а, с. 126—127) были предложены следующие равенства для определения значений γ и k:

$$\gamma = \frac{i_{max}(F)}{F \cdot m} \quad \text{и} \quad k = \frac{F i_{min}(F)}{N},$$

где $i_{min}(F)$ — номер словоформы, образующей левый конец ступеньки, а $i_{max}(F)$ — номер, представляющий ее правый конец (рис. 25).

В связи со всем сказанным возникает вопрос, можно ли рассматривать частотные словари и сформулированный на их основе закон Ципфа—Мандельброта в качестве содержательных моделей, достаточно полно описывающих распределение вероятностей лексических единиц в языке.

Таблица 18

Значения коэффициентов γ и k для разных участков ЧС молдавского публицистического (Маткаш*, 1967, с. 11) и французского научно-технического (Кочеткова*, 1969) текстов

Интервалы	Выборки			
	200 тыс. словоформ французского языка		200 тыс. словоформ молдавского языка	
	γ	k	γ	k
$1 \leqslant i \leqslant 2000$	0.98	0.119 (i=2000)	0.89	0.054
$1 \leqslant i \leqslant 3000$	1.03	0.089 (i=3000)	0.90	0.056
$1 \leqslant i \leqslant 4000$	1.08	0.079 (i=400)	0.91	0.060
$1 \leqslant i \leqslant$конец	1.32	—	1.03	0.134
$1000 \leqslant i \leqslant$конец	1.59	—	1.06	0.190
$1500 \leqslant i \leqslant$конец	1.58	—	1.07	0.194
$2000 \leqslant i \leqslant$конец	1.59	—	1.06	0.190
$3000 \leqslant i \leqslant$конец	1.53	—	1.05	0.165
$4000 \leqslant i \leqslant$конец	1.36	—	1.03	0.128

Каждый частотный словарь, выступающий в роли лексикостатистической модели, фиксирует группы слов и словоформ, употребляющихся ровно один, два, три и т. д. раз. Вероятностная модель Ципфа—Мандельброта также предусматривает появление массивов одно-, двух-, трех- и т. д. кратных лексических единиц. Из (22) и (23) следует, что объемы этих массивов растут по мере роста выборки N. Кроме того, существует массив нульчастотных слов m (F=0) — см. правый нижний угол графиков, изображенных на рис. 25, — реально бытующих в лексике исследуемого языка, но не попавших в ту выборку, на основе которой была построена данная статистическая модель (ср. 25).

Таким образом, имеющиеся статистические распределения лексики в тексте, а также выросшие на их основе теоретические модели слабо отражают вероятностное, а следовательно, и информационное поведение редких слов и словоформ. Но, как уже не раз говорилось, эти лексические единицы не только обладают наибольшей потенциальной информацией (25), а, как правило, входят в те фрагменты, в которых выражена рема («новизна») текста.

Чтобы проанализировать поведение этих редких слов и словоформ представим себе генеральную языковую совокупность N_g, которая охватывает конечное множество устных и письменных текстов, появившихся в фиксированный период времени. Поскольку количество текстов здесь конечно, то конечным является и число лексических единиц в совокупности N_g.

По сравнению с опытными выборочными совокупностями генеральная совокупность имеет ряд принципиальных особенностей. Например, она не содержит нуль-частотных лексических единиц, а однократно использованные слова или словоформы будут встре-

чаться здесь в виде исключения (это относится, вероятно, и к двух-, трех- и четырехкратно употребленным словам и словоформам).[1]

Рассмотрим теперь переход от частотного словаря-модели к словарю, фиксирующему истинные частоты слов или словоформ в генеральной совокупности текстов. Этот переход можно было бы наблюдать путем последовательного увеличения выборки, которая, охватывая все большее и большее количество текстов, приближалась бы к генеральной совокупности, пока окончательно не совпала бы с ней.

Такое последовательное увеличение выборки приводит к тому, что все большее число «нуль-частотных» лексических единиц попадает в выборочную совокупность и фиксируется в частотном словаре. Одновременно увеличивается и частота некоторых редких слов и словоформ, которые передвигаются в средние ярусы частотного словаря. Иными словами, увеличение объемов выборки должно приводить к пополнению среднечастотной зоны ЧС за счет уменьшения доли редких слов. Этот процесс можно наблюдать, сравнивая частотные словари, составленные на выборках разного объема: как показывают данные табл. 15 и рис. 24, доля новых (чаще всего однократных) словоформ в русских и английских прозаических текстах постепенно уменьшается по мере роста длины выборки. В самой же генеральной совокупности однократные, двухкратные слова и словоформы могут оказаться, как уже говорилось, в виде исключения.

Передвижение редких и нуль-частотных слов или словоформ в среднечастотную зону должно повлечь за собой изменение характера зависимости ранг—частота: «ступеньки» в нижних ярусах ЧС должны исчезнуть, зато они будут расширяться в средней части словаря, сдвигая эту часть теоретического графика вправо и увеличивая его кривизну (ср.: Reid, 1944; Алексеев, 1971а, с. 175).

---

[1] Действительно, трудно себе представить такие слова, которые были бы употреблены за более или менее длительный период существования языка только один, два и даже три-четыре раза. Даже если обратиться к примерам словотворчества футуристов, — например к такому четверостишию В. Хлебникова, как

Я смеярышня смехачеств
Смехистеллино беру,
Нераскаянных хохочеств
Кинь зооку — губирю. . .,

— то вряд ли можно утверждать, что создаваемые им «неологизмы» являются однократными словами. Не говоря уже о переизданиях текстов, содержащих это словотворчество, «словоформы» типа *смеярышня*, *хохочеств* повторяются в текстах тех литературоведческих сочинений, в которых исследуется поэтика футуристов (см.: Лейтес, 1973, с. 224).

В результате на билогарифмическом графике вместо прямой Ципфа мы получаем кривую, которая может быть аппроксимирована уравнением окружности вида:

$$\lg [M(F)] = \sqrt{(\lg K)^2 - (\lg i - \lg a)^2} + \lg b. \qquad (25)$$

Переходя от билогарифмического масштаба через натуральные логарифмы к линейному масштабу, имеем:

$$M(F) = b \exp \sqrt{(\ln K)^2 - (\ln i - \ln a)^2}, \qquad (26)$$

где $M(F)$ — математическое ожидание частоты, i — номер лингвистической единицы в ЧС, а K, a, b — коэффициенты, которые могут иметь лингвистическую интерпретацию. Параметр K зависит, вероятно, от длины текста. Величина коэффициента a обусловлена, очевидно, строем языка: можно ожидать, что в аналитических языках величина a будет превосходить его значение для синтетических языков. Величина коэффициента b зависит от того, относительно каких лингвистических единиц строится зависимость ранг—частота. Если количество разных лингвистических единиц невелико и наиболее редкие единицы имеют достаточно высокую частоту, как это имеет место в статистике букв, фонем и слогов, то значение коэффициента b будет выше, чем в лексико-статистических моделях, где последние единицы списка имеют очень малые частоты.

Если алфавит лингвистических единиц невелик, то криволинейность зависимости между рангом и частотой обнаруживается хотя и на больших, но в то же время вполне реализуемых выборках. В качестве примера можно указать на распределение букв в русских текстах по электронике, полученных Е. А. Калининой (1968а, с. 82) на выборке в 113 тыс. буквоупотреблений. Такая выборка оказалась достаточной, чтобы получить распределение букв, описываемое уравнением окружности вида (25) и (26) (ср. рис. 26) с коэффициентами $K=3584$, $a=0.4345$, $b=169.4$. Близость наблюдаемых частот F и их математических ожиданий $M(F)$, вычисленных с помощью (26) (см. табл. 19), была оценена через коэффициент вариации:

$$w = \frac{1}{\bar{F}} \sqrt{\frac{\sum_{j=1}^{s} [F_j - M(F_j)]^2}{S}} \cdot 100\%, \qquad (27)$$

где S — число элементов (букв) алфавита, а $\bar{F}=N/S$. Величина вариации в нашем опыте не превышает 10% (точнее, $w=9.42\%$).

Если число элементов, составляющих данный лингвистический алфавит, велико, то для получения статистической модели, которая была бы достаточно близкой к вероятностной модели, отражающей основные свойства генеральной совокупности, необходимы также выборки, которые позволили бы не только захватить

Таблица 19

Опытные и ожидаемые частоты букв в русских текстах по электронике

i	Буквы	F	$M(F)$	i	Буквы	F	$M(F)$
1	Δ	13350	14345	17	Ы	1919	2113
2	О	11376	11800	18	У	1915	1927
3	Е	8907	9972	19	Ч	1752	1751
4	И	7852	8471	20	З	1563	1608
5	Т	7338	7378	21	Ь, Ъ	1364	1467
6	А	7020	6502	22	Г	1256	1330
7	Н	6889	5699	23	Б	1210	1204
8	Р	5498	5093	24	Х	1200	1101
9	С	5116	4571	25	Й	1032	1000
10	Л	4227	4121	26	Э	789	903
11	В	4104	3741	27	Ж	753	809
12	К	3358	3381	28	Ю	692	718
13	П	3072	3069	29	Ц	477	645
14	М	3047	2788	30	Щ	460	560
15	Д	2641	2542	31	Ф	449	480
16	Я	2302	2307	32	Ш	422	407

нуль-частотные элементы, но и распознать истинные частоты тех редких элементов алфавита, которые при малых выборках формируют группы одно-, двух-, трехкратных и т. д. лингвистических единиц. Эти группы образуют уже известные нам «ступеньки» на графике Ципфа.

Хотя пока еще никому не удавалось построить и обработать выборку, достаточную для этих целей объема, имеющиеся материалы в области грамматической и лексической стилистики позволяют нащупать по мере увеличения выборки тенденции к увеличению кривизны графика Ципфа (ср. графики, приводимые в статье: Алексеев, Навальна, 1971, с. 73 и 77).

Эту тенденцию можно обнаружить иногда и в случаях сокращения лексического алфавита, достигаемого путем перехода от словоформ к словам, при сохранении общей длины выборки (ср. рис. 25).

Итак, использующиеся в настоящее время в лингвистике текста статистические и вероятностные модели дают возможность описать поведение лингвистических единиц, обладающих высокой и средней употребительностью. Такие лингвистические единицы служат чаще всего средством выражения темы текста. Поэтому современные статистико-вероятностные лингвистические модели широко используются при построении простых алгоритмов машинного распознавания темы документа.

Однако эти модели плохо описывают поведение редких лингвистических единиц, участвующих обычно в выявлении новизны — ремы текста. Поэтому для построения машинных алгоритмов, распознающих не только тему, но и рему текста, необходимо разра-

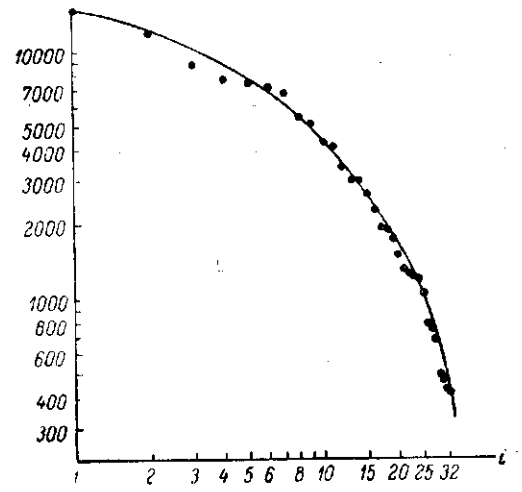


Рис. 26. Аппроксимация опытных данных зависимости ранг—частота для русских букв с помощью уравнения окружности.

ботать более тонкие вероятностно-статистические модели, чем те, которыми в настоящее время располагает лингвистика текста. Само собой разумеется, что эти модели должны описывать не только редкие лексические единицы текста (ср.: Петренко, 1974, с. 4—5), но должны быть также ориентированы на коммуникативную организацию (актуальное членение) фраз текста (ср.: Benešová, Sgall, 1973, с. 29—58).

29. Вероятностно-статистическое выявление в тексте доминантных единиц и единиц заполнения

Рассмотренные выше статистические и вероятностные модели строились исходя из упрощающего допущения о том, что текст представляет собой реализацию стационарного процесса. В роли генератора этого процесса выступает система языка, его норма и узус, которые, согласно этому допущению, остаются неизменными в рамках заданного синхронного среза. В действительности в порождении текста участвуют не только система, норма и узус языка, но также вечно меняющаяся ситуация (ср. 8).

Можно предположить, что те лингвистические элементы, появление которых определяется в основном стационарной системой, нормой и узусом языка, будут иметь в тексте иные статистические характеристики, чем те единицы, появление которых подсказывается нестационарной ситуацией.

Для проверки этого предположения был использован аппарат теории распределений. Идея этой проверки состоит в определении степени схождения эмпирических вариационных рядов отдельных лингвистических единиц (ср. рис. 16, 17) с тремя теоретическими моделями:

а) распределением Пуассона (распределением редких лингвистических элементов)

$$P(F) = \frac{\lambda^F}{F!} e^{-\lambda}, \qquad (28)$$

где $P(F)$ — вероятность появления данной лингвистической единицы ровно F раз в серии из N единиц, а λ — параметр формулы, оцениваемый с помощью средней частоты исследуемого лингвистического элемента;

б) нормальным распределением

$$P(F) \approx \frac{1}{\sqrt{2\pi N f (1-f)}} \exp\left[-\frac{(F-Nf)^2}{2Nf(1-f)}\right], \qquad (29)$$

где f — относительная частота (эмпирическая вероятность) интересующей нас лингвистической единицы (остальные величины имеют тот же смысл, что и в распределении Пуассона);

в) логнормальным распределением

$$P(F) = \frac{1}{F\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln F - \mu)^2}{2\sigma^2}\right], \qquad (30)$$

где

$$\mu = \frac{\sum_{i=1}^{N} \ln F_i}{N},$$

а

$$\sigma = \frac{\sqrt{\sum_{i=1}^{N} (\ln F_i - \mu)}}{N-1}.$$

Подробнее об этих распределениях и их использовании в математической лингвистике см.: Бектаев, Пиотровский, 1973, с. 181—203, 216—217; Бектаев, Зубов и др., 1969.

В современной математической лингвистике делаются попытки применить к текстовому материалу и другие распределения — распределение Шарлие (Джубанов, 1973, с. 15—17), Фукса и Гачечиладзе (Fucks, 1955; Гачечиладзе, Цилосани, 1971, с. 113—133), Джонсона (Рохваргер, Кирюшина, Кудрин, 1974, с. 21—23), а также кривые Пирсона (Каширина, 1974, с. 336—360). Однако ясной лингвистической картины эти эксперименты пока не дали, поэтому мы рассмотрим лингвистическое приложение первых трех моделей.

Начнем с того, что рассмотрим распределение трехсловных немецких сочетаний (триад), образованных от наиболее частых словоформ (Машкина*, 1968, с. 18—28).

Таблица 20

Проверка гипотезы о пуассоновском распределении триады Δdaβ der

№ интервала i	F_i	m_i	$P(F_i)$	$M(m_i)$	$m_i - M(m_i)$	$[m_i - M(m_i)]^2$	$\frac{[m_i - M(m_i)]^2}{M(m_i)}$
1	0	32	0.3104	31.04	0.96	0.9409	0.03
2	1	38	0.3628	36.28	1.72	2.9584	0.08
3	2	19	0.2122	21.22	—2.23	4.9729	0.23
4	3	5	0.0828	8.28	—3.28	10.7584	1.30
	4	4					
5		6	0.0242	2.98	3.02	9.1204	3.03
	5	2					
Суммы:		100	0.9980	100.00			$\chi^2 = 4.67$

Поскольку трехсловные сочетания являются редкими лингвистическими событиями, выдвигается нулевая гипотеза H_0, согласно которой их распределения подчиняются закону Пуассона.

Проверка этой гипотезы осуществляется по следующей схеме.

1. Сначала по формуле Пуассона (28) находят вероятность $P(F)$ появления частот $F = 0, 1, 2 \ldots$ и т. д. раз (ср. табл. 20, стб. 4).

2. Определяются теоретические ожидания частот появлений триады ровно 0, 1, 2 и т. д. раз. Эта величина равна

$$M(m_i) = P(F) \cdot 100.$$

3. Теоретические частоты $M(m_i)$ сравниваются с эмпирическими частотами m_i (табл. 20, стб. 5—8) при помощи критерия χ^2 Пирсона:

$$\chi^2 = \sum_{i=1}^{k} \frac{[m_i - M(m_i)]^2}{M(m_i)}.$$

В тех случаях, когда величины $M(m_i)$ и m_i меньше пяти, они объединяются с соответствующими значениями $M(m_j)$ и m_j.

4. Полученная величина χ^2 сопоставляется с пороговым значением $\chi^2_{q;\nu}$, выбранным из таблиц значений $\chi^2_{q;\nu}$ для фиксированных уровней значимости q и степеней свободы ν (Большев, Смирнов, 1965, с. 228—229; Бектаев, Пиотровский, 1974, с. 306).

Гипотеза H_0 о пуассоновском характере распределения принимается, если имеет место неравенство

$$\chi^2 < \chi^2_{q;\nu}.$$

Если же

$$\chi^2 \geqslant \chi^2_{q;\,\nu},$$

гипотеза отвергается.

По указанной схеме исследовались распределения у большой группы различных по грамматической природе и семантике трехсловных сегментов, взятых из немецких газетных текстов. Ход исследования проследим на примере сегмента Δ daβ der (начало дополнительного придаточного предложения с существительным в начальной позиции, символ Δ указывает здесь на знак препинания).

1. Имеется 100 серий, каждая из которых содержит 1000 трехсловных сочетаний. По формуле (28) определяем вероятность $P(F)$ ($F=0, 1, 2$ и т. д.) появления сегмента Δ daβ der в каждой серии (ср. табл. 20, стб. 4). Вместо параметра λ используем его оценку

$$F=\frac{1}{100}(0\cdot 32+1\cdot 38+2\cdot 19+3\cdot 5+4\cdot 6)=\frac{115}{100}=1.15.$$

2. Затем вычисляются теоретические частоты $M(m_i)$ (см. стб. 5), которые сравниваем с эмпирическими частотами m_i (см. стб. 6—8).

3. Полученная в стб. 8 величина $\chi^2=4.67$ меньше порогового значения $\chi^2_{0.05;\,3}=7.82$ (с учетом укрупнения малых частот имеем $\nu=5-2=3$) и лежит, таким образом, в области принятия нулевой гипотезы о пуассоновском характере интересующего нас лингвистического распределения.

Аналогичным образом исследовано еще 84 трехсловных и двухсловных сегмента. Все эти 85 словосочетаний распадаются на три группы.

В первую группу (34 сегмента) входят словосочетания, состоящие либо из служебных слов (например, Δ daβ ist, Δ es ist 'это—'; Δ daβ die — начало дополнительного придаточного предложения с существительным мн. ч. или ед. ч. женск. рода в начальной позиции), либо представляющие комбинацию служебных и общеупотребительных знаменательных слов (например, in der Welt 'в мире'). Вариационные ряды этих сегментов независимо от количества микровыборок и их объема показывают, подобно только что рассмотренному обороту Δ daβ der, хорошее согласие с законом Пуассона.

Вторая группа (22 сегмента) включает словосочетания, обозначающие политические реалии обоих немецких государств (ср. такие сегменты, как Deutsche Demokratische Republik 'Германская Демократическая Республика', in der Bundesrepublik 'в Федеративной Республике', beiden deutschen Staaten 'оба немецких государства'). Эмпирическое распределение этих оборотов обнаруживает несогласие с распределением Пуассона при любых видах нормировки, т. е. при любых объемах серий (порций, на которые делится выборка).

Третья группа (28 сегментов) занимает промежуточное положение между первыми двумя группировками. Входящие в нее обороты включают служебные слова, а также общеупотребительные и терминологические лексемы. Согласие или несогласие эмпирических распределений этих оборотов с законом Пуассона зависит от способа разбивки выборки на серии.

Тот факт, что триады первой группы четко противопоставлены оборотам второй группы как с точки зрения лексико-грамматической природы, так и в статистическом плане, подтверждает высказывавшиеся предположения о том, что статистическое поведение лексической единицы в тексте может служить формальным признаком для выявления ее лексико-грамматической природы. В частности, можно ожидать, что редкие слова и словосочетания, неподчиняющиеся закону Пуассона, а тем более нормальному закону, будут иметь терминологическую или конкретную семантику, в то время как слова и обороты, показывающие пуассоновское или нормальное распределения, будут нести служебные функции или будут представлять собой лексические единицы общего значения.

Гипотеза о возможности формально-статистического опознания терминологических и семантически доминантных единиц текста была рассмотрена относительно пуассоновского, нормального и логнормального распределений К. Б. Бектаевым и К. Ф. Лукьяненковым (1971, с. 47—112). Исследованию подвергались вариационные ряды двух единиц искусственных языков (буквенные символы, цифры), 298 словоформ и 300 трехсловных сегментов, взятых из английских научно-технических текстов подъязыка судовых механизмов. Общий объем исследованного текста составил 400 тыс. словоупотреблений (около 1200 страниц).

Применяя электронно-вычислительную технику, авторы рассмотрели разные варианты разбивки и нормировки материала, в результате чего для каждой словоформы были построены до 27 вариационных рядов. Затем с помощью ЭВМ по критерию χ^2 при уровне значимости 0.01 проверялось согласие этих рядов с указанными выше распределениями.

Поскольку учесть все совпадения и несовпадения эмпирических и теоретических данных в 27 вариационных рядах каждой лексической единицы невозможно, было введено следующее обобщающее условие. Эмпирическое распределение считалось совпадающим с теоретическим в том случае, если такое согласие показывал вариационный ряд, охватывающий целиком всю выборку, либо несколько вариационных рядов, построенных на частных выборках, составляющих в сумме весь выборочный текст длиной в 400 тыс. словоупотреблений. В других случаях отмечалось несогласие эмпирического и теоретического распределений. Исследование показало, что согласие или несогласие эмпирического распределения с тем или иным теоретическим законом зависит, во-первых, от статистических условий эксперимента — частоты лингвисти-

ческой единицы и способа построения вариационного ряда, а во-вторых, от семантики исследуемой словоформы или словосочетания.

Если говорить о результатах эксперимента с точки зрения статистических условий его поведения, то здесь обнаруживаются следующие тенденции.

1. Если упорядочить лексические единицы в список, построенный по принципу убывания частот исследуемых единиц, то распределение словоформ и отчасти словосочетаний, стоящих в начале такого списка, чаще всего подчиняется нормальному закону. Эта тенденция прослеживается для первых 20 словоформ вне зависимости от объема микровыборки (серии), на которые делится общая выборка. Что касается словосочетаний, то согласие с нормальным законом прослеживается в области первых 15 номеров частотного списка словосочетаний в тех случаях, когда берутся микровыборки длиной не менее 4 тыс. словоупотреблений.

2. В зоне средних частот (номера 60—1500 списка) отмечается постепенное отступление нормального закона перед законом Пуассона. При этом пуассоновское распределение появляется в первую очередь в тех вариационных рядах словоформ, которые обобщают малые микровыборки (одна—две тысячи словоупотреблений), наоборот, нормальность удерживается в рядах, построенных из крупных серий по 4, 8 или 16 тыс. словоупотреблений. Расширяя сферу своего распространения, пуассоновские распределения накладываются на нормальные: начиная с 60-х—70-х номеров списка все чаще попадаются словоформы, вариационные ряды которых при микровыборках объемом в 2—16 тыс. словоупотреблений подчиняются и нормальному, и пуассоновскому законам. Сходная картина наблюдается и в частотном списке словосочетаний с той лишь разницей, что наложение пуассоновского распределения на нормальное обнаруживается в самом начале списка.

3. Распределение редких лексических единиц независимо от разбивки и нормировки выборки подчиняется закону Пуассона, исключая случаи, когда этому препятствует семантика словоформы или словосочетания.

4. Логнормальный закон хорошо описывает эмпирические распределения большинства частых и среднечастотных словоформ при условии использования достаточно больших микровыборок (не менее 8 тыс. словоупотреблений).

Обращаясь к семантическому анализу полученных результатов, разобьем все лексические единицы на две группы. В первую группу войдут словоформы и словосочетания, эмпирические распределения которых не показывают согласия ни с нормальным, ни с логнормальным законами, ни с законом Пуассона. Вторая группа будет включать лексические единицы, вариационные ряды которых подчиняются либо нормальному, либо пуассоновскому распределению, либо одновременно обоим законам. Такая

разбивка экспериментального материала оправдывается тем, что при больших значениях параметра λ распределение Пуассона приобретает форму нормального распределения. Это аппроксимирующее свойство нормального закона обнаруживается, в частности, в наложении этого распределения на закон Пуассона в зоне среднечастотных словоформ и словосочетаний, о чем говорилось выше.

Обратимся к анализу первой группы. Просмотр приведенных в работе К. Б. Бектаева и К. Ф. Лукьяненкова лексических списков показывает, что из 298 словоформ эмпирические распределения 66 единиц не подчиняются ни нормальному, ни пуассоновскому законам, а среди 300 словосочетаний 22 не дают согласия своих вариационных рядов с указанными теоретическими законами.

Какова же лексико-грамматическая природа этих слов и словосочетаний?

Во-первых, сюда относятся именные, глагольные, адъективные словоформы, а также именные обороты явно терминологического значения. Ср.:

bearing(s) ‘подшипник(и)’, blade ‘лопасть’, cylinder(s) ‘цилиндр(ы)’, diesel ‘дизель’, exhaust ‘выхлопная труба, выхлоп, выпуск’, nozzle ‘форсунка’, machinery ‘машинное оборудование, механизм’, piston ‘поршень’, power ‘сила, мощность’, pump ‘насос’, temperature ‘температура’, tubes ‘трубы’, turbine(s) ‘турбина(ы)’, valve ‘клапан’, velocity ‘скорость’ и т. п.;

gas turbine ‘газовая турбина’, marine diesel engine ‘судовой дизельный двигатель’, water level ‘уровень воды’ и др., current ‘текущий’, moving ‘приводя(щий) в движение’, marine ‘морской’, work ‘работать’ и т. п.

Во-вторых, не подчиняются нормальному и пуассоновскому закону эмпирические распределения вспомогательных и модальных глагольных форм was, would, should, must, а также распределение глагольного оборота . . . shown in fig. . . ‘. . . показано на рис. . .’

В-третьих, в рассматриваемую группу попадают сегменты текста, представляющие собой комбинацию словоформ и символов искусственных языков — цифр (X), а также буквенных обозначений, сокращений и формул (Z). Ср.: X and X ‘X и X’, at XZ ‘при XZ’, is XZ ‘равняется XZ’, figure XZ ‘рисунок XZ’, see XZ ‘см. ZX’, value of X ‘значение X’ и т. п.

Вторая группа, как уже говорилось, включает текстовые единицы и словосочетания, вариационные ряды которых дают согласие с распределением Пуассона и нормальным распределением. Сюда входят отдельно употребляемые единицы искусственных языков, большинство служебных слов, все наречия, местоимения, числительные, а также все существительные, прилагательные и глаголы общеупотребительного характера, а также общеупотребительные словосочетания. Вместе с тем здесь обнаруживается

и небольшое количество терминологических слов и словосочетаний.

Метод проверок гипотезы о согласии эмпирического и теоретического распределения с помощью критерия χ^2 применялся не только к научно-техническим текстам, но и к художественной прозе, публицистике, а также к речи душевнобольных. При этом выяснилось, что несогласие эмпирического распределения с нормальным и пуассоновским законами обнаруживают не только терминологические слова и словосочетания, а также близкие к ним по семантической природе комбинации словоупотреблений и элементов искусственных языков, но и другие лексические единицы, обладающие важным, с точки зрения сюжета текста, значением. В художественной речи этим свойством обладают слова и словосочетания, выражающие основные образы произведения, в публицистике — имена собственные и названия особо важных политических, экономико-социальных и географических реалий, в речи душевнобольных — слова и словосочетания, обозначающие объекты навязчивого состояния (Пашковский, Сребрянская, 1971). Короче говоря, несогласие эмпирических и теоретических распределений обнаруживают семантически доминантные (ключевые) единицы текста, напротив, недоминантные единицы текста, обладающие «стертой» семантикой, показывают хорошее согласие своих распределений с теоретическими законами.

Природа этого явления состоит, очевидно, в следующем.

Как уже говорилось, порождение текста определяется, с одной стороны, системой языка и его нормой, а с другой — независимой от языка ситуацией, которую призван описывать этот текст. Появление семантически нагруженных ключевых элементов текста диктуется ситуацией, употребление же семантически «стертых» единиц, так называемых единиц заполнения, служащих средством организации текста и связи ключевых единиц, подчиняется требованиям системы и нормы языка. Отсюда следует, что статистика единиц заполнения подчиняется вероятностным законам нормы языка, в то время как статистика ключевых элементов текстов управляется вероятностью ситуаций.

Статистика нормы имеет иной порядок, чем статистика ситуаций. Текстовые выборки в несколько десятков или сотен словоупотреблений оказываются достаточными, чтобы обеспечить регулярную повторяемость служебных слов, общеупотребительных слов и словосочетаний, не говоря уже о грамматических классах или фонемах, в связи с чем их вариационные ряды начинают сходиться к пуассоновскому и нормальному распределениям. Напротив, такие выборки оказываются слишком малыми, чтобы обеспечить регулярную повторяемость ситуаций: эти малые выборки остаются статистически неоднородными для диктуемых ситуаций ключевых слов и словосочетаний. Вполне естественно, что эмпирические распределения этих ключевых единиц не согласуются

с теоретическими распределениями, действующими только в однородных выборках.

Согласие и несогласие вариационных рядов с тем или иным теоретическим распределением зависит также от валентностных возможностей лингвистической единицы.

Нормальный закон и закон Пуассона хорошо описывают распределения служебных слов, общеупотребительных слов и словосочетаний, в частности, потому, что этим последним присуще сравнительно независимое употребление в тексте. Напротив, лингвистические единицы, характеризующиеся сильными валентностями, статистико-вероятностным выражением которых являются марковские связи текста, дают, как правило, скошенные вправо эмпирические распределения, имеющие иногда несколько вершин. Эти вариационные ряды не подчиняются нормальному закону и закону Пуассона, но зато могут описываться логнормальным распределением или кривыми Пирсона с правой асимметрией (Каширина, 1974).

Вариационные ряды, описываемые логнормальным распределением, дают уже упоминавшиеся английские вспомогательные и модальные глаголы was, would, should, must. Хотя эти глагольные формы не являются ключевыми единицами текста, они, очевидно, тесно связаны с такими знаменательными словоформами, которые несут в тексте смысловую или эмоционально-выделительную нагрузку.

Описанные особенности распределений служебных и общеупотребительных лексических единиц, а также терминологических и вообще доминантных элементов текста могут быть использованы в качестве диагносцирующего аппарата, автоматически распознающего в тексте ключевые слова и выражения текста. Извлечение с помощью ЭВМ из больших массивов текста семантически нагруженных ключевых единиц позволяет, с одной стороны, автоматизировать трудоемкие процессы составления алгоритмов машинного реферирования (56—60, ср.: Каушанская, Лукьяненков, 1972, с. 117—131) и семантического перевода текста (68, ср.: Славина, 1974; Трибис, 1973; Угодчикова*, 1975).

С другой стороны, статистико-автоматическое исследование семантики текста на машине дает богатый материал для решения таких теоретических и прикладных задач, как исследование соотношений «смысл—текст», «норма—речь», построение обучающих алгоритмов, лингвистическая диагностика душевных заболеваний.

Глава 5

ИНФОРМАЦИОННЫЕ ИЗМЕРЕНИЯ В ТЕКСТЕ

30. Машинные алгоритмы и информация. Три подхода к определению понятия «количество синтактической информации»

При построении машинных алгоритмов, функционирующих в системе «человек—машина—человек», а также при выборе оптимального режима работы этой системы необходимы различные информационные характеристики текста. Вероятно, наибольший интерес здесь представляло бы измерение прагматической информации. Ведь именно на этом уровне осуществляется во всей полноте реальный процесс речевого общения, на основе прагматической информации формируется децизивный уровень готовности (4), на котором принимает решение человек, прочитавший текст, который выдала ему ЭВМ.

Однако для расчета прагматической информации относительно приемника информации необходимо иметь вероятностные оценки тех ситуаций, в которых может оказаться интерпретант, принявший и правильно понявший сообщение. Для определения ценности информации с точки зрения передатчика необходимо соответственно оценить ситуации, в которых окажется последний после передачи сообщения. Сложность этой задачи очевидна, и не случайно поэтому, что мы пока еще не располагаем работающей процедурой, с помощью которой можно было бы получить количественные оценки прагматической информации, содержащейся в языковом тексте.

Не менее сложным является измерение семантической информации (Bogodist и др., 1974, с. 19—28). Более доступными являются измерения синтактической и потенциальной информации,[1] и к настоящему времени для многих языков получены ее количественные оценки (ср.: Пиотровский, 1968; Яглом, 1973, с. 236—281). Поэтому мы начнем описание информационной структуры текста с измерения его энтропии и синтактики, чтобы затем, опираясь на них, перейти к количественным оценкам смысловой информации.

При определении понятия «количество информации» используется три подхода: комбинаторный, вероятностный и алгоритмический (Колмогоров, 1965, с. 3—7).

[1] В предыдущей главе осуществлялось измерение потенциальной информации у лингвистических единиц. Однако эта информация менее ценна для нас, поскольку она направлена на наблюдателя информационного процесса, в то время как синтактическая информация воспринимается приемником (2).

При комбинаторном подходе предполагается, что переменное x способно принимать значения, принадлежащие конечному множеству X (в лингвистических терминах — алфавиту), которое состоит из S элементов (в нашем случае — букв, фонем, слогов, морфем, слов и т. д.). Тогда энтропия переменного x будет равна.

$$H(x) = \log S. \tag{1}$$

Определяя через i значение переменного x, снимаем эту энтропию и сообщаем «информацию» (I), количественно равную

$$I = \log S. \tag{2}$$

Обычно в информационно-лингвистических расчетах используются логарифмы с основанием 2, причем энтропия и информация измеряется здесь в двоичных единицах (дв. ед.), или битах. В этом случае выражение (1) можно переписать так:

$$H = \log_2 S \text{ дв. ед.} \tag{3}$$

Если переменные $x_1, x_2, \ldots, x_n$ независимо пробегают множества, состоящие соответственно из $S_1, S_2, \ldots, S_n$ элементов, то тогда

$$H(x_1, x_2, \ldots x_n) = H(x_1) + H(x_2) + \ldots + H(x_n) \text{ дв. ед.}$$

В лингвистике текста комбинаторное измерение информации дает возможность оценить гибкость речи, т. е. определить степень разветвленности ее продолжения на определенном шаге текста. Иными словами, формула (3) позволяет оценить то структурное разнообразие (Урсул, 1971, с. 146—156), которое характеризует либо алфавит языка в целом, либо алфавит лингвистических единиц, употребляемых в данном участке текста. Энтропию алфавита мы будем передавать символом H_0.

Однако комбинаторный подход дает очень грубые завышенные оценки структурно-статистической организации текста. Как уже говорилось выше, норма естественного языка приписывает каждому элементу его системы (фонеме, морфеме, слову и т. д.) определенные вероятности (условные и безусловные). Поэтому более содержательные результаты в информационных исследованиях можно получить, используя вероятностный подход. Так, например, имея распределение только безусловных вероятностей $p_1, p_2, \ldots, p_s$ для элементов, образующих алфавит S, мы можем вычислить среднюю удельную энтропию первого порядка, приходящуюся на один элемент алфавита S. Эта энтропия, согласно второй теореме Шеннона, определяется из выражения: [1]

$$H_I = -\sum_{i=1}^{s} p_i \log_2 p_i \text{ дв. ед.} \tag{4}$$

[1] Доказательство этой теоремы и рассмотрение необходимых для этого допущений см.: Shannon, 1948/1963, с. 394; Фаддеев, 1956, с. 227—231; Feinstein, 1958/1960.

Одним из видов энтропии первого порядка является энтропия начального элемента (буквы, слога, слова) текста или его сегмента, получаемого из равенства

$$H_1 = -\sum_{i=1}^{s} p_i^1 \log p_i^1, \qquad (5)$$

где p_i — вероятность появления в начальной позиции i-того символа алфавита S.

Задача усложняется, когда необходимо учесть и неравномерность в зависимостях между элементами алфавита. Здесь энтропия рассчитывается исходя из следующих соображений.

Обозначим некоторую цепочку лингвистических элементов $l_1, l_2, \ldots, l_{n-1}$ символом b^{n-1}. Цепочка b^{n-1} рассматривается в качестве случайного события, принимающего частный вид (значение) i. Предположим при этом, что $p\,(b_i^{n-1})$ всегда меньше единицы, т. е., грубо говоря, каждый лингвистический алфавит содержит более чем один элемент. Непосредственно за цепочкой n—1 следует позиция l_n. Появление того или иного элемента в этой позиции также рассматривается в качестве случайной величины, принимающей значение j_k $(1 \leqslant k \leqslant S)$. Для каждого значения i, которое может принять b^{n-1}, имеется условная вероятность $p\,(j_{i,k}/b_i^{n-1})$ того, что l_n примет значение j_k.

Средняя условная энтропия $H_n = I_n^1$ для позиции l_n будет получена в результате осреднения энтропии, сосчитанной по всем значениям b_i^{n-1}, с весами, соответствующими вероятностям цепочек n—1. Таким образом, имеем

$$H_n = -\sum_{b_i^{n-1}} p\,(b_i^{n-1}) \sum_{k=1}^{s} p\,(j_{i,k}/b_i^{n-1}) \log p\,(j_{i,k}/b_i^{n-1}). \qquad (6)$$

Эта величина показывает, какова в среднем неопределенность выбора лингвистического элемента n, когда известна цепочка n—1.

Если взаимосвязи элементов распространяются как угодно далеко, то энтропия на один лингвистический элемент будет равна

$$H_\infty = \lim_{n\to\infty} H_n. \qquad (7)$$

Поскольку величины средней условной энтропии зависят от распределения вероятностей элементов на n-м шаге текста и от вероятностей появления b_i^{n-1}, постольку эти величины могут быть определены из статистики многоэлементарных сочетаний. Расчетная формула в этом случае имеет следующий вид:

$$H_n = H\,(j/b_i^{n-1}) = H\,(b_i^n) - H\,(b_i^{n-1}). \qquad (8)$$

Объем работы при вычислении H_I, H_{II}, H_{III}, H_{IV} для буквенного и фонемного алфавитов сравнительно невелик (ср.: Лебедев, Гармаш, 1959; с. 78—80). Существуют также реалистичные способы оценки H_I для слогов, слов, словоформ и даже словосочетаний (Küpfmüller, 1954, с. 368—369; Бектаев, Машкина, Микерина, Ротарь, 1966).[1]

Оценки H, вычисленные с помощью вероятностной методики, находятся гораздо ближе, чем H_0, к истинным количественным значениям синтактической информации, характеризующей структуру текста. Однако использование только что описанных приемов связано с затруднением технологического характера. Дело в том, что число комбинаций из пяти, шести и т. д. букв (фонем), не говоря уже о комбинациях из четырех, пяти и т. д. слов (слогов, морфем) растет так быстро, что для определения условной энтропии этих и более высоких порядков даже при использовании упрощенных методов расчета (Урбах, 1963, с. 112) требуется колоссальная счетная работа. Эту работу нелегко выполнить, даже используя вычислительную технику, возможности которой в смысле обработки больших массивов лингвистической информации весьма ограничены.

Применение комбинаторной и вероятностной методики связано также с трудностями принципиального характера. Эти измерения, как уже говорилось, осуществляются в отвлечении от информационных свойств приемника сообщения.

Действительно, для того чтобы определить количество синтактической информации, содержащееся в сообщении об определенном

---

[1] Если речь идет не о средней удельной энтропии, а о том количестве синтактической информации, которое приносит адресату единичное осуществление лингвистического события i (например, появление слова i), то оно определяется из выражения

$$I\,(i) = -\log_2 p_i \text{ дв. ед.}$$

при условии, что апостериорная вероятность события i в результате проведения опыта стала равной единице. Если же $p_{\text{апост}}\,(i) < 1$, то количество информации, получаемое при осуществлении события i, составит

$$I\,(i) = -\log_2 \frac{p_{\text{апр}}\,(i)}{p_{\text{апост}}\,(i)} \text{ дв. ед.}$$

[1] С помощью этих приемов показано, в частности, что различные языки дают примерно одно и то же структурно-статистическое разнообразие словаря: количество синтактической информации, приходящейся здесь в среднем на одно слово, составляет около 10 дв. ед. Что касается трехсловных сочетаний, то информационная оценка структурно-статистической сложности их словника также, очевидно, колеблется в разных языках вокруг некоторой постоянной величины, которая заметно меньше 30 дв. ед. (т. е. утроенной информационной нагрузки, падающей на одно слово), а это, в свою очередь, говорит о том, что внутри трехсловных сочетаний имеются сильные контекстные связи (подробнее см. в 21).

событии, необходимо знать априорную вероятность этого события с точки зрения приемника информации.

Оценить эту вероятность можно в тех случаях, когда передается большое число не связанных или слабо связанных между собой речевых сообщений, подчиненных определенным закономерностям. Примером таких сообщений может служить последовательность повторяющихся стандартизованных формул, с помощью которых осуществляется связь по каналам «земля—воздух», «земля—вода» и т. п.

Но какой реальный смысл имеет, например, говорить о «количестве информации», заключенном в утверждении, что сейчас 1976 год. Действительно, к какому множеству лет принадлежит этот год? А разве можно определить вероятность наступления этих лет, особенно уже минувших лет, которые не могут уже произойти ни с какой вероятностью (Mazur, 1970/1974, с. 15; Колмогоров, 1965, с. 6—7).

Прямолинейное применение комбинаторного или вероятностного подхода при измерении информации может привести к парадоксальным результатам. Так, например, используя относительные частоты появления букв и их сочетаний в «тексте», представляющем лишенную смысла последовательность букв, удается вычислить содержащееся в ней количество информации, хотя эта последовательность не несет осмысленной информации. Аналогичным образом сочетание *дезоксирибонуклеиновая кислота* количественно несет больше синтактической информации, чем его аббревиатура *ДНК*, хотя с содержательной точки зрения мы имеем дело с одним и тем же объемом информации. Но особенно поучительны в этом отношении эксперименты по восприятию текста людьми с различной степенью владения тем языком, на котором написан текст. Действительно, текст, написанный на данном языке, всегда несет определенное количество синтактической информации, которое может быть вычислено с помощью шенноновской методики. Однако для человека, не знающего языка, эта информация с содержательной точки зрения будет равна нулю. Информация здесь будет больше нуля лишь для такого «приемника», который способен хотя бы частично идентифицировать принимаемую информацию с той, которая уже хранится в его памяти. Эта идентификация представляет собой понимание текста приемником. Максимальное количество смысловой информации извлечет из текста тот приемник сообщения (читатель или слушатель), который будет обладать образцовым знанием языка и тематики. Понимание текста приемником можно рассматривать дифференцированно, отдельно учитывая и оценивая инструктивность и мотивировку, правильность или ложность сообщения, наконец, стоимость и полезность информации. Однако рассмотрение этих вопросов увело бы нас в сторону от вопросов проблемы соотношения между синтактической и семантико-прагматической информацией. Поэтому мы их касаться здесь не будем (к ним придется

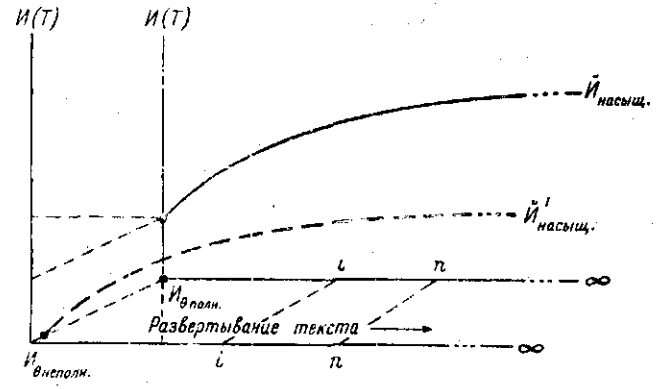


Рис. 27. Исходные состояния тезауруса и его насыщение.

еще вернуться, рассматривая наши эксперименты — см. 42—52). Сейчас нам важно подчеркнуть, что при информационных измерениях текста, написанного на естественном (как, впрочем, и на искусственном) языке целесообразно не только опираться на статистические свойства сообщения, но также учитывать ту априорную смысловую информацию, с помощью которой приемное устройство декодирует сообщение. Эти принципы измерения информации в речи согласуются с алгоритмическим подходом к понятию информации, при котором исследователя интересует количество информации в индивидуальном объекте X относительно индивидуального объекта Y, предложенной при условии достаточно больших количеств информации, содержащихся в объектах (Колмогоров, 1965, с. 7).

Измерение информации при таком подходе можно представить себе следующим образом.

Если приемник сообщения обладает «тезаурусом» сведений, относительно кода (языка) L, с помощью которого декодируется текст T, и величина априорной информации, содержащейся в этом тезаурусе, измеряется величиной $И_\theta$, то количество осмысленной информации, извлекаемой приемником из текста T при изменении $И_\theta$, будет описываться функцией

$$И(T) = f(И_{\theta, i}),$$

имеющей вид, показанный на трехмерном графике (рис. 27).

Здесь $И_{\theta\ неполн.}$ — наименьшее количество априорной информации в тезаурусе, при котором приемное устройство начинает что-то понимать в сообщении T, оно является тем порогом примитивности приемника, ниже которого последний не способен воспринимать информацию, содержащуюся в тексте T; $И_{\theta\ полн.}$ — то исходное количество априорной информации в тезаурусе, при котором приемник может извлечь максимум информации из сообщения; величина $И_{\theta\ неполн.}$ характеризует такое состояние прием-

ника, при котором он, обладая неполным знанием языка (кода), все же способен извлечь из сообщения часть содержащейся в нем информации.

В ходе приема и декодирования сообщения в тезаурусе приемника происходят изменения: к содержащейся в нем априорной информации $И_0$ добавляется информация, извлеченная им из уже прочитанных участков текста 1, 2, . . ., i.

Сумма этих информаций

$$И_i = И_0 + \sum И_i$$

последовательно возрастает по ходу декодирования приемником все новых участков текста. Весь этот процесс насыщения тезауруса угадчика характеризуется равенством

$$И_0 < И_1 < И_2 < \ldots И_{n-1} < И_n < \ldots < И_{\text{насыщ.}}, \qquad (9)$$

где индексы при величинах $И$ обозначают номера равных по длине участков текста, а $И_{\text{насыщ.}}$ указывает на то количество информации, которое будет накоплено приемником при прочтении текста бесконечной длины.

В декодировании начала текста участвует не весь тезаурус приемника. Однако по мере движения приемника по тексту поступающая в него текстовая информация эманирует все большее число «механизмов» тезауруса. Но по мере включения в декодирование все большего числа механизмов количество информации, извлекаемой приемником из каждого участка текста, уменьшается.

Действительно, завязка рассказа, начало научной статьи обычно дают больше информации, чем развязка или заключительные выводы, к которым читатель подходит уже подготовленным всем ходом повествования. С другой стороны, известно, что чем больше знает человек, тем более строгим становится отбор поступающей к нему информации: его тезаурус учитывает лишь минимум ценной информации, малоценная же информация тезаурусом не воспринимается. Напротив, любознательный простак, не умеющий из-за своей слабой осведомленности отличить ценную информацию от малоценной, старается поглотить максимум сведений.

Итак, можно утверждать, что процесс извлечения информации из каждого последующего участка текста характеризуется неравенством

$$И_1 > И_2 > \ldots > И_{n-1} > И_n > \ldots > И_\infty.$$

Достигнув на некотором участке текста полного насыщения своего тезауруса ($И_{\text{насыщ.}}$), приемник перестает извлекать из текста новую информацию (это относится и к приемникам, имеющим полный тезаурус априорной информации, и к приемникам, обладающим неполным тезаурусом). Правда, ниже будет показано, что полное насыщение тезауруса имеет место лишь при декодировании отдельно взятых слов. Что же касается отдельно взятого текста, то при бесконечном его удлинении тезаурус достигает такого порога насыщения, при котором приемник начинает извлекать из текста некоторый постоянный минимум информации $И_\infty$. Достижение этого порога свидетельствует о том, что в процесс декодирования текста включились все «механизмы» тезауруса и поэтому дальнейшая количественная оптимизация в извлечении информации из текста невозможна.

Исходя из всех этих рассуждений, можно предположить, что существуют некоторые зависимости между информационной насыщенностью тезауруса, количеством извлекаемой из текста информации и длиной текста. Однако задать аналитические выражения этих зависимостей и создать реалистическую программу прямого измерения величин $И$ и $И(T)$ не удается из-за невозможности дать численную оценку априорной информации, содержащейся в тезаурусе приемника сообщения. Если бы мы хотели определить величину $И(T)$ с помощью шенноновской меры (а других реалистичных приемов определения количества информации у нас нет), мы должны были бы рассматривать тезаурус приемника как некоторый список событий и их вероятностей. Можно дискутировать о способах составления и описания таких списков для искусственных языков, в частности для языков программирования, используемых в ЭВМ (Шрейдер, 1967). Но как решить эту задачу относительно естественного языка? Ведь нет такого языка, для которого мы располагали бы полным списком элементов и связей системы, экстралингвистических ситуаций одновременно с вероятностями (нормой) их употребления. Система каждого языка, равно как и норма функционирования, остается для нас «черным ящиком».

В этих условиях оценить величину $И(T)$ можно только косвенным путем, наблюдая выходные реакции приемника на подаваемый на его вход текст сообщения.

При решении этой задачи должны выполняться два условия. Во-первых, выдаваемые приемником сигналы о реакциях тезауруса должны иметь такой вид, который был бы пригоден для приложения шенноновской меры. Во-вторых, должно быть фиксировано если не количественно, то хотя бы качественно состояние тезауруса приемника сообщения.

Этим условиям отвечает широко применяемый в информационных измерениях речи эксперимент по угадыванию букв неизвестного текста.

31. Эксперимент по угадыванию текста «образцовым» носителем языка

Выше уже говорилось, что тезаурус (система и норма языка плюс экстралингвистический «житейский» опыт), использующийся носителем языка при декодировании текста, представляет собой

«черный ящик». Устройство и функционирование этого «черного ящика» неодинаково у различных носителей языка: разные носители языка по-разному знают свой родной язык и по-разному будут угадывать текст. Поэтому, чтобы получить для разных языков сопоставимые результаты, необходимо каждый раз использовать угадчика с определенным образом стабилизированным тезаурусом и фиксированной стратегией угадывания. Нетрудно понять, что оптимальным решением этой задачи будет выбор образцового информанта, то есть такого угадчика, тезаурус которого содержит $И_{0\,\text{подл.}}$ информации и который, пользуясь наилучшей стратегией угадывания, будет извлекать из текста $И\,(T)_{\text{макс.}}$ информации. Лингвистическое поведение этого образцового угадчика будет моделировать лингвистические реакции того homo sapiens universalis, на которого ориентировано в будущем построение лингвистического робота (11).

Не задаваясь пока вопросом, как найти образцового угадчика или как приблизить возможности рядового угадчика к уровню идеального информанта (см. об этом ниже, с. 141), рассмотрим в общих чертах ход идеального угадывания.

Опираясь на априорную информацию тезауруса и информацию, извлеченную из уже декодированного текста, наш угадчик формирует спектр индуктивных вероятностей эвентуальных продолжений. Поскольку угадчик использует полный тезаурус и применяет наилучшую стратегию угадывания, ранжирование продолжений, опирающееся на индуктивные вероятности, должно совпадать на каждом шаге текста с их ранжированием по статистическим вероятностям.[1] При обобщении результатов угадывания мы получаем либо относительные частоты появления букв в определенных участках текста, либо частости определенных видов угадывания (см. ниже, с. 134). Рассматривая совокупность этих относительных частот как спектр статистических вероятностей, мы получаем синтактическую информацию I, которая благодаря изоморфности ранжирования индуктивных и статистических вероятностей в условиях идеального угадывания может использоваться для оценки информации $И\,(T)$. Этим путем удается связать синтактическую меру информации с оценкой сообщения со стороны приемника, чего не удавалось сделать при прямом расчете информации из распределения относительных частот букв, словоформ и т. д.

При проведении эксперимента по угадыванию используются разные приемы. Сущность всех этих приемов заключается в следующем. Берется текст, полностью или частично неизвестный для испытуемых. Испытуемые должны восстановить неизвестную им часть, последовательно отгадывая буквы текста. Пробел между словами, условно обозначаемый знаком Δ, а также приравниваемые ему апостроф и черточка считаются буквой слова.[1]

В настоящей работе использованы данные, полученные в результате проведения двух типов экспериментов. Эксперименты первого типа предусматривали угадывание текста одним испытуемым, эксперименты второго типа опирались на коллективное угадывание.

Лингвистической и информационной интерпретации подвергаются в основном статистические данные, которые были добыты с помощью экспериментов первого типа, реализовавшихся в виде трех вариантов: первый вариант мы будем условно называть угадыванием по сокращенной программе, второй — угадыванием по полной программе, третий — по методу Колмогорова.

Начнем рассмотрение этих приемов с индивидуального угадывания по сокращенной и полной программе.

32. Сокращенная программа угадывания

Угадывание по сокращенной программе заключается в том, что испытуемый называет букву, которая, по его мнению, с наибольшей вероятностью стоит на n-м шаге текста (n—1 букв текста уже известны испытуемому). Если предсказание оказалось правильным, экспериментатор сообщает об этом испытуемому, и последний переходит к угадыванию следующей буквы. Если угадывание было неправильным, испытуемому называется правильная буква, и он так же, как и в первом случае, переходит к угадыванию следующей буквы. В каждом случае в протоколе указывается, правильно или неправильно была угадана буква на данном шаге текста (о различении достоверных и недостоверных продолжений см. ниже, с. 131).

[1] У рядового (не идеального) угадчика корреляция между реальными частотами и субъективными их оценками оценивается психологами примерно в 80—90% (Attneave, 1953).

[1] Текст, разумеется, можно угадывать также и по фонемам, слогам, морфемам. Предпочтительное использование буквенного алфавита объясняется его универсальностью, привычностью и понятностью для каждого грамотного носителя языка, чего нельзя сказать о фонемном, слоговом или морфемном алфавитах. В частности, для большинства языков неизвестны однозначные процедуры выделения в потоке речи таких дискретных единиц, как фонемы и слоги. Не заслуживают доверия количественные оценки смысловой информации, получаемые от текста по угадыванию слов (не букв!) неизвестного текста, осуществляемые группой, которая включает конечное число носителей языка. Дело в том, что алфавит слов, которые могут служить продолжениями на каждом участке текста, чаще всего бывает очень велик, поэтому он не только не умещается в непосредственной (кратковременной) памяти угадчика, но некоторые его участки вообще не удается вызвать из долговременной памяти и использовать для идентификации угадываемого слова. Тем более затруднительно распределить на весь алфавит слов индуктивные вероятности или получаемые в эксперименте относительные частоты, которые будут заметно смещены относительно истинных вероятностей появления слов и будут содержать высокую положительную статистическую ошибку. Поэтому получаемые в результате указанного эксперимента распределения относительных частот нельзя рассматривать как приближенное распределение априорных вероятностей словоформ-продолжений.

Рассматривая результаты этого эксперимента в плане кодирования и передачи сообщения, К. Шеннон предложил следующую интерпретацию системы связи, использующей этот вид угадывания.

Имеется исходный текст и его сокращенный вариант. Последний составлен таким образом, что угаданные с первой попытки буквы отмечены черточкой, а в случае неправильного угадывания выписаны буквы исходного текста. Сокращенный текст хотя и передает ту же смысловую информацию, что и исходный, но менее избыточен, чем этот последний (см. табл. 21).

Таблица 21

Исходный русский текст и текст, компрессированный по сокращенной программе угадывания

Шаги текста (n)	1	2	3	4	5	6	7	8	9	10	11
Исходный текст	м	н	о	г	п	е	Δ	д	е	л	е
Сокращенный текст	м	н	—	—	—	—	—	д	е	—	е
Шаги текста (n)	12	13	14	15	16	17	18	19	20	21	22
Исходный текст	г	а	т	ы	Δ	а	з	н	а	т	с
Сокращенный текст	—	—	—	—	—	а	з	—	—	—	—
Шаги текста (n)	23	24	25	26	27	28	29	30	31	32	33
Исходный текст	к	и	х	Δ	и	Δ	а	ф	р	и	к
Сокращенный текст	—	—	—	—	и	—	—	—	—	—	—
Шаги текста (n)	34	35	36	37	38	39	40	41	42	43	44
Исходный текст	а	н	с	к	и	х	Δ	с	т	р	а
Сокращенный текст	—	—	—	—	—	—	—	—	—	—	—
Шаги текста (n)	45	46	47	48	49	50	51	52	53	54	55
Исходный текст	н	Δ	с	о	г	л	а	с	н	ы	Δ
Сокращенный текст	—	—	с	—	—	—	—	—	—	—	

Будем рассматривать сокращенный текст как закодированную и одновременно компрессированную форму исходного текста. Предположим также, что наш информант (им может быть не только человек, но и машина) предсказывает буквы наилучшим образом, т. е. в соответствии с истинными вероятностями их появления в тексте. Предположим, наконец, что существует двойник нашего информанта, который пользуется той же стратегией угадывания. При этих условиях можно построить систему связи, преобразующую на входе исходный текст в сокращенный, а затем на выходе декодирующую этот последний. Блок-схема этой системы дана на рис. 28.

Система работает следующим образом. На входное сравнивающее устройство подается n-я буква исходного текста. Сюда же с входного предсказывающего устройства, которое работает, опираясь на знание $(n-1)$-й буквенной цепочки, подается предсказание n-й буквы (правильное или неправильное). Результат предсказания, будучи записанным так, как это было указано выше, образует n-ю букву сокращенного текста, которая передается

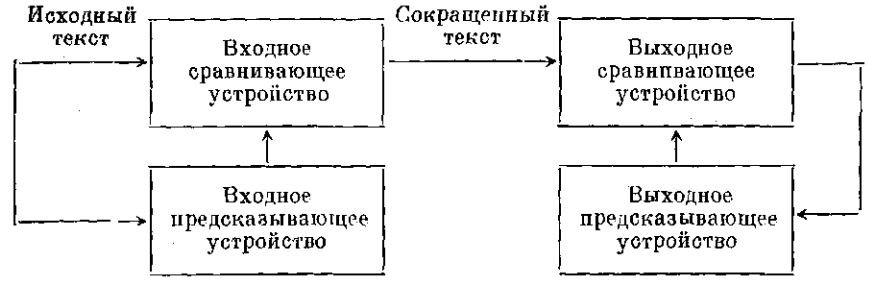


Рис. 28. Система связи, использующая сокращенный текст.

дальше в канал связи. Аналогичным путем обрабатываются и следующие буквы исходного текста.

Выходное сравнивающее устройство, приняв символ сокращенного текста, запускает выходное предсказывающее устройство. Поскольку это последнее является двойником входного предсказателя, оно должно с первой же попытки восстановить отсутствующие буквы сокращенного текста, — буквы, которые были правильно предсказаны первым предсказателем. Те же буквы, которые были неправильно угаданы на входе, обозначены в сокращенном тексте. Восстановление исходного текста из сокращенного, которое мы будем называть обратимостью текста, указывает на отсутствие потерь смысловой информации в описанной системе связи.

33. Полная программа угадывания

Проведение полной программы угадывания также дает сокращенный текст, который мы будем называть цифровым текстом. Отличие полной программы от сокращенной заключается в том, что после неудачной попытки угадывания испытуемому не сообщается правильная буква, а он продолжает попытки предсказания вплоть до получения правильного результата. В протоколе каждый раз фиксируется число попыток, понадобившихся для того, чтобы определить, какая буква стоит на данном шаге текста. При проведении как полной, так и сокращенной программ угадывания особо выделяются достоверные продолжения. Достоверными продолжениями считаются такие стоящие на n-ом шаге текста буквы, появление которых полностью предопределено с точки зрения современных орфографических и стилистических норм литературного языка предшествующими $n-1$ буквами текста (ср. конечное *о* и пробел в русском слове *которого* Δ или буквы е, г, Δ в английском слове other Δ). Достоверными продолжениями мы будем также считать и такие буквы, появление которых на данном шаге текста имеет очень большую вероятность (т. е. при $p > 0.889$, об этом см. в 38).

Протокол угадывания по полной программе приводившегося выше отрывка показан в табл. 22. Здесь в строке «Цифровой текст»

Таблица 22

Исходный текст и сокращенный (цифровой) текст, полученный в результате проведения полной программы угадывания

Шаги текста (n)	1	2	3	4	5	6	7	8	9	10	11
Исходный текст	м	н	о	г	и	е	Δ	д	е	л	е
Цифровой текст	8	3	1	1	1	1	0	5	3	1	2
Шаги текста (n)	12	13	14	15	16	17	18	19	20	21	22
Исходный текст	г	а	т	ы	Δ	а	з	и	а	т	с
Цифровой текст	1	1	1	0	0	4	7	1	0	0	0
Шаги текста (n)	23	24	25	26	27	28	29	30	31	32	33
Исходный текст	к	и	х	Δ	и	Δ	а	ф	р	и	к
Цифровой текст	0	0	0	0	4	1	1	1	0	0	0
Шаги текста (n)	34	35	36	37	38	39	40	41	42	43	44
Исходный текст	а	н	с	к	и	х	Δ	с	т	р	а
Цифровой текст	0	0	0	0	0	0	0	1	0	0	0
Шаги текста (n)	45	46	47	48	49	50	51	52	53	54	55
Исходный текст	н	Δ	с	о	г	л	а	с	н	ы	Δ
Цифровой текст	0	0	4	1	1	0	0	1	1	1	0

указывается число попыток, понадобившихся для правильного отгадывания соответствующей буквы исходного текста.

Достоверные продолжения отмечаются с помощью цифры 0.

Прежде чем приступить к решению указанных задач, рассмотрим некоторые свойства угадывания и порождаемого им сокращенного текста.

34. Сокращенный текст

Если рассматривать сокращенный текст как закодированную форму буквенного текста, то весь процесс кодирования можно представить следующим образом.

Имеем текст, написанный на языке L, который использует алфавит, состоящий из букв (A, B, C. . . Z, пробел). Этот текст преобразуется в цифровой текст, который, в свою очередь, написан на языке L_1, использующем алфавит, символами которого являются натуральные числа 1, 2. . ., k. . ., S, а также число нуль. Символы этого алфавита, кроме нуля, упорядочены по убыванию их вероятностей. Это значит, что символ «1» появляется в позиции l_n i-того текста всякий раз, когда за цепочкой b_i^{n-1} следует такая буква (обозначим ее j_1), для которой $p\,(j_1/b_i^{n-1})$ является максимальной. Соответственно символ «2» ставится в позиции l_n тогда, когда за b_i^{n-1} идет буква (j_2), для которой $p\,(j_2/b_i^{n-1})$ является второй по величине и т. д. Достоверные продолжения (т. е. случаи, когда $p\,(j/b_i^{n-1}) \approx 1$, см. 36) отмечаются символом «0».

Что же касается сокращенной программы, то она дает текст, который записывается в алфавите, состоящем только из трех символов: 0 — достоверное продолжение, 1 — правильное угадывание, 2 — неправильное угадывание.

Цифровой текст, так же как и сокращенный текст первого типа, может рассматриваться в качестве закодированной формы исходного текста и может быть передан по каналу с двумя идеальными

предсказывающими устройствами. Первый предсказатель будет угадывать неизвестную букву в порядке убывания условных вероятностей букв, употребление которых допустимо на данном шаге. Предсказатель-двойник, получив цифру, соответствующую числу попыток, которые понадобились первому предсказателю, и используя распределение условных вероятностей букв, безошибочно восстановит букву исходного текста. Полная обратимость цифрового текста снова говорит об отсутствии потерь смысловой информации в нашей системе связи.

Преобразование исходного текста в сокращенный упрощает статистическую структуру текста. Это упрощение, связанное с ослаблением взаимозависимостей между символами, позволяет предположить, что вероятности символов цифрового текста (эти вероятности мы будем обозначать символом q) перестают зависеть или слабо зависят от i (т. е. от вида предшествующей буквенной цепочки b^{n-1}), а зависят от k и n. Поэтому энтропию сокращенного текста, полученного в ходе идеального угадывания, можно оценить с помощью одномерного распределения его символов (Циотровский, 1968, с. 24—26). С точки зрения объема счетной работы такая задача может быть легко решена.

При исследовании как отдельных слов, так и связных текстов обычно используются выборки в 100 цепочек (подробнее см. в 41).

Результаты угадывания связных текстов и отдельных слов обобщаются в матрицы. Часть таких матриц, суммирующих результаты угадывания 100 стобуквенных русских текстов по полной и сокращенной программам, приведены в табл. 23 и 24. Каждый столбец в таблицах указывает на определенный шаг текста (n). Номер строки k соответствует числу попыток при угадывании. Иными словами, номер строки соответствует символу цифрового текста (табл. 23). Для полной программы число строк равно числу букв алфавита исходного языка (для русского языка $S=32$, для английского $S=27$, для польского $S=33$ и т. п.) плюс одна строка для достоверных продолжений. Для сокращенной программы задаются три строки соответственно трем символам сокращенного текста (ср. выше).

Число, стоящее на пересечении n-го столбца и k-той строки, показывает условную вероятность того, что буква будет угадана с k-той попытки. Эта условная вероятность получена как частное от деления общего числа угадываний с k попыток на общее количество отрывков. В нижних строках табл. 23 указываются верхние с учетом достоверных продолжений оценки синтактической информации ($\bar{I}_n$), здесь же показаны нижние экспериментальные оценки $\underline{I}_n$, а также нижние оценки, полученные путем приближения к идеальному угадыванию I_n.

Данные по первому шагу текста не приводятся, поскольку его информация получена во всех выборках из распределения частот первых букв слова.

Сходным образом строится и табл. 24.

Таблица 23

Результаты угадывания ста русских текстов по полной программе

Относительные частоты предсказания букв с определенного числа попыток на каждом шаге текста

Число попыток (k)	Шаги текста (n)										
	2	3	4	5	6	7	8	9	10	...	100
1	0.43	0.41	0.46	0.36	0.40	0.42	0.42	0.42	0.39	...	0.40
2	0.13	0.12	0.17	0.17	0.18	0.18	0.15	0.15	0.16	...	0.06
3	0.07	0.03	0.05	0.08	0.07	0.08	0.07	0.07	0.05	...	0.03
4	0.07	0.07	0.06	0.06	0.05	0.02	0.02	0.02	0.03	...	0.01
5	0.07	0.05	0.04	0.03	0.03	0.04	0.05	0.05	0.01	...	0.01
6	0.05	0.05	0.05	0.04	0.06	0.03	0.01	0.01	0.02	...	—
7	0.03	0.03	—	0.02	0.01	0.02	0.01	0.01	—	...	0.01
8	0.04	0.02	0.01	0.03	0.02	—	0.01	0.02	0.02	...	—
9	—	0.04	—	—	0.03	0.04	0.01	0.01	0.02	...	0.01
10	0.04	0.03	—	—	—	—	—	0.01	0.02	...	0.01
11	—	0.01	0.02	0.01	—	—	0.02	—	—	...	0.01
12	0.03	—	0.03	0.01	0.01	0.01	0.01	—	0.01	...	—
13	0.02	0.01	0.01	—	—	0.01	—	0.01	—	...	—
14	—	0.01	0.01	0.01	—	—	—	0.01	—	...	—
15	0.01	—	0.01	—	0.01	—	—	—	—	...	—
16	—	0.01	—	0.02	0.01	—	0.01	0.01	—	...	—
17	—	0.01	—	—	0.01	—	0.01	0.01	—	...	—
18	0.01	0.01	—	—	—	—	—	0.01	—	...	—
19	—	0.01	—	0.01	—	—	—	0.01	—	...	—
20	—	0.01	0.02	—	0.01	0.01	0.01	—	—	...	0.01
21	—	0.02	0.01	—	0.01	—	0.01	—	—	...	—
22	—	0.01	—	—	—	—	0.01	—	—	...	—
23	—	0.01	—	—	—	—	—	—	—	...	0.01
24	—	—	—	—	—	—	0.01	0.01	0.01	...	—
25	—	—	—	—	—	—	—	0.01	—	...	—
26	—	—	—	—	—	—	—	—	—	...	0.01
27	—	—	—	—	—	—	—	—	—	...	—
28	—	—	—	—	—	—	—	—	—	...	—
29	—	—	—	—	—	—	—	—	—	...	—
30	—	0.01	—	—	—	—	—	—	—	...	—
31	—	—	—	—	—	—	—	—	—	...	—
32	—	—	—	—	—	—	—	—	—	...	—
0	—	0.02	0.05	0.15	0.10	0.14	0.15	0.18	0.26	...	0.42
Σ_{gk}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00

Оценки информации цифрового текста

I_n	2.85	3.16	2.45	2.28	2.54	2.12	2.22	2.02	1.65	...	1.02
I'_n	1.97	2.18	1.63	1.54	1.60	1.33	1.44	1.34	1.04	...	0.63
I_n	1.94	2.11	1.55	1.52	1.56	5.28	1.34	1.30	1.02	...	0.58

Таблица 24

Результаты угадывания ста русских текстов по сокращенной программе

Число попыток (k)	Эмпирические условия вероятности										
	Шаги текста (n)										
	2	3	4	5	6	7	8	9	10	...	100
0	—	0.02	0.05	0.15	0.10	0.14	0.15	0.18	0.26	...	0.42
1	0.43	0.41	0.46	0.36	0.40	0.42	0.42	0.42	0.39	...	0.40
2	0.57	0.57	0.49	0.49	0.50	0.44	0.43	0.40	0.35	...	0.18
Σq_k	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00

Верхние оценки синтактической информации сокращенного текста

I'_n (дв. ед.)	2.70	2.67	2.42	2.21	2.38	2.18	2.14	2.02	1.79	...	1.06

35. Идеальное предсказывание

Для того чтобы оценить информацию цифрового текста и соотнести ее с информацией исходного текста, необходимо рассмотреть закономерности идеального предсказывания букв на n-м шаге текста в некоторой репрезентативной выборке отрывков при условии, что $n-1$ предшествующих букв всех отрывков известны предсказателю.

Поскольку идеальное предсказывающее устройство угадывает буквы на каждом шаге текста и после каждой цепочки в порядке убывания их условных вероятностей, постольку вероятность появления угадываний с первой попытки на n-м шаге матрицы будет равна

$$q_1^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,1}/b_i^{n-1}).$$

Величину q_1^n можно рассматривать и как суммарную условную вероятность появления символа 1 языка на n-м шаге цифрового текста. Аналогичным образом задаются суммарные частоты символов 2, 3, . . ., k, . . ., S, обозначаемые как q_2^n, q_3^n, . . ., q_k^n, . . ., q_s^n. При этом j выбирается так, что дает второе, третье, . . ., k-тое, . . ., S-тое по величине значения p. В общем случае имеем

$$q_k^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}/b_i^{n-1}). \qquad (10)$$

Поскольку после каждой b_i^{n-1} буквы исходного алфавита заново упорядочиваются так, что

$$p(j_{i,1}/b_i^n) \geqslant p(j_{i,2}/b_i^n) \geqslant p(j_{i,s}/b_i^n), \qquad (11)$$

постольку каждому символу цифрового алфавита L_1 соответствует несколько букв алфавита L. В результате этого усреднения статистические связи между символами цифрового текста оказываются более слабыми, чем связи между буквами исходного текста.

При осуществлении идеального угадывания букв на следующем, $(n+1)$-м шаге текста получаем новое распределение вероятностей цифровых символов: q_1^{n+1}, q_2^{n+1}, . . ., q_k^{n+1}. Угадывание опирается здесь уже на более насыщенный тезаурус, чем это имело место на n-м шаге текста: действительно, цепочка в n символов дает больше знаний, чем цепочка в $n-1$ символов. Можно ожидать, что условные вероятности наиболее частых, а также достоверных продолжений будут возрастать, а условные вероятности редких продолжений будут уменьшаться. Таким образом, предсказание $(n+1)$-й буквы облегчается по сравнению с угадыванием n-й буквы. В связи с этим имеет место следующее неравенство:

$$\sum_{k=1}^{g} q_k^{n+1} \geqslant \sum_{k=1}^{g} q_k^n \quad g = 1, 2 \ldots \qquad (12)$$

Смысл этого неравенства состоит в том, что вероятность правильного предсказания буквы за первые g попыток при условии, что известны предшествующие n букв, больше или равна той же вероятности, когда известны $n-1$ букв для всех g (Пиотровский, 1968, с. 20—22).

Может случиться, что список событий в тезаурусе на $n+1$ шаге угадывания расширится по сравнению со списком, который был на n-м шаге. Тогда условные вероятности наиболее вероятных продолжений на $(n+1)$-м шаге текста оказываются меньше, чем соответствующие вероятности на n-м шаге, и неравенство (12) превращается в неравенство противоположного смысла. Эта ситуация, не предусмотренная Шенноном, характерна для естественного текста, имеющего квантовую («зернистую») структуру. Кванты текста, в первую очередь слова (отчасти морфемы и устойчивые словосочетания), дают значительное увеличение условных вероятностей наиболее вероятных продолжений на конечных буквах; этот рост особенно ощутим на пробелах. Поэтому последовательность величин:

$$\sum_{k=1}^{g} q_k^1, \quad \sum_{k=1}^{g} q_k^2, \quad \ldots, \quad \sum_{k=1}^{g} q_k^n, \quad \sum_{k=1}^{g} q_k^{n+1} \qquad (13)$$

в выборке реальных текстов не образует монотонно возрастающий ряд, а будет характеризоваться перепадами. Эти перепады будут обнаруживаться между теми шагами выборки, на которых стоит большое число пробелов между словами и следующим шагом, на который соответственно падает большее число начал слов.

Однако кванты текста (слова, а также морфемы и словосочетания) неравновероятны и обладают разной длиной. Поэтому при

достаточно больших выборках отрывков можно ожидать, что пробелы окажутся разнесенными по разным буквенным позициям и число перепадов будет сравнительно невелико. Вероятно, они будут ощутимы лишь в начале суммарного текста, примерно вплоть до пятнадцатой буквенной позиции. Таким образом, утверждение Шеннона о том, что неравенство (12) выполняется на всем протяжении цифрового текста, может быть использовано в качестве рабочей гипотезы при выведении формулы, описывающей распределение информации в усредненном тексте языка.

Если рассматривать идеальное предсказывание как преобразователь, переводящий буквенный текст в последовательность чисел от 1 до S, то нетрудно заметить, что входной сигнал n-го шага текста (определенная буква исходного текста) может быть мгновенно восстановлен путем анализа выходного цифрового символа и предшествующей цепочки из $n-1$ букв.

Представим себе таблицу, в которой по вертикали расположены все b_i^n, а по горизонтали даны цифровые символы от единицы до S. В каждой клетке, образованной пересечением столбца и строки, указана вероятность, которая одновременно соответствует и данному цифровому сигналу (т. е. выходному сигналу, выдаваемому входным сравнивающим устройством; см. рис. 28) и некоторой букве (входному сигналу) j_k, являющейся последним символом цепочки $b_i^n j_k$. При этом следует помнить, что вероятности расположены слева направо в порядке убывания их величин. Путем соотнесения через величину вероятности цифрового символа с цепочкой $b_i^n j_k$, получаем букву j_k исходного текста. Задача эта имеет неоднозначное решение лишь в том случае, если в строке появляется несколько одинаковых чисел (вероятностей). Тогда при декодировании цифрового текста возникает неопределенность в выборе между несколькими возможностями в восстановлении входной буквы.

Таблица 25

Значения $H_{\text{Ш}}$

Языки	$H_{\text{Ш}}$ дв. ед.
Русский	3.00
Польский	3.02
Английский	2.84
Немецкий	2.95
Французский	2.83
Румынский	2.69

36. Оценки значений синтактической информации

Исходя из приведенных выше рассуждений было показано, что при осуществлении полной программы угадывания, учитывающей вероятности достоверных продолжений q_0^n, истинное значение синтактической информации, которое несет лингвистическая единица, стоящая на n-м шаге текста, оценивается двойным неравенством:

$$\sum_{k=2}^{s} k\,(q_k^n - q_{k+1}^n) \log_2 k \leqslant I_n \leqslant (1 - q_0^n) \log_2 (1 - q_0^n) - \sum_{k=1}^{s} q_k^n \log_2 q_k^n, \qquad (14)$$

где левая часть неравенства представляет собой нижнюю границу ($\underline{H}_n = \underline{I}_n$), а правая часть указывает на верхнюю ($\overline{H}_n = \overline{I}_n$) границу интервала, в котором заключено истинное значение информации I_n (Shannon, 1951/1963, Пиотровский, 1968, с. 27—33).

Для сокращенной программы угадывания верхняя оценка I_n имеет вид:

$$I' = H_{III}(1 - q_0^n - q_1^n) + (1 - q_0^n)\log_2(1 - q_0^n) - q_1^n \log_2 q_1^n - \\ - (1 - q_0^n - q_1^n)\log_2(1 - q_0^n - q_1^n), \qquad (15)$$

где H_{III} — та неопределенность, которую несет третья буква текста при условии, что известны две предшествующие буквы (Пиотровский, 1968, с. 35, 40—41). Значения H_{III} для разных языков показаны в табл. 25.

Вывод равенств, а также математический анализ соотношений верхних и нижних оценок информации с истинным ее значением в тексте см.: Пиотровский, 1968, с. 27—41.

37. Угадывание по методу Колмогорова

Угадывание по методу, предложенному А. Н. Колмогоровым, происходит следующим образом. Испытуемому сообщается n—1 букв отрывка и предлагается последовательно угадывать n-ю, $(n+1)$-ю, . . . , $(n+\nu)$-ю буквы того же отрывка. При угадывании буквы, стоящей на соответствующем шаге текста, испытуемый должен дать один из следующих ответов:

а) назвать предлагаемую букву с большой степенью уверенности;

б) назвать предлагаемую букву с малой степенью уверенности;

в) назвать две возможные в данном участке отрывка буквы, каждая из которых обладает, по его мнению, одинаковой вероятностью;

г) назвать две возможные буквы, указав при этом, что одна из них более вероятна, чем другая;

д) отказаться от угадывания.

Называя букву с большой степенью уверенности, испытуемый записывает ее на пересечении столбца, соответствующего угадываемой букве, и первой строки. Если буква названа с малой степенью уверенности, она записывается во второй строке. Предложение о двух равновероятных буквах записывается в третьей строке, а предложение о неравновероятных буквах заносится в четвертую строку. Отказ от угадывания отмечается вопросительным знаком в первой строке. В нижней (пятой) строке записывается правильная буква, сообщаемая информанту после того, как он определит свое предсказание. Образец протокола показан в табл. 26.

Таблица 26

Образец протокола угадывания слова *ребячьи* по методу Колмогорова

Строки	Шаги текста						
	n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$
1			?	[я]	[ч]		и
2		а					
3	д/в						
4						[ь] л	
5	р	е	б	я	ч	ь	и

Примечание. Правильные предсказания отмечены □. Контрольному слову предшествует следующий контекст: *И вот, приехав в тот город, где друг мой в свое время справлял свадьбу, я встретил его жену и сына, который учится уже в пятом классе. И сын его меня расспрашивает, правда ли, что я служил с его отцом, на каких кораблях плавали, где бывали. Я отвечаю на эти как будто нехитрые на первый взгляд...* («Известия», № 261, от 4 ноября 1965 г.).

Результаты угадывания обрабатываются по формуле:

$$\bar{I}(\nu+1) = -\frac{\gamma_a}{\nu+1}[f_0 \log_2 f_0 + (1-f_0)\log_2(1-f_0)] - \\ - \frac{\gamma_б}{\nu+1}[f_1 \log_2 f_1 + (1-f_1)\log_2(1-f_1)] - \frac{\gamma_в}{\nu+1}[f_2 \log_2 f_2 + \\ + (1-f_2)\log_2(1-f_2)] + \frac{\gamma_б}{\nu+1} f_2 - \frac{\gamma_г}{\nu+1}[f_3 \log_2 f_3 + f_4 \log_2 f_4 + \\ + (1-f_3-f_4)\log_2(1-f_3-f_4)] - \frac{\Sigma a_{10} + \Sigma a_2 a_4 + \Sigma a_6 a_9}{\nu+1}, \qquad (16)$$

где $\bar{I}$ — средняя энтропия на букву на участке текста от n-й до $(n+\nu)$-той буквы (общее число угадываемых букв равно $\nu+1$); γ_x — число предсказаний того или иного типа, например γ_a — число предсказаний с большой степенью уверенности (правильных и неправильных), $\gamma_б$ — число предсказаний с малой степенью уверенности (правильных и неправильных), и т. д.

Символы f имеют следующие значения:

1) f_0 — вероятность не угадать в случае предсказания с большой степенью уверенности;

2) f_1 — вероятность не угадать в случае предсказания одной буквы с малой степенью уверенности;

3) $1-f_2$ — вероятность не угадать в случае, когда испытуемый предлагает две равновероятных буквы;

4) f_3 — вероятность угадать наиболее вероятную букву из двух предлагаемых испытуемым неравновероятных букв;

5) f_4 — вероятность угадать из двух неравновероятных букв маловероятную букву.

Значения f получены следующим образом. Для каждого из перечисленных выше случаев угадывания (1—5) возможны такие исходы:

а) a_1 — правильное предсказание, a_2 — неправильное предсказание;

б) a_3 — правильное предсказание, a_4 — неправильное предсказание;

в) a_5 — правильное предсказание, a_6 — неправильное предсказание;

г) a_7 — правильное предсказание для более вероятной буквы, a_8 — правильное предсказание для менее вероятной буквы, a_9 — неправильное предсказание;

д) a_{10} — отказ от предсказания.

Условимся, что число исходов каждого типа равно a_i, тогда

$$f_0=\frac{a_2}{a_1+a_2};\quad f_1=\frac{a_4}{a_3+a_4};\quad f_2=\frac{a_5}{a_5+a_6};\quad f_3=\frac{a_7}{a_7+a_8+a_9}.$$

Далее будем считать, что

$$f_4=\frac{a_8}{a_7+a_8+a_9};$$

$\sum a_{10}$ — сумма отказов, $\sum a_2a_4$ — сумма ошибок в случаях «а» и «б», $\sum a_8a_9$ — сумма ошибок в случаях «в» и «г» (Яглом и Яглом, 1973, с. 258—260).

Схема Колмогорова позволяет получить значения информации, приближающиеся к истинному среднему значению в блоке из $\nu+1$ букв для данного участка текста. С помощью схемы Колмогорова были получены оценки информации для русских и французских беллетристических текстов, а также для французских деловых текстов (Петрова*, 1968, с. 13—17, табл. 8). Эти оценки попадают в интервал между верхней и нижней оценками, полученными по методам, описанным в 36.

38. Информант и идеальное предсказание

Обрабатывая результаты индивидуального угадывания, мы получаем количественную оценку той информации текста, которую воспринял и соотнес со своим тезаурусом угадчик. Измерить количество информации, содержащееся в каждом индивидуальном тезаурусе, а тем более учесть его колебания невозможно. Однако можно попытаться фиксировать тезаурус и поведение угадчика, приблизив его к уровню идеального предсказания, которое осуществлялось бы распознающим устройством, опирающимся на глобальный лингвистический опыт коллектива (10).

Существует ряд лингвистических, психологических и математических приемов, с помощью которых можно приблизить индивидуальное угадывание к уровню идеального предсказания. Наиболее эффективными являются следующие приемы.

1. Подбор испытуемого, обладающего хорошим чутьем языка, высокой лингвистической культурой и знанием общей тематики текста.

2. Угадывание небольших объемов текста с целью предупредить притупление лингвистической реакции испытуемого.

3. Использование справочного и словарно-статистического аппарата. При угадывании букв незнакомого текста информант обычно пользуется:

а) специально составленными таблицами частотности начальных букв и двухбуквенных сочетаний, а также таблицами частотности двухбуквенных и трехбуквенных сочетаний независимо от их позиций в слове; таблицы частотности начальных букв используются при угадывании первой буквы, а остальные при распознавании второй и третьей букв;

б) энциклопедическими и двуязычными словарями, которые используются при угадывании четвертой и последующих букв, а также при определении достоверных продолжений. Применение статистических таблиц и словарей не только облегчает и ускоряет процесс угадывания, но в значительной степени нивелирует отклонения в результатах эксперимента у разных испытуемых.

Кроме того, применялись три вида вероятностно-лингвистической коррекции протокола.

Первый вид коррекции, осуществляемый экспериментатором с испытуемым, имеет целью устранить «лишние попытки» при угадывании отдельных букв. Смысл устранения лишних попыток можно проиллюстрировать на следующем примере. В тексте *он не литератор* (газ. «Известия» от 20 декабря 1960 г.) следующая за *т* буква была угадана со второй попытки, что и было указано в протоколе. Однако проверка по словарям показывает, что в данном случае имеет место система из двух выборов: *литерату* (*-ра*, *-рный* и т. п.) и *литерато* (*-р*, *-ра* и т. д.).

Отсюда ясно, что уже после первой попытки угадчик получает полную информацию о следующей за *т* букве: если не *у*, то *о*. Вторая попытка угадывания является здесь избыточной. В связи с этим в протоколе должно быть отмечено не две, а одна попытка.[1] Такие же поправки вводились в протокол во всех случаях, когда число сделанных испытуемым попыток превышало их предельное количество, необходимое для отгадывания данной буквы.

В тех же случаях, когда экспериментатор или испытуемый

[1] Аналогичные примеры коррекции протокола при угадывании английского, немецкого и французского текста (см.: Bogusławskaja, Korzeniec, Piotrowski, 1972, с. 10; Boguslawskaja, Koženec, Piotrowski, 1971, с. 461; Петрова*, 1968, с. 32).

сомневается в достоверности того или иного продолжения или степени его вероятности, вопрос может быть решен с помощью языковедов-экспертов.

Второй вид коррекции имеет целью проверить, насколько объективно выделил испытуемый достоверные продолжения. Эта коррекция должна также выявить те «нули информации», которые не были учтены в процессе эксперимента.

Выше уже указывалось (33), что к числу достоверных продолжений следует относить наряду с абсолютно достоверными и такие продолжения, которые обладают достаточно высокой вероятностью. Рассмотрим этот случай на конкретном примере.

Предположим, что испытуемый, определив все буквы цепочки *мать*, переходит к угадыванию следующей буквы. Будем считать, что вероятность (p) появления после этой цепочки пробела равна 0.999, а вероятность ($1-p$) появления буквы *е* (ср. *Матье*, *Матьез*) составляет всего 0.001.

В этом случае условная энтропия рассматриваемой буквенной позиции равна:

$$H = -p \log_2 p - (1-p) \log_2 (1-p) = -0.999 \log_2 0.999 - \\ - 0.001 \log_2 0.001 = 0.0114 \text{ дв. ед.}$$

Если информант и экспериментатор объявят пробел после цепочки *мать* единственным достоверным продолжением, энтропия которого равна нулю, они сделают ошибку в 0.0114 дв. ед. Наоборот, признав, что рассматриваемая буквенная позиция дает два продолжения, они припишут ей в условиях эксперимента энтропию, равную одной дв. ед. В этом случае ошибка возрастет до 0.9886 дв. ед. и предсказание оказывается неэффективным. Отсюда ясно, почему к числу достоверных продолжений, отмечаемых цифровым символом «нуль», должны быть отнесены не только абсолютно достоверные продолжения, но и такие продолжения, условная вероятность которых выше 0.889. В этом случае ошибка, возникающая при отнесении рассматриваемых буквенных позиций к числу достоверных продолжений, будет меньше ошибки, появляющейся в том случае, если считать эти продолжения недостоверными (ср.: Bogusławskaja, Korzeniec, Piotrowski, 1972, с. 12; Boguslavskaja, Koženec, Piotrowski, 1971, с. 463).

Само собой разумеется, что в большинстве случаев степень достоверности продолжения может быть определена только исходя из лингвистической интуиции информанта, экспериментатора или экспертов.

Третий вид коррекции предусматривает расположение экспериментальных значений q_k^n в порядке убывания q и возрастания k, с тем чтобы приблизить опытное значение I_n к той оценке нижней границы информации, которая могла бы быть получена в результате идеального предсказания (35). Впрочем, как это видно из

табл. 23, указанный вид корреляции дает незначительное уменьшение величины I_n. О других приемах коррекции см.: Башарин, 1959, с. 361—364.

39. Коллективное угадывание

Коллективное угадывание осуществляется по следующей схеме. Достаточно большой группе испытуемых — носителям языка — предлагается предсказать букву, находящуюся на n-м шаге связного текста или слова, причем $n-1$ букв этого отрывка или слова испытуемым известны.

Каждый из испытуемых независимо от других записывает наиболее вероятное, с его точки зрения, буквенное продолжение для n-го шага текста. После этого экспериментатор сообщает действительную букву и испытуемые переходят к следующему угадыванию.

Коллектив испытуемых рассматривается как предсказывающее устройство, осуществляющее глобальное распознавание. Лингвистический тезаурус этого устройства приближается к идеальному тезаурусу, содержащему $H_{\text{идеал.}}$ информации. Поэтому на каждом шаге текста наибольшее число продолжений должно падать на наиболее вероятную букву, следующее количество продолжений даст вторая по вероятности буква и т. д. Частное от деления числа испытуемых, предложивших данную букву, на общее число испытуемых можно рассматривать как условную вероятность p_k появления некоторой буквы k в данном участке текста или слова. Само собой разумеется, что коллектив испытуемых лишь условно может рассматриваться в качестве идеального предсказывающего устройства. В действительности реальные предсказания коллектива обычно довольно далеки от идеального угадывания. В частности, испытуемые обычно не называют редкие продолжения, в связи с чем оценки истинных вероятностей через относительные частоты будут смещены. Чтобы избежать этого, необходимо приблизить результаты угадывания к идеальному предсказанию. Здесь можно предложить два приема. Во-первых, перед угадыванием буквы на n-м шаге текста испытуемым сообщаются все возможные для этого шага буквенные продолжения (без указания их вероятностей). Во-вторых, не названные испытуемыми продолжения после завершения эксперимента вносятся в протокол и им приписываются сколь угодно малые вероятности (весь спектр вероятностей соответственным образом пересчитывается).

Протокол коллективного угадывания с только что указанной коррекцией показан в табл. 27.

Имея распределение условных вероятностей, можно получить из выражения

$$I'_n = -\sum_{k=1}^{s} p_k \log_2 p_k \qquad (17), \text{ ср. } (5)$$

Таблица 27

Образец протокола и расчетная таблица протокола коллективного угадывания внеконтекстного русского слова *знаменитый*

Алфавит	З	Н	А	М	Е	Н	И	Т	Ы	Й	А
а		13 0.220	52 0.961				6 0.109		2 0.036		
б											
в		2 0.035									
г		0.010									
д		10 0.170									
е		15 0.255		2 0.036	13 0.481			5 0.092	0.010	12 0.215	
ж											
з											
и		2 0.035				40 0.726					
й										42 0.755	
к				4 0.072							
л		0.010		0.010							
м		0.010		24 0.441						0.010	
н		10 0.170		16 0.295		54 1.000	0.010				
о		2 0.035	2 0.039		2 0.037		6 0.109		14 0.254		
п											
р		0.010									
с				6 0.110				49 0.908			
т											
у		0.010							0.010		
ф										0.010	
х											
ц											
ч											
ш							0.010				
щ											
ы		0.010							36 0.654		
ъ, ь											
э				2							
ю		0.010		0.036							
я		0.010			13 0.482					0.010	
Δ							2 0.036		2 0.036		54 1.000

Таблица 27 (*продолжение*)

Алфавит	З	Н	А	М	Е	Н	И	Т	Ы	Й	Δ
Двоичных единиц информации											
	4.220	2.891	0.238	2.075	1.191	0.000	1.338	0.433	1.382	0.983	0.000
	I'_1	I'_2	I'_3	I'_4	I'_5	I'_6	I'_7	I'_8	I'_9	I'_{10}	I'_{11}

Примечание. Целое число указывает количество испытуемых, предложивших данное продолжение (всего в эксперименте участвовало 54 испытуемых). Ниже дана эмпирическая вероятность данного продолжения. Для тех продолжений, которые не были предложены информантами, дается их приблизительно взятая вероятность, равная 0.01. Синтактическая информация первой буквы получена из расчета статистического спектра первых букв слов, взятых вне контекста.

условную информацию, оценивающую среднюю синтактическую информацию I'_n рассматриваемой буквенной позиции. Возможна и более общая задача. Можно, например, попытаться определить среднюю условную информацию буквенных позиций в некоторой модели текста либо в слове определенной длины или морфемно-слогового строения. В этом случае коллективному угадыванию последовательно подвергается ряд отрывков или слов, составляющих достаточно репрезентативную выборку. Расчет средней условной энтропии осуществляется по формуле:

$$I''_n = -\sum_{d_i^{n-1}} p(d_i^{n-1}) \sum_{k=1}^{s} p(j_{i,k}/d_i^{n-1}) \log_2 p(j_{i,k}/d_i^{n-1}) \quad (18), \text{ ср.} \quad (6)$$

Если угадываются отдельно взятые слова, то вероятность цепочки d_i^{n-1} может быть получена с помощью данных частотного словаря. В этом случае I''_n близко к истинному значению I_n.

Если же эксперимент проводится на выборке связанных текстов и вероятность $p(d_i^{n-1})$ определить невозможно, то приходится считать все цепочки d_i^{n-1} равновероятными. Расчет средней условной энтропии осуществляется здесь по формуле (17).

40. Угадывание букв текста на основе иного языкового кода или неполного владения тем языком, на котором написан текст

До сих пор угадывание букв текста осуществлялось при условии, что информант, являясь носителем того языка, на котором написан текст, декодирует этот текст наилучшим образом. Однако ничто нам не мешает поставить такой эксперимент, в котором угадчик либо использует другой языковый код, либо опирается на неполное или недостаточное владение языком, на котором написан текст.

В первом случае речь идет об угадывании текста таким информантом, который не знает языка текста и угадывает буквы текста,

исходя из кода другого языка. Этот язык может быть близкородственным, далекородственным или вообще неродственным языку, на котором написан контрольный текст. Совершенно очевидно, что угадывание в этом случае будет давать иные результаты, чем угадывание на родном языке. Больше всего отклонений даст опыт с носителем неродственного языка. Несколько меньше расхождений дает опыт с носителем далекородственного языка, еще большее схождение покажет, очевидно, опыт с носителем близкородственного языка.

Во втором случае речь идет об угадывании текста лицом, недостаточно владеющим иностранным языком. Чем хуже владеет угадчик языком, на котором написан контрольный текст, тем заметнее будут отличаться результаты угадывания от того эталона, который показал на том же тексте идеальный угадчик.

И в том и в другом случае испытуемый, располагая в своем тезаурусе $И_{\theta\,\text{неполн.}}$ априорной информации, все же способен извлечь из сообщения часть содержащейся в нем информации (см. 30 и рис. 27).

Попытаемся построить процедуру количественной оценки этой информации.

Будем считать, что идеальный угадчик (индивидуальный или коллективный) показывает на k-том шаге текста истинное распределение вероятностей появления 1-й, 2-й, 3-й, . . . , S-той букв алфавита:

$$p_1,\ p_2,\ p_3,\ \ldots\ p_s.$$

Имеется угадчик А, либо недостаточно владеющий языком, либо вообще не знающий данного языка и декодирующий текст с помощью другого языка. Этот угадчик дает следующие относительные частоты 1-й, 2-й, 3-й, . . . , S-той букв:

$$f_1,\ f_2,\ f_3,\ \ldots\ f_s.$$

Расхождение между образцовым знанием языка и его владением у угадчика A относительно k-того шага оценивается величиной:

$$D_k(A) = \sum_{i=1}^{s} |\, p_i - f_i \,|. \tag{19}$$

Если все $f_i = p_i$, то $D_k(A)_{\min} = 0$. В этом случае субъективное знание языка у угадчика A ничем не отличается от образцового владения языком. Максимума величина $D_k(A)$ достигает тогда, когда угадчик A, не зная языка текста и опираясь при угадывании на систему другого языка, называет только такие продолжения, вероятности которых (p_j) для k-того шага текста равны нулю, но не называет те буквы, вероятность которых (p_i) здесь

будет больше нуля. Тогда величина максимального расхождения будет равна

$$D_k(A)_{\text{макс.}} = \sum |\, p_i - f_i \,| + \sum_j |\, p_j - f_j \,|, \tag{20}$$

поскольку $\sum p_i = 1$ и $\sum f_j = 1$, а все f_i и p_j равны нулю, то $D_k(A)_{\text{макс.}} = 2$.

Рассмотрим также случай, когда на k-том шаге имеет место достоверное продолжение (т. е. $p_1 = 1$, а вероятность остальных букв равна нулю), в то время как угадчик A находится в состоянии полной неопределенности. В этом случае априорная вероятность каждой буквы для информанта A равна $\frac{1}{S}$ (к этой величине будут приближаться полученные в ходе эксперимента относительные частоты f). При этом

$$D_k(A) = \left(1 - \frac{1}{S}\right) + (S-1)\frac{1}{S} = 2 - \frac{2}{S}. \tag{21}$$

При полном незнании языка угадчику A ничего не известно о количестве букв в алфавите, поэтому здесь $S \to \infty$. В этих условиях имеем

$$\lim_{s\to\infty} D_k(A) = D_k(A)_{\text{макс.}} = 2. \tag{22}$$

Отношение равенств (21) и (22) показывает долю отклонения реального угадывания у информанта от идеального угадывания. Эта доля равна

$$\frac{D_k(A)}{D_k(A)_{\text{макс.}}} = \frac{\sum_{i=1}^{s} |\, p_i - f_i \,|}{2}. \tag{23}$$

Разность же

$$\delta = \left(1 - \frac{\sum_{i=1}^{s} |\, p_i - f_i \,|}{2}\right) 100\% \tag{24}$$

показывает долю близости идеального угадывания и угадывания, осуществленного информантом A.

Если угадчик A недостаточно владеет языком, то количество извлекаемой им из участка k синтактической информации $I_k(A)$ будет оцениваться произведением

$$I_k(A) = I_k\left(1 - \frac{\sum_{i=1}^{s} |\, p_i - f_i \,|}{2}\right) = I_k - I_k \frac{\sum_{i=1}^{s} |\, p_i - f_i \,|}{2}, \tag{25}$$

где дробный член указывает на то количество синтактической информации, которое было потеряно испытуемым из-за недостаточного знания языка.[1] Другие приемы оценки информации, извлекаемой информантом из текста в зависимости от его владения языком, см.: Garner, Hake, 1951, с. 446—459; Невельский, Розенбаум, 1971, с. 134—148.

41. Подбор экспериментального материала и оценка достоверности результатов

Чтобы получить усредненные вероятностные и информационные характеристики сокращенного текста, необходимо провести эксперимент по угадыванию на достаточно репрезентативной выборке текстов. Наблюдения показывают, что уже выборка в 75 текстах дает сравнительно небольшой разброс, а дальнейшее добавление числа исследуемых отрывков или отдельных слов незначительно сглаживает статистические колебания. Поэтому при исследовании как отдельных слов, так и связных текстов обычно используются выборки, содержащие не менее 100 цепочек[2] по 100—200 букв каждая.

При исследовании энтропии текста в русском, польском, азербайджанском, английском, немецком, а также румынском и молдавском языках использовалось 100—120 связных отрывков длиной от ста до двухсот букв. В каждом тексте угадыванию подвергаются буквы от второй до сотой включительно. Затем угадываются также 150-я и 200-я буквы (отрывки от 101-й до 149-й и от 151-й до 199-й букв предварительно сообщаются информанту). Экспериментально тексты обычно представляют собой просто распространенные или сложные предложения средней длины.

Кроме того, для каждого из языков угадывается и обрабатывается от пятисот до шестисот отдельных слов, взятых вне контекста.

Для языкознания интерес представляют не столько абсолютные величины синтактической информации, сколько относительное распределение этих величин в различных участках текста.

[1] В русле описанной методики был проведен эксперимент немецкого научно-технического текста студентами-немцами и русскими студентами I—IV курсов одного из технических вузов Ленинграда (Целиковская, 1969, с. 336). Расчет полученных данных показал, что пассивное владение немецким языком (научно-технический стиль) у испытуемых русских колеблется в пределах 45—65%.

[2] Исключения составляют длинные слова (от 9—10 букв и выше). Поскольку их число в каждом языке обычно не очень велико и они имеют сходную морфологическую структуру, мы ограничивались выборками, состоящими из пятидесяти слов. Кстати говоря, нижний предел объема выборки для получения достоверных информационных данных относительно английского языка равен пятидесяти текстам (Богуславская, Новак, 1968, с. 21).

Абсолютные величины информации и избыточности могут быть оценены путем применения полной программы угадывания. Распределение этих величин может быть показано с помощью сокращенной программы. Обычно по полной программе угадываются 5-я, 10-я, 20-я, 30-я, 40-я, 100-я, 150-я и 200-я буквы, остальные буквы предсказываются по сокращенной схеме.

Комбинаторное использование обеих программ значительно сокращает объем эксперимента и дает два взаимоконтролируемых ряда числовых данных.

В большинстве математических работ, посвященных информационным измерениям языка, исследуются тексты, извлеченные из одного и того же произведения. Например, К. Шеннон использовал отрывки из книги «Виргинец Джефферсон» Дюма Малона, ученики А. Н. Колмогорова проводили угадывание по повести «Детские годы Багрова-внука» С. Т. Аксакова. Подбирая однородные с точки зрения жанра и индивидуального стиля тексты, авторы этих исследований стремились сократить разброс выходных данных. Однако при перенесении получаемых результатов на язык в целом возникают серьезные затруднения: дело в том, что, как показали исследования У. Пейсли (Paisley, 1966, с. 28—34), на значения синтактической информации значительное влияние оказывают авторская манера, тема повествования и даже время написания произведения. Поэтому выбор экспериментального материала осуществляется из разных источников с соблюдением следующих жанрово-стилистических пропорций:

1) устно-разговорная речь	— 30%
2) беллетристика	— 30
3) публицистика и научно-деловая речь	— 30
4) поэзия	— 10
Всего	— 100%

При определении информационных характеристик языка в целом из общего числа исследованных отрывков отбирается 100 текстов указанной длины, которые распределялись в только что названных соотношениях.

Информационные характеристики по каждой из первых трех жанрово-статистических разновидностей вычисляются путем обработки 33 текстов, принадлежащих данной разновидности. Поэтические тексты в особую выборку не выделяются, поскольку их информационные характеристики не используются при автоматической обработке текста.

Оценка достоверности данных, получаемых в ходе информационных измерений текста, является важным методическим вопросом лингвистики текста (Altmann, 1974, с. 125). Достоверность результатов угадывания может быть оценена путем их сравнения

с соответствующими характеристиками, получаемыми на аналогичном языковом материале с помощью других приемов угадывания (ср. табл. 26—27). Кроме того, можно сравнить результаты угадывания с аналогичными информационными оценками, которые получены либо из расчета вероятностных спектров слов, слогов, букв (см. 18), либо путем исследования дальних корреляционных связей в тексте (Пиотровский, 1968, с. 67—69; Урбах, 1963). Наконец, делаются попытки применить для этих целей нормированный критерий χ^2, а также критерий знаков (Богуславская, Новак, 1968, с. 21).

Глава 6

ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ ТЕКСТА

42. Осмысленная информация сообщения и ее синтактико-информационные оценки

Прежде чем говорить о распределении синтактической и осмысленной информации в тексте, необходимо определить, что представляет собой та осмысленная информация, которую наш информант получает в ходе угадывания текста.

Считается, что индивидуум находится в целеустремленном состоянии, если выполняются два условия:

а) существует, по крайней мере, один исход O_1 его поведения, который имеет для индивидуума M определенную значимость в данной ситуации;

б) существует, по крайней мере, два поведения C_1 и C_2, которые приводят к результату O_1 с различной эффективностью.

В общем виде целеустремленное состояние определяется:

а) индивидуумом M;

б) альтернативными поведениями (последовательностями действий) $C_1, C_2, C_3, \ldots, C_i, \ldots, C_s$;

в) возможными результатами поведения.

К переменным относятся:

а) p_i — вероятность того, что M выберет путь (поведение) C_i;

б) E_{ij} — вероятность того, что путь C_i приведет к результату O_j (E_{ij} измеряет здесь эффективность поведения C_i по отношению к результату O_j);

в) V_j — относительная значимость результата O_j для M.

Вероятность каждого результата равна

$$P(O_j) = p_i E_{ij}. \tag{1}$$

Умножая эту вероятность на относительную значимость результата O_j, а затем суммируя все аналогичные произведения

по i и j, получаем математическое ожидание целеустремленного состояния нашего индивидуума, которое равно

$$V(S/M) = \sum_i^s \sum_j^m p_i E_{ij} V_j. \tag{2}$$

Имеется два «собеседника» — M_1 и M_2. Связь между ними в смысле передачи сообщения имеет место тогда, когда сообщение, передаваемое M_2, меняет переменные p_i, E_{ij}, V_j целеустремленного состояния собеседника M_1, и наоборот. Если сообщение воздействует на вероятность p_i выбора поведения M_1, то считается, что оно информирует. Если сообщение меняет E_{ij}, то оно инструктирует собеседника M_1. Если же сообщение влияет на V_j, оно мотивирует. Изменение величины $V(S/M)$ отражает изменение ценности сообщения.

Количество осмысленной информации, содержащейся в переданном сообщении, будет равно

$$И_c = ИS_2/M_1 - ИS_1/M_1 \tag{3}$$

(S_1 — состояние M_1 до приема сообщения, а S_2 — после приема сообщения); количественная оценка инструкции, содержащейся в сообщении, будет равна

$$B_c = B(S_2/M_1) - B(S_1/M_1), \tag{4}$$

а степень мотивировки в переданном сообщении может быть оценена разностью

$$C_c = C(S_2/M_1) - C(S_1/M_1). \tag{5}$$

Наконец, ценность сообщения может быть определена как разность целеустремленных состояний после и до передачи сообщения:

$$V_c = V(S_2/M_1) - V(S_1/M_1) = \sum_i^s \sum_j^m (p_i + \Delta p_i)(E_{ij} + \Delta E_{ij})(V_j + \Delta V_j) - \sum_i^s \sum_j^m p_i E_{ij} V_j \tag{6}$$

(подробнее об этом см.: Ackoff, Emery, 1972/1974, с. 43—69).

Теперь рассмотрим эту схему применительно к эксперименту по угадыванию. В качестве собеседника M_1' здесь выступает наш идеальный угадчик (индивидуальный или коллективный). На каждом шаге угадываемого текста имеет место ряд альтернативных поведений угадчика, представляющий собой альтернативный выбор букв $C_1, C_2, \ldots, C_i, \ldots, C_s$. Вероятности этих альтернативных выборов соответственно равны $p_1, p_2, \ldots, p_i, \ldots, p_s$. Вероятности меняются в зависимости от того, какую смысловую информацию извлек угадчик из предшествующего текста.

Так, например, если угадчик ничего не знает о содержании, заключенном в цепочке b^{n-1}, то, приступая к угадыванию лингвистического элемента, стоящего на n-м шаге, он имеет дело с распределением вероятностей $p_1, p_2, \ldots, p_i, \ldots, p_s$, дающих неопределенность H_n. Однако, если содержание цепочки в b^{n-1} ему известно, то угадчик будет исходить из вероятностей $p'_1, p'_2, \ldots, p'_i, \ldots, p'_s$, совокупность которых будет характеризоваться энтропией H'_n. Разность

$$H_n - H'_n = И(b^{n-1}) \qquad (7)$$

и будет количественно оценивать всю или часть (см. 52) осмысленной информации $И$, содержащейся в сегменте текста b^{n-1}.

Обратимся теперь к другим аспектам обмена информации.

Поскольку эксперимент по угадыванию предусматривает всего лишь один результат O_n, представляющий собой правильное декодирование угадываемого текста, становится очевидным, что угадчик не способен менять вероятности исходов эксперимента. Поэтому исход опыта является неконтролируемым со стороны M'_1. Отсюда вытекает, что инструктивный аспект, а вместе с ним и его количественные оценки нашим экспериментом не предусмотрены.

Сложнее обстоит дело с относительной значимостью V. Дело в том, что каждое предлагаемое угадчиком продолжение, верное или неверное, обладает для него определенной значимостью с точки зрения его жизненного опыта. Значимость V присуща как уже полученному результату (содержание угаданного сегмента b^{n-1}), так и конечному результату (содержание всего текста). К сожалению, объективных приемов, которые помогли бы измерить или оценить величину V, у нас нет. Поэтому приходится отказаться от измерения мотивировки и ценности сообщения.[1]

Короче говоря, эксперимент по угадыванию дает возможность оценить осмысленную информацию, содержащуюся в сообщении. Еще раз напоминаем, что эта осмысленная информация количественно оценивается не шенноновской мерой информации этого же сообщения, но разностью значений энтропии, падающих на следующий за данным сообщением участок текста (одно значение получено при условии, что содержание сообщения неизвестно, а второе — что это содержание уже известно угадчику).

Теперь посмотрим, что представляет собой оцениваемая таким образом информация. Для этого обратимся к условиям подбора текстового материала и выбора угадчика.

Для угадывания используются тексты, содержащие истинные и осмысленные предложения. Поэтому информация $И$ имеет семантическую природу. Наш угадчик не только идеально владеет системой и нормой языка, но является по условию знатоком того

[1] К вопросу о соотношении ценности, мотивировки, инструктивности и смысла в сообщении мы вернемся в 50—52.

специального предмета, который трактуется в угадываемых текстах. Поэтому извлекаемая им из текста информация оптимальным образом используется для угадывания следующих участков текста (иными словами, в смысловую информацию каким-то образом включаются некоторые оценки V). Отсюда следует, что информации $И$ присущ и прагматический аспект. Таким образом, оцениваемая с помощью величины I осмысленная информация является семантико-прагматической и сигматической информацией. Мы будем называть ее **смысловой информацией**.

Как уже говорилось (30), количество информации в тезаурусе и тексте не остается неизменным.

Начиная угадывать текст, образцовый информант оперирует тезаурусом, содержащим $И_{тезаур.}$ смысловой информации. При этом неопределенность начального участка (в нашем случае начальной буквы) для нашего угадчика равна H_1. Узнав первую букву текста, угадчик увеличивает количество информации в тезаурусе, переводя его в состояние $И_1$, угадывание второго, третьего и т. д. участков текста еще более насыщает тезаурус [гл. 5, (9)].

По мере накопления тезаурусом сведений о тексте неопределенность при угадывании букв последовательно убывает так, что

$$H_1 > H_2 > \ldots > H_{n-1} > H_n > \ldots > H_\infty. \qquad (8)$$

Помня, что $H=I$, выражение (8) можно переписать в виде:

$$I_1 > I_2 > \ldots > I_{n-1} > I_n > \ldots > I_\infty. \qquad (9)^1$$

Последний член неравенства (9) количественно равен той информации, которую извлекает угадчик при отгадывании участка (буквы), сколь угодно далеко отстоящего от начала текста. Величину I_∞ мы будем называть предельной синтактической **информацией связного текста**.

Предельная информация в идеальной схеме текста всегда будет больше нуля. Это и понятно. Всякий текст, будучи образован из сложных знаков (слов, словосочетаний, предложений), обладающих практически неограниченной комбинаторной способностью, имеет несколько продолжений или, иначе говоря, всегда обладает неопределенностью выбора. Даже в тех случаях, когда данный шаг конкретного текста предусматривает единственно возможное продолжение, всегда найдутся последующие шаги, которые дадут несколько возможных продолжений. Если же рассматривать некоторую совокупность текстов, на основе кото-

[1] Разумеется, неравенства (8) и (9) следует рассматривать как идеальную схему, усредняющую большое число реальных текстов, каждый из которых, имея квантовую («зернистую») информационную структуру, характеризуется перепадами информации, т. е. участками, где $I_{k-1} < I_k$ (см. 35). Эти перепады обнаруживаются и в полученных нами в результате угадывания обобщенных схемах (см. рис. 29).

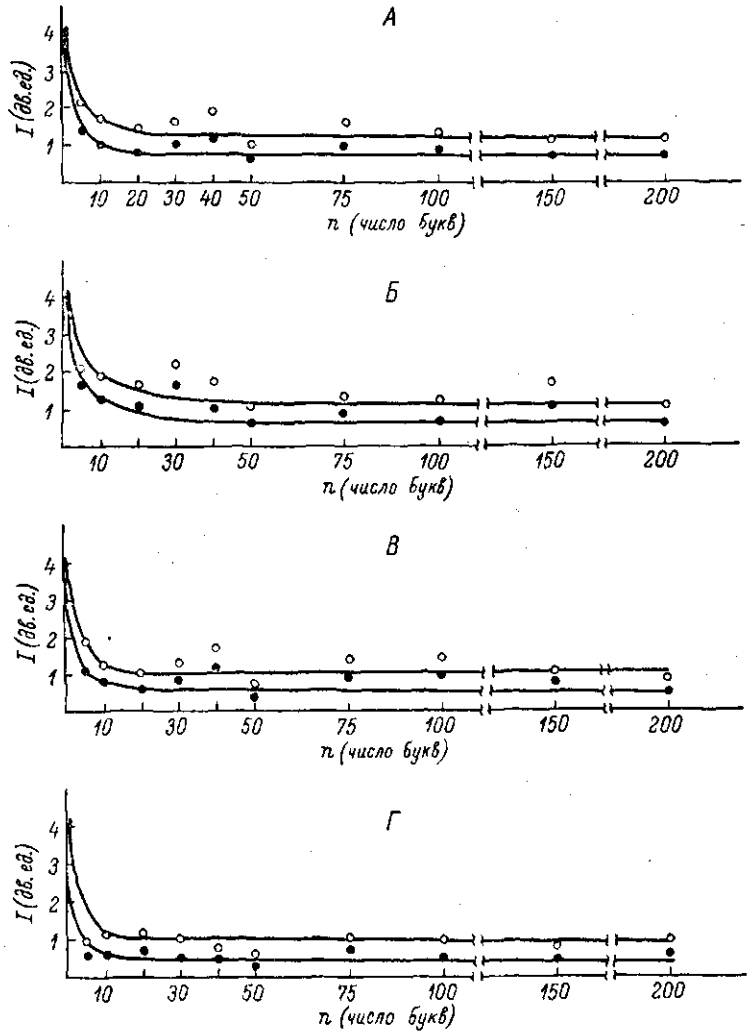


Рис. 29. Распределение информации в английских текстах (ср.: Богуславская, Новак, 1968, с. 14—16).

А — язык в целом; *Б* — разговорная речь; *В* — беллетристические тексты; *Г* — публицистика.

рой строится усредненная схема, то наряду с текстами, дающими на данном шаге достоверное продолжение, всегда обнаружатся другие тексты, которые дадут на этом шаге несколько продолжений.

Величину I_∞ можно рассматривать как суммарную оценку статистической информации, получаемой идеальным носителем для данного языка или его разновидности от одной буквы связ-

ного текста, — буквы, находящейся достаточно далеко от начала текста (в этом случае идеальный носитель уже учитывает контекстные связи сообщения).

43. Информация, избыточность и контекстная связанность в индоевропейских и тюркских языках

Наш эксперимент построен на индивидуальном угадывании, поэтому мы получаем две оценки предельной информации $\bar{I}_\infty$ и $\underline{I}_\infty$ (см. табл. 28). Эти значения по русскому языку получены как среднее арифметическое соответствующих значений для 30-й, 31-й, 32-й, . . ., 100-й букв; для польского, английского, немецкого и казахского языков — как среднее арифметическое для 30-й, 40-й, 50-й, 75-й, 100-й, 150-й, 200-й букв; для французского и румынского — как среднее арифметическое для 30-й, 40-й, 50-й, 75-й, 100-й букв и, наконец, для азербайджанского — как среднее арифметическое для 30-й, 40-й, 50-й, 60-й, 70-й, 80-й, 90-й, 100-й, 150-й, 200-й букв. Кроме того, имеются оценки I_∞, полученные в результате проведения коллективного угадывания болгарских и испанских текстов носителями языка (табл. 29). Значения I_∞ рассчитаны по формуле (17) из гл. 5 на интервале 30÷100 буква текста.

Такие интервалы выбраны исходя из тех соображений, что начиная с тридцатой буквенной позиции значения I_n практически перестают зависеть от n. Использование буквенных позиций, лежащих ближе к началу текста, дает явное завышение значений I_∞, $\bar{I}_\infty$ и $\underline{I}_\infty$. Дело в том, что по мере развертывания текста количество извлекаемой из него по мере угадывания информации последовательно убывает (ср. рис. 29).

Убывание информации I по ходу угадывания текста легко объяснимо. Действительно, если бы в тезаурус угадчика был бы заложен один лишь перечень букв, составляющих алфавит данного языка, а ограничения на сочетаемость этих букв оставались бы на всем протяжении эксперимента неизвестными угадчику, то угадывание каждой буквы текста, независимо от того, на каком отдалении от его начала находится эта буква, давало бы I_0 статистической информации, равную величине энтропии алфавита H_0.

Тот факт, что значения I_n заметно меньше величины $I_0=H_0$ и постепенно убывают по ходу чтения текста, объясняется тем, что образцовый информант производит угадывание, учитывая все комбинаторно-статистические связи и ограничения, налагающиеся на употребления конкретных букв в данном участке текста. Эти ограничения можно оценить относительно n-й буквенной позиции текста с помощью разности

$$K_n = I_0 - I_n \text{ дв. ед.}, \qquad (10)$$

Таблица 28

Информация (I, $\underline{I}$) и избыточность (R, $\bar{R}$) в индоевропейских и тюркских языках

Разновидности языка	Русский				Польский				Английский				Немецкий			
	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$
Разговорная речь	1.40	0.83	72.0	83.4	1.18	0.69	76.3	86.3	1.47	0.90	69.4	81.2	1.24	0.74	73.9	84.4
Беллетристика	1.19	0.70	76.3	86.0	1.29	0.83	74.5	83.6	1.10	0.65	77.1	86.5	1.36	0.83	71.4	82.5
Научно-технические и публицистические тексты	0.83	0.49	83.4	90.1	0.83	0.53	83.6	89.5	0.82	0.37	82.9	92.1	0.97	0.56	79.6	88.2
Язык в целом	1.37	0.82	72.1	83.6	1.28	0.76	74.7	85.0	1.35	0.74	71.9	84.5	1.36	0.71	71.4	85.1

Таблица 28 (*продолжение*)

Разновидности языка	Французский				Румынский				Казахский				Азербайджанский			
	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	$\bar{I}$	$\underline{I}$	$\underline{R}$	$\bar{R}$	I_∞	I_∞	$\underline{R}$	$\bar{R}$
Разговорная речь	1.32	0.81	68.5	82.9	1.24	0.71	74.2	85.4	1.27	0.81	76.5	85.0	—	—	—	—
Беллетристика	1.36	0.78	71.0	83.6	1.26	0.78	73.8	83.8	1.35	0.81	75.0	85.0	—	—	—	—
Научно-технические и публицистические тексты	0.77	0.45	83.9	90.4	1.23	0.68	74.4	85.7	1.16	0.65	78.5	88.0	—	—	—	—
Язык в целом	1.38	0.79	70.6	83.4	1.34	0.72	72.1	85.0	1.52	0.81	71.9	85.0	1.17	1.07	65.2	79.0

Примечание. Оценки энтропии и избыточности заимствованы из следующих работ: Пиотровский, 1968, с. 58; Petrova, Piotrowski, Giraud, 1964; Piotrowski, 1969; Bogusławskaja, Korzeniec, Piotrowski, 1972, с. 13; Байтанаева, Бектаев, 1973, с. 691.

Таблица 29

Оценки I_∞ по результатам коллективного угадывания

Разновидности языка	Болгарский язык (Георгиев*, 1973, с. 7)		Испанский язык (по данным В. Я. Шабеса)	
	I_∞	R_∞	I_∞	R_∞
Разговорная речь	1.00	79.8	—	—
Беллетристика	0.88	82.2	—	—
Научно-деловая речь	0.65	86.8	—	—
Язык в целом	0.91	81.6	1.05	77.8

которую мы назовем контекстной связанностью текста (Ingarden, Urbanik, 1962, с. 135; Пиотровский, 1968, с. 64—65; Nauta, 1972, с. 260). Пределом ее будет разность

$$K_\infty = I_0 - I_\infty \text{ дв. ед.}, \qquad (11)$$

которую мы будем называть предельной контекстной связанностью текста. Численные ее значения относительно разных языков и стилей даны в табл. 30. Величины K_n и K_∞ измеряют синтактическую информацию, лингвисту же

Таблица 30

Предельная контекстная связанность по оценкам I для семи языков

Разновидности языка	Русский	Польский	Английский	Немецкий	Французский	Румынский	Казахский
Разговорная речь	3.60	3.64	3.31	3.51	3.39	3.51	4.13
Беллетристика	3.31	3.85	3.68	3.29	3.38	3.49	4.05
Деловые тексты	4.17	4.21	3.96	3.78	3.92	3.52	4.24
Язык в целом	3.63	3.67	3.43	3.28	3.36	3.41	3.88

интересно знать, имеют ли они какой-либо семантико-прагматический смысл.

Чтобы ответить на этот вопрос, обратимся снова к механизму угадывания. Образцовый угадчик опирается при декодировании текста как на априорную информацию тезауруса $И_0$, так и на смысловую информацию $И(T)$, извлекаемую им из уже прочитанной части текста. Чем длиннее этот участок, тем большее число элементов и связей своего тезауруса использует он при угадывании последующей буквы. Если при угадывании первой буквы текста используется статистика начальных букв слова, то при угадывании второй буквы угадчик учитывает правую дистрибуцию уже известной ему первой буквы. Переходя к отгадыванию третьей буквы, информант опирается на правую сочетаемость **начальных двухбуквенных сочетаний. Знание трех, четырех пер-**

вых букв слова дает возможность угадчику распознать значение этого слова или его начальной морфемы. В этом случае информант стремится использовать уже смысловую информацию и строит свои прогнозы уже не на комбинаторике фигур (букв, слогов), но на сочетаемости знаков (морфем, слов). Позднее вступают в действие заложенные в тезаурусе правила сочетаемости словосочетаний, а также схемы и нормы построения предложения, оценки возможных ситуаций и т. д. Таким образом, по мере движения по тексту в его декодирование вовлекаются все новые «механизмы» тезауруса угадчика.

Выше уже говорилось, что чем больше сведений о статистической структуре и содержании текста имеется у угадчика, тем меньше неопределенность при выборе им продолжений. Иными словами, чем более жестки комбинаторно-статистические связи текста и чем ближе его тематика к житейскому опыту угадчика, тем больше «механизмов» своего тезауруса может использовать образцовый информант. Поэтому величины I_n и K_n не только оценивают комбинаторно-статистические связи уже угаданной части текста, но и характеризуют одновременно эффективность предсказания. А эта эффективность определяется взаимодействием между уже извлеченной из текста информацией и приведенными в действие «механизмами» тезауруса. Что же касается величины K_∞, то она количественно характеризует оптимальное использование этих механизмов во взаимодействии со статистической структурой и тематикой декодируемого текста. Таков смысл предельной контекстной связанности.

Абсолютные величины I_∞ и K_∞ зависят от числа букв в алфавите, использующемся в данном языке. Поэтому при сравнении информационных свойств разных языков удобнее пользоваться величиной и з б ы т о ч н о с т и (R), в которой $I_\infty = H_\infty$ и K_∞ соотнесены с энтропией алфавита.

$$R = \frac{H_0 - H_\infty}{H_0} 100\%. \qquad (12)$$

Это выражение, исходя из (11), может быть записано как

$$R = \frac{K_\infty}{H_0} 100\%. \qquad (13)$$

Данные об избыточности, точнее, об ее нижних ($\underline{R}$) и верхних ($\bar{R}$) оценках в некоторых индоевропейских и тюркских языках см. в табл. 28.

Нетрудно заметить, что значения R, равно как и величины I_∞, K_∞, характеризуют недифференцированную совокупность статистических и смысловых связей, а также информаций, заложенных в тексте. Вместе с тем поскольку избыточность является отношением связанности и статистической упорядоченности текста

к той неопределенности, которая существовала бы при угадывании текста, если бы приемник не имел жизненного опыта и ему ничего не было бы известно о структуре и норме языка, кроме перечня равновероятных букв алфавита, постольку величина R может рассматриваться, подобно K_∞, в качестве меры оптимального взаимодействия тезауруса угадчика со статистической структурой и тематикой текста.

Теперь рассмотрим вопросы, связанные с достоверностью наших информационных оценок и сопоставлением их по разным языкам и стилям. Без оценки этой достоверности приводимые цифры остаются всего лишь цифрами (ср.: Altmann, 1974, с. 125—126).

Нетрудно заметить, что все девять языков обнаруживают большую близость в общих оценках избыточности. Как уже говорилось (см. 41), эти оценки для языка в целом получены путем анализа достаточно большого объема материала (исключение представляет лишь азербайджанский материал).

Коэффициент вариации

$$W = \frac{\sigma}{I_\infty} \cdot 100\%$$

для значений I_∞ и $\underline{I}_\infty$ относительно языка в целом сравнительно невелик (в среднем он равен примерно 23%). Контрольные значения I_∞, $\underline{I}_\infty$ и $\underline{\underline{I}}_\infty$, заимствованные из работ других авторов и полученные в результате проверочного эксперимента (Пиотровский, 1968, с. 60—61), близки к нашим данным. Все это дает нам право считать оценки информации, избыточности и предельной контекстной связности, приводимые в таблицах, достаточно надежными. Иными словами, избыточность развитого языка при условии, что текст на этом языке воспринимается носителем, идеально владеющим языком, находится, очевидно, в пределах между 70 и 85%.

Что касается отдельных стилей, то материала по ним было мало (33 текста на один стиль). Поэтому статистический разброс здесь значительно выше, чем для языка в целом (в среднем $w = 32.3\%$, а пиковые его значения достигают 60%). В связи с этим появляется необходимость определить существенность расхождений в количественных оценках информационных величин по этим стилям. Возьмем средние избыточности ($\tilde{R}_р$, $\tilde{R}_б$, $\tilde{R}_х$) по указанным трем стилям для шести европейских языков (см. табл. 31) и оценим степень их расхождений с помощью критерия Стьюдента, имеющего следующий вид:

$$t_{ij} = \frac{\tilde{R}_i - \tilde{R}_j}{S_{ij}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}. \qquad (14)$$

Здесь t_{ij} — значение критерия Стьюдента, оценивающее расхождения избыточностей $\tilde{R}_i$ и $\tilde{R}_j$ для стилей i и j, а $N_1 = N_2$ — ко-

личество языков, использующихся для сравнения (см.: Бектаев, Пиотровский, 1974, с. 210—215).

Таблица 31

Оценки расхождений избыточности в шести европейских языках

Разновидности языка	Избыточность $R=\frac{\bar{R}+\tilde{R}}{2}$						Средняя избыточность по стилю $\tilde{R}_i$ ($\bar{R}_j$)	Значения стандартов	
	русский	польский	английский	немецкий	французский	румынский		s_i	s_{ij}
Разговорная речь (р)	0.777	0.813	0.753	0.792	0.757	0.798	0.782	0.024	$S_{рб}=0.0255$
Беллетристика (б)	0.812	0.791	0.818	0.769	0.773	0.788	0.791	0.027	$S_{рд}=0.0265$
Деловая речь (д)	0.868	0.866	0.875	0.839	0.872	0.800	0.853	0.029	$S_{бд}=0.0280$

Пользуясь выражением (14), вычисляем значение t_{ij}, которое:

а) для пары разговорная речь—беллетристика равно $t_{р,\,б}=-0.41$;

б) для пары разговорная речь—деловая речь составляет $t_{р,\,д}=-4.59$;

в) для пары беллетристика—деловая речь равняется $t_{б,\,д}=-4.02$.

Критическая область t при 5%-м уровне значимости и при числе степени свободы $\nu=6+6-2=10$ ограничена значением $t_{0.05;\,10}=2.23$.

Таким образом,

$$|t_{р,\,б}|=0.41<2.23<4.02=|t_{б,\,д}|<4.59=|t_{р,\,д}|.$$

Отсюда следует, что различия в величинах избыточности для беллетристической речи и разговорной речи несущественны, в то время как расхождения в значениях $\tilde{R}$ для деловой речи, с одной стороны, и разговорно-беллетристической разновидности — с другой, существенны.[1] Причины этих информационно-статистических расхождений ясны: большое число лексико-синтаксических штампов и хорошее знакомство угадчика-специалиста с общей тематикой значительно повышает предсказуемость специального текста. Высокая избыточность делового текста обеспечивает правильность его понимания приемником сообщения.[1]

Следует также всегда помнить, что значения I_∞ и $\underline{I}_\infty$ очень чувствительны по отношению к методике проведения эксперимента.[2]

Значения избыточности выступают в качестве того эталона структурной связанности текста, с которым можно сравнивать внутреннюю структурную упорядоченность (избыточность) более мелких лингвистических единиц.

Пользуясь данными табл. 32, сопоставим значения R в научно-технических и публицистических текстах трех индоевропейских языков с соответствующими значениями избыточности словоформ и трехсловных сочетаний. Такое сопоставление показывает, что разрыв между внутренней структурно-статистической упорядоченностью слова и общей избыточностью текста достаточно велик (соотношение этих избыточностей составляет 0.7 : 1.0). Разрыв заметно уменьшается, когда мы переходим к трехсловным сочетаниям (здесь соотношение избыточности равно 0.85 : 1.0).

Таблица 32

Избыточность словоформ и словосочетаний сравнительно с избыточностью текста в трех индоевропейских языках

№ пп.	Язык и тип текста	Словоформы		Текст	Язык и тип текста	Словосочетания	
		$\bar{H}_c^*$	R_c%	$\bar{R}\div\tilde{R}$%		$R_{с/с}$%	$\bar{H}_{с/с}^*$
1	Английские тексты по судостроению (Микерина*, 1967, с. 14)	9.37	56.1	82.9÷92.1	Английские тексты по электронике (триады разного типа) (Нехай*, 1968, с. 10)	69.6÷72.3	15.77÷17.54
2	Немецкие тексты по публицистике (Ротарь*, 1971, с. 21)	10.82	60.0	79.6÷88.2	Немецкие тексты по публицистике (триады разного типа) (Машкина*, 1968, с. 7)	70.8÷73.4	16.41÷18.01
3	Французские тексты по электронике (Кочеткова*, 1969, с. 7)	9.37	63.5	83.6÷90.4	Французские тексты по публицистике (глагольные триады) (Ионицэ*, 1971, с. 12)	74.9	15.50

Примечание. Избыточность словоформы вычисляется с помощью формулы:

$$R_c=\frac{\lambda H_0-\bar{H}_c^*}{\lambda H_0}\,100\%.$$

[1] Следует помнить, что статистическому анализу были подвергнуты не сами разговорные тексты, а лишь их письменная запись. В этом случае не могла быть учтена та информация, которая передается обстановкой беседы, интонацией, мимикой и другими коммуникативными средствами, характеризующими устное общение. Если бы удалось учесть всю эту информацию, то оценки избыточности устной речи были бы значительно выше.

[1] В ряде случаев избыточность деловой речи достигает 95—96%. Этим путем гарантируется, например, надежность в приеме команд аэродрома экипажем идущего на посадку самолета (ср.: Frick, Sumby, 1952, с. 595—596).

[2] Мы использовали очень простые методы оценки достоверности получаемых из эксперимента количественных данных. К сожалению, более тонкого и в то же время надежного математического аппарата для оценки надежности информационных оценок пока еще не создано (в этом плане ср.: Башарин, 1959, с. 361—364).

Все это говорит о том, что при построении лексического индексирования, аннотирования (см. 56—62) целесообразнее опираться не на ключевые слова, но на словосочетания, в которых заложено заметно больше структурно-статистических связей текста. Естественным образом можно ожидать, что пообротный перевод на ЭВМ окажется гораздо более эффективным, чем пословный МП (ср. 66).

44. Общая информационная схема текста

Рассмотрим цепочки частных значений I_n (соответственно $\bar{I}_n$ или $\underline{I}_n$), оценивающие величины той информации, которую извлекает угадчик из текста. Как и предполагалось (см. 30), эти значения обнаруживают тенденцию к убыванию в зависимости от роста значений n. Представим каждую из этих цепочек не как последовательности дискретных значений энтропии, но как непрерывные кривые (точнее, как непрерывные функции непрерывного аргумента). Если отвлечься на время от непериодических колебаний, относя их за счет разброса, то эти цепочки значений могут быть аппроксимированы в первом приближении теоретической показательной кривой (экспонентой) вида:

$$I_\xi = (I_0 - I_\infty)\, e^{-s\xi} + I_\infty, \qquad (15)$$

где I — верхний или нижний теоретические пределы информации в данном участке текста; ξ — непрерывный аргумент функции, заменяющий дискретные величины n; $I_0 = H_0$ — информация алфавита, а I_∞ — предельная информация языка или его разновидности, величина которой служит асимптотой кривой I_ξ; e — основание натуральных логарифмов; s — специально рассчитываемый для каждой кривой контекстный коэффициент.[1]

Заменив величину $I_\xi = I_n$ (10) правой частью выражения (15) и произведя некоторые упрощающие преобразования, получим общее выражение контекстной связанности в данном участке текста:

$$K_\xi = (I_0 - I_\infty)(1 - e^{s\xi}). \qquad (16)$$

По существу, для каждого языка или стиля мы имеем по две формулы относительно выражений (15) и (16). Одну для $\underline{I}_\xi$ (соответственно $\underline{K}_\xi$), другую — для $\bar{I}_\xi$ ($\underline{K}_\xi$) или $\bar{I}'_\xi$ ($\underline{K}'_\xi$). Фиксируем в угадываемом тексте участок $\xi = A$. Тогда площадь, ограничен-

[1] Показательная кривая I_ξ рассчитывается на участке с 31-й по 100-ю букву с помощью метода наименьших квадратов. На участке от 1-й до 30-й буквы этот метод не дает минимализации линейных (и квадратичных) отклонений. Поэтому начальную часть графика целесообразно получать путем подбора с таким расчетом, чтобы сумма линейных отклонений экспериментальных точек от экспоненты стремилась к нулю.

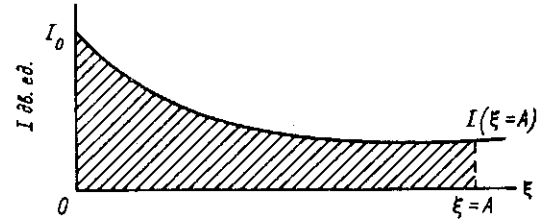


Рис. 30. Общее количество синтактической информации, извлекаемой угадчиком из текста на участке OA.

ная кривой I_ξ, осью OA, ординатами OI_0 и $AI(\xi = A)$, представляет общее количество синтактической информации I_{OA}, извлеченной образцовым угадчиком из декодированного им участка текста OA (рис. 30). Расчет этой информации можно осуществить с помощью интегрального выражения:

$$I_{OA} = \int_0^{\xi=A} I_\xi d\xi = \int_0^{\xi=A} [(I_0 - I_\infty)\, e^{-s\xi} + I_\infty)\, d\xi] = I_\infty A + \frac{I_0 - I_\infty}{S}(1 - e^{-sA}).$$

Так, например, если наш образцовый информант прочел или отгадал русский деловой текст длиной в 100 букв, то он извлек из него, считая по верхней границе информации, $\bar{I}_{OA} = 0.83 \cdot 100 + \frac{5 - 0.83}{0.24}(1 - e^{-0.24 \cdot 100}) = 100.8$ дв. ед. синтактической информации.

Обращаясь к общей интерпретации только что выведенных зависимостей, можно утверждать, что кривая I_ξ показывает ход извлечения из текста статистической и смысловой информации. Кривая K_ξ отражает динамику взаимодействия «механизмов» тезауруса угадчика и извлекаемой из текста информации.

В выражениях (15) и (16) используется контекстный коэффициент s, выступающий в качестве показателя скорости изменения величин I_ξ и K_ξ. Чем больше величина s, тем скорее идет увеличение значений K_ξ (соответственно уменьшение I_ξ). Иначе говоря, коэффициент s является показателем темпа роста контекстных связей. Характерно, что наибольшую величину s дают тексты делового стиля (см. табл. 33; рис. 31), в которых благодаря использованию большого количества устойчивых словосочетаний и ограниченного круга лексики контекстные связи языковых единиц устанавливаются быстрее, чем в текстах других стилей.

Лингвистическая природа коэффициента s достаточно сложна. Самый начальный ход экспоненты от $n=0$ до $n=1$ обусловлен введением заложенных в тезаурусе приемника статистических ограничений на употребление первых букв слова. Затем, при n порядка нескольких единиц ход кривой определен закономерностями в норме сочетаемости букв (фонем), слогов и отчасти морфем. Позднее вступают в действие статистические ограничения в сочетаемости слов с их грамматическими формами, затем, как

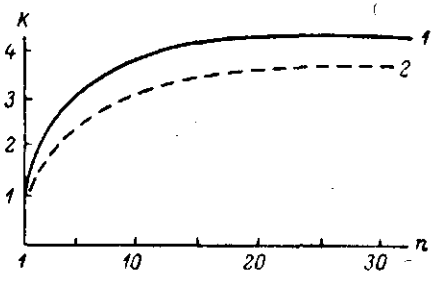


Рис. 31. Рост контекстной обусловленности в деловых текстах и в смешанной выборке текстов русского языка.

1 — деловой текст; 2 — смешанная выборка.

уже говорилось, появляются ограничения, связанные с комбинаторикой более крупных единиц и содержанием текста (43). Поэтому коэффициент s следует рассматривать, подобно величи-

Таблица 33

Значения контекстного коэффициента s

Разновидности языка	Русский язык		Французский язык	
	верхняя граница	нижняя граница	верхняя граница	нижняя граница
Разговорная речь	0.22	0.31	0.22	0.31
Беллетристика	0.21	0.29	0.26	0.29
Деловые тексты	0.24	0.32	0.34	0.42
Язык в целом	0.19	0.31	0.30	0.36

нам K_ξ и I_ξ, как суммарный показатель скорости взаимодействия заложенных в тезаурусе угадчика парадигматических ограничений и синтагматической связанности текста. Этот коэффициент, по всей вероятности, достаточно хорошо отражает действие буквенных (фонемных), слоговых, морфемных сочетаемостей, а также сочетаемости слов. Что касается синтактико-композиционных связей, действие которых распространяется на отрезки текста длиной в несколько десятков букв (фонем), то эти связи плохо описываются контекстным коэффициентом. Ведь для порядка 30 и более букв показатель s дает ничтожное изменение в ходе экспоненты.

45. Пословная информационная схема текста

Рассмотренные до сих пор информационные величины являлись суммарными оценками разнородных лингвистических и экстралингвистических явлений.

При построении алгоритмов автоматического анализа текста для нас важны не столько суммарные данные, сколько раздельные оценки информационного веса отдельных слов, лексико-грамматических связей между ними, морфологии и т. п.

Таблица 34

Участок матрицы, использующейся для расчета пословной схемы текста (угадывание начала румынского текста по сокращенной программе)

№ текста	Столбцы											
	1	2	3	4	5	6	7	8	9	10	11	12
	первое слово						второе слово					
	1-я буква	2-я буква	середина слова	предпоследняя буква	последняя буква	пробел	1-я буква	2-я буква	середина слова	предпоследняя буква	последняя буква	пробел
1	a	c	e	s	t	Δ	f	a	—	p	t	Δ
	—	1	1	1	0	1	2	2	—	1	0	1
2	î	—	—	—	n	Δ	p	r	eze	n	t	Δ
	—	—	—	—	1	2	2	1	—	1	1	1
3	v	—	—	—	—	Δ	a	—	—	—	m	Δ
	—	—	—	—	—	2	1	—	—	—	1	1

В связи с этим попробуем построить такую усредненную информационную схему, которая отражала бы членение текста на слова (эту схему мы будем называть лексической схемой текста).

Чтобы построить такую схему, необходимо, во-первых, определить средние длины первого, второго и т. д. слов в текстах данного языка или стиля, во-вторых, получить схемы распределения информации для первого, второго, третьего и т. д. слов усредненного текста. Первая задача решается обычным статистическим путем, вторая требует особой группировки того материала, который был получен из эксперимента по угадыванию букв. Принципы этой группировки иллюстрирует табл. 34, представляющая собой участок матрицы, по которой рассчитывалась лексическая схема русского текста. В первом столбце собраны первые буквы слов из каждого текста. Во втором столбце указываются вторые буквы первых слов при условии, что эти слова имеют длину не менее трех букв. Для одно- и двухбуквенных слов во втором столбце ставится прочерк. В четвертый столбец попадают предпоследние буквы слов длиной в четыре, пять и т. д. букв. Для коротких слов в четвертом столбце ставится прочерк. В пятый столбец помещаются последние буквы слов, имеющих длину две и более букв.[1] В шестой столбец записываются все пробелы между первым и вторым словами. При каждой букве и пробеле указывается количество попыток, понадобившихся для их угадывания. В третий столбец заносятся середины длинных слов без указания коли-

[1] Нетрудно заметить, что число букв, попадающих в столбец, который обобщает последние буквы слов (соответственно вторые или предпоследние), может быть меньше общего количества текстов, представленных в матрице.

чества попыток. Таким же образом расписывается второе и последующие слова текста.

Аналогичное построение имеет матрица для расчета лексической схемы текста, отгадывающегося по сокращенной программе. Ее отличие от первой матрицы состоит лишь в том, что угадывание буквы с двух и более попыток отмечается одним цифровым символом «2» (ср. табл. 35).

Таблица 35

Участок матрицы, использующейся для расчета пословной схемы текста (угадывание русского текста по полной программе)

№ текста	Столбцы 1	2	3	4	5	6	7	8	9	10	11	12
	Первое слово						Второе слово					
	1-я буква	2-я буква	середина слова	предпоследняя буква	последняя буква	пробел	1-я буква	2-я буква	середина слова	предпоследняя буква	последняя буква	пробел
1	м	н	ог	и	е	Δ	д	е	лега	т	ы	Δ
	8	3		1	1	0	5	3		1	0	0
2	д	—	—	—	0	Δ	в	е	че	р	а	Δ
	9	—		—	3	2	6	2		1	1	0
3	я	—	—	—	—	Δ	п	е	—	—	л	Δ
	11	—		—	—	1	1	2			1	1

Только что описанная матрица содержит распределения числа попыток, понадобившихся для угадывания первых, вторых, предпоследних и последних букв первого, второго, третьего и т. д. слов текста, а также аналогичные распределения относительно пробелов между словами.

Обработка этих распределений по формулам дает верхние и нижние оценки информации (энтропии) для первых двух и для последних двух букв слов нашего усредненного текста, а также оценки информации на пробелах между этими словами. Информация, падающая на середины слов (ср. стб. 3 и 9) оценивается как среднее арифметическое от информаций, приходящихся на первую, вторую, предпоследнюю и последнюю буквы усредненного слова. В итоге мы получаем схему распределения информации в тексте, учитывающую его членение на слова. Полученная путем обработки результатов угадывания по полной программе схема русского текста длиной в 15 слов показана на рис. 32. Аналогичные схемы были получены по сокращенной программе угадывания для английского и румынского (Богуславская, Новак, 1968, с. 31—32), немецкого (Boguslavskaja, Koženec, Piotrowski, 1971, с. 469) и болгарского (Георгиев,* 1973, с. 13) текстов. Обработка результатов коллективного угадывания текста [гл. 5, (17)] дает лексическую информационную схему конкретного текста

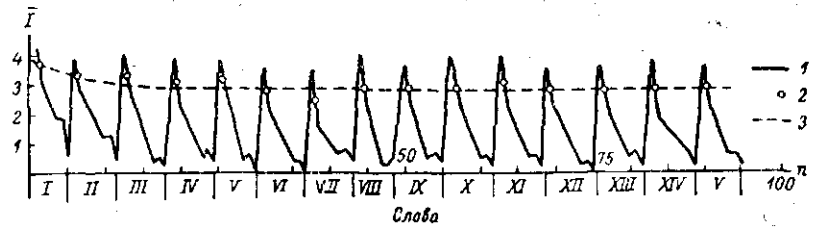


Рис. 32. Лексическая схема усредненного русского текста.

Условные обозначения: *1* — полигон распределения информации по верхней границе с учетом достоверных продолжений; *2* — значения $\frac{I_1^6 + I_2^6}{2}$; *3* — график значений $\frac{I_1^6 + I_2^6}{2}$.

(см. рис. 33). Используя описанные выше приемы, можно суммировать статистически достаточное число таких частных схем и получить новую усредненную лексическую схему, которая может быть использована для оценки достоверности той схемы текста, которая была построена путем обработки результатов индивидуального угадывания.

В заключение отметим, что, как показывают данные исследованных языков, распределение в тексте статистической информации имеет квантовый характер. Начало слов несут максимумы информации, в то время как середины и особенно следующие за ними пробелы оказываются либо мало информативными, либо вообще избыточными.

Что касается конечных букв, то они несут небольшое количество информации. Чтобы проверить, насколько квантовое распределение синтактической информации соответствует размещению смысловой информации в письменном тексте, был поставлен опыт по восстановлению утраченных букв в связном тексте. Так, например, группам студентов, аспирантов и преподавателей филологических и математических специальностей Московского и Ленинградского университетов, а также Калининского пединститута (общее число испытуемых — 80 человек) были пред

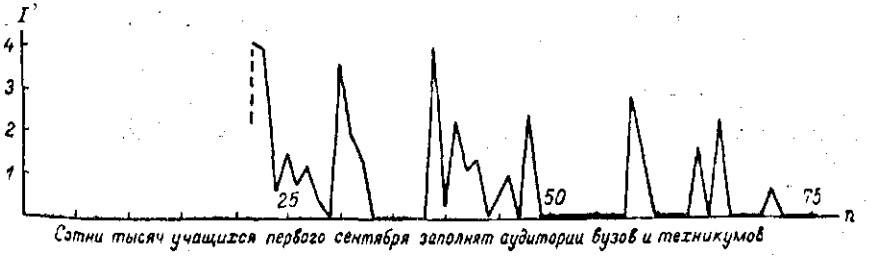


Рис. 33. Лексическая схема конкретного русского текста.

При построении графика использовались данные коллективного угадывания газетного текста («Комсомольская правда» от 20 VIII 1966 г.) при условии, что испытуемым были сообщены первые три слова (21 буква) отрывка.

ложены русские газетные тексты (газ. «Известия» за 15 II 65 и 26 III 66), первая часть которых давалась полностью, а во второй части опускалось 50% букв. Пропуски букв осуществлялись тремя способами:

а) пропускались конечные буквы слова;

б) пропускались начальные буквы слова;

в) пропускалась каждая вторая буква.[1]

Обработка результатов проводилась по следующей схеме. Правильное восстановление буквы оценивалось в два очка, за неправильно восстановленную букву давалось одно очко, отказ от восстановления буквы оценивался в нуль очков. Отношение суммы полученных всеми информантами очков к оптимальному количеству очков (т. е. к сумме пропущенных букв, помноженных на два) оценивает эффективность восстановления утраченной части текста.

Результаты эксперимента следующие:

а) при опущении концов слов текст восстанавливается на 96.4%;

б) при пропуске начальных букв слова информанты восстанавливают текст на 69.8%;

в) при опущении каждой второй буквы восстановление текста равно 20.4%.

Эти результаты еще раз говорят о том, что смысловая информация распределена в тексте дискретными «сгустками»: наиболее информационно нагруженными являются обычно начальные морфемы слов, меньше информации несут конечные морфемы. Что же касается «разрушения» морфемной структуры текста путем пропуска каждой второй буквы, то здесь восстановление смысла текста оказывается наиболее затруднительным.[2]

---

[1] Вот образец текста с пропуском концов слов: *О классиках писать легче, тем более о тех, которые стано- - - - Δ класс- - - - н-Δтв- - - Δ гла- - - Δ пра- - - . . .* и т. д. Тот же текст с пропусками начала слова: *. . . о тех, которые - - - вятся - - -иками Δ-а- - -оих- - - -зах Δ- - -вда. . .* и т. д.; или с пропуском каждой второй буквы: *. . . о тех, которые с-а-о-я-с- Δ-л-с-и-а-и-н-Δ-в-и-Δ. . .* и т. д.

[2] Сходные данные получили в свое время Э. Фридман и Дж. Миллер при проведении экспериментов по восстановлению искаженного английского текста (см.: Miller, Friedman, 1957). Труднее всего оказалось восстановить текст, в котором была произведена хаотическая замена букв. Вместе с тем выяснилось, что пропуск пробелов и гласных в меньшей степени искажает смысл текста. Это говорит о большей информационной нагруженности английских согласных по сравнению с гласными и пробелами. Аналогичное явление наблюдается и в других языках (Пиотровский, 1962, с. 56).

Сходные результаты получены относительно узбекского текста, который восстанавливался носителями языка на 95—99% при сохранении только лексических основ и практически не мог быть реконструирован по одним лишь словообразовательным и морфологическим аффиксам.

46. Лексическая и грамматическая обусловленность единиц текста

В ходе эксперимента было замечено, что по мере продвижения от начала текста испытуемый все чаще угадывает вторую, а иногда первую букву слова, опираясь не на комбинаторику букв, а на предшествующий лексический контекст. Это дало возможность предположить, что, исследуя убывание сумм информаций, падающих на 1-ю и 2-ю буквы слов (ср. рис. 32), можно оценить нарастание лексических связей в тексте. Нарастание этих связей, которое мы будем обозначать как рост лексической связанности L_ξ, было оценено в первом приближении относительно русского и французского языков с помощью показательной кривой, имеющей вид:

$$L_\xi = (I_I^{\text{л}} - I_\infty^{\text{л}})(1 - e^{-l\xi}), \tag{17}$$

где $I_I^{\text{л}}$ — среднее арифметическое информаций, вычисленных для верхней или нижней границ, которые падают на 1-ю и 2-ю буквы первого слова текста, $I_\infty^{\text{л}}$ — предел лексической обусловленности текста (ср. величины I_∞ в выражениях), ξ — непрерывный аргумент функции, заменяющий дискретные величины n, причем здесь $\xi = n - 1.5$, поскольку абсцисса начала кривой равна 1.5 «буквы», l — лексический коэффициент, характеризующий темп нарастания лексических связей в тексте (ср. коэффициент s в формулах (15) и (16), остальные обозначения имеют тот же смысл, что и в предшествующих выражениях.

Ход кривой $I_\xi^{\text{л}}$ для русского текста показан на рис. 32. При построении этого графика использованы данные табл. 36.

Для того чтобы охарактеризовать тот предел, к которому стремится лексическая обусловленность в данном языке и его стилях, мы введем понятие «предельная лексическая связанность» (L_∞), аналогичное понятию «предельная контекстная связанность» (см. выше, с. 157). При этом

$$L_\infty = I^{(\text{л})} - I_\infty^{(\text{л})}. \tag{18}$$

Соотнося значения L_∞ с предельной контекстной связанностью K_∞, мы узнаем ту долю, которую занимает величина L_∞ в общей сумме контекстных связей:

$$Л = \frac{L_\infty}{K_\infty} 100\%. \tag{19}$$

В табл. 36 показаны величины $Л$ для русского, английского и французского языков, а также для стилевых разновидностей последнего.

Хотя величины $I^{\text{л}}$, L_∞, $Л$ являются по своей природе синтактико-информационными величинами, они количественно харак-

Таблица 36

Распределение информации (дв. ед.) в лексической схеме русского текста по верхней границе информации

№ слова и его средняя длина (цифры в скобках) без учета пробела	Количество информации, приходящееся на		$\frac{I_1^* + I_2^*}{2} = I_\xi^{(n)}$	Количество информации, приходящееся на		$\frac{I_m^* + I_n^*}{2} = I_\xi^{(r)}$	Количество информации, приходящееся на пробел
	первую букву слова (I_1^*)	вторую букву слова (I_2^*)		предпоследнюю букву слова I_m^*	последнюю букву слова I_n^*		
I (4.40)	4.22	3.12	3.67	1.87	1.83	1.85	0.55
II (5.50)	3.77	3.00	3.38	1.07	1.23	1.15	0.40
III (5.67)	3.98	2.66	3.32	0.41	0.47	0.44	0.25
IV (5.16)	3.92	2.14	3.02	0.51	0.81	0.66	0.45
V (5.60)	3.79	2.49	3.14	0.29	0.61	0.45	0.13
VI (5.90)	3.58	1.98	2.78	0.35	0.39	0.37	0.06
VII (5.52)	3.43	1.43	2.43	0.57	0.66	0.61	0.35
VIII (4.51)	3.95	1.96	2.96	0.23	0.28	0.25	0.45
IX (5.28)	3.59	2.24	2.92	0.48	0.68	0.59	0.40
X (5.64)	3.83	2.06	2.94	0.47	0.60	0.54	0.18
XI (5.69)	3.93	2.17	3.05	0.47	0.68	0.58	0.29
XII (5.65)	3.52	2.21	2.86	0.28	0.38	0.33	0.07
XIII (5.60)	3.58	1.99	2.78	0.53	0.80	0.66	0.11
XIV (5.71)	3.79	1.84	2.82	0.92	0.62	0.77	0.21
XV (5.55)	3.58	2.21	2.88	0.62	0.57	0.60	0.15

теризуют взаимодействие текста с лексическими механизмами и с «житейским» опытом образцового угадчика, заложенными в его тезаурусе. Действительно, чем больше угадчик знает о содержании текста и чем лучше чувствует он правые лексические валентности отдельных словоформ, тем лучше он угадывает первые буквы каждого слова текста.

По условиям эксперимента все угадчики, независимо от языка, находятся в равных условиях как с точки зрения знакомства с тематикой, к которой относился угадываемый текст, так и с точки зрения справочного аппарата. Поэтому заметные различия по языкам в значениях как L_∞, так и $Л$ можно относить не за счет экстралингвистических факторов, но объяснять их особенностями лексической структуры того или иного текста.

Так, например, высокие значения L_∞ и $Л$ во французском языке и его разновидностях следует относить за счет более регламентированного здесь по сравнению с другими языками употребления лексических и фразеологических единиц.

Что касается самого французского языка, то наиболее высокий процент лексической предсказуемости букв дает деловой текст. Причину этого снова следует искать, во-первых, в использовании большого количества устойчивых словосочетаний, связанных с той или иной тематикой, во-вторых, в сравнительно ограниченном круге лексики, значительную часть которой образует

терминология данной специальности, в-третьих, в логическом и строго нормализованном построении предложений.

Более низкая избыточность беллетристического стиля является результатом большей по сравнению с деловой речью неопределенностью в выборе языковых элементов. Хотя беллетристический стиль также использует строго нормализованный синтаксис, лексические связи здесь значительно слабее: языковые штампы применяются гораздо реже, вместе с тем в этом стиле образуется большое количество неожиданных словосоединений (метафоры и другие «фигуры стиля»), а круг лексики здесь гораздо шире, чем имеет место в деловой речи.

Высокие значения коэффициентов L_∞ и $Л$ для французского делового текста, свидетельствующие о сильной регламентации употребления лексико-фразеологических единиц, дают здесь определенные перспективы для использования приемов предсказуемостного анализа (Rhodes, Alt, 1962/1971) при решении на ЭВМ семантических и синтаксических задач (ср.: Угодчикова*, 1975).

Если при угадывании начальных букв слова в тексте образцовый информант использует свой «житейский» опыт и лексические механизмы тезауруса, то при угадывании последних букв изменяемых слов используются знания морфологии. Поэтому для языков, широко использующих флективную и агглютинативную технику, подсчет синтактической информации, падающей на последние буквы слов, взятых из разных участков текста, может дать количественную оценку взаимодействия грамматических «механизмов» тезауруса и морфолого-синтаксической структуры текста.

Таблица 37

Предельная контекстная и лексическая связанность текста в английском и французском языках

Языки	Предельная обусловленность. Верхняя граница по сокращ. прогр. K'_∞	Предельная лексическая обусловленность, L_∞ в дв. ед.	$Л$ в %
Английский	3.41	0.69	20.2
Французский:			
разговорная речь	3.29	1.07	32.5
беллетристический стиль	3.38	1.22	36.1
деловой стиль	3.56	1.49	41.9
язык в целом	3.36	1.17	34.8

Исходя из этих соображений, было исследовано убывание полусумм информаций, падающих на последнюю и предпоследнюю буквы слова в лексической схеме русского текста (см. табл. 36; рис. 32).

Таблица 38

Данные о лексической и грамматической связанности русского текста, полученные путем обработки значений I

$I_I^{(д)}$ в дв. ед.	$I_\infty^{(д)}$ в дв. ед.	l	K_∞ в дв. ед.	L_∞ в дв. ед.	$Л$ в %
3.67	2.87	0.06	3.63	0.80	22
1.85	0.50	0.10	3.63	1.35	37
$I_I^{(г)}$ в дв. ед.	$I_\infty^{(г)}$ в дв. ед.	g	K_∞ в дв. ед.	G_∞ в дв. ед.	$Г$ в %

Нарастание грамматической связанности G было оценено относительно «общеязыкового» усредненного русского текста с помощью выражения:

$$G_\gamma = (I_I^{(г)} - I_\infty^{(г)})(1 - e^{-g\gamma}), \qquad (20)$$

аналогичного выражениям (16) и (17). Здесь $I_I^{(г)}$ — полусумма информаций, падающих на предпоследнюю и последнюю буквы первого слова, $I_\infty^{(г)}$ — предел, к которому стремятся полусуммы информаций последних букв слова, g — грамматический коэффициент, характеризующий скорость нарастания морфолого-синтаксических связей в тексте (ср. коэффициенты s и l в формулах (16), (17) и др.), γ — непрерывный аргумент функции, заменяющий дискретные величины n (если учесть сдвиг начала кривой G_γ вправо, то

$$\gamma = n - \lambda_I + 0.5,$$

где λ_I — число букв в первом слове текста).

Чтобы определить предел, к которому стремится грамматическая связанность текста, введем понятие предельной грамматической связанности (ср. аналогичные величины K_∞, L_∞). Предельная грамматическая связанность будет равна:

$$G_\infty = I_I^{(г)} - I_\infty^{(г)}. \qquad (21)$$

По аналогии с выражением (19) мы можем определить ту долю, которую занимает в общей сумме контекстных связей предельная грамматическая связанность. Эта доля равна

$$Г = \frac{G_\infty}{K_\infty} 100\%. \qquad (22)$$

Все величины, характеризующие динамику грамматических связей в русском тексте, приведены в табл. 38 вместе с аналогичными характеристиками роста лексической предсказуемости.

Сравнение лексических и грамматических коэффициентов показывает, что доля грамматических связей и темп их роста в русском тексте более чем в полтора раза выше доли лексической предсказуемости и скорости ее нарастания в том же тексте (ср. рис. 34). Это говорит о том, что поиск снимающих неоднозначность лексических детерминант должен осуществляться в ходе автоматической переработки на бо́льшую глубину текста, чем поиск грамматических признаков, которые обнаруживаются обычно в непосредственном окружении анализируемого словоупотребления.

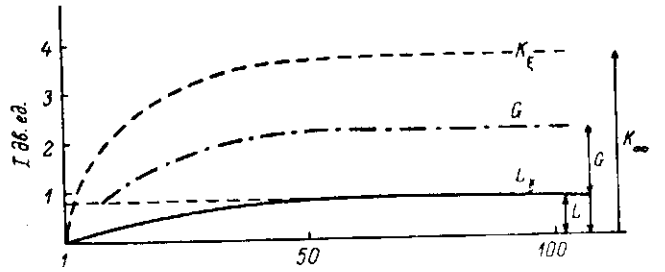


Рис. 34. Рост лексической (L_ξ) и грамматической (G) связанности текста в сравнении с ростом общей контекстной связанности K_ξ.

47. Информационные схемы слова

При оценке общего количества и размещения информации в слове методика эксперимента и расчетов остается в целом той же, что и при исследовании текста. Здесь снова применяется индивидуальное или коллективное побуквенное угадывание, осуществляющееся либо относительно слов, взятых вне контекста (эти слова мы будем называть внеконтекстными), либо при условии, что испытуемому известен контекст, предшествующий угадываемому слову (такие слова мы будем называть контекстными). В выборку внеконтекстных слов включаются обычно слова, стоящие в начале тех связных текстов, которые были подвергнуты угадыванию. Наборы контекстных слов пополняются обычно словами, стоящими на пятом, шестом и т. д. местах тех же текстов.

В тех случаях, когда необходимо получить информационную схему слова на буквенном уровне, для выборки слов определенной длины строятся обобщающие матрицы, в которых все результаты угадывания для первых букв слова попадают в первый столбец, для вторых букв — во второй и т. д. (ср. табл. 39). Каждый столбец обрабатывается по формулам (14) и (15).

В тех случаях, когда нужно определить распределение информации в слове на слоговом и морфемном уровнях, результаты угадывания группируются таким образом, чтобы получить распределения частот попыток для угадывания первой и последней букв слога или морфемы. Эта процедура, аналогичная перегруппировке экспериментального материала при построении лексической

Таблица 39

Распределение информации по $\bar{I}$ в обобщающей схеме русского текстового слова, соответствующей средней длине русского слова

№ пп.	λ	P_λ	1-я буква $\bar{I}$	P'_λ	2-я буква $\bar{I}$	Средние буквы (1.37 буквы) $\bar{I}$	P''_λ	Предпоследняя буква $\bar{I}$	P'''_λ	Последняя буква $\bar{I}$	Пробел $\bar{I}$
1	2	0.11	3.10	—	—		—	—	—		0.84
2	3	0.11	2.98	—	—		—	—	0.12	1.65	0.62
3	4	0.10	3.63	0.13	1.83	—	—	—	0.11	1.04	0.31
4	5	0.08	3.52	0.10	2.05		0.12	0.61	0.10	0.47	0.05
5	6	0.13	3.66	0.16	1.46		0.19	0.47	0.14	0.39	0.06
6	7	0.11	3.61	0.14	1.88		0.16	0.30	0.12	0.25	0.05
7	8	0.11	3.49	0.14	2.33		0.16	0.29	0.12	0.20	0.00
8	9	0.09	3.38	0.12	1.80		0.13	0.07	0.10	0.20	0.00
9	10	0.05	2.82	0.07	1.85		0.08	0.25	0.06	0.11	0.00
10	11	0.04	3.59	0.05	2.25		0.06	0.24	0.05	0.00	0.00
11	12 и более	0.07	3.50	0.09	1.90		0.10	0.26	0.08	0.15	0.05
12	$\bar{I}_{\text{п}}^{(\text{с})}$		3.45		1.90	1.75		0.33		0.51	0.21

схемы текста (см. 46 и табл. 36), дает возможность не только определить общее количество информации, приходящееся на слог или морфему, но позволяет также численно оценить информацию на слоговых и морфемных границах, выясняя при этом особенности слогового и морфемного членения слова.

При построении побуквенных распределений пробел считается последней буквой слова. В слоговых схемах пробел не учитывается (как известно, пробел не входит в состав фонетического слога). В морфемных схемах учитываются лишь те пробелы, которые выступают в функции нулевых морфем.

Следует иметь в виду, что исследование информационного членения слова может быть успешно осуществлено лишь при условии, что предшествующее расчетам выделение слогов и морфем внутри слова проводится на основе строгой и последовательной лингвистической процедуры.

Побуквенное, слоговое и морфемное распределения информации в словах определенной длины относятся лишь к частным случаям, не дают полного представления об общем распределении информации и мало что говорят об общих принципах слогового и морфемного членения слова. Поэтому возникает необходимость построить обобщающую информационную схему слова на всех трех уровнях.

Учитывая эмпирические вероятности появления слов определенной длины, можно свести их побуквенные схемы в обобщающую схему распределения информации в слове. Используется два вида такого обобщения.

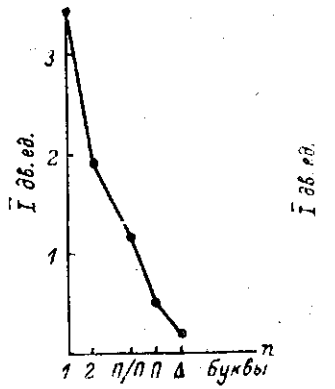


Рис. 35. Обобщающая схема русского текстового слова, соответствующая его средней длине, построенная по $\bar{I}$.

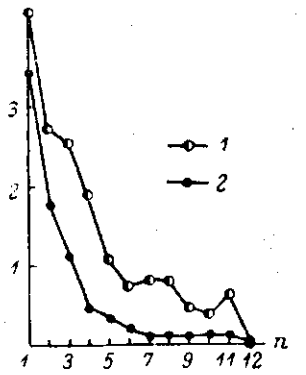


Рис. 36. Обобщающая схема русского слова без ограничения длины, построенная по $\bar{I}$.

Условные обозначения: *1* — внетекстовое слово; *2* — текстовое слово.

Во-первых, может быть построена схема, соответствующая средней длине слова в данном языке. При построении этой схемы буквы, или, точнее, информации, падающие на эти буквы, группируются так, как это делалось при расчете лексической схемы текста (см. 46). Информации ($I_n^{(\text{с})}$), приходящиеся на первую, вторую, предпоследнюю, последнюю буквы и пробел в обобщающей схеме, получаются как суммы взвешенных информаций, приходящихся на соответствующие буквы в частных схемах. Иными словами

$$I_n^{(\text{с})} = \sum P_\lambda I_n^{(\text{с})}, \qquad (23)$$

где P_λ — эмпирическая вероятность слова определенной длины (λ); $I_n^{(\text{с})}$ — информация, приходящаяся на n-ю букву в частной схеме. Количество информации, содержащееся в буквах, которые находятся в интервале между второй и предпоследней буквами усредненной схемы, получается как среднее арифметическое первой, второй, предпоследней и последней букв слова, умноженное на количество букв в этом интервале. Расчет обобщающей схемы русского текстового слова по $\bar{I}$ дан в табл. 39, а на рис. 35 представлен ее чертеж.

Во-вторых, может быть построена обобщающая схема, на которую не накладываются ограничения, связанные со средней длиной слова в текстах данного языка. Поскольку предел длины слова ничем не ограничен, то длина обобщающей схемы в этом случае стремится к бесконечности. Однако практически оказывается, что

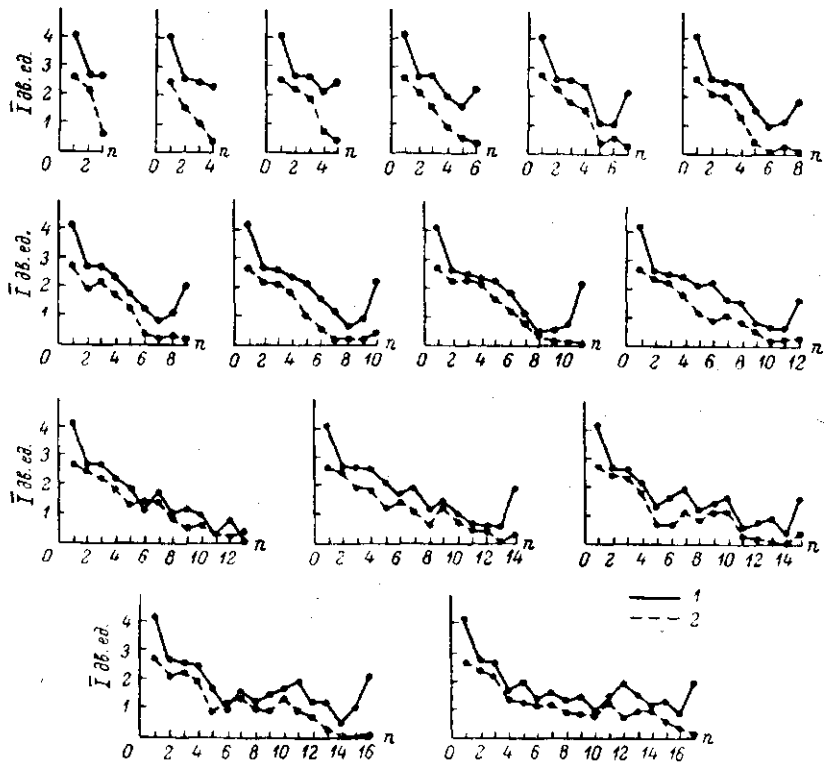


Рис. 37. Побуквенные распределения информации в немецких словах разной длины (по данным: Коженец*, 1973, рис. 2—3).

Условные обозначения: *1* — внетекстовое слово; *2* — текстовое слово.

очень длинные слова встречаются в языке сравнительно редко. Так, например, в русском языке слова длиной более одиннадцати букв, а во французском и английском языках более десяти букв составляют в среднем не более 5—7% общего числа слов. Поэтому целесообразно ограничить рассматриваемую схему некоторым пределом. В русском языке, например, эта схема доведена до двенадцатой буквы (11 букв+пробел). Расчет этой схемы для русского текстового слова по $\bar{I}$ и $\underline{I}$ дан в табл. 40, ее чертеж относительно $\bar{I}$ см. на рис. 36. На этом же рисунке дана схема распределения информации внетекстового слова. Аналогичные схемы для словацкого языка см.: Majernik, Krútel', 1972, с. 286.

Иначе обстоит дело в немецком языке, в текстах которого встречается достаточно большое число длинных слов. Поэтому здесь приходится учитывать распределение информации и в словах от 12 до 16 букв (ср. рис. 37).

Таблица 40

Распределение информации в русских словоформах различной длины, взятых в тексте (в дв. ед.)

№ пп.	Длина словоформы в буквах (λ), включая пробел	Вероятность словоформы данной длины $P(\lambda)$	Информация на букву																						Общее количество информации, приходящееся на словоформу определенной длины		Среднее количество информации, приходящееся на словоформу		Общее количество грамматической информации, приходящееся на словоформу		Среднее количество грамматической информации, приходящееся на букву	
			1-я		2-я		3-я		4-я		5-я		6-я		7-я		8-я		9-я		10-я		11-я		12-я									
			I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	2	0.11	3.10	3.10	0.38	0.84	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3.48	3.94	1.74	1.97	0.00	0.00	0.00	0.00
2	3	0.11	2.83	2.98	0.95	1.65	0.28	0.62	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4.06	5.25	1.35	1.75	0.24	0.35	0.08	0.12
3	4	0.10	3.50	3.63	1.00	1.83	0.57	1.04	0.14	0.31	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5.21	6.81	1.30	1.70	0.59	1.20	0.15	0.30
4	5	0.08	3.26	3.52	1.25	2.05	0.24	0.61	0.20	0.47	0.03	0.05	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5.08	6.70	1.01	1.32	0.61	1.05	0.12	0.21
5	6	0.13	2.94	3.66	0.82	1.46	0.43	0.81	0.24	0.47	0.20	0.39	0.02	0.06	—	—	—	—	—	—	—	—	—	—	—	—	4.65	6.85	0.78	1.14	0.36	0.72	0.06	0.12
6	7	0.11	2.57	3.61	1.14	1.88	0.55	1.08	0.21	0.44	0.13	0.30	0.12	0.25	0.02	0.05	—	—	—	—	—	—	—	—	—	—	4.74	7.61	0.68	1.09	0.43	0.75	0.06	0.11
7	8	0.11	2.90	3.49	1.46	2.33	0.77	1.43	0.17	0.38	0.08	0.17	0.14	0.29	0.10	0.20	0.00	0.00	—	—	—	—	—	—	—	—	5.62	8.29	0.70	1.04	0.40	0.67	0.05	0.08
8	9	0.09	2.79	3.38	1.02	1.80	0.56	1.06	0.25	0.52	0.16	0.30	0.10	0.21	0.02	0.07	0.07	0.20	0.00	0.00	—	—	—	—	—	—	4.97	7.54	0.55	0.84	0.37	0.72	0.04	0.08
9	10	0.05	3.20	3.82	1.04	1.85	0.93	1.66	0.18	0.42	0.19	0.41	0.04	0.11	0.04	0.10	0.12	0.25	0.04	0.11	0.00	0.00	—	—	—	—	5.78	8.73	0.58	0.87	0.42	0.49	0.04	0.05
10	11	0.04	2.94	3.59	1.43	2.25	1.63	2.24	0.51	0.88	0.10	0.24	0.10	0.22	0.00	0.00	0.12	0.21	0.12	0.24	0.00	0.00	0.00	0.00	—	—	6.95	9.82	0.63	0.90	0.30	0.56	0.03	0.05
11	12	0.07	2.88	3.50	1.12	1.90	0.91	1.58	0.44	0.84	0.28	0.56	0.24	0.40	0.14	0.27	0.00	0.00	0.04	0.10	0.12	0.26	0.07	0.15	0.04	0.0	6.28	9.61	0.52	0.80	0.60	0.98	0.05	0.08
12	Среднее количество информации, приходящееся на букву в обобщающей схеме словоформы		2.98	3.45	1.02	1.75	0.59	1.11	0.22	0.46	0.14	0.30	0.11	0.20	0.06	0.12	0.05	0.11	0.04	0.09	0.05	0.11	0.04	0.10	0.04	0.0	5.99	8.15	0.94	1.28	0.38	0.70	0.06	0.11

Таблица 41

Распределение информации в русских словоформах различной длины, взятых вне текста (в дв. ед.)

№ пп.	Длина словоформы в буквах (λ), включая пробел	Вероятность словоформы данной длины P (λ)	Информация на букву																						Общее количество информации, приходящееся на словоформу определенной длины		Среднее количество информации, приходящееся на одну букву словоформы		Общее количество грамматической информации, приходящееся на словоформу		Среднее количество грамматической информации, приходящееся на букву	
			1-я		2-я		3-я		4-я		5-я		6-я		7-я		8-я		9-я		10-я		11-я		12-я									
			$I=\bar{I}$		I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$	I	$\bar{I}$
			1—2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	2	0.11	4.22		0.55	1.09	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4.77	5.31	2.38	2.66	0.00	0.00	0.00	0.00
2	3	0.11	4.22		1.53	2.25	0.81	1.34	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6.56	7.81	2.19	2.60	0.20	0.38	0.07	0.13
3	4	0.10	4.22		2.13	2.80	1.26	1.95	0.29	0.58	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	7.90	9.55	1.98	2.39	1.85	3.10	0.46	0.78
4	5	0.08	4.22		2.60	3.26	1.75	2.56	1.43	2.21	0.00	0.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	10.00	12.27	2.00	2.45	2.25	3.29	0.45	0.66
5	6	0.13	4.22		2.40	2.98	2.27	2.95	1.51	1.89	0.85	1.33	0.16	0.31	—	—	—	—	—	—	—	—	—	—	—	—	11.41	13.68	1.90	2.28	2.30	3.52	0.38	0.59
6	7	0.11	4.22		3.23	3.26	1.66	2.31	1.15	1.71	1.02	1.36	0.52	0.98	0.23	0.43	—	—	—	—	—	—	—	—	—	—	12.03	14.27	1.72	2.04	1.43	2.27	0.20	0.32
7	8	0.11	4.22		2.83	3.14	3.18	3.28	1.53	2.19	0.73	0.92	0.73	1.10	0.77	1.26	0.02	0.26	—	—	—	—	—	—	—	—	14.01	16.37	1.75	2.05	1.27	2.05	0.16	0.26
8	9	0.09	4.22		2.12	2.88	2.88	3.08	1.37	1.92	0.88	1.13	0.34	0.61	0.62	1.35	0.30	0.66	0.00	0.00	—	—	—	—	—	—	12.93	15.85	1.44	1.76	1.70	2.87	0.19	0.32
9	10	0.05	4.22		2.30	2.89	2.66	3.19	2.21	2.60	0.93	1.43	0.20	0.58	0.32	0.51	0.96	1.42	0.96	1.32	0.00	0.00	—	—	—	—	14.76	18.16	1.48	1.82	3.50	4.78	0.35	0.48
10	11	0.04	4.22		2.51	2.95	2.61	3.08	1.38	2.15	0.69	1.18	0.74	1.21	0.46	0.72	0.40	0.64	0.29	0.56	0.58	0.89	0.32	0.64	—	—	13.93	18.14	1.27	1.65	1.81	2.88	0.17	0.26
11	12 и более	0.07	4.22		2.60	2.99	2.43	2 .62	1.65	2.14	0.75	1.20	0.66	0.69	0.18	0.33	0.38	0.62	0.32	0.53	0.20	0.40	0.54	0.71	0.00	0.00	13.93	16.45	1.16	1.37	1.41	2.12	0.12	0.18
12	Среднее количество информации, приходящееся на букву в обобщающей схеме словоформы		4.22		2.20	2.73	2.01	2.52	1.31	1.87	0.74	1.06	0.46	0.75	0.45	0.81	0.37	0.81	0.33	0.50	0.23	0.40	0.46	0.65	0.00	0.00	11.59	13.50	1.82	2.12	1.46	2.30	0.23	0.36

Таблица 42

Слоговая схема русского слова по *I*

№ пп.	Длина слова в слогах	Вероятность появления слова данной длины вне текста	Информация, в дв. ед.											
			слоги											
			I		II		III		IV		V		VI	
			первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква
1	1	0.339	4.22	2.25										
2	2	0.303	4.22	2.85	2.60	1.35								
3	3	0.214	4.22	3.00	3.26	1.09	0.82	1.05						
4	4	0.097	4.22	2.16	3.12	1.66	1.58	1.12	0.88	1.07				
5	5	0.035	4.22	2.15	2.57	1.23	0.74	0.62	0.49	0.35	0.71	0.57		
6	6 и более	0.012	4.22	2.56	2.32	1.48	1.24	0.38	0.60	0.18	0.69	1.23	0.75	0.40
7	Обобщающая слоговая схема слова вне текста		4.22	2.59	2.89	1.31	1.03	1.08	0.77	0.82	0.74	0.74	0.75	0.40
8	Обобщающая слоговая схема текстового слова		3.34	1.17	1.28	0.34	0.23	0.19	0.12	0.02	0.07	0.00	0.00	0.20
9	Средняя длина слога в буквах		2.30		2.20		2.39		2.30		1.95		1.55	

Таблица 43

Морфемная схема русского слова по $\bar{I}$

№ пп.	Длина слова в морфемах	Вероятность появления слов данной длины	Информация, в дв. ед.									
			морфемы									
			I		II		III		IV		V	
			первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква
1	1	0.260	4.22	1.87	2.24	1.05	1.72	0.73				
2	2	0.417	4.22	2.28	2.30	1.96	1.31	0.00				
3	3	0.165	4.22	1.80	2.93	0.99	1.49	0.00				
4	4	0.118	4.22	2.02	2.84	0.39			1.18	0.56	1.11	0.79
5	5 и более	0.040	4.22	2.17					1.77	0.80		
6	Обобщающая морфемная схема слова вне текста		4.22	2.03	2.39	1.19	1.54	0.35	1.39	0.63	1.11	0.79
7	Обобщающая морфемная схема текстового слова		3.49	0.88	1.27	0.31	0.30	0.07	0.36	0.00	0.11	0.00
8	Средняя длина морфемы в буквах		3.24		2.07		1.70		1.70		1.44	

Рис. 38. Обобщающие слоговая (*а*) и морфемная (*б*) схемы распределения информации русского слова.

Условные обозначения: *1* — внетекстовое слово; *2* — текстовое слово.

Обобщающие схемы слов строятся также на слоговом и морфемном уровнях. Ниже в качестве образца приводятся табл. 41, 42, 43, содержащие данные для построения обобщающих слоговой и морфемной схем русского внетекстового и текстового слова. Чертежи этих схем см. на рис. 38.

При оценке достоверности частных и общих информационных схем слова используются, как уже говорилось (41), контрольные угадывания на других испытуемых, коллективное угадывание, а также сопоставление с данными, полученными путем расчета вероятностных спектров слов (частотных словарей).

Теперь попробуем проследить, как меняется распределение информации в зависимости от изменения длины слова, а также от того, рассматривается ли буквенная, слоговая или морфемная его структура.

Распределение информации в коротких (до четырех букв), средних (от пяти до семи букв) словах, с одной стороны, и длинных словах (от восьми букв и выше) — с другой, имеет разный характер. Короткие и средние слова, как правило, дают монотонное убывание информации от начала слова к его концу. Убывание это обычно происходит гладко, и полигоны (графические схемы) этих слов имеют компактный вид (ср. рис. 37). Такое распределение связано с тем, что среди коротких и длинных слов достаточно часто употребляются неизменяемые формы (ср. русск. *о*, *да*, *еще*, *если*, *очень*, *только*; нем. *ab*, *und*, *wenn*; англ. *a*, *of*, *and*, *over*, *alone*, *almost*; франц. *à*, *on oui*, *très*, *aussi*, *encore*; рум. *o*, *de*, *mai*, *dacă*, *afară*, *pentru*), которые в суммирующей схеме сглаживают максимумы информации на конечных аффиксах изменяемых слов.

У длинных слов (в первую очередь у внетекстовых) полигоны принимают постепенно так называемую *U*-образную форму: максимумы информации сосредоточены здесь в начале и в конце

словоформы, буквы же, находящиеся в средней части словоформы, несут наименьшую информацию (ср. рис. 37). *U*-образный характер распределения информации в длинных словоформах отражает определенные вероятностно-статистические и лингвистические закономерности. Присутствие максимума информации в начале слова отражает тот факт, что начальные буквы письменного слова содержат наибольшее количество неопределенности в выборах.

Появление максимума информации в конце слова обусловлено, очевидно, морфологической структурой слова в исследованных языках, в которых бóльшая часть грамматической информации концентрируется в присоединяемых к концу слова аффиксах (окончаниях и суффиксах).

Тот факт, что морфологическая структура изменяемого слова обнаруживается яснее всего в информационных схемах длинных словоформ, объясняется тем, что в этих схемах обобщаются почти исключительно словоформы, имеющие приставки и конечные аффиксы.

Буквенные распределения, в которых обобщаются формы слов разных структур (изменяемые и неизменяемые, односложные и многосложные), дают весьма приближенную и грубую схему распределения информации. Для того чтобы глубже проникнуть в информационное строение слова, были проанализированы распределения информации в русских, болгарских, английских, немецких, французских и румынских словоформах на слоговом и морфемном уровнях (Пиотровский, 1968, с. 86—89; Novak, Piotrowski, 1968, с. 225—236; Bogusławskaja, Korzeniec, Piotrowski, 1972, с. 26—29; Георгиев*, 1973, с. 15).

Построение слоговых и морфемных схем таково, что позволяет наблюдать межслоговые «швы» и границы между морфемами, образующими слово.

Схемы слогового распределения информации показывают, что во всех языках слоговое деление слова обнаруживается лишь на границе первого и второго слова или, иначе говоря, в начале хода кривой распределения (ср. в этом смысле чертежи обобщающих слоговых схем, приведенных на рис. 38). По мере продвижения кривой вправо слоговые границы все более затушевываются, а начиная с 4-го слога полностью исчезают.

Иную картину дает морфемное построение слова. Морфемные схемы внетекстовых слов во всех трех языках характеризуются последовательным чередованием максимумов и минимумов информации. Максимумы информации падают на первую, а минимумы на последнюю букву морфемы. Иными словами, на всем протяжении внетекстового письменного слова оказываются четко обозначенными границы между морфемами. Эти границы совпадают с водоразделом между последней буквой предшествующей морфемы (минимум информации) и первой буквой последующей морфемы (максимум информации). Если обратиться к текстовым

словам, то окажется, что они также обнаруживают морфемное членение, хотя оно и выражено здесь значительно слабее (ср. рис. 38).

Итак, можно утверждать, что с точки зрения синтактической информации слову в европейских языках присуща отчетливо выраженная морфемная структура, которая подавляет слоговое членение слова. Так обстоит, очевидно, дело во всех тех языках, в которых слог не совпадает с морфемой.

Тот факт, что знаковая (морфемная) структура слова подавляет синтагматическую систему фигур (т. е. слогов), проливает свет на взаимодействие различных механизмов тезауруса нашего идеального угадчика. Как известно, речь представляет собой сложный марковский процесс линейного следования различных элементов языка — как фигур, так и знаков-символов. Иными словами, вероятностно-статистические закономерности, характеризующие сочетаемость фигур, взаимодействуют с вероятностью сочетаемости символов (морфем, слов, словосочетаний и т. д.)

Исследованный материал показывает, что вероятностно-статистические связи и ограничения, характеризующие сочетаемость фигур (букв, слогов) на коротких начальных участках текста (в нашем случае — на первых буквах словоформы), не превышают длины кратчайшего символа (чаще всего морфемы), в который входят данные фигуры. Как только следующие друг за другом по синтагматической оси фигуры сформируют символ, на сцену выступают вероятностно-статистические закономерности их сочетаемости.

Руководствуясь заложенной в его тезаурусе априорной информацией о сочетаемости символов, идеальный угадчик накладывает ее на вероятностные спектры сочетаемости последующих фигур, отбирая из этих спектров лишь такие сочетания, которые соответствуют правилам сочетаемости знаков-символов.

Все это приводит к тому, что в обобщающей схеме слова и текста вероятностная комбинаторная статистика фигур оказывается подавленной вероятностной комбинаторикой символов. Здесь результаты нашего эксперимента согласуются с современными представлениями о непосредственной и оперативной памяти человека, которая обобщает такие дискретные единицы информации, как буквы, фонемы, слоги в более крупные оперативные единицы — лингвистические знаки-символы (Miller, 1956/1964; Невельский, 1965; Репкина, 1965).

Из сказанного следует, что широко проводимые сейчас работы по статистике и комбинаторике букв, буквосочетаний и слогов дают мало сведений о механизме текстообразования. Гораздо большую ценность для лингвистики текста и инженерной лингвистики представляют статистические исследования в области комбинаторики значащих единиц текста.

Вместе с тем сведения о зернистости строения текста, а также об синтактико-информационной нагруженности начал слов и избыточности середин длинных словоформ используются при

решении задач, связанных с кодированием и компрессированием текстовой информации в ЭВМ. Так, например, при формировании сжатых кодов (сверток) словоформ, т. е. таких кодов, с помощью которых удается уменьшить массивы вводимой в ЭВМ информации, целесообразно сохранять коды первых букв слова. В этом случае однозначная идентификация свертками текстовых словоупотреблений приближается к 100% (Михлин, Пиотровский, Фрумкин, 1974, с. 30).

При построении автоматических словарей, ориентированных на машины с малой памятью, применяются и другие методы свертки, учитывающие распределение информации в слове. Так, например, при построении англо-русского автоматического словаря у длинных словоформ (более 18 букв) в целях компрессии устранялись не несущие информации буквы, находящиеся в середине слова (Вертель и др., 1971, с. 46). Разумеется, этим методом компрессии следует пользоваться с большой осторожностью. Он вряд ли может быть применен к длинным немецким словам построенным обычно по методу словосложения. Поскольку эти слова представляют собой композицию нескольких основ, средние части сложных слов могут нести достаточно большую информационную нагрузку (ср. рис. 37; Коженец*, 1973, с. 16).

48. Информационные оценки контекстных связей слова

Используя информационные оценки слова, а также понятие контекстной связанности, можно оценить те лексико-грамматические ограничения, которые предшествующий контекст накладывает на употребление слова в тексте.

Предположим, что в языке не существует комбинаторно-статистических ограничений и что все буквы и другие единицы обладают неограниченными возможностями сочетаемости. Мы можем определить то оптимальное количество информации, которое несет слово длиной в λ букв. Эта информация равна:

$$I_0^{(c)} = \lambda I_0 = \lambda H_0. \qquad (24)$$

В действительности слово несет гораздо меньше информации, что является, как уже указывалось, результатом разного вида статистических ограничений. Разность между оптимально возможным и реальным количеством информации, содержащимся в слове ($I^{(c)}$), мы будем называть общей контекстной связанностью слова ($K^{(c)}$). При этом

$$K^{(c)} = I_0^{(c)} - I^{(c)}. \qquad (25)$$

Общая контекстная связанность слова является функцией разных лингвостатистических ограничений. У внетекстового слова величина этой связанности характеризует сумму статистических ограничений, связанных с сочетаемостью букв, слогов и морфем.

Эту величину мы будем называть внутрисловной связанностью ($K^{(вс)}$). У текстовой словоформы, содержащей $I^{(ст)}$ дв. ед. информации, контекстная связанность ($K^{(ст)}$) представляет собой сумму внутрисловной связанности и той обусловленности, которую вносят лексико-грамматические связи между словами. Эту последнюю мы будем называть лексико-грамматической связанностью ($K^{(лг)}$) слова. Нетрудно показать, что

$$K^{(лг)} = I^{(с)} - I^{(ст)}. \qquad (26)$$

Все получаемые до сих пор измерения не могли служить непосредственными оценками смысловой информации, поскольку в эти измерения обязательно включались ограничения, связанные с комбинаторикой фигур (букв и слогов). Вычисление величины лексико-грамматической связанности свободно от этих ограничений. Заменяя в выражении (26) величину $I^{(с)}$ на неопределенность внетекстового слова $H^{(с)}$, а $I^{(ст)}$ на энтропию n-го слова в тексте ($H_n^{(ст)}$), нетрудно показать, что разность

$$K^{(лг)} = H^{(с)} - H_n^{(ст)} \qquad (27)$$

является оценкой той доли смысловой информации, несомой предшествующим контекстом, b^{n-1}, которая падает на n-е слово текста.

Попытаемся теперь определить, какую долю в общей связанности слова $K^{(с)}$ составляют статистические ограничения, вносимые лексико-грамматическими связями в предшествующем слову тексте b^{n-1}. Доля этих ограничений в процентах может быть определена из следующего выражения:

$$C = \frac{I^{(с)} - I^{(ст)}}{I_0^{(с)} - I^{(ст)}} 100\% = \frac{K^{(лг)}}{K^{(ст)}} \cdot 100\%. \qquad (28)$$

Соответственно доля внутрисловных статистических ограничений будет составлять:

$$C_w = 100 - C\%. \qquad (29)$$

Значения C и C_w для пяти индоевропейских языков показаны в табл. 44.

Таковы общие количественные оценки функциональной отдачи лексико-грамматического контекста. Теперь попытаемся выяснить некоторые детали воздействия указанного контекста на информационную структуру слова.

Первое, что бросается в глаза при сопоставлении текстовых (нижние кривые на рис. 36—38) и внетекстовых (верхние кривые на тех же рисунках) графиков, это не только значительное уменьшение общего количества информации у текстового слова, но и изменение общей конфигурации распределения. Обобщающие полигоны морфемного распределения информации (см. рис. 38) показывают, что лексико-грамматический контекст, срезая максимумы информации на первых буквах морфем, расположенных

Таблица 44

Доля ограничений, вносимых лексико-грамматическими связями между словами в тексте в общую контекстную связанность слова (по *I*)
(см.: Пиотровский, 1968, с. 90; Коженец*, 1973, с. 18; Бектаев*, 1975, с. 16)

Язык	C %	$C_{ю}$ %
Английский	35.0	65.0*
Немецкий	29.2	70.8
Румынский	27.1	72.9
Французский	23.5	76.5
Русский	22.5	77.5
Казахский (по данным К. Б. Бектаева)	18.0	82.0

* Аналогичные данные приведены в работе: Sommers, 1961.

в середине и конце слов, в значительной степени сглаживает границы между морфемами.

Сходную картину дают полигоны буквенных распределений (см. рис. 36—37). *U*-образные распределения информации, характерные для длинных русских и немецких слов, употребленных вне контекста, сменяются в контексте *J*-образной конфигурацией (аналогичную картину дают английский, французский и румынский языки). Превращение *U*-образных конфигураций в *J*-образные схемы связано с тем, что лексико-грамматический контекст срезает максимумы информации, расположенные в конце слова. А эти максимумы, как уже указывалось, концентрируют в себе основную часть грамматических характеристик, содержащихся в слове.

49. Информационные оценки морфологии

Начнем с того, что попытаемся оценить с помощью синтактической информации те морфологические сведения, которые содержатся в префиксах, суффиксах, внутренних флексиях и окончаниях слова. Что же касается информации, передаваемой словообразовательными аффиксами и средствами аналитической морфологии, то она в расчет приниматься не будет.

Грамматические сведения, содержащиеся в слове (точнее, в схеме слова определенной длины), оцениваются суммой синтактических информаций, падающих на каждую буквенную позицию в этой схеме.

Для определения грамматической информации, которую несет данная буквенная позиция, обобщенные в этой позиции буквы экспериментальных слов делятся на два разряда. В первый включаются те буквы, которые либо входят в состав грамматического аффикса (окончание, внутренняя флексия, приставка), либо хотя и не составляют грамматической части слова, но имеют альтернативой другую букву, входящую в грамматический аффикс данного или другого конкретного слова. Второй разряд составляют буквы, не входящие в грамматический аффикс и не имеющие грамматических альтернатив.

Буквы, входящие в первый разряд, группируются по количеству попыток, понадобившихся для их отгадывания. Все буквы второго разряда относятся к группе достоверных продолжений (угадывания с «нулевой» попытки) независимо от того, сколько попыток понадобилось, чтобы угадать каждую букву. Это делается из тех соображений, что «неграмматические» буквы, равно как и буквы первого разряда, угаданные с «нулевой» попытки, не несут грамматической информации. Полученное этим путем распределение вероятностей рассчитывается с помощью формул.

Данные о количестве синтактико-грамматической информации, содержащейся в русских внетекстовых и текстовых словах определенной длины, в усредненных внетекстовых и текстовых словах, а также сведения о среднем количестве грамматической информации, приходящемся на букву, даны в табл. 41, 42, 43. Аналогичные данные получены для английского и французского языков (Богуславская, Новак, 1968, с. 54—55; Петрова, Пиотровский, 1966, с. 120—122; Petrova, Piotrowski, Giraud, 1970).

Из всех этих данных видно, что, попадая в контекст, русская словоформа теряет в среднем 1.60 дв. ед. грамматической информации (считая по верхней границе). Это составляет около 69% от всей грамматической информации, содержащейся в формах внетекстового слова (для нижней границы информации эта величина составляет 67%).

Французское слово, попадая в текст, теряет 1.71 дв. ед. информации (считая по I'), что составляет около 66% грамматической информации внетекстового слова. Для английского языка эти величины соответственно равны 0.605 дв. ед. и 76%.

Уже много раз говорилось о том, что концы отдельно взятых слов, не говоря уже об окончаниях слов, находящихся на достаточном удалении от начала текста, распознаются идеальным угадчиком с помощью лексико-грамматических механизмов тезауруса и смысловой информации, извлеченной из текста. Учитывая это, можно предположить, что в русском, французском и, вероятно, в английском текстах значительная часть грамматических аффиксов является избыточной. Если это так, то грамматические значения, присущие этим аффиксам, должны передаваться с помощью других средств, в первую очередь с помощью служебных слов.

В этой связи было бы интересно рассмотреть с информационно-синтактической точки зрения то взаимодействие, в которое вступают в тексте служебные и знаменательные слова в целом, с одной

стороны, и служебные слова, основы знаменательных слов и их флексии — с другой.

То, что в ряде европейских языков служебные слова функционально эквивалентны флексиям и, наоборот, окончания знаменательных слов дублируют значения служебных слов, не подлежит сомнению. Об этом можно судить, в частности, по французской фразе типа nous ne le conaissons pas, которая легко преобразуется в разговорный вариант conaissons pas (Laffitte, 1951, с. 100), или по такому русскому примеру, как *завтра я выезжаю в Москву*, который в «телеграфном» стиле выглядел бы: *завтра выезжаю Москву*; ср. также польск. teraz (ja) mam do pana pytanie.

Всякий текст имеет линейный характер. Вместе с тем развертывание текста осуществляется на синтагматической оси, которая имеет либо временную (устная речь), либо пространственную (письменная речь) направленность. Поскольку служебные слова в рассматриваемых языках предшествуют на этой оси функционально эквивалентным им аффиксам знаменательных слов, нетрудно догадаться, что первые несут информационную нагрузку (в данном случае речь идет как о синтактической, так, вероятно, и о смысловой информации), в то время как вторые оказываются избыточными.

Об этих общеизвестных фактах приходится еще раз говорить по двум причинам.

Во-первых, в большинстве индоевропейских языков обнаруживается тенденция к развитию аналитизма и к ослаблению флексий. Во-вторых, при построении алгоритмов для автоматической переработки текста аналитический способ передачи грамматических отношений оказывается иногда более удобным, чем флективно-фузионная техника (ср. с. 286—287).

В этой ситуации интересно было бы рассмотреть информационно-синтактический механизм отношений, существующих между служебным и знаменательным словом, с одной стороны, и основой знаменательного слова и его окончанием — с другой.

Будем считать мерой связи данного лингвистического знака с предшествующим контекстом величину, обратную сумме информаций, падающих на первые две буквы воплощающей его цепочки

$$B = \frac{1}{I_1 + I_2}. \qquad (30)$$

Используя величину I для русского языка и I' для французского, получаем оценки C для служебных и знаменательных слов. а также для флексий знаменательных слов в тексте (см. табл. 45),

Приведенные данные показывают, что служебное слово в значительно меньшей степени зависит от предшествующего слова, чем флексия. Невелика степень предсказуемости предшествующим контекстом и самого знаменательного слова в целом. Это дает нам

Таблица 45

Оценки B для служебных, знаменательных слов и флексий слов в тексте

(Пиотровский, 1968, с. 94; Богуславская, 1969, с. 270)

Язык	Служебное слово	Знаменательное слово	Флексия знаменательного слова
Русский	0.21	0.19	0.46
Английский	0.22	0.17	0.50
Французский	0.24	0.19	0.66

право предполагать, что высокую степень обусловленности флексий следует относить не за счет всего предшествующего им контекста, а за счет лексической основы знаменательного слова. Иными словами, можно утверждать, что служебные слова благодаря своему положению на синтагматической оси обладают меньшей контекстной обусловленностью и одновременно несут больше синтактической информации, чем флексии.

Вполне возможно, что преимущества аналитической морфологии перед морфологией флективной обусловлены еще одной информационно-статистической особенностью служебных слов. Дело в том, что короткие слова, значительную долю которых составляют служебные слова, с точки зрения количества передаваемой ими информации оказываются менее подверженными воздействию контекста, чем средние и длинные слова: в русском языке короткие слова, попадая в контекст, теряют от 30 до 33% несомой ими информации, в то время как у длинных слов и слов средней длины контекст снимает от 47 до 58% информации (для французского языка эти величины соответственно равны 25—30 и 62—87%).

Чтобы разобраться в механизме этого явления, рассмотрим рост контекстной связанности в усредненной схеме текстового и внетекстового слова, а также ход этой обусловленности в связном тексте.

Как уже говорилось, ход кривой контекстной связанности (K_z) описывается зависимостью:

$$K_z = (I_0 - I_\infty)(1 - e^{-\beta z}), \qquad (16)$$

в которой I_∞ характеризует тот предел, к которому стремится информация данного типа сообщения. Следует подчеркнуть, что предельная информация связного текста, при $n \to \infty$, будет всегда больше нуля.

Иное дело слово. Слово состоит из фигур (букв, фонем, слогов) и простых знаков (морфем), обладающих ограниченной комбинаторикой, и — что самое главное — слово выступает в тексте в виде кванта информации. Поэтому при $n \to \infty$, т. е. при бес-

конечном удлинении слова, информация отдельных составляющих ее фигур и знаков будет стремиться к нулю.

В связи с этим выражение, описывающее рост контекстной связанности внутри слова, принимает вид:

$$K_{\xi}^{(c)} = I_0 - I_0 e^{-s\xi}. \qquad (31)$$

Сравнение распределений общей контекстной связанности в усредненной схеме текстового и внетекстового слов, проведенное на русском и французском материале, показало, что лексико-грамматический контекст значительно ускоряет рост контекстной связанности букв внутри слова. Коэффициент s текстового слова в два раза превышает по величине аналогичный коэффициент для нетекстового слова.

Особенно круто возрастает контекстная связанность в текстовом слове на участке от первой до четвертой букв. После четвертой буквы кривая связанности значительно приближается к своему пределу K_∞ (для русского слова $K_\infty = I_0 = 5$ дв. ед., для французского — $K_\infty = I_0 = 4.76$ дв. ед.).

Что же касается внетекстового слова, то здесь нарастание контекстных связей происходит более плавно. Кривая связанности достигает своего предела лишь после двенадцатой буквы (ср. рис. 39).

Быстрое нарастание общей контекстной связанности букв на участке между первой и четвертой буквами, — нарастание, постоянно наблюдаемое в процессе самого эксперимента, имеет важные последствия для информационной структуры текстового слова.

Как уже говорилось, основная часть грамматической информации длинных и средних слов концентрируется на пятой, шестой и т. д. буквах. Когда слова этого типа попадают в текст, общая контекстная связанность их начальных букв растет настолько быстро, что несущие грамматическую информацию конечные буквы, равно как и буквы, находящиеся в центральной части слова, оказываются почти полностью предопределенными предшествующим контекстом. Поэтому они теряют значительную часть своей грамматической информации.

Иное дело короткие слова. Контекстная связанность при употреблении их в тексте также растет довольно быстро. Однако она не успевает к концу слова достичь своего предела (максимальная длина коротких слов — четыре буквы). В связи с этим все буквы короткого слова еще сохраняют свой информационный вес. Поэтому короткие слова, основную массу которых составляют служебные слова, оказываются менее подверженными воздействию контекста по сравнению с длинными и средними словами, бо́льшая часть которых является знаменательными.

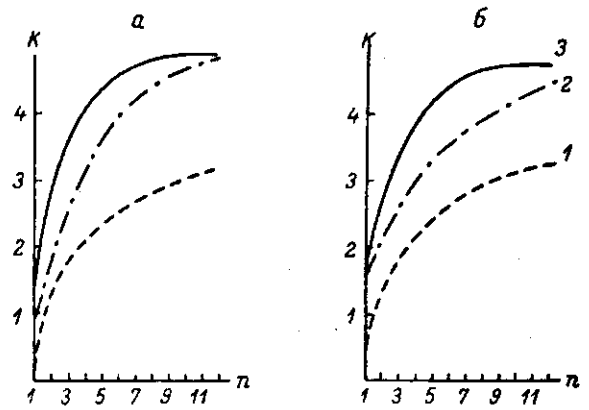


Рис. 39. Рост контекстной обусловленности в связном тексте и внутри слова в русском (*а*) и французском (*б*) языках (по оценкам I).

Условные обозначения: *1* — связный текст; *2* — внетекстовое слово; *3* — текстовое слово.

В свое время были предприняты попытки оценить числом степень аналитичности некоторых языков. Эти оценки были получены из выражения:

$$A = \frac{L_c}{L_{сф}}, \qquad (32)$$

где L_c — количество разных слов, встретившихся в исследованном корпусе текстов, а $L_{сф}$ — общее количество порожденных этими словами разных словоформ (табл. 46).

Как показывают данные табл. 46, английский, французский, испанский, румынский языки показывают более высокие оценки аналитизма, чем русский язык.

Однако эти оценки, хотя они и учитывают степень насыщенности текста флективными фонемами, еще ничего не говорят о функциональной отдаче этих форм. Эксперимент по угадыванию позволяет получить информационно-статистические информации, содержащиеся в этой схеме.

Процентные доли этой грамматической информации по оценкам $\bar{I}$ и $\bar{I}'$ показаны в табл. 47. Сравнение долей грамматической информации приводит к неожиданным результатам.

Информационная доля флективной морфологии во французском и румынском тексте в несколько раз превосходит информационный вес английских флексий. Более того, в аналитических французском и румынском языках указанная доля флективной морфологии примерно в два раза превышает информационный вес морфологических аффиксов синтетического русского языка.

Таблица 46

Коэффициенты аналитизма для индоевропейских и тюркских языков (Ешан*, 1966, с. 10; Киссен, 1972; Михайлова*, 1972, с. 6)

Язык	A
Английский	0.68
Французский	0.56
Испанский	0.56
Румынский	0.40
Русский	0.32
Узбекский	0.11

Таблица 47

Доля информации, передаваемая флективной морфологией, в общей сумме статистической информации, падающей на слово

Язык	Внетекстовое слово, в %	Текстовое слово, в %
Русский (по I')	17.0	8.6
Английский (по I')	6.3	3.5
Французский (по I')	22.0	16.0
Румынский (по I')	23.0	13.0

Объяснение *этим соотношениям следует искать, очевидно,* в следующих особенностях романских языков.

1. Глагольные парадигмы французского и румынского языков характеризуются значительным количеством флективных графических форм (ср. французские Conditionnel, Présent, Passé simple, Imparfait, Imparfait du Subjonctif и соответствующие им румынские времена) и превосходят в этом отношении не только английский, но и русский язык. Употребление романских глагольных окончаний не всегда предопределено формой существительного или местоимения, ср. фр. **je (tu)** chant**ais**, **nous** chant**ons** (chant**ions**); рум. **eu** cînt (cînt**am**, cînt**ai**).

2. Романские служебные слова, играющие большую роль в передаче грамматической информации, часто сами имеют флексии (ср. формы разных видов артикля: фр. le, l', la, les; un, une; рум. un, unui, unei). Что же касается русских и английских служебных слов, то они чаще всего оказываются неизменяемыми.

Отсюда следует, что романские языки характеризуются особым видом аналитизма, при котором служебные слова имеют разветвленные флективные парадигмы. Этот флективно-аналитический характер романского текста требует особого подхода при построении автоматического словаря оборотов (67) и при формировании морфологических алгоритмов (69).

Глава 7

ИЗМЕРЕНИЕ СМЫСЛОВОЙ ИНФОРМАЦИИ [1]

50. Измерение семантической информации в тексте, порождаемом закрытым языком

Проведенные в предыдущей главе информационные измерения позволили получить разнообразные количественные оценки синтактической информации. Эта информация выступает в качестве меры комбинаторно-статистического разнообразия и упорядочения лингвистических единиц в тексте в том виде, в каком это разнообразие и упорядочение отражается в лингвистическом сознании адресата, принимающего текстовые сообщения.

Однако для семиотики, лингвистики текста и прикладного языкознания важны измерения не только синтактической, но и смысловой (семантической, сигматической, прагматической) информации.

За последние десять лет интерес к смысловому аспекту информации неизменно рос, причем обсуждались как вероятностные, так и невероятностные подходы к измерению семантической и прагматической информации.

Впервые в более или менее завершенном виде концепция измерения смысла была сформулирована в начале 50-х годов Р. Карнапом и И. Бар-Хиллелом. Последующие семантические теории обычно развивали и усовершенствовали основные идеи этой концепции. Поэтому мы начнем ее рассмотрение применительно к лингвистическому материалу (см.: Bar-Hillel, Carnap, 1970, с. 18—23).

Строится искусственный закрытый язык L_n^π с конечным числом существительных («индивидуалов»), взаимооднозначно соотнесенных с определенными объектами и понятиями. В качестве примера таких индивидуалов можно указать на существительные *взлет, ВПП, (взлетно-посадочная полоса), посадка, радиосвязь, снижение, шасси* и т. п. в подъязыке «земля—воздух».

В символической записи индивидуалы образуют ряд $a_1, a_2 \ldots, a_i, \ldots, a_n$. Имеется также одна связка со значением «есть» (*является, имеет свойство*) и π прилагательных/причастий — одноместных предикатов, обозначенных символом $P_1, P_2 \ldots, P_j, P_\pi$. Каждое прилагательное имеет свое отрицание (ср. пары: *занятый—не занятый, разрешен—не разрешен, скользкий—не скользкий* и т. д.), в связи с чем образуется новый ряд предикатов

[1] *Настоящая глава написана на основе следующих работ:* Георгиев, Богодист, Коженец, Пестунова, Пиотровский, Райтар, 1972; Пиотровский, Байтанаева, Бектаев, Богодист, Пестунова, Георгиев, Райтар, 1973; Богодист, Георгиев, Пестунова, Пиотровский, Райтар, 1975.

$\bar{P}_1, \bar{P}_2 \ldots, \bar{P}_j \ldots, \bar{P}_\pi$. Наконец, используется четыре союза, указывающих на логические связи: «или» (дизъюнкция — $\cup$), «и» (конъюнкция — $\cap$), «если. . ., то» (импликация — $\supset$), «если и только если» (равнозначность — $\equiv$).

Сочетания прилагательных-предикатов и существительных-индивидуалов $P_j a_i$, $P_l a_k$ и т. д. (ср. *посадка разрешена*, *взлетно-посадочная полоса* (*ВПП*) *не занята*) дают атомарные предложения, из которых с помощью союзов можно строить более длинные предложения например: *если ВПП занята, то посадка не разрешена* (в символах это можно записать так: $P_l a_k \supset P_j a_i$).

Вводится понятие описания состояния D, которое представляет собой предложение, состоящее из конъюнкций n атомарных предложений и включающее каждое из n существительных-индивидуалов, сочетающихся с π прилагательных-предикатов или их отрицаниями (таким образом, в конъюнкцию может входить либо само атомарное предложение $P_j a_j$, либо его отрицание $\bar{P}_j a_j$, но не оба вместе). Число состояний конечно, оно составляет в нашем искусственном языке L_n^π всего $2^{\pi n}$ возможных состояний. Так, например, если имеется язык L_2^2, состоящий только из двух существительных (*ВПП*, *посадка*), двух прилагательных (*разрешен*, *о*, *а*; *занят*, *о*, *а*) и одной связки, то мы будем иметь в языке L_2^2

$$2^{2\cdot 2} = 16$$

возможных описаний состояний (ср. табл. 48), из которых только некоторые (D_1, D_2, D_3, D_4) имеют смысл, описывая такие состояния, которые могут иметь место в реальной действительности. Таким образом, в каждом языке L_n^π имеются истинные и ложные (противоречивые) предложения, соответственно описывающие логически истинные и логически ложные (невозможные) состояния. Вся эта процедура позволяет давать формальное представление конечному числу состояний (ситуаций), полностью описывающих некоторый участок объективной действительности (назовем его системой) и одновременно указывает, какие из предложений, входящих в эти описания являются истинными, а какие ложными. Однако эта процедура еще не дает возможности получать количественные оценки информации. Для того чтобы получить такие оценки, вводится числовая функция меры m, применяемая как к описаниям состояний, так и к предложениям. Эту функцию меры Карнап называет абсолютной логической вероятностью. Функция m обладает следующими свойствами.

1. Для каждого описания состояния D_k имеется логическая вероятность $0 \leqslant m(D_k) \leqslant 1$, которую авторы рассматривают как априорную вероятность описания состояния D_k.

2. Сумма $\sum_{k=1}^{\pi n} m(D_k) = 1$.

Таблица 48

Описание состояний в языке L_2^2

Код описания	Описание ситуации
D_1	*Взлетно-посадочная полоса занята и посадка разрешена;*
D_2	*Взлетно-посадочная полоса занята и посадка не разрешена;*
D_3	*Взлетно-посадочная полоса не занята и посадка разрешена;*
D_4	*Взлетно-посадочная полоса не занята и посадка не разрешена;*
D_5	*Взлетно-посадочная полоса разрешена и посадка разрешена;*
. . . .	. . . . .
D_9	*Взлетно-посадочная полоса занята и посадка занята;*
. . . .	. . . . .
D_{16}	*Взлетно-посадочная полоса не занята и посадка не занята.*

3. Если некоторое предложение A, входящее в описание состояния, является ложным, то вероятность $m(A)=0$.

4. Если предложение A из языка L_n^π не является ложным, то его абсолютная логическая вероятность $m(A)$ определяется как сумма $m(D_k)$ по всем описаниям состояний D_k, в которых выполняется предложение A.

В приведенных выше лингвистических примерах предложения *ВПП разрешена*, *ВПП не разрешена* являются ложными, поэтому их абсолютная логическая вероятность будет равна нулю. Напротив, предложение *ВПП занята* является неложным, поэтому априорная вероятность будет равна сумме вероятностей всех описаний состояний, в которых это предложение присутствует. Пользуясь данными табл. 48, легко показать, что m (*ВПП занята*)$=m(D_1)+m(D_2)+m(D_9)+\ldots$

По аналогии со статистической теорией информации информация предложения A будет равна:

$$I(A) = -\log_2 m(A) = -\log_2 \sum m(D_k), \qquad (1)$$

где сумма берется по всем D_k.

Методика семантических измерений Карнапа—Бар-Хиллела может быть успешно применена к таким коммуникативным системам, которые состоят из фиксированного числа семиотических (языковых) измерений, каждое из которых взаимооднозначно соотнесено с неязыковым измерением таким образом, что выбор некоторой точки (сигнала) в языковом измерении задает некоторую точку в соответствующем неязыковом измерении (объект реальной действительности). Этим условиям отвечают искусственные коммуникативные системы типа рассмотренного выше языка «земля—воздух», а также «языки» животных. Однако процедура Карнапа—Бар-Хиллела не может быть применена к естествен-

ному языку, который, являясь открытой системой, порождающей бесконечное число текстовых состояний, предусматривает многомногозначные отношения между точками языкового и неязыкового измерений. Однако и при измерении семантики закрытых языков применение процедуры Карнапа—Бар-Хиллела наталкивается на ряд технологических трудностей.

Во-первых, используемый здесь язык слишком беден, чтобы с его помощью можно было бы формализовать и привести к атомарным предложениям неэлементарные фразы типа *берите посадку на себя, когда увидите ВПП* из уже упоминавшегося языка «земля—воздух».

Во-вторых, неясны те критерии, с помощью которых можно присваивать априорные вероятности отдельным предложениям и состояниям.

Попытка преодолеть некоторые из указанных технологических и теоретических трудностей и тем самым приспособить концепцию семантической информации Карнапа—Бар-Хиллела к измерению смысла в текстах естественного языка сделана в работах Ю. А. Шрейдера (1967, с. 39; 1974, с. 6—10).

С одной стороны, автор расширяет формальный язык L_n^π, используя в нем не только одноместные, но и многоместные предикаты, вводя наряду с объектами (индивидуалами) понятие события, а также применяя дополнительные логические связи (союзы).

С другой стороны, — и это самое главное — семантическая информация истолковывается и измеряется не относительно самого языка и вероятностей появления его единиц, как это имело место у Карнапа—Бар-Хиллела, но по отношению к тому запасу знаний о внешнем мире, которым располагает получатель сообщения.

Этот запас знаний описывается с помощью тезауруса θ, который представляет собой модель, включающую конечное множество состояний внешнего мира, каждое из которых формулируется в терминах искусственного языка. В отличие от модели Карнапа—Бар-Хиллела тезаурус предполагается строить в виде нежесткой системы, допускающей расширение и пополнение под воздействием перерабатываемого с ее помощью текста T_i (в качестве текста может рассматриваться любой отрезок речи — предложение, словосочетание, словоупотребление, морфема).

Тезаурус должен строиться на основе языка L_n^π, причем объекты (а также события) и предикаты распределяются на множество смысловых категорий, так что каждое слово из L_n^π оказывается отнесенным к тому или иному подмножеству, имеющему свой номер. Тогда смысловые связи между словами характеризуются набором номеров смысловых рубрик. То же самое можно сделать и по отношению к более крупным лингвистическим единицам — словосочетаниям, предложениям и т. д.

Если бы удалось построить только что описанный тезаурус, то мы получили бы семантическое пространство, на которое могла бы быть наложена метрика. Действительно, определив число

смысловых связей в θ, можно количественно оценить объем заключенной в нем информации.

Чтобы получить возможность определять количество информации, содержащейся в тексте T_i, относительно тезауруса θ, вводится понятие алгоритма смыслового анализа текста. Этот алгоритм фиксирует систему правил (операторов) $A\ (T_i)$, пользуясь которыми предполагается приводить реальные тексты к конечному числу канонических высказываний, допускаемых и воспринимаемых тезаурусом. Если канонизированный таким образом текст T_i вносит в тезаурус θ некоторое число новых смысловых связей, увеличивая тем самым его объем и переводя θ в новое состояние θ', то количество семантической информации $I\ (T_i, \theta)$, содержащейся в тексте T_i, оценивается как разность

$$\theta' - \theta = I\ (T_i,\ \theta).$$

Процедура Шрейдера может быть использована для измерения информации, содержащейся в компрессированных и закрытых лингвистических системах типа машинной грамматики или автоматического словаря (ср.: Вертель, Вертель, Крисевич, Пиотровский, Трибис, 1971, с. 34—51).

Что же касается измерения смысловой информации в текстах естественного языка, то здесь процедура Шрейдера встречает ряд затруднений теоретического и технологического порядка.

Во-первых, остается неясным, что следует считать стандартным тезаурусом носителя языка и как, преодолевая антиномию синхронии и диахронии, фиксировать исходное состояние θ этого тезауруса. Во-вторых, нет уверенности в том, что можно в обозримый срок создать искусственный формализованный язык, который, будучи более богатым, чем естественный язык, мог бы служить средством исчерпывающего описания и измерения семантики естественного языка (13). Наконец, остаются неясными лингвистические правила задания операторов, различения объектов, событий и т. п.

Однако, несмотря на эти неясности, лингвистическая ценность концепции Ю. А. Шрейдера состоит в том, что здесь четко сформулирована идея не прямого, но косвенного измерения семантики текста, — измерения, осуществляющегося через количественную оценку тех измерений и реакций, которые обнаруживает языковая система под воздействием перерабатываемого ею текста.

51. Прагматическая информация и ее измерение в тексте, порождаемом закрытым языком

Язык функционирует в социальной среде, поэтому реакции лингвистической системы на текст измеряют не только семантическую информацию, но включают также оценку полезности со-

общения для принимающего его носителя языка или коллектива носителей. На это обстоятельство обращали внимание многие авторы, в частности Р. Уэлз, который ввел важное для лингвистики понятие «информативность сообщения», а также «метрический объем информативности», включающие, очевидно, как саму семантику истинных предложений из языка L_n, так и ту эвристичность, полезность, которые определяются следствиями, вытекающими из этих высказываний (Wells, 1961/1965, с. 175).

Оставляя в стороне не находящий применения в лингвистике теоретико-игровой подход (Гавурин, 1963, с. 27—34), а также произвольное использование «весов полезности» в шенноновской методике (Guiaşu, 1971, с. 165—179), обратимся к работам Я. Хинтикки и Р. Хилпинена (Hintikka, 1970, с. 3—28; Hilpinen, 1970, с. 114), пытающихся получить оценки полезности информации (прагматическую информацию $I_{\text{пр}}$) из сравнения семантической информации Карнапа—Бар-Хиллела с опытной статистической информацией:

$$I_{\text{пр}} = \inf(A) - I(A/E) = \log_2 p' - \log_2 p,$$

где inf (A) — информация предложения A в языке L_n, имеющего априорную вероятность p; $I(A/E)$ — статистическая информация, полученная из вероятности p', которая, в свою очередь, добыта путем наблюдения E над появлением события, описываемого предложением A.

Полезность информации будет положительной, если $p'_i > p_i$; она будет равна нулю, если $p'_i = p_i$, и эта полезность будет отрицательной, если $p'_i < p_i$ (в последнем случае речь идет о дезинформации).

Подобно концепции Карнапа—Бар-Хиллела процедура Хинтикки—Хилпинена применима лишь к закрытым языкам, дающим конечное число высказываний. Измерить прагматическую информацию в естественном языке, порождающем бесконечное число высказываний, с помощью описанного метода нельзя.

Более гибкую процедуру, не требующую измерения семантической информации, сформулировал в свое время А. А. Харкевич (1960, с. 53—57), предлагавший оценивать полезность информации в рамках шенноновского подхода при условии точной фиксации цели (задачи) деятельности получателя сообщения. Ценность информации определялась им как приращение вероятности решения задачи после получения некоторого сообщения. Иными словами,

$$I_{\text{пр}} = -\log_2 \frac{p}{p'} = \log_2 p' - \log_2 p.$$

Если исходы опыта, направленного на решение задачи, равновероятны, то

$$I_{\text{пр}} = \log_2 S_0 - \log_2 S_1.$$

Здесь S_0 — число исходов опыта до получения информации, а S_1 — сокращенное в результате получения информации число исходов.

Уточняя процедуру А. А. Харкевича, М. М. Бонгард (1963, с. 71) показал, что изменение трудности решения задачи зависит не только от полученного сообщения, но и от того, как к этому сообщению относится сам получатель. Поэтому предлагается учитывать запас полезной информации, содержащейся в тезаурусе приемника относительно данной задачи, и рекомендуется различать неопределенность самой задачи и ее неопределенность для конкретного получателя. Эта последняя может быть равна или быть больше неопределенности самой задачи, т. е.

$$H(p/q) \geqslant H(p).$$

Здесь $H(p/q)$ — неопределенность задачи с истинным распределением верностей ответа

$$p_1, p_2, \ldots p_i, \ldots, p_s \tag{2}$$

при условии, что наблюдатель исходит из гипотезы, согласно которой указанное распределение равно

$$q_1, q_2, \ldots q_i, \ldots, q_s. \tag{3}$$

Величина $H(p)$ представляет собой истинную неопределенность задачи, определяемую распределением (2).

При этих условиях незнание задачи будет равно разности

$$H(p/q) - H(p),$$

которая будет равна нулю в том случае, когда распределения (2) и (3) будут тождественны.

Информация, полученная человеком в новом сообщении a', может уменьшить незнание задачи. В этом случае возникает новое распределение вероятностей:

$$q'_1, q'_2, \ldots, q'_i, \ldots, q'_s,$$

более близкое к распределению (2), чем распределение (3). Полезность этой информации вычисляется как разность исходной неопределенности $H(p/q)$ задачи и неопределенности $H'(p/q')$, образовавшейся после получения дополнительного сообщения, т. е.

$$I_{\text{пр}}(a') = H(p/q) - H'(p/q').$$

При всей изящности отдельных ее решений процедура Харкевича—Бонгарда, подобно другим рассмотренным выше концепциям семантической и прагматической информации, оставляет открытым вопрос о том, как осуществлять количественную оценку вероятности достижения цели или решения задачи живым приемником информации и как измерить изменения этой вероятности под влиянием того или иного речевого сообщения.

Только что проанализированные подходы предусматривают измерение смысловой информации относительно закрытых, нединамических систем с конечным числом состояний. Хотя язык представляет собой открытую, динамическую систему, порождающую бесконечное число текстов, рассмотренные приемы смысловых измерений представляют большой интерес и для лингвистики. Во-первых, с помощью их могут осуществляться измерения семантики и прагматики внутри «закрытых» компрессированных лингвистических систем типа языков «земля—воздух», «земля—вода» или словарей и тезаурусов, предназначенных для нужд машинного индексирования, аннотирования и перевода. Во-вторых, опираясь на аппарат указанных приемов, можно попытаться построить такую методику измерения смысла, которая была бы применима к тексту, написанному на естественном языке.

52. Измерение семантической, сигматической и прагматической информации в тексте естественного языка

Система прагматических оценок собственной деятельности, а также семантическая система естественного языка и ее функционирование в мозгу человека продолжают оставаться «черным ящиком», доступ к которому практически закрыт. Поэтому при построении работающей процедуры измерения смысла текста нужно отказаться от попыток предварительного количественного описания всей семантической системы языка или создания системы оценок полезности различных житейских задач и ситуаций, описываемых в речи. Измерение смысла текста должно опираться на количественные оценки реакций лингвистического сознания живого приемника информации на предъявляемые ему речевые сообщения без того, чтобы пытаться измерить семантику языка или лингвистический тезаурус приемника в целом.

Этим требованиям отвечает тест по угадыванию букв, использовавшийся для измерений синтактической информации. С одной стороны, эта методика не требует измерения системы языка, а также тезауруса угадчика. С другой стороны, оказывается, что уже при угадывании четвертой, пятой и т. д. букв текста испытуемый опирается не на комбинаторику фигур (букв, слогов), но на смысл уже известного ему отрывка и на ситуацию (ср. 43).

На этих особенностях угадывания строится следующая процедура измерения смысловой (семантико-прагматической и сигматической) информации. Имеется осмысленный текст, состоящий из цепочки слов w_1, w_2, w_3, . . ., w_k. Необходимо количественно оценить смысловую информацию, содержащуюся в слове w_1. Для этого выделим слово w_1 из текста, а оставшийся сегмент $w_2 \div w_k$ подвергнем дважды коллективному угадыванию. Первый раз будем угадывать наш сегмент, не сообщая испытуемым слова w_1. Второй раз проведем угадывание того же сегмента, предварительно сообщив испытуемым, что перед ним стоит слово w_1.

Приступая к угадыванию сегмента $w_2 \div w_k$, угадчики в обоих случаях имеют дело с разной неопределенностью (энтропией). Если слово w_1 известно, то неопределенность угадывания меньше неопределенности сегмента $w_2 \div w_k$ при неизвестном w_1. Как известно, снятая в результате угадывания энтропия H сегмента $w_2 \div w_k$ количественно равна той синтактической информации I, которую получил коллектив угадчиков от указанного сегмента. Отсюда следует, что информация первого угадывания при неизвестном w_1 — $I\,(w_2 \div w_k)$ больше информации $I\,(w_2 \div w_k/w_1)$, извлекаемой при втором угадывании, когда w_1 известно испытуемым. Разность же

$$И\,(w_1) = I\,(w_2 \div w_k) - I\,(w_2 \div w_k/w_1) \qquad (4)$$

представляет собой количественную оценку той прогнозирующей информации, содержащейся в слове w_1, которая облегчила второе угадывание сегмента $w_2 \div w_k$. Аналогичным образом измеряется прогнозирующая информация, содержащаяся в словосочетании $w_1 \div w_2$ или вообще в текстовой цепочке $w_1 \div w_{n-1} = b^{n-1}$. В первом случае эта информация будет оцениваться с помощью диагностирующего сегмента $w_3 \div w_k$, а во втором на контексте $w_n \div w_t$.[1]

Опыт показывает, что сегмент длиной в четыре-пять слов оказывается достаточным для полной реализации правых прогнозирующих возможностей оцениваемого слова или словосочетания. Так, например, для измерения смысла союза *чтобы* использовались результаты двойного угадывания следующего за ним сегмента *успешно развивать наши промыслы. . .*, а значение словосочетания *органы печати* количественно оценивалось на идущем за ним сегменте *с беспокойством отмечают, что страна. . .* (оба примера взяты из газеты «Правда», 1966, № 215 и 1969, № 3).

Первое коллективное угадывание сегмента *успешно развивать наши промыслы. . .* (то, что перед этим сегментом стоит союз *чтобы*, испытуемым неизвестно) дало 54, 58 дв. ед. синтактической информации, второе угадывание (с известным *чтобы*) принесло 47, 59 дв. ед. Отсюда следует, что

$$И\,(чтобы) = 54.48 - 47.59 = 6.89 \text{ дв. ед.}$$

Аналогичным образом имеем:

$$И\,(органы\ печати) = 58.40 - 44.84 = 13.56 \text{ дв. ед.}$$

По описанной методике измерялся смысл 42 русских, 34 французских, 9 болгарских, 2 немецких, 8 эстонских слов и словосочетаний. Угадывание осуществлялось на материале современных газетных текстов.[2]

[1] Если информация $И_{(w_1)}$ будет отрицательной величиной, т. е. $I\,(w_2 \div w_k/w_1) > I\,(w_2 \div w_k)$, то слово w_1 содержит не полезную информацию относительно сегмента $w_2 \div w_k$, но дезинформацию, равную по величине $|И\,(w_1)|$.

[2] Эксперимент по русскому языку осуществлен В. Н. Пестуновой и частично автором; угадывание французских текстов проведено В. И. Бого-

Прежде чем перейти к лингвистической интерпретации результатов, попытаемся рассмотреть природу получаемых нами информационных оценок.

Как уже говорилось, каждый текст создается в результате взаимодействия двух величин: во-первых, системы языка с ограничивающими ее нормой и узусом и, во-вторых, независимой от языка ситуации (совокупности объектов внешнего мира), которая должна быть описана текстом. Отсюда следует, что измеряемые нашим экспериментом прогнозирующие способности, заключенные в слове w_1, включают в себя разные виды информации.

Во-первых, они включают семантическую информацию, представляющую отношение между словом w_1 и обозначаемым им понятием (3).

Во-вторых, они, очевидно, содержат сигматическую информацию, характеризующую отношение между словом w_1 и обозначенным им конкретным объектом. Об этом свидетельствует, в частности, и тот факт, что разные актуальные (текстовые) значения слова показывают разные величины смысловой информации, при этом более специальные (терминологические) значения несут, вероятно, больше информации, чем значения более общего характера (ср. табл. 49, 51, 52).

Как семантическая, так и сигматическая информация, содержащаяся в слове w_1, подсказывает информантам общую стратегию поиска наиболее вероятных лексических продолжений в диагностирующем сегменте $w_1 \div w_k$.

В-третьих, в них заложены сведения о закрепленных нормой и узусом, а также задаваемых системой правых лексико-грамматических валентностях слова w_1. Неслучайно, поэтому, что получаемые в результате осуществления отдельного эксперимента величины информации зависят от вида того контекста $w_2 \div w_k$, на котором измеряется значение слова w_1. Если лексико-грамматические связи слова w_1 действуют достаточно далеко вправо, то эксперимент обнаружит у слова w_1 бóльшую смысловую информацию. Если же правые валентности имеют малую глубину, то эксперимент покажет у w_1 заметно меньшую смысловую нагрузку (ср. табл. 50).

Наконец, результаты нашего эксперимента включают оценку прагматики испытуемых. Действительно, смысловая стратегия угадывания и правые валентности w_1 выбираются испытуемыми исходя из их жизненного и лингвистического опыта. Каждый угадчик или коллектив угадчиков выбирает со своей точки зрения наилучшие для достижения цели (т. е. для расшифровки текста)

дистом с учащимися одного из парижских лицеев и французами-учителями; болгарская часть эксперимента осуществлена Х. Ц. Георгиевым в Софии, казахская — К. Б. Бектаевым в Алма-Ате, эстонская — С. В. Райтар в Тарту, эксперимент по немецкому языку проведен в Минске Т. Ф. Коженец со стажерами из ГДР.

Таблица 49

Смысловая и синтактическая информация русских слов и словосочетаний по данным В. Н. Пестуновой (Георгиев, Богодист, Коженец, Пестунова, Пиотровский, Райтар, 1972, с. 9—11)

№ пп.	Контрольное слово или словосочетание	$I(w_2 \div w_k)$	$I(w_2 \div w_k/w_1)$	$H(w_1)$	$I(w_1)$
1	печать	47.44	31.70	15.74	13.29
2	комментатор	52.73	35.44	17.29	15.01
3	корреспондент	50.70	34.82	15.88	13.38
4	многотиражка	57.73	42.79	14.94	15.74
5	«Юманите»	50.47	35.68	14.79	11.69
6	редактор	46.88	33.78	13.10	13.57
7	рабкоры	56.84	40.94	15.90	13.47
8	материалы	46.72	33.75	12.97	16.33
9	пропагандистам	58.64	42.67	15.97	22.23
10	рупор	51.02	37.82	13.20	11.95
11	еженедельник	53.02	40.46	12.56	13.92
12	коммюнике	56.46	44.25	12.21	14.18
13	сообщения	61.28	48.35	12.93	13.98
14	правдисты	51.40	40.42	10.98	15.62
15	обозреватели	52.72	42.00	10.72	22.76
16	репортеры	65.86	53.04	12.82	14.58
17	пресса	49.06	39.80	9.26	16.06
18	«Правда»	54.66	44.35	10.31	13.46
19	ТАСС	62.73	51.72	11.00	13.82
20	газета	—	—	5.27	—
21	газеты	49.64	40.86	8.78	13.66
22	писанина	51.41	42.29	9.12	13.93
23	журналисты	59.40	49.53	9.87	14.34
24	телевидение	62.31	51.57	10.74	16.07
25	«Известия»	65.63	54.72	10.91	15.50
26	наблюдатель	53.11	44.69	8.42	16.73
27	радиостанция	49.35	41.38	7.97	19.10
28	читатели	45.79	38.93	6.86	13.79
29	периодика	60.55	52.39	8.16	17.91
30	документ	47.09	40.79	6.30	15.26
31	газетчики	54.46	47.14	7.32	12.22
32	журнал	53.61	47.33	6.28	11.67
33	пропаганда	47.55	42.23	5.32	19.61
34	агентство	54.49	48.85	5.64	14.22
35	автор	70.77	65.28	5.49	13.80
36	номер	65.24	61.25	3.81	12.42
37	шутиха	63.67	59.81	3.86	12.82
38	центральные (газеты)	45.92	33.13	12.79	13.23
39	зарубежная (печать)	34.87	26.25	8.62	15.84
40	французская (печать)	51.90	42.68	9.22	15.39
41	органы печати	58.40	44.84	13.56	21.15
42	президент	—	—	8.44	—
43	народ	—	—	8.85	—
44	правительство	—	—	13.36	—
45	государство	—	—	17.66	—
46	класс	—	—	18.33	—
47	эксплуатация	—	—	19.84	—
48	я	38.35	32.62	5.73	4.22
49	чтобы	54.48	47.59	6.89	8.12

Таблица 50

Оценки смысловой информации слов, полученные в разных текстах и в разных коллективах угадчиков.
По данным В. Н. Пестуновой (см.: Bogodist, Guéorguiev, Pestunova, Piotrowski, Raitar, 1974, с. 37)

№ пп.	Слово (w_1)	Контекст ($w_2 \div w_k$)	Количество слов, испытывающих влияние слова w_1	H (w_1) Студенты	H (w_1) Учащиеся техникумов
1	агентство	а) *сообщает, что* южновьетнамские войска	2	12.26	10.44
		б) *описало* праздично украшенные улицы столицы	1	5.64	—
2	«Правда»	а) *рассказала вчера о социалистических обязательствах*	5	10.31	8.05
		б) *напечатала* письмо ревизора финансовой службы	1	2.93	—
3	правительство	а) *представит в распоряжение НАТО* свои. . .	4	13.36	—
		б) *объявило девальвацию* франка	2	3.88	—
4	печать	*широко комментирует опубликованное* заявление	3	15.74	9.77

Примечание. Курсивом отмечены те слова, на которые распространяются правые лексико-грамматические валентности слова w_1.

продолжения. Поэтому смысловые измерения w_1 включают в себя также оценки полезности слова w_1 для угадчика или коллектива. Об этом можно судить, в частности, по результатам измерений смысла одних и тех же слов, произведенных на одинаковых контекстах $w_2 \div w_k$, но осуществленных относительно разных коллективов угадчиков. Как видно из табл. 50 учащиеся техникумов, обладающие по сравнению со студентами более низкой лингвистической культурой и меньшим житейским опытом, извлекают из контрольных слов меньше смысловой информации, чем студенты вуза.[1]

[1] Хотя угадчики, обладающие богатым тезаурусом, извлекают из отдельного слова больше смысловой информации, чем угадчики, располагающие более бедным тезаурусом, соотношение величин синтактической информации на слово здесь обратное. Эксперимент, осуществленный В. Н. Пестуновой, показал, что студенты извлекают в среднем из русского слова 11.36 дв. ед. смысловой информации и 14.46 дв. ед. синтактической информации (энтропии), для учащихся техникумов это соотношение равно 9.75 и 15.03 дв. ед. Это и понятно: малообразованный угадчик плохо ориентируется в контексте и поэтому с большим трудом и неопределенностью угадывает неизвестное для него слово, извлекая из него одновременно меньше смысловой информации, чем более образованный информант. Сходную картину дает французский материал. Эксперимент с учащимися лицея оценивает значение слов ministre 'министр' в 6.04 дв. ед., а опыт, проведенный в группе учителей, показывает для этого слова уже 9.53 дв. ед. смысловой информации, для слова président имеем соответственно 7.44 и 12.46 дв. ед. (Богодист*, 1974, с. 16).

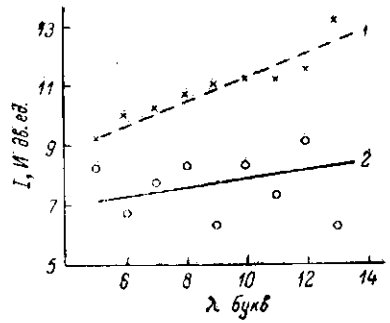


Рис. 40. Зависимость синтактической (I) и смысловой (H) информации французского слова от его длины (λ).

Условные обозначения: *1* — синтактическая информация; *2* — смысловая информация.

Из всего сказанного следует, что величина смысловой информации, получаемой от отдельного текстового слова или словосочетания, определяется его виртуальным и актуальным значениями, конкретной правой дистрибуцией и житейско-лингвистическим опытом угадчика. Чтобы получить количественную оценку значения слова или словосочетания относительно всего языка или его разновидности, необходимо, очевидно, с одной стороны, провести измерения разных текстовых значений в достаточно представительном количестве разных контекстов. С другой стороны, нужно освободить нашу смысловую оценку от прагматики отдельного угадчика или необходимо провести эксперимент в достаточно представительном множестве коллективов, которое обеспечило бы усреднение индивидуальных прагматик до некоторого эталона стандартного носителя языка.

Получаемые с помощью описанного эксперимента оценки смысловой информации могут быть использованы при решении некоторых теоретических и прикладных вопросов. Например, сопоставление по разным языкам слов, имеющих одно и то же (или очень близкое) значение, употребленных в сходных контекстах и выполняющих одну и ту же синтаксическую функцию, показывает, что в аналитических языках слово, очевидно, несет меньшую смысловую нагрузку, чем в языках синтетических (флективных и агглютинирующих) (см. табл. 53). Это и понятно, аналитическое слово лишь частично включает в свою смысловую структуру грамматические значения (в данном случае в него включается артикль). Смысловая структура флективного или агглютинативного слова богаче: она имеет в своем составе наряду с лексической основой несколько морфем, указывающих на грамматические отношения

Таблица 51

Смысловая и синтактическая информация французских слов и словосочетаний по данным В. И. Богодиста
(см.: Георгиев, Богодист, Коженец, Пестунова, Пиотровский, Райтар, 1972, с. 12—14; Богодист*, 1974, с. 21)

№ пп.	Контрольное слово или словосочетание	$I(w_2 + w_k)$	$I(w_2 + w_k/w_1)$	$H(w_1)$	$I(w_1)$
1	l'exploitation 'разработка, добыча'	50.96	38.88	12.08	15.54
2	l'exploitation 'эксплуатация, угнетение'	43.88	34.27	9.55	15.00
3	l'exploitation 'участок земли'	38.80	32.53	6.27	15.96
4	la classe 'класс' (социальный)	39.26	30.00	9.26	12.95
5	la classe 'классная комната'	38.39	28.70	9.69	12.65
6	la classe 'урок'	39.64	32.21	7.43	12.70
7	l'armistice 'перемирие'	46.56	36.56	10.00	13.96
8	la sécession 'отпадение, отделение'	47.48	37.17	10.31	14.53
9	le régime 'режим' (реки, работы)	46.64	37.40	9.24	14.72
10	le régime 'государственный строй'	32.60	25.93	6.67	13.25
11	le régime 'условия жизни'	52.09	46.82	5.27	14.48
12	la moratoire 'мораторий'	53.22	43.45	9.77	16.57
13	la défense 'защита' (юридическая)	47.73	38.25	9.48	15.44
14	le défense 'защита'	47.83	40.81	7.02	15.08
15	la défense 'запрет'	55.93	46.38	9.55	15.43
16	l'état 'состояние'	55.21	45.69	9.52	13.31
17	l'état 'государство'	49.19	42.17	7.02	14.05
18	le président 'президент'	42.82	45.00	7.68	14.41
19	le président 'президент'	42.45	35.16	7.29	14.80
20	le président 'председатель'	42.64	37.49	5.15	14.10
21	le gouvernement 'тип государственного управления'	36.54	30.62	5.92	13.10
22	le gouvernement 'правительство'	35.88	30.44	5.44	13.78
23	le gouvernement 'управление'	39.21	34.53	4.68	13.91
24	le peuple 'народность'	48.87	40.09	8.78	12.76
25	le peuple 'племя'	36.87	31.04	5.83	12.20
26	le peuple 'народ'	44.31	37.48	6.83	12.20
27	le ministre 'министр'	39.11	33.24	5.87	13.95
28	le ministre 'министр'	39.01	32.79	6.30	13.23
29	le ministre 'министр'	43.65	37.70	5.95	13.23
30	l'ordre 'государственный порядок'	44.54	38.47	6.07	14.52
31	l'ordre 'порядок (в квартире)'	50.85	43.54	7.34	13.52
32	l'ordre 'приказ'	54.05	47.37	6.68	14.03
33	la délégation 'делегация'	47.75	40.93	6.82	13.67

Таблица 52

Смысловая и синтактическая информация болгарских, немецких, казахских и эстонских слов
(см.: Георгиев, Богодист, Коженец, Пестунова, Пиотровский, Райтар, 1972, с. 15; Коженец*, 1973, с. 23; Пиотровский, Байтанаева, Бектаев, Богодист, Георгиев, Пестунова, Райтар, 1973, с. 720)

Язык	Контрольное слово или словосочетание и его значение	$I(w_2 + w_k)$	$I(w_2 + w_k/w_1)$	$H(w_1)$	$I(w_1)$
Болгарский	коментаторите	52.16	33.33	18.83	16.08
	правителството	51.95	34.21	17.74	15.51
	защитата	51.23	35.84	15.39	12.19
	кореспондентът	38.21	26.06	12.15	15.79
	режимът	45.70	34.40	11.30	14.45
	председателят	32.35	25.63	6.72	18.70
	печатът	32.50	27.65	4.85	15.18
	журналистите	33.09	30.70	2.39	13.89
Немецкий	der Krieg 'война'	50.28	45.23	5.05	—
	der Mann 'человек'	52.54	50.31	2.23	—
Казахский	газет 'газета'	—	—	10.64	7.48
	білу үшін 'чтобы знать'	—	—	10.64	13.34
	мен 'я'	—	—	5.88	9.24
Эстонский	riik 'государство'	79.85	59.22	20.63	—
	klass 'класс'	83.77	65.48	18.29	—
	rahvas 'народ'	76.32	58.69	17.63	—
	ekspluateerimine 'эксплуатация'	52.18	35.04	17.14	—
	valitsus 'правительство'	72.85	55.75	17.10	—
	president 'президент'	36.80	24.11	12.69	—

этого слова в предложении. Отсюда следует, что при построении алгоритмов автоматической переработки текстов языков фузионного и агглютинативного строя основное внимание должно быть обращено на анализ и синтез слова, в то время как в языках с сильно развитым аналитизмом значительное внимание должно быть обращено на автоматический анализ контекста (ср. 67).

Сравнение изменений величин смысловой и синтактической информации (см. табл. 49, 51 и рис. 40) в зависимости от длины слова показывает, что если между длиной слова и несомым им количеством синтактической информации отчетливо прослеживается прямая зависимость, то в отношении смысловой информации такая зависимость имеет менее явный характер, хотя короткие буквенные и фонемные цепочки являются чаще всего семантически опустошенными служебными словами или терминами общего значения, а длинные цепочки выступают в качестве полнозначных знаменательных слов конкретного значения. Таким образом, с опре-

Таблица 53

Величины смысловой информации, приходящиеся на слова-эквиваленты в пяти языках
(см.: Пиотровский, Байгазиаева, Бектаев, Богодист, Георгиев, Пестунова, Райтар, 1973, с. 720; Bogodist, Guéorguiev, Pestiunova, Piotrowski, Raitar, 1974, с. 37)

№ п.п.	Строй языка: фузионно-аналитический с остаточным синтетизмом — французский ЛЕ	$H(w_1)$	фузионно-аналитический с элементами синтетизма — болгарский ЛЕ	$H(w_1)$	фузионный с преобладанием синтетизма — русский ЛЕ	$H(w_1)$	агглютинирующий — казахский ЛЕ	$H(w_1)$	эстонский ЛЕ	$H(w_1)$
1	le gouvernement	5.44	правителство	17.74	правительство	13.36	—		valitsus	17.10
2	le peuple	6.83	—		народ	8.85	—		rahvas	17.64
3	l'état	7.02	—		государство	17.66	—		riik	20.63
4	le président	7.44			президент	8.44	—		president	12.69
5	la classe	9.24			класс{социальный}	18.33	—		klass	18.29
6	l'exploitation	9.56			эксплуатация	19.84	—		ekspluateerimine	17.14
7	le président 'председатель'	5.15	председателят	6.72					—	
8	—		печатът	4.95	печать	15.74	—		—	
9	—		—		газета	5.27	газет 'газета'	10.64	—	
10	—		—		я	5.73	мен 'я'	5.88	—	
	Среднее количество смысловой информации на слово	7.24		9.80		12.58				17.25

деленными оговорками и поправками длина слова может быть использована в автоматических системах переработки текста в качестве диагносцирующего признака при выделении значащих (ключевых или доминантных) лексических единиц (ср. 56).

Наконец, зависимость количественных оценок смысловой информации от тезауруса угадчика указывает на то, что при построении автоматических систем переработки текста необходимо также учитывать состояние «типичных тезаурусов» тех групп потребителей, на которые ориентирована данная автоматическая информационная система.

Глава 8

ТЕОРИЯ НЕЧЕТКИХ МНОЖЕСТВ И ЕЕ ПРИЛОЖЕНИЕ К ОПИСАНИЮ ЯЗЫКА

53. Квантитативная и алгебраическая лингвистика и описание системы языка с помощью нечетких множеств

Проведенные в предшествующих главах статистико-информационные исследования проливают свет на некоторые важные с точки зрения теоретической и инженерной лингвистики вопросы образования и структуры текста.

В частности, обнаруживается, что текст имеет зернистое квантовое строение. Вводимые в текст нормой языка малоинформативные лексико-грамматические элементы заполнения, выполняющие организующую функцию, перемежаются здесь с обусловленными ситуацией доминантными (ключевыми) лексическими единицами, которые несут основную смысловую нагрузку в тексте. При этом выясняется, что большое количество понятийных единиц в научно-технических текстах воплощается не в отдельных словах или словоформах, но чаще всего передается с помощью словосочетаний. Это значит, что в современных развитых языках — языках как аналитическо-корневого, так и фузионно-синтетического строя — намечается тенденция к превращению словосочетания в самостоятельную лексическую единицу, особым образом записывающуюся в лингвистической памяти человека. Это предположение находит косвенное подтверждение в наших информационно-статистическом и инженерно-лингвистическом экспериментах (ср. гл. 5—7; 9—11).

Зернистая информационная структура характерна также для слов и словосочетаний. Начала слов, т. е. их лексические компоненты, несут обычно бóльшую информационную нагрузку, в то время как середины и концы слов, которые чаще всего заняты словообразовательными и словоизменительными морфемами, имеют гораздо более скромный информационный вес. Такое распределе-

ние информации наблюдается не только в фузионных, но и в агглютинирующих языках (ср.: Байтанаева, Бектаев, Чалабаева, 1973, с. 151).

Неслучайно поэтому, именно лексика текста несет основную информацию, в то время как на морфологию и семантико-синтаксические связи между словами приходится сравнительно скромная информационная доля (ср. 48).

Зернистое распределение информации в тексте, фразе, словосочетании и слове обусловливает возможность компрессии сообщения как по линии широкого использования аббревиатур, так и путем удаления из текста элементов заполнения и сохранения в нем только ключевых слов и словосочетаний (ср. 56—60).

Само собой разумеется, что применение информационно-статистических методов анализа текста не может дать ответ на все вопросы, связанные с проблематикой ЛТ и инженерного языкознания.

Решение конкретных лингвистических и инженерно-лингвистических задач должно опираться не только на вероятностное описание нормы языка, статистику ситуаций или информационное построение текста, но также на описание производящей системы языка.

Как уже не раз говорилось, пришедшая на смену традиционным описаниям системы языка алгебраическая лингвистика предложила недостаточно адекватный аппарат дискретной математики. Этот аппарат, ориентированный на «хорошие» структуры (Newell, Simon, 1958), либо опирается на двузначную логику с правилом исключенного третьего (соответственно, четырехзначную логику с исключенным четвертым), либо использует многозначную логику (Зиновьев, 1968).

Неадекватность дискретной математики проистекает из парадокса несовместимости ее аппарата с языковыми и другими гуманистическими системами, имеющими «плохие» (Newell, Simon, 1958) или, как принято сейчас говорить, «мягкие» (Налимов, 1974, с. 53) или нежесткие структуры.

Существо этого парадокса можно сформулировать следующим образом: чем сложнее исследуемая система, тем менее мы способны дать точные строго формальные и в то же время имеющие практическую пользу суждения о ее поведении (ср.: Zadeh, 1973/1974, с. 7).

Преодоление этого парадокса лежит, очевидно, на пути создания особого математического аппарата для описания лингвистических объектов, их связей и процессов. Ниже, опираясь на разработки Л. Заде и его соавторов (Zadeh, 1973/1974; Bellmann, Zadeh, 1970; De Luca, S. Termini, 1972; Santos, 1970), исследования Лакоффа, Сапса, Левелта (Lakoff, 1971, 1972; Suppes, 1970, с. 95—116; Levelt, 1974, с. 35—68; Thomason, Marinos, 1974), работу В. В. Налимова (1974) и собственные исследования (Пиотровская, Пиотровский, 1974, с. 366—274), мы попытаемся изло-

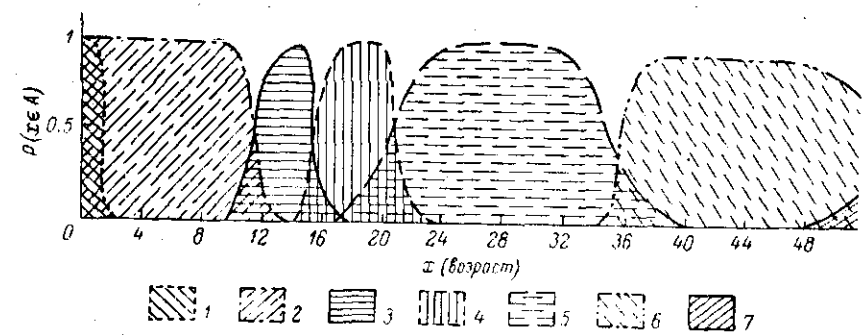


Рис. 41. Вероятностно-семантические поля русских прилагательных (и адъективальных словосочетаний) *младенческий (1)*, *детский (2)*, *отроческий (3)*, *юношеский (4)*, *молодой (5)*, *среднего возраста (6)*, *старый (старческий) (7)*.

жить основные принципы, на которых должен строиться этот аппарат.

Как уже говорилось (15), оперативными единицами человеческого мышления являются не дискретные математические объекты, но элементы некоторых нечетких множеств (fuzzy sets). При этом переход от принадлежности элемента x к множеству A, к «непринадлежности» этого элемента к указанному множеству осуществляется не скачкообразно, но непрерывно, характеризуясь убыванием вероятности $P(x \in A)$ (ср. рис. 41).

Использование нечетких множеств дает возможность человеку приближенно оценивать сведения, непрерывно поступающие в его мозг, выбирая из них те, которые необходимы для приближенно правильного принятия решений.

В этих условиях для анализа лингвистических и других систем человеческого поведения необходим такой подход, при котором допускаются частичные истины, а строгий математический формализм не является чем-то категорически необходимым. Этот подход характеризуется четырьмя признаками:

1) в нем используются нечеткие множества, обладающие «размытыми» границами;

2) в нем применяются нечеткие («лингвистические») переменные;

3) простые отношения между лингвистическими переменными характеризуются нечеткими высказываниями;

4) сложные отношения оформляются здесь в виде нечетких алгоритмов.

Рассмотрим каждое из этих только что введенных понятий.

1. Описание внешнего мира задается в сознании носителей того или иного языка в виде суммы пространств рассуждений (семантических пространств) ΣS_i. Каждое из пространств S_i можно рассматривать как сумму нечетких подмножеств (смысло-

вых областей) A_j (w_j), где w_j есть лингвистическая метка (название) области, а само A_j есть значение w_j.

Рассмотрим в этой связи семантическое пространство, отражающее возраст человека. Исходя из семасиологической системы и норм функционирования словаря современного русского языка в этом пространстве можно выделить области, обозначаемые прилагательными (и определительными словосочетаниями) *младенческий*, *детский*, *отроческий*, *юношеский*, *молодой*, *среднего возраста*, *старый*.

Чтобы определить границы этих областей, точнее, границы семантических полей указанных слов и словосочетаний, произведем следующий эксперимент.

Предложим большой группе разного пола и возраста испытуемых — носителей русского языка — относить к той или иной возрастной группе мужчин различного возраста x ($0 < x$). При этом выясняется, что интервал от 10 до 14 лет одними испытуемыми будет квалифицироваться как *детский*, а другими как *отроческий* возраст. Аналогичным образом период от 17 до 23 лет будет обозначаться либо как *юношеский*, либо как относящийся к *молодому возрасту*.

Если нанести результаты этого эксперимента на график, где по оси абсцисс будет отмечаться конкретный возраст x, а по оси ординат вероятность его отнесения к той или иной понятийной области — $P(x \in A)$, то мы получим картину распределения семантических полей указанных терминов. При этом выясняется, что каждое из рассмотренных семантических полей представляет собой нечеткое подмножество с размытыми краями, которые накладываются на края соседних множеств.

2. Если понятию «возраст» задавать каждый раз точные числовые значения, то мы получим обычную переменную (в нашем случае — аргумент x). Однако в качестве значений этой переменной можно рассматривать также термины *младенческий*, *детский*, *отроческий* и т. д. В этом случае обозначение возраста будет выступать в качестве нечеткой лингвистической переменной. Ср. в этом плане априорную (персональную) функцию распределения смыслового содержания слова у В. В. Налимова (1974, с. 100).

Имплицитно понятия нечетких множеств и лингвистических переменных давно использовались в структурной семантике и диалектологии: ср. примеры разбивки на области семантического пространства хроматической гаммы и пространства «дерево—лес» (Hjelmslev, 1953/1960, с. 311—312; Piotrowski, 1964, с. 105—106), а также в грамматической диахронии (Пиотровская, Пиотровский, 1974, с. 266—374).

3. В классической математике зависимость между двумя переменными X и Y может быть описана с помощью набора высказываний типа:

если X равен 1, то Y равен тоже 1;

если X равен 2, то Y равен 4;
если X равен 3, то Y равен 9 и т. д.

Этот же прием может быть использован и в рассматриваемом подходе с той лишь разницей, что вместо четких переменных X и Y мы будем употреблять нечеткие лингвистические переменные, например значения слов *дети* и *родители*. В этом случае мы получим нечеткие высказывания:

если дети молоды, то родители стары;
если дети отроческого возраста, то родители среднего возраста;
если дети младенческого возраста, то родители молоды.

4. Подобно тому как классическую функцию можно определить алгоритмически, т. е. задать программу ее вычисления, лингвистическую функцию также можно задать с помощью нечеткого марковского алгоритма или понятия нечетких машин Тьюринга (Santos, 1970, с. 326—339; Zadeh, 1973/1974, с. 35; Berztiss, 1971/1974, с. 157 и сл.). Пример такого нечеткого алгоритма показан на рис. 42.

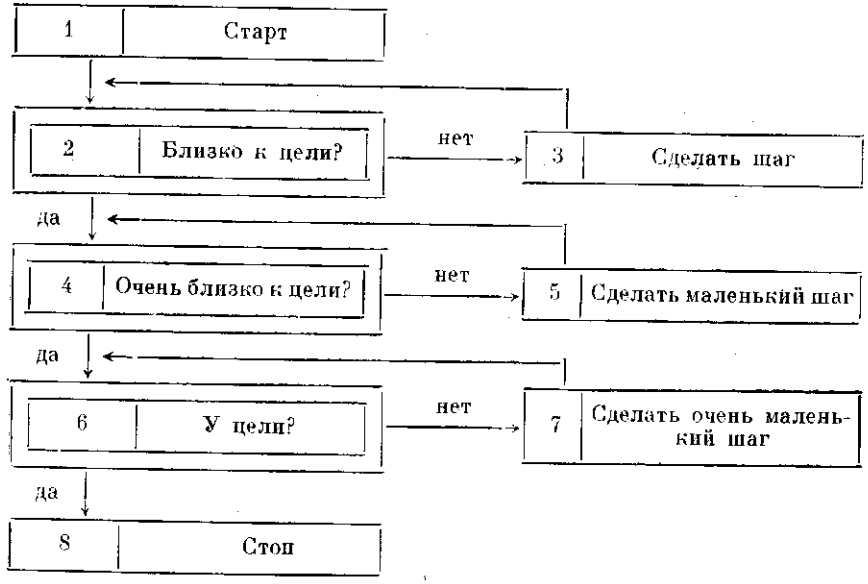


Рис. 42. Нечеткий алгоритм подвода слепого к цели (Raphael, Duda и др., 1971).

54. Свойства нечетких множеств и производимые на них операции

Согласно Л. Заде (Zadeh, 1973/1974, с. 12—19), нечеткие множества обладают рядом свойств, которые позволяют осуществлять на них некоторые важные для нас операции. Перечислим основные из этих свойств.

1. Нечеткое подмножество A_j из семантического пространства S_i характеризуется функцией принадлежности $\mu A : S_j$ (0.1), которая каждому элементу x из множества S_i соотносит число $\mu A_j(x)$ из отрезка (0.1). Это число указывает на вероятность принадлежности x подмножеству A_j. При этом A_j рассматривается как множество элементов x, для которых $\mu A_j(x)$ положительно. Элемент x из пространства S_i, для которого $\mu A_j(x)=0.5$ называется точкой перехода.

2. Нечеткое подмножество A'_j, состоящее из одного элемента x, называется нечетким одноточечным множеством:

$$A'_j = \mu/x,$$

где μ — вероятность принадлежности x множеству A'_j (четкое одноточечное множество определяется как $A_j = 1/x$).

3. Нечеткое множество A_j можно представить в виде объединения с помощью союза *или* одноточечных множеств:

$$A'_j + A''_j + A'''_j \ldots = \int_S (\mu A'_j(x_1) \cup \mu A''_j(x_2) \cup \mu A'''_j(x_3))/x_i = \int_S \mu A_j(x_i)/x_i, \quad (1)$$

где знак интегрирования обозначает операцию объединения одноточечных нечетких множеств $\mu A_j(x_i)/x_i$. В тех случаях, когда A_j состоит из конечного числа элементов, интегрирование заменяется суммированием:

$$A_j = \mu_1/x_1 + \mu_2/x_2 + \ldots + \mu_n/x_n = \sum_{i=1}^{n} \mu_i/x_i, \quad (2)$$

где число μ_i ($i=1, 2, \ldots, n$) — вероятность принадлежности элемента x_i множеству A_j.

Полное четкое множество (пространство) $S_i\{x_1, x_2, \ldots, x_n\}$ можно представить как сумму

$$S_i = 1/x_1 + 1/x_2 + \ldots + 1/x_n = \sum_{i=1}^{n} 1/x_i. \quad (3)$$

Предположим теперь, что пространство S_i есть интервал (0.100) с переменной x_i, указывающей на возраст. Тогда, согласно данным рис. 41, нечеткое множество, лингвистической меткой которого является слово *детский*, можно записать, например, так:

$$\textit{ребенок} = \int_0^{10} 1/x + \int_{10}^{14} \left(1 + \left(\frac{x-10}{2}\right)^{10}\right)^{-1} \Big/ x.$$

Над нечеткими лингвистическими множествами могут быть произведены следующие операции.

1. Дополнение множества A, которое определяется как

$$\neg A = \int_S (1 - \mu A(x))/x. \quad (4)$$

Эта операция соответствует отрицанию *не* $A = \bar{A}$, *не* $x = \bar{x}$.

2. Пересечение лингвистических множеств A и B, которое обозначается как

$$A \text{ и } B = A \cap B = \int_S (\mu A(x) \cap \mu B(x))/x. \quad (5)$$

3. Объединение множеств, которое соответствует союзу *или* и представлено в выражениях (1) и (2).

4. Концентрация лингвистического множества, которое представляет собой операцию, соответствующую значению наречия *очень*, и определяется равенством

$$C(A) = A^2. \quad (6)$$

В итоге осуществления этой операции уменьшается вероятность принадлежности элементов этому множеству A, причем для элементов с высокой вероятностью принадлежности это уменьшение относительно мало, а для элементов с малой вероятностью оно относительно велико.

5. Деконцентрация множества, которая соответствует наречию *приблизительно*, определяется выражением

$$Dc(A) = \sqrt{A}. \quad (7)$$

6. Сгущение $\mathrm{Int}(A)$ нечеткого множества, которое отличается от операции концентрации тем, что она увеличивает вероятности $\mu A(x)$, которые больше 0.5, и уменьшает те, которые меньше 0.5. Таким образом, сгущение уменьшает нечеткость множества A, превращая его в идеале в обычное (четкое) множество M (см. 58).

7. Размывание множества $F(A)$, которое является операцией, обратной сгущению. С помощью размывания, которому соответствуют выражения: *более или менее*, *слабо*, *много* и т. п., обычное (четкое) множество может быть превращено в нечеткое множество.

Выше уже говорилось, что каждому множеству A_j соответствует лингвистическая метка — термин (слово, словосочетание) w_j. Тогда язык или подъязык L следует рассматривать как соответствие между множеством (алфавитом) терминов

$$W = \{w_1, w_2, \ldots, w_i, w_j \ldots\}.$$

и семантическим пространством:

$$S_i = \{A_1, A_2, \ldots, A_i, A_j \ldots\}$$

(см.: Zadeh, 1971, с. 159—176).

Это соответствие характеризуется нечетким отношением N из W в S_i. Указанное отношение между w в W, с одной стороны, и x в A_j — с другой, характеризуется каждый раз вероятностью $\mu_N(w, x)$ применимости w к x. Так, если w=*отроческий*, а x=12 лет, то μ_N (*отроческий*, 12)=0.5 (ср. рис. 41).

Таким образом, значением термина w_k в языке (подъязыке) L служит нечеткое множество $A_j(x)$ из семантического пространства S_i. Степень применимости $A_j(x)$ к w_k измеряется вероятностью $\mu_N(w_k, x)$, диктуемой нормой языка (подъязыка) L.

Теория нечетких множеств, высказываний и алгоритмов включает также описание лингвистических переменных и приемы вычисления их значений, а также анализ нечетких условных предложений и составное правило вывода (Zadeh, 1973/1974, с. 21—34). Всех этих вопросов мы касаться не будем, поскольку они не имеют непосредственного отношения к задачам инженерного языкознания и лингвистики текста.

Описанные приемы теории нечетких множеств представляют пока еще пробный шаг на пути создания адекватного математического аппарата для описания системы языка и построения текста. Хотя в настоящее время алгоритмы переработки текста опираются обычно либо на традиционные описания системы языка, либо на ее формализацию, выполненную с помощью методов дискретной математики (ср. главы 9—11), некоторые идеи теории нечетких множеств уже сейчас могут быть применены для решения инженерно-лингвистических задач.

Итак, мы рассмотрели три подхода, использующиеся в современной ЛТ: статистико-дистрибутивный, информационный, а также подход, использующий теорию нечетких множеств и алгоритмов. Первые два позволяют извлекать информацию непосредственно из текста, третий должен помочь построить на основе статистико-информационных данных, результатов психолингвистического эксперимента и, наконец, данных традиционного описания такую модель, которая адекватно описывала бы производящую текст систему и норму языка.

Часть третья

ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ АВТОМАТИЗИРОВАННЫХ СИСТЕМ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

Глава 9

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СМЫСЛА ТЕКСТА И ОБРАЗУЮЩИХ ЕГО ЕДИНИЦ

55. Основные понятия теории распознавания

Все виды машинной переработки текста связаны с распознаванием смысла таких лингвистических единиц, как морфема, словоупотребление, словосочетание, предложение, сам текст. Машинное распознавание смысла обычно предусматривает переход от некоторого нечеткого лингвистического множества A_j к обычному четкому (машинно-лингвистическому) множеству. При описании этого перехода можно воспользоваться некоторыми приемами и понятиями, использующимися в теории распознавания образов.

Теория распознавания смысла оперирует следующими понятиями.

1. Поле рассуждений S_i, отражающее некоторый участок объективной действительности, называется *семантическим пространством* (Piotrowski, Palibina, 1974; ср. в этом плане такие понятия, как «предметная область», «классификационное поле» — Панова, Шрейдер, 1974, с. 4). Примером такого пространства может служить семантическое поле «возраст». Если речь идет о научно-технических текстах, то семантическое пространство (СП) охватывает чаще всего определенную отрасль знаний.

2. СП определяется признаками. Обычно в качестве семантических признаков выступают некоторые единицы смысла. Если пространство строится алгоритмическим конструктивным путем, то в качестве его семантических признаков выступают дифференциальные семантические элементы, задаваемые в виде конструкторов или выводимые экспериментальным путем. Для выявления этих дифференциальных элементов используются различные приемы структурной лингвистики, в частности компонентный анализ, в основе которого лежит представление о десигнативном значении лингвистического символа как совокупности конструкторов, например «тем», семантических множителей и компонентов. Каждый семантический признак может быть снабжен вероятностно-статистическим весом (вторичным признаком).

Если же СП задается аксиоматически, то в качестве его семантических признаков выступают значения отдельных морфем, основ, исходных форм слова, словоформ, а также словосочетаний.

3. СП распадается на смысловые области, в качестве которых могут выступать разделы или подразделы отрасли знаний, содержание книги, статьи, главы, абзаца, наконец, значение отдельного предложения, словосочетания и даже отдельного слова, словоформы, основы или морфемы. Области СП представляют собой нечеткие множества A_j.

4. Следует различать смысловые области, функционирующие в сознании отправителей и адресатов сообщений, и модели этих областей, задаваемые машине. В первом случае мы имеем дело с нечеткими множествами A_j, во втором — со сгущенными, а иногда и четкими множествами лингвистических объектов, которые мы будем называть классами M_j.

5. Каждый класс имеет свою метку w_j, в качестве которой выступает слово, словосочетание, предложение и т. д. Полный перечень этих меток называется алфавитом меток W.

6. Каждый класс имеет недвусмысленное описание — эталон, в котором обычно используются признаки, характеризующие СП.

7. Автоматическое распознавание смысла заключается в отнесении машиной того или иного лингвистического объекта к одному или нескольким заложенным в ее памяти классам с одновременной выдачей ЭВМ лингвистической метки этого класса.

Совокупность признаков, классов, на которые разбито СП, эталоны и алфавит меток, а также правила их функционирования выступают в качестве машинного информационно-поискового языка (МИПЯ). Введение этого языка в ЭВМ рассматривается как обучение кибернетического автомата.

Прежде чем говорить о деталях этого обучения, напомним, что система «человек—машина—человек» ($Ч_1МЧ_2$), в рамках которой осуществляется автоматическое распознавание смысла, характеризуется следующими особенностями.

1. Отправитель ($Ч_1$) и получатель ($Ч_2$) сообщения работают по принципу черного ящика: их лингвистические реакции не поддаются прямому наблюдению и полному эксплицированию.

2. Предполагается, что отправитель и получатель работают в одном и том же или близких семантических пространствах, а также пользуются близкими разбиениями этого пространства (областями и алфавитами смысловых образов).

3. При интерпретации машинных результатов адресат формирует новый неформализованный образ, который в той или иной степени совпадает с неформализованным образом входного текста. Степень совпадения этих образов оценивается с помощью психолингвистического эксперимента.

Хотя различия в структуре и функционировании мозга человека и ЭВМ достаточно велики, автомат М выступает в роли аналога лингвистического сознания человека. Поэтому при построении распознающего автомата М необходимо использовать лингвистические знания человека. Оптимизировать эту распознающую модель можно при условии, что в автомат М будут заложены наиболее типичные и наиболее информативные признаки, использующиеся человеком при построении СП и при работе в этом пространстве. При этих условиях удается создать искусственное семантическое пространство, достаточно надежно моделирующее СП, заложенное в сознании отправителей и получателей текста.

Обучение автомата осуществляется на основе либо вероятностно-статистической, либо детерминистской (невероятностной) методики. Лингвистическая информация, необходимая для обучения ЭВМ, может быть получена трояким путем.

Во-первых, можно воспользоваться теми семантическими признаками и объектами, которые даны в словарях или на которые указывает, опираясь на свою интроспекцию, носитель языка и специалист в интересующей нас области знаний.

Во-вторых, можно построить и ввести в машину искусственное пространство, используя конструктивные смысловые единицы, например дифференциальные семантические признаки. Оба этих подхода носят, по преимуществу, детерминистский характер.

В-третьих, нужную для «обучения» автомата М информацию можно получить путем статистико-дистрибутивного анализа документов, осуществляемого в рамках лингвистики текста. Этот путь, лежащий в русле вероятностно-статистической методики, хотя и является весьма трудоемким с точки зрения машинной реализации, тем не менее широко используется в работах группы «Статистика речи», являясь наиболее надежным приемом получения той лингвистической информации, которая необходима для построения распознающего автомата (ср. **16—18**).

Само машинное распознавание смысла также осуществляется в двух ключах — вероятностном и детерминистском. В первом случае указывается вероятностный вес принятого автоматом решения: численные значения этого веса лежат в промежутке от нуля до единицы.

Во втором случае результаты распознавания принимают только два значения вероятности — 0 и 1. Распознавание смысла может в одних случаях достигаться путем простого отождествления текста и составляющих его единиц с эталонами, заложенными в памяти ЭВМ. Такой подход мы будем называть *иконическим*. В других случаях распознавание достигается с помощью сложного анализа входного текста и его сегментов. Здесь мы имеем дело с *алгоритмическим* подходом.

Выбор вероятностного или детерминистского, алгоритмического или иконического подхода зависит, в первую очередь, от природы тех лингвистических объектов, смысл которых распознается автоматом М. В тех случаях, когда распознается смысл мелких текстовых единиц, например словоформ или словосоче-

таний, то в настоящее время чаще всего применяется детерминистско-иконический подход (именно на этой основе строятся автоматические словари слов и оборотов; см. **66**).

Если же речь идет о целом тексте, то при определении его смыслового образа целесообразно применять вероятностно-алгоритмическую процедуру.

56. Вероятностное распознавание смысла (машинная атрибуция документа)

Наиболее типичным приемом вероятностного распознавания является автоматическая индексация, представляющая собой отнесение документа к той или иной области (тематической рубрике) СП. Вероятностное распознавание может быть использовано также для сравнения поискового предписания с поисковым образцом документа.

Методику и технологию вероятностного распознавания рассмотрим на примере систем автоматического индексирования английских, французских, немецких, а также русских научно-технических текстов — систем, построенных в группе «Статистика речи» (см.: Попескул*, 1975; Попеску, Колесникова, Палибина, Рахубо, Тарасова, Чайковская, 1974, с. 30—31; Перцовая*, 1975; Колева, 1972). Ср. также работы: Climenson, Hardwick, Jacobson, 1961; Earl, 1970.

Обучение автомата вероятностному распознаванию предполагает выполнение следующих операций:

1) выявление признаков того СП, модель которого мы собираемся строить, снабжение этих признаков вероятностными весами;

2) разбиение СП на области (классы) и присвоение им лингвистических меток;

3) накопление признаков и формирование из них эталонов для каждого класса;

4) разработка алгоритма распознавания.

Рассмотрим каждую из этих операций в отдельности.

1. В качестве семантических признаков выступают единицы смысла (семантемы), отражающие те объекты реального мира, которые специфичны для текстов данного макрополя. Эти единицы могут быть воплощены в основах доминантных слов, исходных формах или словоформах их парадигм и, наконец, в словосочетаниях.

В качестве вероятностно-статистических признаков используются либо абсолютные частоты указанных основ, либо их частости.

2. Разбиение СП на классы чаще всего осуществляется исходя из коллективного опыта. При этом используются разного вида научные рубрикаторы и классификации типа УДК, библиотечно-библиографической классификации Библиотеки им. В. И. Ленина

(ср.: Михайлов, Черный, Гиляревский, 1968, с. 316—366; Тезаурус. . ., 1972). В группе «Статистика речи» делаются попытки производить разбиение СП алгоритмическим путем на ЭВМ, используя статистические свойства текста. Ниже, в табл. 54, показаны разбиения тех СП, в рамках которых осуществляется автоматическое распознавание смысла текста.

3. Накопление семантических и вероятностно-статистических признаков осуществляется в русле методики лингвистики текста. Источником, откуда выбираются эти признаки, служат чаще всего отраслевые ЧС, составленные по текстам каждой из областей исследуемого пространства. Кроме того, делаются попытки выделять семантические и вероятностно-статистические признаки путем обработки ПП (запросов) (Иванкин, 1974, с. 27—31; 1975, с. 21—26).

Если семантические признаки воплощаются в основах, то построение эталонов осуществляется следующим образом.

Сначала с помощью словарей, путем обращения к информантам-специалистам в данной области знаний или с помощью математической процедуры (**29**) из ЧС выделяются доминантные словоформы вместе с их статистическими характеристиками.

Эти доминантные словоформы приводятся либо к машинным основам, либо к каноническим формам. Приведение к канонической форме производится в зависимости от информационно-морфологической структуры слова в рассматриваемом языке (см. **61**). Что же касается выделения машинных основ, проводимого либо вручную, либо программно (Беляева*, 1974), то в ходе его осуществления должны соблюдаться следующие условия:

а) машинная основа не должна быть омографична машинному постфиксу или другой основе;

б) конец машинной основы не должен совпадать с машинным постфиксом (такое совпадение может привести к выделению ложной основы);

в) при выделении основы должны учитываться чередования (внутренняя флексия) (см.: Колесникова*, 1974).

Выделяемые в результате отсечения постфиксов основы могут совпадать с традиционными основами или отдельными словоформами: например, словоформы abandonement(s) 'ликвидация скважины', 'ликвидации скважин' и их производные abandoning-, abandoned и т. д. (ГНГ) приводятся к основе abandon-, выступающей одновременно в качестве самостоятельной словоформы to abandon, имеющей значение 'ликвидировать скважину'.

Эти основы могут и не совпадать с традиционной основой: например, словоформы accumulation(s) 'залежь', 'залежи' и их производные accumulated, accumulates и т. д. (ГНГ) приводятся во избежание омонимии ion(s) (постфикс) — ion, ions 'ион', 'ионы' к машинной основе accumul-.

Словоформы парадигмы могут быть приведены к двум основам, отражающим чередования гласных или согласных; например,

Таблица 54

Разбиения исследуемых СП на области (классы)

№ пп	Язык	Семантическое пространство	Алфавит меток областей (классов)	Объем обучающей выборки (словоупотреблений)
1	Английский	Стоматология — Ст (Дунаевский, Перцовая, 1972).	1) анатомия полости рта, 2) кариес, 3) заболевания пульпы, 4) заболевания периодонта, 5) заболевания слизистой полости рта, 6) анестезиология в стоматологической практике, 7) экстракция зубов, 8) оперативная и восстановительная челюстно-лицевая хирургия, 9) новообразования в стоматологической практике, 10) общие заболевания, вызывающие патологические изменения в полости рта и зубах, 11) фармакология в стоматологической практике, 12) ортодонтия и протезирование.	200 тыс.
2	Английский	Геология нефти и газа — ГНГ (Колесникова*, 1974)	1) происхождение нефти и природного газа и формирование залежей, 2) битуминозные породы, 3) методы поисков и разведки, 4) нефтепромысловая геология, 5) геология нефтегазоносных территорий и месторождений, 6) геология нефтегазоносных акваторий, 7) подсчет запасов нефти и газа, 8) воды нефтяных и газовых месторождений, 9) состав и свойства нефти и газа.	235 тыс.
3	Английский	Полупроводниковая техника — ПТ (Тарасова*, 1974)	1) материаловедение, 2) технология, 3) схемы и устройства с применением полупроводниковых приборов, 4) физика полупроводников,	300 тыс.

Таблица 54 (*продолжение*)

№ пп	Язык	Семантическое пространство	Алфавит меток областей (классов)	Объем обучающей выборки (словоупотреблений)
			5) полупроводниковая электроника и микроэлектроника.	
4	Английский и немецкий	Виноградарство и виноделие — ВВ (Тарасова*, 1974; Колева, 1972)	1) соки, 2) фитопатология, 3) обработка виноградников, 4) виноделие (для немецкого яз. различается первичное и вторичное виноделие).	100 тыс. для английского яз. 200 тыс. для немецкого яз.
5	Английский и русский	Морские тексты — МТ (Палибина, 1971)	7 областей.	100 тыс. для английского языка 100 тыс. для русского яз.,
6	Французский	Тракторное и сельскохозяйственное машиностроение — ТСХМ (Рахубо*, 1974)	1) трактора, 2) машины для посева, посадки и внесения удобрений, 3) машины для уборки зерновых, очистки и сортировки зерна, 4) уборочные машины, 5) почвообрабатывающие машины, 6) оборудование по механизации животноводческих ферм, 7) машины по борьбе с вредителями и болезнями сельскохозяйственных культур, 8) погрузочно-разгрузочные и транспортные средства, 9) двигатели сельскохозяйственных машин, 10) хранение сельскохозяйственных продуктов, 11) лесопилки, ирригация и пластмассы в сельском хозяйстве.	300 тыс.
7	Французский	Сушка лакокрасочных поверхностей (СЛП) (Попеску, Хажинская, 1973)		

при построении эталонов классов ГНГ имеет shelf—shelv 'шелф(ы)', foot—feet 'основание—основания'.

Каждая выделенная основа получает весовой признак, представляющий собой сумму частот, или частостей, обобщенных ею словоформ. Так, например, при составлении эталона для класса «виноделие» (ВВ) слоформы:

alcohol ($F = 79$),
alcoholic ($F = 20$),
alcohols ($F = 8$)

приводятся к основе alcohol='алкогол', имеющей абсолютную частоту $F=107$ (см.: Тарасова*, 1974, с. 14).

В эталон не должны включаться очень редкие лексические единицы, а также частые термины слишком общего значения. У неоднозначных терминов должно отмечаться то значение, которое характерно для данного СП (ср.: Salton, 1968/1973, с. 45).

Машинные основы и канонические формы мы будем называть ключевыми (КС) и полиключевыми (ПКС) словами. Первые являются диагностирующими признаками для одного класса (области), вторые — для нескольких классов СП (ср. табл. 55).

Таблица 55

Ключевые и полиключевые слова из СП «Полупроводниковая техника» (Тарасова*, 1974)

Словоформа	Частоты в классах (областях) СП «Полупроводниковая техника»					Вид лексической единицы
	технология	материаловедение	схемы	физика	электроника	
spark 'искра'	0	0	0	0	21	КС
cesium 'цезий'	31	0	0	0	0	КС
gate 'строб-импульс', 'затвор', 'вентильный провод', 'управляющий электрод'	0	10	147	19	59	ПКС

Весь арсенал лингвистических средств, использующийся для построения СП и эталонов образующих его классов, можно рассматривать как некоторый машинный информационно-поисковый язык. Этот МИПЯ имеет морфологию и лексику. Морфология представляет собой список машинных постфиксов, которые используются при выделении машинных основ.[1] Лексика же включает набор признаков (КС и ПКС), а также перечень «антипризнаков», к которым относятся служебные слова и общеупотребительная лексика, не использующаяся в качестве диагностирующих признаков при вероятностном распознавании.

В группе «Статистика речи» построено два варианта алгоритма вероятностного распознавания смысла (автоматической индексации) текста.

[1] В разных ИПЯ используются различные по объему списки машинных постфиксов. Так, например, морфология английской версии ИПЯ ГНГ включает 198 машинных постфиксов (Колесникова*, 1974), в то время как французский вариант ИПЯ ТСХМ включает всего два постфикса *-s* и *-x* (Рахубо*, 1974, с. 6).

57. Первый вариант алгоритма вероятностной атрибуции документа

В этом алгоритме, построенном А. Н. Попескулом * (1975), формируются документы, каждый из которых включает КС или ПКС (обозначим их символом R) и их частоту (F). Наборы этих документов образуют массивы (файлы). При реализации алгоритма на ЕС-ЭВМ эти массивы хранятся на запоминающих устройствах прямого доступа, например на МД. Поскольку в указанных ЗУ каждый документ имеет определенное местоположение и адрес, имеется возможность прямого программного обращения к любому документу без просмотра всей предыдущей информации внутри файла.

Первая часть документа (байты 1—15) отведена для кода i-той КС или ПКС (R_i), вторая часть документа (байты 16—19) включает j-тый номер эталона области A_j и частоту F_{ij} встречаемости R_i в A_j (ср. табл. 56). Если записи документов перенесены на МД, то используется индексно-последовательный метод доступа к этим записям (см.: Germain, 1967/1971, с. 261, 268—273).

Таблица 56

Машинные документы, образующие эталоны классов СП

№ пп.	КС и ПКС	Классы (области)			
		A_1	$A_2 \ldots$	$A_j \ldots$	A_n
1	R_1	F_{11}	$F_{12} \ldots$	$F_{1j} \ldots$	F_{1n}
2	R_2	F_{21}	$F_{22} \ldots$	$F_{2j} \ldots$	F_{2n}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
i	R_i	F_{i1}	$F_{i2} \ldots$	$F_{ij} \ldots$	F_{in}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
m	R_m	F_{m1}	$F_{m2} \ldots$	$F_{mj} \ldots$	F_{mn}

Массивы КС (ПКС) и их частот (Φ_1), характеризующие данный класс СП, мы будем считать поисковыми образами (ПО) класса.

Вторую группу (файл Φ_2) образуют машинные документы, сформированные программно при вводе в ЭВМ индексируемого

текста. Каждый документ включает код текстового словоупотребления w_k (байты 1÷15) и его порядковый номер k (байты 16÷17).

Весь массив Φ_2 пропускается через фильтр списка антипризнаков (Φ_3), в результате чего из Φ_2 удаляются все служебные слова и общеупотребительные словоформы, не выполняющие диагностирующей функции. Затем оставшиеся словоформы путем уже описанных морфологических процедур приводятся к основам или каноническим формам, которые вместе с их суммарными частотами образуют поисковое предписание (ПП) исследуемого текста.

На следующем шаге устанавливается степень смысловой близости между ПП текста и ПО класса СП. Степень этой близости может быть оценена с помощью таких статистических критериев, как корреляционные методы, критерии Пирсона χ^2 для проверки гипотезы о законе распределения, критерий для проверки однородности распределений и сравнения двух вероятностей (частостей или частот), критерий Колмогорова—Смирнова для проверки гипотезы о законе распределения и гипотезы об однородности распределений (Бектаев, Пиотровский, 1974, с. 224—275). А. Н. Попескул * (1975) использует для этих целей критерий согласия χ^2, который имеет следующий вид:

$$\chi_i^2 = \frac{(N_T + N_j - 1)(F_i N_j - F_{ij} N_T)^2}{N_T N_j (F_i + F_{ij})[(N_T - F_{iT}) + (N_j - F_{ij})]}, \qquad (1)$$

где N_T — объем последующего текста; E_{iT} — частота употребления i-той основы в исследуемом тексте; N_j — объем выборки j-той области; F_{ij} — частота употребления i-той основы в j-той области.

Таблица 57

Вероятностное отнесение текста $[P(T_j)]$ к областям семантического пространства «геология нефти и газа» (Колесникова*, 1974)

№ пп.	Ключевые и полиключевые слова	Значения χ^2 для отдельных КС и ПКС классов (областей) СП ГНГ								
		1	2	3	4	5	6	7	8	9
1	accumul-	12.9	1.8	29.9	25.7	33.6	24.5	12.8	23.5	45.2
2	absorpt-	1000.0	289.0	221.0	221.0	171.0	131.0	110.0	49.2	53.7
. . .	. . . . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
23	sediment	3.7	6.2	9.8	2.4	5.9	10.1	9.3	15.1	17.3
24	solut	50.5	-26.1	107.0	221.0	11.2	35.4	2.0	15.1	9.3
G	—	6	10	3	5	8	4	5	2	3
L	—	24	24	24	24	24	24	24	24	24
$P(T, j)$	—	0.250	0.417	0.125	0.208	0.333	0.167	0.208	0.083	0.125

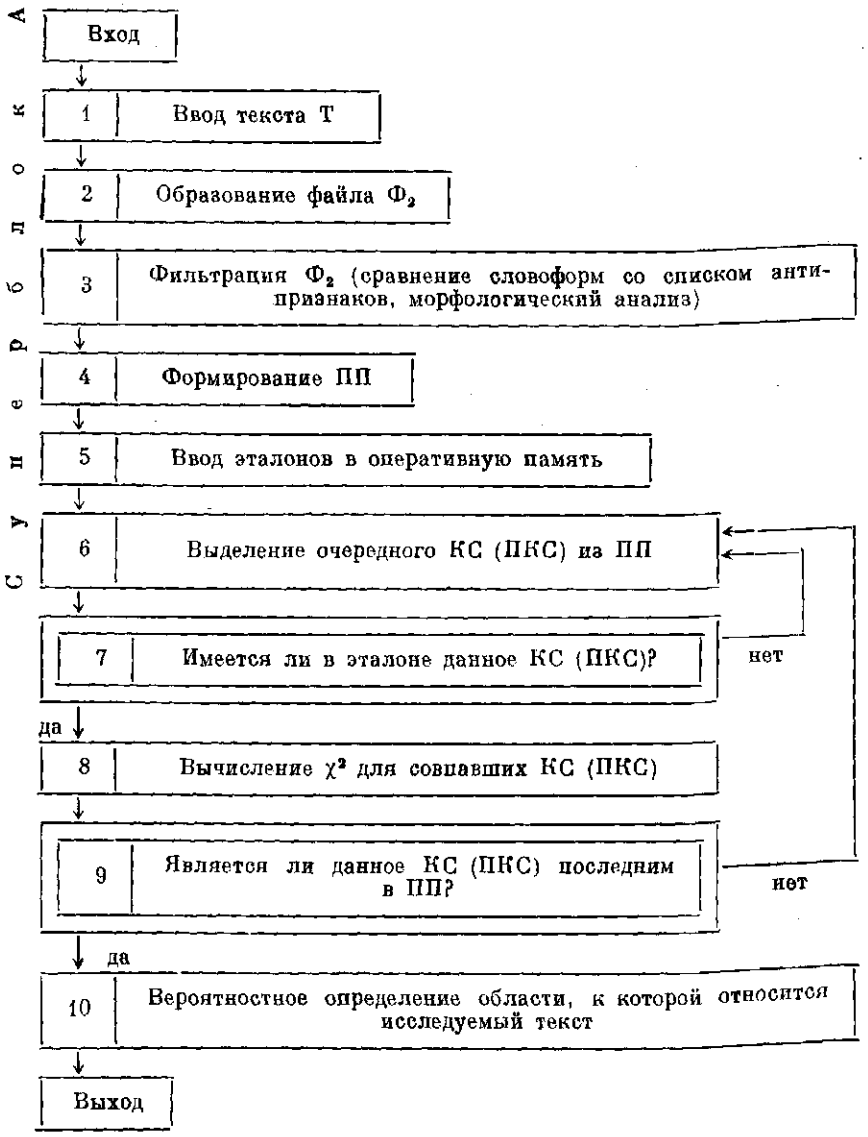


Рис. 43. Принципиальная блок-схема алгоритма вероятностного распознавания смысла текста.

Полученные для каждой КС и ПКС значения χ^2 (ср. табл. 57) сравниваются с заданной пороговой величиной $\chi^2_{0.01;\,1} = 6.64$ (см.: Бектаев, Пиотровский, 1974, с. 306). В результате из общего количества L КС и ПКС в тексте T получаем G случаев хорошего согласия статистики КС и ПКС в тексте с их вероятностными признаками в эталоне (т. е. $\chi^2 < \chi^2_{0.01;\,1}$).

Общая вероятность отнесения индексируемого текста к каждой из областей (классов) СП определяется из формулы:

$$P(T,\ A) = \frac{G}{L}. \qquad (2)$$

В качестве примера (Колесникова*, 1974) рассмотрим индексацию статьи «Geological Aspects of the Origin of Oil» (автор — F. E. Wellings, статья опубликована в журнале «Journal of the Institute of Petroleum», vol. 52, N 508, April 1966, p. 124):

THE ORIGIN OF OIL IS ONE OF THE MOST BAFFLING PROBLEMS IN NATURAL SCIENCE BECAUSE OIL DOES NOT CARRY WITH IT DEFINITE TRACES OF THE ENVIRONMENT OF ITS ORIGIN AND BECAUSE ITS MOBILE NATURE SUGGESTS THAT THE PRESENT ACCUMULATIONS IN POROUS LIMESTONE AND SANDSTONE RESERVOIRS ARE NOT NECESSARILY WHERE IT COULD HAVE ORIGINATED. DURING MIGRATION IT MAY ALSO HAVE UNDERGONE PHYSICAL AND CHEMICAL CHANGES BY NATURAL PROCESSES SUCH AS PRESSURE, TEMPERATURE, SOLUTION, FILTRATION, OR ADSORPTION. TWO METHODS OF INVESTIGATION HAVE BEEN ADOPTED, THE CHEMICAL ONE OF PRODUCING OIL IN THE LABORATORY FROM ORGANIC MATERIALS, COAL, ANIMAL, AND VEGETABLE FATS, AND THE GEOLOGICAL ONE OF STUDYING THE RELATIONSHIP BETWEEN THE DISTRIBUTION OF OIL IN THE EARTH AND THE SEDIMENTARY CONDITIONS UNDER WHICH IT COULD HAVE BEEN FORMED . . . и т. д.

Текст обрабатывался по блок-схеме, показанной на рис. 43. Из текста извлечено 24 КС и ПКС, которые последовательно сопоставлялись с основами, заданными в эталонах. Результаты этого сопоставления, обработанные по формуле (1), показаны в табл. 57. Опираясь на эти результаты, можно утверждать, что текст «Geological Aspects of the Origin of Oil» может быть отнесен:

— к области «Происхождение нефти и природного газа и формирование залежей» — с вероятностью 41.7%;

— к области «Нефтепромысловая геология» — с вероятностью 33.3%;

— к области «Геология нефтегазоносных территорий и месторождений» — с вероятностью 15%;

— к области «Геология нефтегазоносных акваторий» — с вероятностью 10%.

Поскольку вероятности отнесения рассматриваемого текста к другим областям СП ГНГ очень малы, ими можно пренебречь.

58. Второй вариант алгоритма вероятностной атрибуции документа

Этот вариант алгоритма вероятностного распознавания (Перцовая*, 1975) опирается на Бейесову стратегию и строится по следующей схеме.

Пусть $P(A_j)$ — априорная вероятность принадлежности текста T области A_j. В тексте обнаружено ключевое слово R_i, тогда вероятность принадлежности нашего текста к j-той области определяется из формулы апостериорной вероятности:

$$P(A_j/R_i) = \frac{P(A_j)\,P(R_i/A_j)}{\sum_{j=1}^{n} P(A_j)\,P(R_i/A_j)}. \qquad (2)$$

Обычно при решении практических задач распознавания все области (классы) СП считаются равновероятными. В этом случае (2) принимает вид:

$$P(A_j/R_i) = \frac{P(R_i/A_j)}{\sum_{j=1}^{n} P(R_i/A_j)}. \qquad (3)$$

Распознавание смысла текста нельзя строить на статистической характеристике одного ключевого слова, оно должно учитывать поведение всей совокупности КС и ПКС, обнаруженных в индексируемом тексте. Поэтому вероятность $P(T, A_j)$ отнесения текста T к j-той области (классу) СП рассматривается как взвешенная сумма апостериорных вероятностей (3), полученных для всех КС и ПКС, обнаруженных в тексте Т. Иными словами,

$$P(T,\ A_j) = \frac{1}{m} \sum_{i=1}^{m} P(A_j/R_i).$$

С помощью этой методики Г. М. Перцовой осуществлено вероятностное распознавание смысла английских стоматологических текстов на ЭВМ семейства «Минск».[1] Так, например, текст: R. K. Stellmach. Bone Grafting of the Alignment of the Alveolar Arch in Infants with Complete Cleft Lip and Cleft Palate. «Oral Surgery, Oral Medicine and Oral Pathology», vol. 16, no 8, 1963, с. 897—913 (см.: Дунаевский, Перцовая, 1972, с. 124—125) — был отнесен машиной к области «Оперативная и восстановитель-

[1] Алгоритм вероятностного распознавания, использованный Г. М. Перцовой, сходен с алгоритмом А. Н. Попеску (см. рис. 43). Исключение составляет блок 8, в котором вместо величины χ^2 вычисляется вероятность $P(A_i/R_i)$.

ная челюстно-лицевая хирургия» — с вероятностью 26%; к области «Особенности анатомического строения полости рта и зубов» — с вероятностью 16%, к области «Ортодонтия и протезирование» — с вероятностью 12%.

Остальными вероятностями атрибуции в виду их малости можно пренебречь.

Описанные выше приемы вероятностного распознавания смысла интересны не столько с технологической, сколько с методологической точки зрения. Осуществляя вероятностное распознавание смысла текста, автомат имитирует здесь поведение человека, который производит атрибуцию текста к той или иной области знаний, исходя не из двухпозиционной логики (да—нет), но опираясь на вероятностную логику, т. е. принимая каждое решение с определенной долей сомнения. Это не случайно: ведь каждая область СП не образует собой четко отграниченного от других областей множества документов (точнее, содержаний этих документов), но выступает в качестве нечеткого смыслового множества, к которому данный текст T можно отнести лишь с определенной долей достоверности.

Однако при массовой машинной индексации бывает удобно дать потребителю однозначное решение. В связи с этим алгоритмом предусматривается такой режим работы, при котором текст относится к одной, самой вероятной области СП. Остальные области, к которым текст может быть отнесен с меньшими вероятностями, либо объединяются в виде аннотации, вводимой словами: *Кроме того, в статье рассматриваются вопросы, связанные с темами*: (рис. 44), либо не указываются вовсе.

И в том, и в другом случае осуществляется операция сгущения нечеткого множества. В первом случае наибольшая вероятность атрибуции текста к области A_j

$$P(T, A_j) = \mu A_j(x) \qquad (4)$$

значительно увеличивается, в то время как остальные вероятности соответственно уменьшаются. Во втором случае первая вероятность увеличивается до единицы, а остальные малые вероятности уменьшаются до нуля. При этом нечеткое множество — область A_j — превращается в четкое множество M_j — детерминированный класс исследуемого СП.

59. Последовательное вероятностное распознавание смысла

Если распознаваемый текст достаточно велик, то его переработка на ЭВМ с начала до конца может требовать такого количества машинного времени и настолько больших затрат на перфорацию, что экономический эффект автоматического индексирования будет сведен на нет.

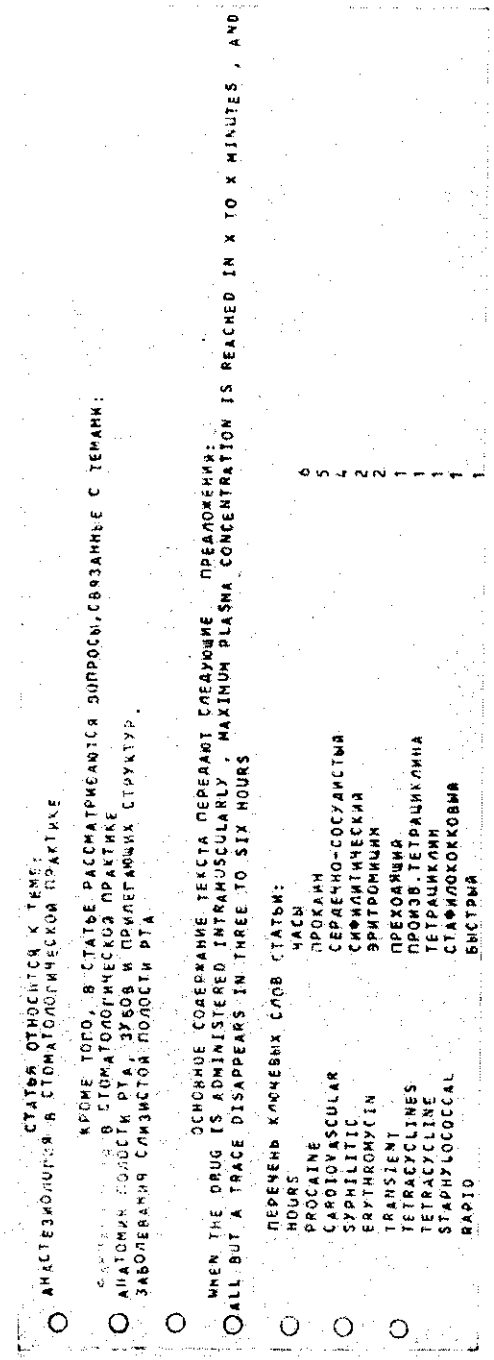


Рис. 44. Машинная индексация английского текста по стоматологии с указанием наиболее вероятной области атрибуции и объединением менее вероятных областей в аннотацию (результат Г. М. Перцовой и И. И. Чайковской).

Чтобы избежать необходимости полной машинной обработки документа, А. Н. Попескул* (1975) использует процедуру последовательного распознавания смысла отдельных порций документа до тех пор, пока ЭВМ с достаточно большой вероятностью не остановится на одной из двух конкурирующих атрибуций ($T \in A_i$, $T \in A_j$). При этом выбранное множество (область) подвергается операции сгущения.

Эта процедура, основанная на известной методике последовательного контроля А. Вальда (Wald, 1947/1960), предусматривает следующие операции.

1. Текст T делится на участки с шагом λ словоупотреблений.

2. Осуществляются испытания, состоящие в определении темы для порций длиной в λ, 2λ, . . ., $N\lambda$ словоупотреблений (каждая последующая порция включает предыдущую).

3. Вводится допустимый риск отнесения текста к областям СП A_i и A_j, который определяется вероятностями P_0, P_1, α, β. Вероятность принять атрибуцию ($T \in A_j$) при $P \leqslant P_0$ должна быть не больше α, а вероятность принять $T \in A_i$ при $P \geqslant P_1$ не должна превышать β. Отсюда в качестве граничных условий следует, что вероятность принять для T область A_j при $P = P_0$ равна α, а вероятность принять область A_i при $P = P_0$ составляет β.

4. Правило последовательного анализа при вероятностной атрибуции текста определяется неравенством

$$\frac{\ln\frac{1-\beta}{\alpha}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}} + N\,\frac{\ln\frac{1-P_0}{1-P_1}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}} > M > \frac{\ln\frac{\beta}{1-\alpha}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}} + N\,\frac{\ln\frac{1-P_0}{1-P_1}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}}, \qquad (5)$$

где N — общее количество порций текста, которые подвергались вероятностной атрибуции, M — количество порций, которые были отнесены к области A_j.

Введем коэффициенты:

$$a = \frac{\ln\frac{1-\beta}{\alpha}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}}; \quad b = \frac{\ln\frac{\beta}{1-\alpha}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}}; \quad k = \frac{\ln\frac{1-P_0}{1-P_1}}{\ln\frac{P_1(1-P_0)}{P_0(1-P_1)}}, \qquad (6)$$

что позволяет переписать неравенство (5) в сокращенном виде:

$$a + kN > M > b + kN, \qquad (7)$$

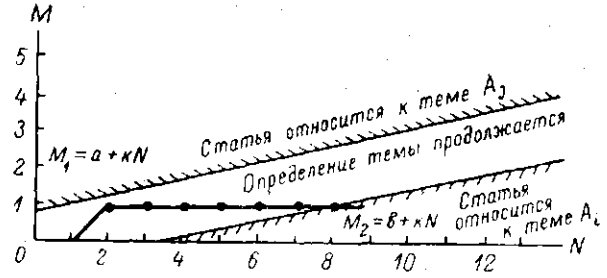


Рис. 45. Отнесение текста T к области A_j путем его последовательного анализа (см.: Попеску, 1972, с. 21).

Из (7) следует, что зона взаимного наложения множеств A_i и A_j лежит между двумя прямыми:

$$M_1 = a + kN \quad \text{и} \quad M_3 = b + kN.$$

5. Предположим теперь, что проведено N атрибуций, причем принадлежность T к области A_j засвидетельствована M раз, а тема A_i зафиксирована $N - M$ раз ($j \neq i$). Тогда при M, удовлетворяющем (7), проводится новое испытание, имеющее целью осуществить атрибуцию порции из T длиной в $\lambda(N+1)$. При $M \geqslant a + kN$ атрибуция выходит из критической области наложения A_i и A_j, и текст относится к классу A_j. При $M \leqslant b + kN$ он попадает в область A_i.

В тех же случаях, когда вся статья просмотрена и M по-прежнему удовлетворяет неравенству (7), текст T остается в критической зоне и должен быть отнесен машиной к обеим накладывающимся областям A_i и A_j.

Однозначная атрибуция текста к области A_i (или A_j) указывает на сгущение множества A_i (соответственно A_j). При отнесении текста одновременно к двум областям этого сгущения не происходит.

В качестве примера воспользуемся приводимой А. Н. Попеску (1972, с. 12) последовательной атрибуцией английского текста «The Browning Problem in Wines» к текстам A_i («виноделие») и A_j («соки») при допустимых рисках $P_0 = 0.1$; $P_1 = 0.4$; $\alpha = 0.2$; $\beta = 0.2$.

Отнесение порции текста к теме A_i (+) и отнесение ее к теме A_j (—) выражаются последовательностью + — +++++++, которой соответствуют следующие значения N и M:

N	1	2	3	4	5	6	7	8
M	0	1	1	1	1	1	1	1

На рис. 45 дано графическое изображение области, ограниченной прямыми $M_1 = 0.773 + 0.226$ и $M_2 = -0.773 + 0.226$, а также

ломаной линии $M=f(N)$, представляющей последовательный ход испытаний до пересечения ее с прямой $M_2=b+kN$, когда становится возможным принять решение, что статья «The Browning Problem in Wines» относится к теме «Виноделие».

60. Детерминистское распознавание смысла (машинное аннотирование документа)

При индексировании текста читатель получает самую общую его семантическую характеристику — указание на принадлежность документа к определенной отрасли знания. Читатель-специалист нуждается, разумеется, в более глубоком эксплицировании содержания документов. Этой задаче отвечает автоматическое аннотирование документа, построенное на детерминистском распознавании текста.

Здесь понятия теории распознавания интерпретируются следующим образом.

1. В качестве СП снова выступает поле рассуждений S_i, отражающее некоторую область знаний.

2. В роли семантических признаков выступают отдельные КС (ПКС) и ключевые (полиключевые) словосочетания (КСс, ПКСс). Эти слова и словосочетания представлены на языке аннотируемого документа. Что же касается вероятностных признаков (p_i), то они принимают только два значения — единицу и нуль. Это значит, что либо имеет место полное совпадение текстовой словоформы (словосочетания) с ключевой лексической единицей, заданной в эталоне (здесь $p_i=1$), либо этого совпадения не наблюдается (тогда $p_i=0$).

3. Областями СП являются единицы смысла — семантические доминанты. При моделировании СП области представлены в виде машинных классов, метками которых выступают простые или сложные термины, воплощающие указанные выше семантические доминанты.

4. Словарь терминов МИПЯ, использующегося для автоматического аннотирования текстов, считается алфавитом областей (классов).

5. Если ЭВМ выдает аннотацию на русском языке, то в качестве меток классов используются формы творительного падежа указанных терминов (ср. ниже, с. 234).

6. Эталон области представлен классом условной эквивалентности (КУЭ), т. е. синонимическим рядом ключевых и полиключевых словоформ и словосочетаний, который формируется на основе смысловой общности и частичной взаимозаменяемости КС, ПКС, КСс, ПКСс. В КУЭ могут входить грамматические формы одного и того же слова (ср. табл. 58, КУЭ 1), лексические и синтаксические синонимы (табл. 58, КУЭ 2—5).

Если обрабатываются русские документы, то в виде метки КУЭ используется один из синонимов.

Например, при аннотировании русских документов, относящихся к СП «Морские тексты» (Палибина, 1971), падежные парадигмы *атомная энергетическая установка, атомной энергетической установки. . .* и т. д., а также аббревиатуры этих словосочетаний АЭУ, ЯЭУ выступают все вместе в виде эталона машинного класса. Меткой этого класса служит словосочетание *ядерной энергетической установкой.*

При аннотировании иностранных текстов роль эталона играет набор синонимических лексических единиц, объединенных одной смысловой доминантой, которая выступает в роли семантической области A_j. Моделирующий эту область машинный класс M_j представлен эталоном — указанным выше набором КС, ПКС, КСс, ПКСс, а его меткой является русский эквивалент всего класса, в роли которого может выступать как целое словосочетание, так и отдельная словоформа в творительном падеже (ср. табл. 58).

Таблица 58

Классы условной эквивалентности при аннотировании английских и французских научно-технических текстов (Колесникова*, 1974; Рахубо*, 1974; Тарасова*, 1974)

№ КУЭ	Язык	Семантическое пространство	Признаки, образующие эталон класса	Метка класса
1	Английский	Геология нефти и газа	abandoned abandoning abandonment abandonments	*ликвидацией скважины*
2	»	То же	after-production secondary recovery	*вторичной добычей*
3	»	Виноградарство и виноделие	spray treatment spray application prebloom sprays *bloom spray treatment*	*обработкой опрыскиванием*
4	»	То же	prebloom treatment prebloom thinning treatment	*обработкой виноградников до цветения*
5	Французский	Тракторное и сельскохозяйственное машиностроение	refroidissement par eau refroidissement à eau	*водяным охлаждением*

Оформление метки класса с помощью творительного падежа определено тем, что все метки вводятся в аннотацию с помощью заранее задаваемого машине штампа: (*статья*) *освещает вопросы, связанные с. . .* (ср. рис. 46).

Рис. 46. Машинная индексация, аннотирование и компрессия (квазиреферирование) английского текста по нефти и газу (результат В. В. Колесниковой, А. Н. Попеску и И. И. Чайковской).

Смысловой образ текстового сегмента считается распознанным при совпадении его с КС, ПКС, КСс или ПКСс эталона.

Совокупность распознанных смысловых образов в виде перечня меток — русских слов и словосочетаний — представляет собой машинную аннотацию, по которой человек-адресат может составить себе представление о содержании обработанного машиной документа (рис. 46).

Исходный материал для построения детерминистских моделей аннотирования черпается из ЧС словоформ и словосочетаний (см. 18—23), из двуязычных и энциклопедических словарей, а также отбирается путем «ручного» анализа контекстов (Колесникова*, 1975). Кроме того, делаются попытки выбирать КС и ПКС программно на ЭВМ с помощью разного вида статистических критериев (Чапля, Чапля, Зубов, 1973, с. 76—93; Каушанская, Лукьяненков, 1972).

Выбранные указанным путем КС и ПКС образуют лексику МИПЯ, а фразеология формируется из КСс и ПКс.

Рассмотрев вопросы отбора лингвистического материала, обратимся теперь к построению алгоритмов.

ЕС-алгоритм аннотации текста с помощью набора словоформ (Попескул*, 1975) предусматривает следующую схему размещения информации в эталоне. Запись каждой ключевой словоформы (R_1) занимает 20 байтов. Первые 15 байтов отведены для кода КС (ПКС), 16-й байт для названия машинного класса M_j (русского эквивалента в творительном падеже), а номер дорожки (Z_j) на МД и номер записи на дорожке (E_j) расположены соответственно в байтах 17÷18 и 19÷20. Все эти записи образуют для каждой из областей СП файл Φ_4.

Для получения аннотации на уровне словоформ необходимо вначале провести вероятностное распознавание смысла документа (см. блок-схему на рис. 43), далее осуществить выбор из Φ_4 либо всех ключевых и полиключевых лексических единиц, которые обнаружены в документе, либо только таких, которые имеют наибольший статистический вес. Русские эквиваленты-метки этих словоформ, точнее КУЭ, куда входят эти словоформы, образуют аннотацию, раскрывающую потребителю содержание статьи. Более подробно формирование аннотации на уровне словоформ представлено в блок-схеме на рис. 47.

Как уже говорилось (21), основные профессиональные сведения, содержащиеся в научно-техническом тексте (особенно в новых областях знаний), заложены чаще всего не в отдельных словах, но в сложных терминах (словосочетаниях). Поэтому более содержательную аннотацию документа можно получить путем детерминистского распознавания смысла ключевых и полиключевых словосочетаний.

Алгоритм такого распознавания ориентированный на ЕС-ЭВМ, использует следующую запись информации в эталоне. Каждое

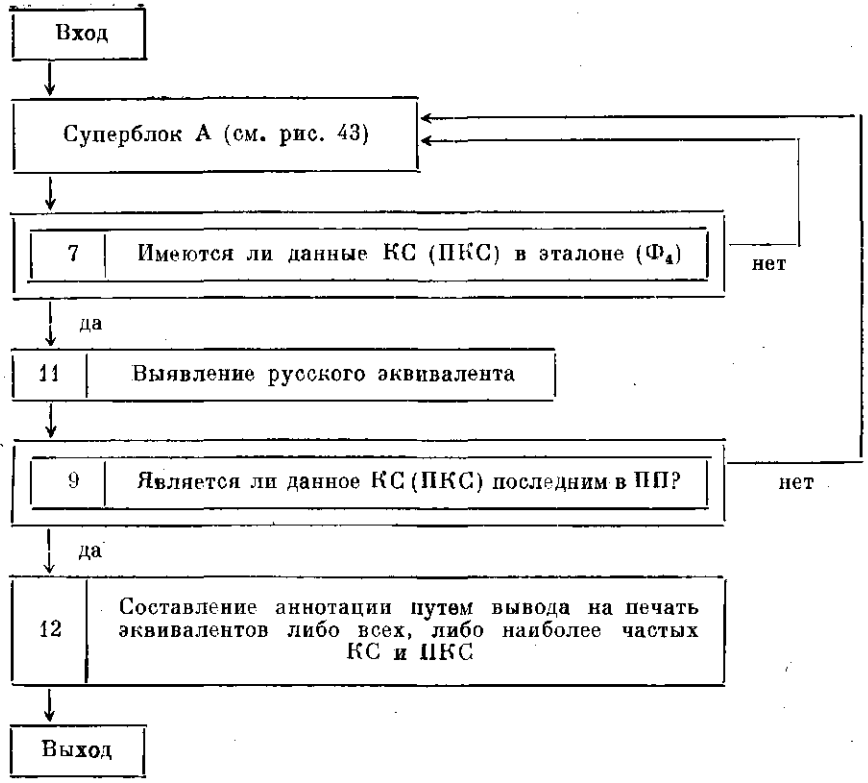


Рис. 47. Принципиальная блок-схема детерминистского распознавания смысла на уровне словоформ (см.: Попескул*, 1975; Тарасова*, 1974; Рахубо*, 1974).

КСс и ПКСс записывается на МД в 20 байтах. Первые 10 байтов отводятся на запись словосочетания, а в байтах 11—15 содержится сжатый код символов словосочетания, не вместившихся в первые десять байтов. Остальные 5 байтов (16÷20) содержат ту же информацию, что и записи КС и ПКС. Совокупность эталонных КСс и ПКСс образует файл $Ф_5$.

Выбор из текста словосочетаний и сопоставление их с эталонными КСс, и ПКСс осуществляется с помощью ядерных лексических единиц, собранных в $Ф_6$. Выбор словосочетания из текста может быть осуществлен при следующих условиях:

— если задано ядро;

— если указано направление поиска (влево или вправо) по тексту от ядра;

— если определено количество шагов поиска, соответствующее числу словоформ, находящихся влево или вправо от ядра.

Эти условия определяют организацию записи ядер в файле $Ф_6$.

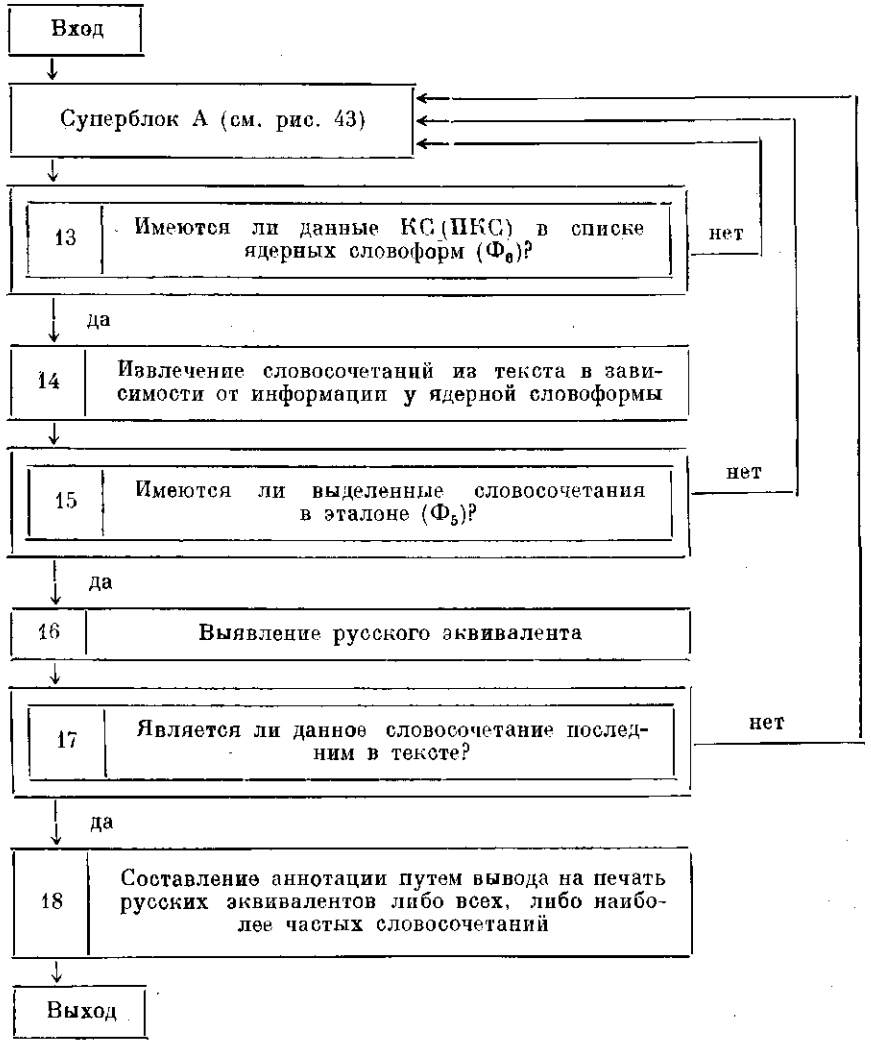


Рис. 48. Принципиальная блок-схема детерминистского распознавания смысла на уровне словосочетаний (см.: Попескул*, 1975; Колесникова*, 1974; Рахубо*, 1974).

Для ядерной лексической единицы w_j отводятся байты 1÷15, а для кода П (Л), указывающего на поиск вправо (влево) от w_j дается 16-й байт. Каждую половину байтов 17÷20 занимает символ w_{ij} (i=1.8), указывающий на длину словосочетания.

Машинный алгоритм составления аннотации с помощью терминологических словосочетаний (см. рис. 48) работает следующим образом.

Лексические единицы, попавшие после фильтрации текста (см. блок 3 на рис. 43), в ПП, сравниваются с ядрами в $Ф_6$ (если фильтрация не проводится, то сравнению подвергаются все словоупотребления текста подряд). Если текстовая лексическая единица совпала с одной из ядерных КС или ПКС, то с помощью информации, заложенной в байтах 16÷20 — записи П (Л) и w_{ij}, из текста выбираются словосочетания-заготовки, представленные записями в 15 байтов (словосочетания, не поместившиеся в эти 15 байтов подвергаются сжатию — см. выше, с. 236). Эти заготовки сопоставляются с КСс и ПКСс, заданными в $Ф_5$. Для совпавших словосочетаний определяется метка КУЭ (русский эквивалент в творительном падеже), которая вместе с другими найденными метками выводится на печать, образуя машинную аннотацию текста.

61. Машинное приведение словоупотреблений текста к их канонической форме

В ходе автоматического распознавания смысла документа или запроса и формирования их поисковых образов каждая из входящих в ПОД или ПОЗ лексическая единица должна быть приведена к некоторой заранее установленной канонической форме (КФ). Без такого приведения невозможно, в частности, осуществить сравнение ПП с фондом документов АСНТИ.

В качестве КФ может выступать либо основа (или даже корень — ср.: Salton, 1968/1973, с. 58; Белоногов, Шемакин и др., 1973; Wenzel, 1973; Беляева*, 1974), либо традиционные исходные формы слова, либо другие заранее фиксированные формы парадигмы. Например, в русском выходе используются формы творительного падежа, используемые в алгоритмах детерминистского распознания смысла (см.: Попескул*, 1975; Рахубо*, 1974; Тарасова*, 1974).

Для массового потребителя информационной продукции АСНТИ на русском языке наиболее привычными и удобными являются традиционные исходные формы КС, КСс — именительный падеж единственного (реже множественного) числа для именных классов, инфинитив несовершенного вида для глаголов.

Если речь идет об аналитических западноевропейских языках, имеющих слабо развитую именную флексию, то приведение КС и КСс к исходной форме может быть осуществлено иконически (см.: Рахубо*, 1974). Что же касается обладающих богатой флексией синтетических языков, например русского, то здесь приведение текстового словоупотребления к КФ представляет собой достаточно сложную алгоритмическую задачу.

Эта задача может быть решена либо путем сравнения текстового словоупотребления со словником словоупотреблений, заранее заложенным в памяти ЭВМ (ср.: Вертель, Вертель, Крисевич и др., 1971, с. 51 и сл.) [1] с тем, чтобы при совпадении текстовой и словарной единицы перейти от этой последней к исходной форме парадигмы (гнезда); — либо с помощью алгоритма, анализирующего комбинаторику конечных морфем и букв словоформы.

В первом случае правильное решение задачи приведения целиком зависит от состава и объема словника. Если словоформа, с которой должно быть произведено отождествление текстового словоупотребления, отсутствует в машинном словаре, то приведение к КФ оказывается принципиально невозможным.

Более гибким и конструктивным является второй — комбинаторно-морфологический подход, который опирается на конечное число разрешенных в данном языке сочетаний букв и морфем в конце слова. Построенные этим путем алгоритмы в значительно меньшей степени зависят от состава словаря, чем алгоритмы первого типа.

Комбинаторно-морфологический подход может быть ориентирован либо на вероятностное, либо на детерминистское решение задачи приведения. В первом случае редкие варианты не учитываются (ср.: Белоногов, Богатырев, 1973; Фатовская*, 1974), во втором случае приходится учитывать все типовые комбинаторные варианты, а также лексику, дающую отклонения от типовой комбинаторики.

Рассмотрим в этом плане детерминистский алгоритм приведения текстовых существительных к форме именительного падежа единственного числа, которая и рассматривается как каноническая форма (Пиотровская, 1974а, 1974б).

Этот алгоритм построен на исследовании комбинаторики последней буквы словоформы с одной или двумя предшествующими буквами, в нем используется также сравнение со списками канонических словоформ — исключений. Схему поиска, который осуществляется справа налево, можно представить в виде иерархии, началом (I) которой является пробел, в качестве второго яруса иерархии (II) выступает последняя буква именной словоформы, а третьим уровнем (III) являются две (в отдельных случаях одна) предпоследние буквы (см. рис. 49). В системе поиска участвуют также фильтры (Фл), которыми являются списки КФ-исключений. Образование самой КФ представляет собой последний ярус иерархии (IV).

Анализ парадигм склонения существительных (Волоцкая, Молошная, Николаева, 1964, с. 55—63) и материалов словарей (Калугина, 1965; ОС, 1974) позволяет обнаружить те формальные критерии, по которым выделяются независимые, частично зависимые и зависимые ветки алгоритма приведения.

---

[1] Парадигмы не обязательно держать целиком в памяти ЭВМ, их можно при необходимости алгоритмически порождать от основы или исходной формы гнезда (Пиотровская, 1973, с. 260—277).

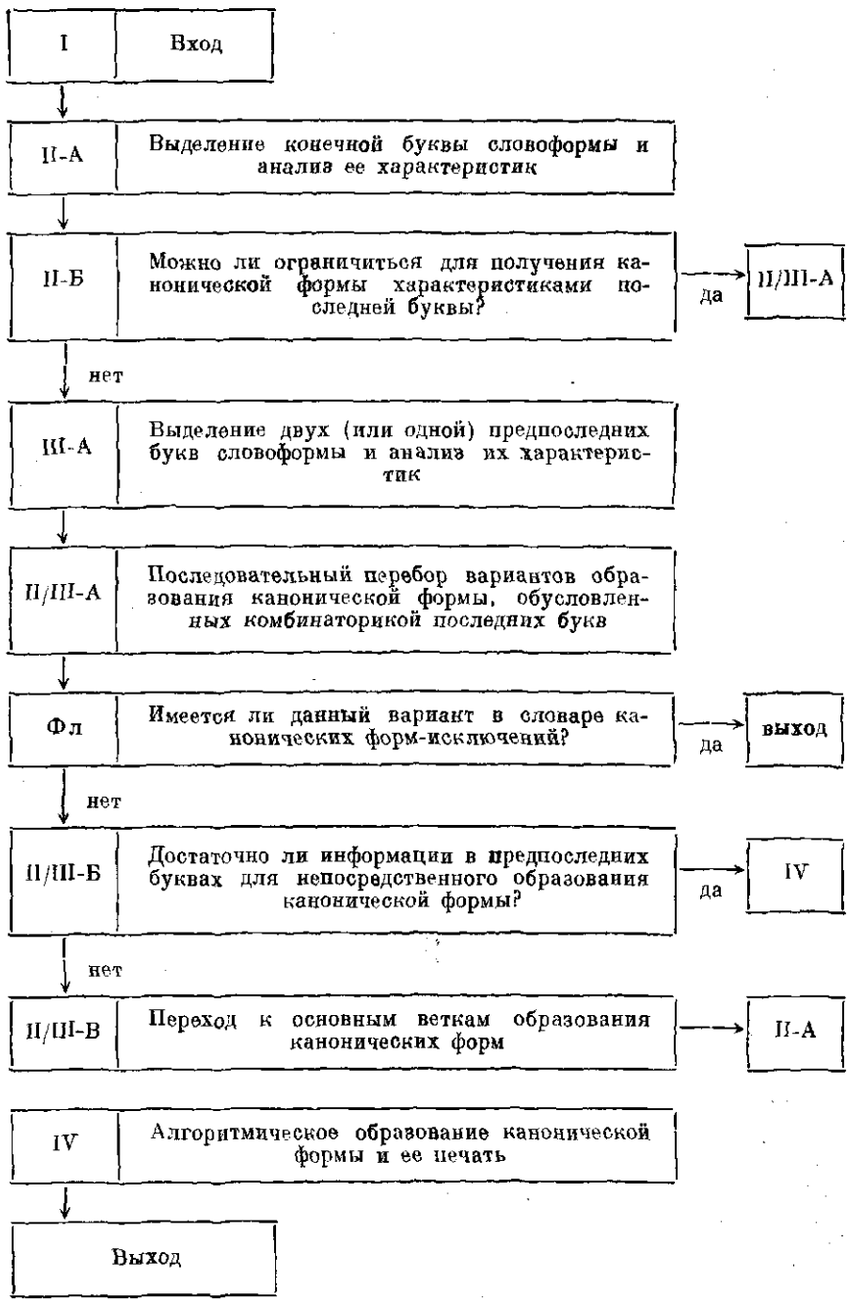


Рис. 49. Принципиальная блок-схема порождения канонических форм русского существительного.

Среди независимых веток алгоритма приведения центральное место занимает «подалгоритм» образования КФ от существительных, оканчивающихся на *а*. (Эта ветка охватывает 23 схемы и большое число вариантов). Особое положение ветки А объясняется тем, что конечное *а* выступает в качестве флексии в 1-м (род., вин. пад. ед. и мн. ч.) и 2-м (им. пад. ед. ч.) склонениях существительных с твердой основой.

В связи с этим на ветку А замыкаются зависимые ветки приведения словоформ, оканчивающихся на гласные *у* и *ы*, которые входят в ту же парадигму окончаний, что и *а*, а также согласные *б*, *г*, *ж*, *з*, *п*, *р*, *с*, *т*, *ф*, *ч*, *ш*, *щ*, которые выступают в качестве концов твердых основ у существительных 1-го (им. пад. ед. ч.) или 2-го (род. пад. мн. ч.) склонений. Кроме того, ветка А может обслуживать словоформы, оканчивающиеся на гласные *е*, *и* и согласные *в*, *й*, *к*, *л*, *м*, *н*, *х*, *ц* (см. ниже, с. 243).

Ветка А работает следующим образом. Если в результате операций, проведенных в блоке II-А (см. блок-схему на рис. 49), выясняется, что словоформа оканчивается на *а*, то выделяются либо одна, либо две предпоследние буквы словоформы (см. блок III-А). Затем определяется адрес одной из 23 схем образования КФ-той схемы, которая соответствует полученной в блоке III-А комбинаторике. Одновременно фиксируется номер принадлежащего к этой комбинаторике списка словоформ-исключений (блоки III-А, II/III-А, Фл).

В каждой схеме учтены возможные варианты «конструирования» потенциальных КФ и их сравнения с КФ-исключениями в списке-фильтре (ср. ветку приведения словоформ, оканчивающихся на ь). Если это сравнение дает полное совпадение списковой и порожденной словоформы, то такая словоформа считается канонической и выводится на печать. В случае несовпадения осуществляется переход к следующему варианту.

Потенциальные КФ, построенные в последнем варианте схемы, со словарем исключений не сравниваются, но непосредственно рассматриваются как канонические формы (эти КФ обычно относятся к наиболее продуктивному классу формообразования существительных).

Независимые ветки алгоритма приведения дают также существительные, оканчивающиеся на *о* и ь.

Конечное *о* дает однозначную схему приведения: все субстантивные словоформы, оканчивающиеся на *о*, являются либо неизменяемыми формами типа *радио*, либо формами имен. и вин. пад. ед. ч. (*железо*, *масло*) и поэтому выступают в качестве КФ.

Что же касается словоформ, оканчивающихся на ь, то их приведение осуществляется по одной схеме, использующей следующие варианты перебора (ср. блоки II/III-А, Фл, IV):

1) стереть 3 последние буквы, приписать *-ля*, сравнить со списком, включающим существительные *земля*, *капля*, *кровля*,

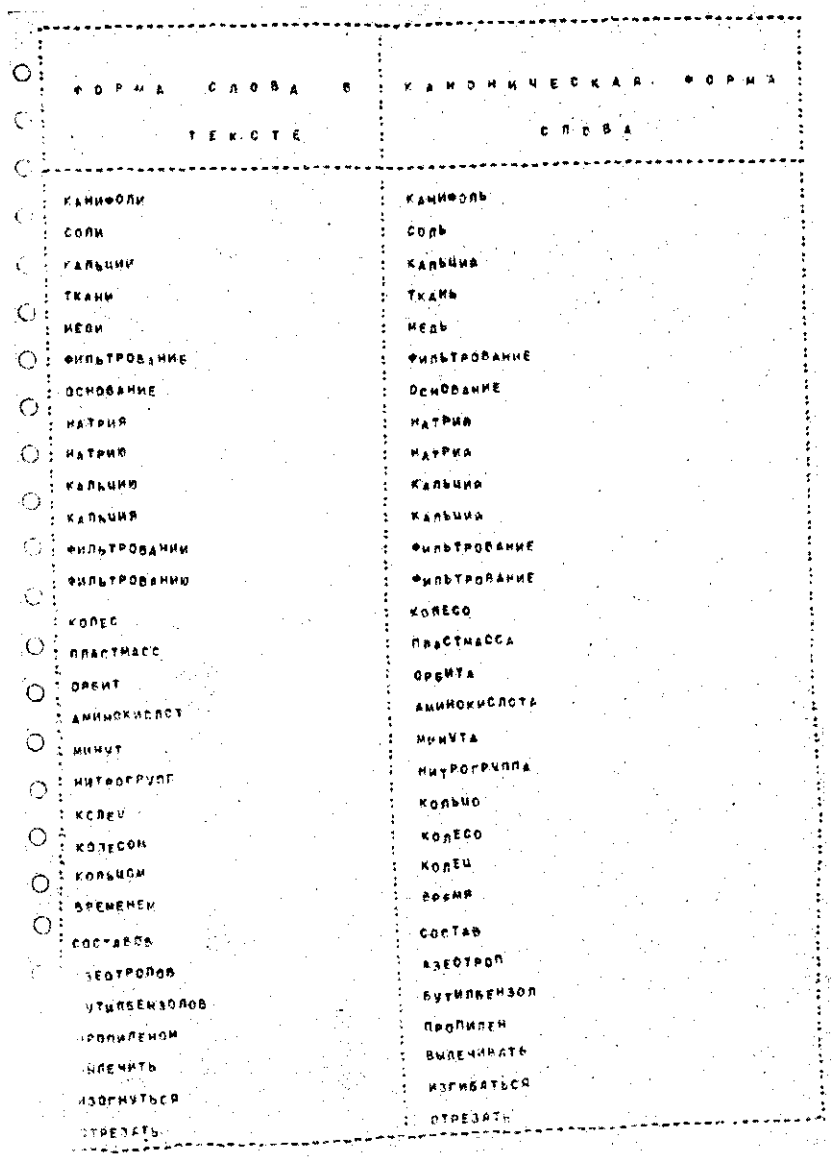


Рис. 50. Машинное приведение текстовых существительных и глаголов к их каноническим формам (см.: Пиотровская, Рихтер, 1970, с. 231—266; Пиотровская, 1974а, 1974б).

петля и др., при совпадении потенциальной КФ с одной из форм списка считать первую истинной КФ, в случае несовпадения анализировать исходную словоформу по следующему варианту;

2) стереть 3 последние буквы, приписать *-ня*, сравнить с существительным *барышня*, при совпадении вывести КФ на печать, при несовпадении перейти к следующему варианту;

3) стереть последнюю букву, приписать *-я*, сравнить со списком, включающим существительные *пуля*, *баня*, *яблоня*, *буря*, *гиря*, *потеря* и др., при совпадении вывести к КФ на печать;

4) при несовпадении со списком исключений считать исходную словоформу канонической (ср. *вероятность*, *кабель*, *кровь*, *медь*, *никель*, *ткань* и т. д.).

По частично зависимым веткам осуществляется приведение словоформ, оканчивающихся на:

1) буквы *е*, *к*, *н*, *ц* — эти ветки либо замыкаются на ветку А, либо дают самостоятельные выходы на КФ;

2) гласную *ю* — эта ветка либо замыкается на ветку И, либо имеет самостоятельный выход на КФ;

3) согласные *в*, *й*, *м*, *х* — эти ветки либо могут замыкаться на ветки А и И, либо давать самостоятельный выход;

4) гласную *и* — эта ветка может замыкаться на ветки А, Е, М, но может давать и самостоятельный выход.

Ветка Я является зависимой веткой, замыкающейся на ветку И.

Результаты машинной реализации алгоритма на ЭВМ М-222 показаны на рис. 50.

Рассмотренный алгоритм приведения КФ построен по принципу открытой системы, допускающей введение новых, ранее не учтенных схем и вариантов, а также неограниченное пополнение фильтров — списков исключений. Такое построение позволяет достигнуть в процессе эксплуатации системы стопроцентной правильности образования канонических форм.

Хотя в основе всех систем машинного аннотирования, реферирования и перевода текста в группе «Статистика речи» лежит вероятностный принцип, только что описанный алгоритм, как, впрочем, и другие алгоритмы морфологического анализа, строятся на детерминистском подходе. Это не случайно. Опыт эксплуатации больших автоматизированных систем переработки текста говорит о том, что анализирующие сервисные программы, к которым относится, в частности, программа приведения к КФ, должны обеспечивать получение правильных лингвистических результатов с надежностью около 100%.

При более низкой надежности эти программы вносят в общую систему такой информационный шум, который может вызвать последовательную цепь ошибок, нарушающую работу всей системы.

62. Машинное реферирование текста путем его компрессии

С вопросами автоматического распознавания смысла тесно связаны приемы машинного свертывания текста. Наиболее простым приемом является автоматическое извлечение из текста тех предложений, которые содержат одно или более ключевых слов или словосочетаний, являющихся «пиками» в распределении смысловой информации текста. Эти предложения, расположенные в порядке их следования, выводятся на печать, образуя квазиреферат текста (ср. рис. 46). Машинный «подалгоритм» квазиреферирования, являясь составной частью общего алгоритма распознавания, подключается к блокам 12 и 14 этого алгоритма (ср. рис. 47, 48; см. также: Попеску, 1972, с. 19; Рахубо*, 1974, с. 10).

При всей его простоте этот подход имеет серьезные недостатки, состоящие в том, что предложения в квазиреферат выбираются без всякого учета существующих между ними смысловых связей. В тех случаях, когда критерием извлечения предложения из текста является присутствие в нем хотя бы одного КС или КСс, компрессия текста оказывается очень незначительной. Если же выбирать только те предложения, которые имеют не менее двух или трех доминантных единиц, квазиреферат будет заметно меньше исходного текста, но в нем может оказаться потерянной значительная часть смысловой информации, а образующие его предложения будут слабо связаны по смыслу друг с другом.

Связный реферат можно получить с помощью алгоритма свертывания, учитывающего сверхфразовые (суперсегментные) связи перерабатываемого на ЭВМ текста.

Попытка построить такой алгоритм для русского текста с реализацией на ЭВМ «Минск-22» была предпринята В. Е. Берзоном* (1972; 1974; см. также: НТИ, серия 2, 1971, № 8, 10, 12).

Алгоритм строится на использовании двух типов экспликации суперсегментных связей: повторений в тексте некоторых доминантных лексических единиц (ср.: Севбо, 1969) и выражении этих связей с помощью реляторов, т. е. единиц заполнения связывающих отдельные предложения и сегменты текста.

В качестве реляторов могут выступать:

1) личные местоимения 3-го лица (ср. в приведенном ниже иллюстративном тексте: *Термин «математическая лингвистика» вошел. . . Однако . . . он понимается по-разному*);

2) указательные местоимения, прилагательные и причастия, выполняющие функцию репризы (ср.: *Термин «математическая лингвистика» . . . как следует употреблять этот термин. Подобное употребление . . ., там же*);

3) союзы, стоящие в начале предложения (*но, поэтому, однако* — ср. пример в п. 1);

4) вводные слова и словосочетания типа *итак, кроме того*, выполняющие связочную функцию (см. иллюстрирующий текст);

5) слова и словосочетания, выполняющие перечислительную функцию, типа *с одной стороны, с другой стороны, во-первых, во-вторых*;

6) глагольные формы типа *рассмотрим*;

7) существительные, стоящие в самом начале предложения (ср. в иллюстрирующем тексте: *Как обычно употребляется термин «математическая лингвистика». Термин «математическая лингвистика» вошел. . .*;

8) некоторые знаки препинания.

Чтобы избежать ложного отнесения реляторов к несвязанным между собой сегментам, вводится статистический критерий среднего и максимального расстояния между сегментами, связанными данными реляторами. В частности, статистическое исследование русских текстов показало, что для местоименных реляторов со стертой семантикой типа *он, такой* это расстояние равно, в среднем, трем словоупотреблениям при максимальном удалении связанных сегментов на 26 словоупотреблений. Для полнозначных реляторов (*аналогичный, другими словами*) эти расстояния соответственно равны 1, 4 и 3, 5 словоупотреблений.

Наконец, вводится понятие проективности сверхфразовых связей, получающее формальное выражение в заранее задаваемой иерархии реляторов.

Идея этой иерархии состоит в том, что одни реляторы — например вводные слова: *итак, кроме того, наконец*, или реляторы-перечисления типа: *с одной стороны, с другой стороны* — вводят наиболее важные с точки зрения содержания текста сегменты; в то же время реляторы, относящиеся к группам 1, 2, 6, 7, 8 привязывают к этим стержневым сегментам второстепенные сегменты, выполняющие пояснительную функцию.

Работа алгоритма Берзона строится по схеме, сходной с той схемой компрессии, которая была рассмотрена в начале параграфа.

Сначала происходит автоматическое распознавание реляторов в тексте, после чего на печать выводятся предложения или сегменты текста, в которые входят эти реляторы.

Рассмотрим работу алгоритма на примере свертки русского текста по математической лингвистике (Гладкий, Мельчук, 1969, с. 15—16).

§ 1. СОДЕРЖАНИЕ ПОНЯТИЯ «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» (ПРЕДВАРИТЕЛЬНЫЕ СООБРАЖЕНИЯ).

КАК ОБЫЧНО УПОТРЕБЛЯЮТ ТЕРМИН «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА».

ТЕРМИН «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ВОШЕЛ В УПОТРЕБЛЕНИЕ В СЕРЕДИНЕ 50-Х ГОДОВ И К НАСТОЯЩЕМУ ВРЕМЕНИ ПОЛУЧИЛ ШИРОКОЕ РАСПРОСТРАНЕНИЕ. ОДНАКО ДО СИХ ПОР РАЗНЫМИ ЛЮДЬМИ ОН ПОНИМАЕТСЯ ПО-РАЗНОМУ.

С ОДНОЙ СТОРОНЫ, СЛОВА «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ЧАЩЕ ВСЕГО УПОТРЕБЛЯЮТСЯ В ОЧЕНЬ ШИРОКОМ И ВЕСЬМА РАСПЛЫВЧАТОМ СМЫСЛЕ, А ИМЕННО ИХ ПРИМЕНЯЮТ К САМЫМ РАЗЛИЧНЫМ ЛИНГВИСТИЧЕСКИМ ИССЛЕДОВАНИЯМ, ЕСЛИ В НИХ

ХОТЯ БЫ В НЕЗНАЧИТЕЛЬНОЙ СТЕПЕНИ ИСПОЛЬЗУЕТСЯ МАТЕМАТИКА ИЛИ ДАЖЕ ЕСЛИ МАЛОИСКУШЕННЫМ ЧИТАТЕЛЯМ ТОЛЬКО КАЖЕТСЯ, БУДТО ОНА ИСПОЛЬЗУЕТСЯ. ТАК, СЮДА ОТНОСЯТ И РАБОТЫ ПО СОЗДАНИЮ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ АППАРАТА АЛГЕБРЫ, МАТЕМАТИЧЕСКОЙ ЛОГИКИ, ТЕОРИИ АЛГОРИТМОВ; И РАБОТЫ, СВЯЗАННЫЕ СО СТАТИСТИКОЙ; И РАБОТЫ, ГДЕ ДЛЯ ФОРМУЛИРОВАНИЯ ТЕХ ИЛИ ИНЫХ ЛИНГВИСТИЧЕСКИХ ПОЛОЖЕНИЙ ПРИВЛЕКАЮТСЯ ПРОСТЫЕ ПОНЯТИЯ И СПОСОБЫ ВЫРАЖЕНИЯ (ИЛИ ОБОЗНАЧЕНИЯ), ЗАИМСТВОВАННЫЕ ИЗ МАТЕМАТИКИ. СЮДА ЖЕ ЗАЧИСЛЯЮТ ВСЕ ЛИНГВИСТИЧЕСКИЕ РАБОТЫ, ПРЕДПОЛАГАЮЩИЕ ПРИМЕНЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ МАШИН, В ТОМ ЧИСЛЕ РАБОТЫ ПО АВТОМАТИЧЕСКОМУ ПЕРЕВОДУ, ДАЖЕ ЕСЛИ В НИХ МАТЕМАТИКА НЕ ИСПОЛЬЗУЕТСЯ НИ ПО СУЩЕСТВУ, НИ ПО ФОРМЕ. КРОМЕ ТОГО, ПОД НАЗВАНИЕ «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ЧАСТО ПОДВОДЯТСЯ ВСЕВОЗМОЖНЫЕ РАБОТЫ ПРИКЛАДНОГО ХАРАКТЕРА И ВООБЩЕ ЛИНГВИСТИЧЕСКИЕ СОЧИНЕНИЯ, НОСЯЩИЕ ЧЕТКО ВЫРАЖЕННЫЙ НЕТРАДИЦИОННЫЙ ХАРАКТЕР.

КАК СЛЕДУЕТ УПОТРЕБЛЯТЬ ЭТОТ ТЕРМИН.

ПОДОБНОЕ СЛОВОУПОТРЕБЛЕНИЕ, Т. Е. ПРИМЕНЕНИЕ ТЕРМИНА «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ДЛЯ ОБОЗНАЧЕНИЯ ЧАСТИ ЛИНГВИСТИКИ, ПРЕДСТАВЛЯЕТСЯ АВТОРАМ НЕУДАЧНЫМ. ОНО СОЗДАЕТ ОШИБОЧНОЕ ВПЕЧАТЛЕНИЕ, БУДТО СУЩЕСТВУЮТ ДВЕ РАЗНЫЕ ЛИНГВИСТИКИ: ОДНА — ПРИНЦИПИАЛЬНО НЕ МАТЕМАТИЧЕСКАЯ, ДРУГАЯ — ОСОБАЯ, «МАТЕМАТИЧЕСКАЯ». В ДЕЙСТВИТЕЛЬНОСТИ ЖЕ ЛИНГВИСТИКА ЕСТЬ ЕДИНАЯ НАУКА СО СВОИМИ ЗАДАЧАМИ И СВОИМ ОБЪЕКТОМ, КОТОРАЯ ПОЛЬЗУЕТСЯ МАТЕМАТИЧЕСКИМИ МЕТОДАМИ ТАМ, ГДЕ ОНИ НУЖНЫ, И НЕ ПОЛЬЗУЕТСЯ ИМИ ТАМ, ГДЕ ОНИ НЕ НУЖНЫ.

В НАСТОЯЩЕЙ КНИГЕ ПОД «МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКОЙ» ПОНИМАЕТСЯ НЕЧТО СОВСЕМ ДРУГОЕ; А ИМЕННО — ОПРЕДЕЛЕННЫЙ КРУГ МАТЕМАТИЧЕСКИХ ПО СУЩЕСТВУ РАБОТ, ВОЗНИКШИХ ИЗ ПОПЫТОК СТРОГО ОПИСАТЬ ФАКТЫ ЕСТЕСТВЕННЫХ ЯЗЫКОВ И СОДЕРЖАЩИХ РЕЗУЛЬТАТЫ, КОТОРЫЕ МОГУТ ОКАЗАТЬСЯ ПОЛЕЗНЫМИ ДЛЯ ЛИНГВИСТИКИ. ПРИ ЭТОМ К УКАЗАННОМУ КРУГУ НЕ ОТНОСЯТСЯ ТЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ, ДЛЯ КОТОРЫХ ЯЗЫК ЯВЛЯЕТСЯ ЛИШЬ ОДНИМ ИЗ ВОЗМОЖНЫХ ПРИЛОЖЕНИЙ (В ЧАСТНОСТИ, РАБОТЫ ЧИСТО КОЛИЧЕСТВЕННОГО ХАРАКТЕРА, Т. Е. ЛИНГВИСТИЧЕСКАЯ СТАТИСТИКА И Т. Д.); ДЛЯ МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКИ В ДАННОМ ПОНИМАНИИ ХАРАКТЕРНО ИСПОЛЬЗОВАНИЕ ЛИШЬ ТЕХ МАТЕМАТИЧЕСКИХ МЕТОДОВ, КОТОРЫЕ В ОПРЕДЕЛЕННОМ СМЫСЛЕ (СМ. НИЖЕ) СПЕЦИФИЧНЫ ДЛЯ ЯЗЫКА КАК ТАКОВОГО. ИТАК, МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА ЕСТЬ МАТЕМАТИЧЕСКАЯ ДИСЦИПЛИНА, «ОБРАЩЕННАЯ» В СТОРОНУ ЕСТЕСТВЕННЫХ ЯЗЫКОВ И ЛИНГВИСТИКИ.

ЗАДАЧА НАСТОЯЩЕЙ КНИГИ СОСТОИТ В ТОМ, ЧТОБЫ ПОПЫТАТЬСЯ ОХАРАКТЕРИЗОВАТЬ ОБЪЕКТ И МЕТОДЫ МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКИ, А ТАКЖЕ ЕЕ СООТНОШЕНИЕ С ЛИНГВИСТИКОЙ, ПО ВОЗМОЖНОСТИ ПОЛЬЗУЯСЬ ПРИ ЭТОМ ПРИВЫЧНЫМИ ДЛЯ ЛИНГВИСТОВ ПРЕДСТАВЛЕНИЯМИ.

(Ср.: Берзон*, 1972, с. 25).

При переработке этого текста используется трехуровневая иерархия реляторов (ср. рис. 51), которая дает возможность получить три вида (эшелона) свертки.

Предположим, что ЭВМ распознает в перерабатываемом тексте все три уровня. В этом случае на печать будет выведен первый эшелон свертки, состоящий из 2-го и 4-го абзацев статьи.

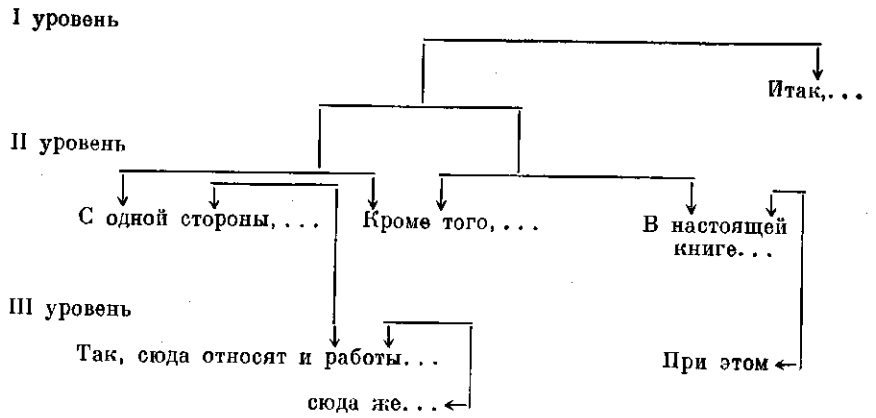


Рис. 51. Трехуровневая иерархия сверхфразовых связей в компрессируемом русском тексте, приведенном на с. 245.

У концов стрелок зависимостей указаны реляторы, начинающие соответствующий сегмент текста.

Второй эшелон свертки, образующийся при условии автоматического распознавания реляторов I и II уровня, имеет следующий вид.

С ОДНОЙ СТОРОНЫ, СЛОВА «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ЧАЩЕ ВСЕГО УПОТРЕБЛЯЮТСЯ В ОЧЕНЬ ШИРОКОМ СМЫСЛЕ, А ИМЕННО: ИХ ПРИМЕНЯЮТ К САМЫМ РАЗЛИЧНЫМ ЛИНГВИСТИЧЕСКИМ ИССЛЕДОВАНИЯМ, ЕСЛИ В НИХ ХОТЯ БЫ В НЕЗНАЧИТЕЛЬНОЙ СТЕПЕНИ ИСПОЛЬЗУЕТСЯ МАТЕМАТИКА ИЛИ ДАЖЕ ЕСЛИ МАЛОИСКУШЕННЫМ ЧИТАТЕЛЯМ ТОЛЬКО КАЖЕТСЯ, БУДТО ОНА ИСПОЛЬЗУЕТСЯ.

КРОМЕ ТОГО, ПОД НАЗВАНИЕ «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА» ЧАСТО ПОДВОДЯТСЯ ВСЕВОЗМОЖНЫЕ РАБОТЫ ПРИКЛАДНОГО ХАРАКТЕРА И ВООБЩЕ ЛИНГВИСТИЧЕСКИЕ СОЧИНЕНИЯ, НОСЯЩИЕ НЕТРАДИЦИОННЫЙ ХАРАКТЕР.

В НАСТОЯЩЕЙ КНИГЕ ПОД МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКОЙ ПОНИМАЕТСЯ НЕЧТО СОВСЕМ ДРУГОЕ, А ИМЕННО — ОПРЕДЕЛЕННЫЙ КРУГ МАТЕМАТИЧЕСКИХ ПО СУЩЕСТВУ РАБОТ, ВОЗНИКШИХ ИЗ ПОПЫТОК СТРОГО ОПИСАТЬ ФАКТЫ ЕСТЕСТВЕННЫХ ЯЗЫКОВ И СОДЕРЖАЩИХ РЕЗУЛЬТАТЫ, КОТОРЫЕ МОГУТ ОКАЗАТЬСЯ ПОЛЕЗНЫМИ ДЛЯ ЛИНГВИСТИКИ.

ИТАК, МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА ЕСТЬ МАТЕМАТИЧЕСКАЯ ДИСЦИПЛИНА, ОБРАЩЕННАЯ В СТОРОНУ ЕСТЕСТВЕННЫХ ЯЗЫКОВ И ЛИНГВИСТИКИ.

(См.: Берзон*, 1972, с. 26).

Третий эшелон свертки может быть реализован в результате распознавания только релятора I уровня ИТАК. В этом случае ЭВМ выдаст наиболее лаконичный ответ:

ИТАК, МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА ЕСТЬ МАТЕМАТИЧЕСКАЯ ДИСЦИПЛИНА, ОБРАЩЕННАЯ В СТОРОНУ ЕСТЕСТВЕННЫХ ЯЗЫКОВ И ЛИНГВИСТИКИ.

Образцы машинных результатов проективных сверток только что рассмотренного и других научно-технических текстов приведены в работах: Берзон*, 1972, с. 28; Берзон, 1974, с. 158.

Рассматривая индексирование, аннотирование и квазиреферирование текста в рамках теории автоматического распознавания смысла, нужно строго различать «машинно-семиотическую» природу выдаваемых ЭВМ результатов и их восприятие человеком.

В случае автоматической атрибуции выдаваемый машиной индекс можно рассматривать и в том и в другом случае как распознанный смысловой образ текста.

Машинная аннотация может выступать в качестве единого смыслового образа текста только с точки зрения интерпретирующего ее человека. Если не рассматривать аннотацию в плане машинной семиотики, то нужно будет признать, что эта аннотация представляет собой совокупность никак не связанных между собой смысловых образов (в нашем случае русских эквивалентов) ключевых слов и словосочетаний входного текста.

Наконец, если обратиться к реферированию, то выдаваемая ЭВМ свертка воспринимается адресатом как смысловой образ текста, в то время как для машины распознавание смысла ограничивается здесь выделением в тексте дескрипторов и реляторов.

Глава 10

ТЕЗАУРУСНОЕ РАСПОЗНАВАНИЕ СМЫСЛА

63. Построение тезауруса с помощью методов лингвистики текста

Слабость рассмотренных в 56—62 алгоритмов автоматического распознавания смысла текста состоит в том, что эти алгоритмы используют такие МИПЯ, в которых задается лишь перечень лексических единиц (терминов-дескрипторов) без указания на существующие между ними смысловые связи.

Получаемые аннотации представляют собой простое перечисление наиболее частых дескрипторов, о синтаксико-семантическом взаимодействии которых в тексте документа читатель аннотации должен догадываться сам, опираясь на собственные знания предмета, а также на профессиональную и языковую интуицию. Все это ограничивает возможности вероятностного и детерминистского распознавания, сводя его лишь к определению темы документа. Использование этой методики для машинного выявления ремы (новизны) документа имеет мало шансов на успех. Что же касается свертывания документа, опирающегося на анализ реляторов, которые воплощают сверхфразовые связи, то его применение ограничено лишь одноязычной ситуацией. Кроме того, получаемые

в результате свертки квазирефераты не могут быть использованы в качестве поисковых образов запросов и самих документов.

Более эффективное распознавание смысла можно получить, опираясь на тезаурусный подход, в котором объединяются дескрипторная и реляторная методики. Термином «тезаурус» обычно обозначается список лексических единиц (словоформ, канонических форм слова, основ, словосочетаний), между которыми заданы смысловые связи как иерархического (родо-видового), так и неиерархического (синонимия, антонимия, ассоциации) типов (ср.: Masterman, 1957/1964; Salton, 1968/1973; Бектаев, Пиотровский, Шабес, 1974, с. 28; Шемакин, 1974). Таким образом, тезаурус выступает как некоторая парадигматическая модель плана содержания, — модель, которая должна включать основные семантические единицы и связи описываемого языка или его разновидности. Эту модель обычно представляют в виде древовидного (чаще всего бинарного) графа.

Построение тезауруса предусматривает:

1) выбор СП и его разбиение на области (предметные классы), каждой из которых приписывается лингвистическая метка;

2) подбор терминологических слов и словосочетаний, выступающих в роли дескрипторов;

3) формирование классов условной эквивалентности;

4) формирование семантических полей (СПо), охватывающих определенные классы условной эквивалентности, а также слова и словосочетания, связанные с ними по смыслу;

5) задание родо-видовых и ассоциативных связей как внутри КУЭ и семантических полей, так и между отдельными КУЭ и разными СПо;[1]

6) формирование общей схемы тезауруса путем соотнесения КУЭ и СПо с областями СП.

Построенный этим путем тезаурусный словарь выступает не только как семантическая модель лексики языка или его разновидности. Тезаурус можно рассматривать также как иерархическую классификацию областей и подобластей внутри исследуемого СП. Лексические единицы, стоящие в вершинах древовидного графа, выступают в этих случаях в качестве названий (меток) соответствующих областей СПо, КУЭ. Наконец, тезаурусный граф можно использовать и как инструмент семантического и семантико-синтаксического кодирования словоформ, слов и словосочетаний (ср. 67).

Выделение СП, его областей, меток и дескрипторов проводится по тем схемам, по которым осуществлялось построение МИПЯ для вероятностного и детерминистского распознавания (55—56, 60). Более сложной задачей является формирование КУЭ и СПо, а также задание связей и структурализация тезауруса. Чаще всего

[1] Эти связи могут быть представлены либо в виде кодов, либо средствами естественного языка — в форме текстовых сегментов-реляторов.

Образцы машинных результатов проективных сверток только что рассмотренного и других научно-технических текстов приведены в работах: Берзон*, 1972, с. 28; Берзон, 1974, с. 158.

Рассматривая индексирование, аннотирование и квазиреферирование текста в рамках теории автоматического распознавания смысла, нужно строго различать «машинно-семиотическую» природу выдаваемых ЭВМ результатов и их восприятие человеком.

В случае автоматической атрибуции выдаваемый машиной индекс можно рассматривать и в том и в другом случае как распознанный смысловой образ текста.

Машинная аннотация может выступать в качестве единого смыслового образа текста только с точки зрения интерпретирующего ее человека. Если не рассматривать аннотацию в плане машинной семиотики, то нужно будет признать, что эта аннотация представляет собой совокупность никак не связанных между собой смысловых образов (в нашем случае русских эквивалентов) ключевых слов и словосочетаний входного текста.

Наконец, если обратиться к реферированию, то выдаваемая ЭВМ свертка воспринимается адресатом как смысловой образ текста, в то время как для машины распознавание смысла ограничивается здесь выделением в тексте дескрипторов и реляторов.

Глава 10

ТЕЗАУРУСНОЕ РАСПОЗНАВАНИЕ СМЫСЛА

63. Построение тезауруса с помощью методов лингвистики текста

Слабость рассмотренных в 56—62 алгоритмов автоматического распознавания смысла текста состоит в том, что эти алгоритмы используют такие МИПЯ, в которых задается лишь перечень лексических единиц (терминов-дескрипторов) без указания на существующие между ними смысловые связи.

Получаемые аннотации представляют собой простое перечисление наиболее частых дескрипторов, о синтаксико-семантическом взаимодействии которых в тексте документа читатель аннотации должен догадываться сам, опираясь на собственные знания предмета, а также на профессиональную и языковую интуицию. Все это ограничивает возможности вероятностного и детерминистского распознавания, сводя его лишь к определению темы документа. Использование этой методики для машинного выявления ремы (новизны) документа имеет мало шансов на успех. Что же касается свертывания документа, опирающегося на анализ реляторов, которые воплощают сверхфразовые связи, то его применение ограничено лишь одноязычной ситуацией. Кроме того, получаемые

в результате свертки квазирефераты не могут быть использованы в качестве поисковых образов запросов и самих документов.

Более эффективное распознавание смысла можно получить, опираясь на тезаурусный подход, в котором объединяются дескрипторная и реляторная методики. Термином «тезаурус» обычно обозначается список лексических единиц (словоформ, канонических форм слова, основ, словосочетаний), между которыми заданы смысловые связи как иерархического (родо-видового), так и неиерархического (синонимия, антонимия, ассоциации) типов (ср.: Masterman, 1957/1964; Salton, 1968/1973; Бектаев, Пиотровский, Шабес, 1974, с. 28; Шемакин, 1974). Таким образом, тезаурус выступает как некоторая парадигматическая модель плана содержания, — модель, которая должна включать основные семантические единицы и связи описываемого языка или его разновидности. Эту модель обычно представляют в виде древовидного (чаще всего бинарного) графа.

Построение тезауруса предусматривает:

1) выбор СП и его разбиение на области (предметные классы), каждой из которых приписывается лингвистическая метка;

2) подбор терминологических слов и словосочетаний, выступающих в роли дескрипторов;

3) формирование классов условной эквивалентности;

4) формирование семантических полей (СПо), охватывающих определенные классы условной эквивалентности, а также слова и словосочетания, связанные с ними по смыслу;

5) задание родо-видовых и ассоциативных связей как внутри КУЭ и семантических полей, так и между отдельными КУЭ и разными СПо;[1]

6) формирование общей схемы тезауруса путем соотнесения КУЭ и СПо с областями СП.

Построенный этим путем тезаурусный словарь выступает не только как семантическая модель лексики языка или его разновидности. Тезаурус можно рассматривать также как иерархическую классификацию областей и подобластей внутри исследуемого СП. Лексические единицы, стоящие в вершинах древовидного графа, выступают в этих случаях в качестве названий (меток) соответствующих областей СПо, КУЭ. Наконец, тезаурусный граф можно использовать и как инструмент семантического и семантико-синтаксического кодирования словоформ, слов и словосочетаний (ср. 67).

Выделение СП, его областей, меток и дескрипторов проводится по тем схемам, по которым осуществлялось построение МИПЯ для вероятностного и детерминистского распознавания (55—56, 60). Более сложной задачей является формирование КУЭ и СПо, а также задание связей и структурализация тезауруса. Чаще всего

[1] Эти связи могут быть представлены либо в виде кодов, либо средствами естественного языка — в форме текстовых сегментов-реляторов.

здесь используется два подхода, опирающихся либо на индивидуальную, либо на коллективную интуицию. Во-первых, составитель может строить тезаурус, опираясь на собственное «виденье» объективной действительности без обращения к приемам остранения (ср.: Masterman, 1957/1964 и др. — см. также 13). Во-вторых, при составлении тезауруса используются научно-библиографические и терминологические иерархии типа УДК (ср.: Тезаурус. . ., 1972, с. 3; Словарь. . . , 1966, с. 4—5), опирающиеся на различные классификационные подходы, каждый из которых имеет свой принцип построения, отражающий научные позиции его создателей и специфику описываемого им материала.

Субъективизм этих подходов пытаются иногда корректировать методом экспертных оценок. Однако этот прием оказывается слишком громоздким и недостаточно эффективным (Salton, 1968/1973, с. 67). Поэтому более целесообразным является введение таких приемов остранения, которые опираются на массовые психолингвистические тесты, а также на методы лингвистики текста (ср.: Garvin, 1962, с. 138). Среди них наиболее перспективными являются следующие приемы.

1. Использование данных частотных словарей и информационных измерений семантики. Основная идея статистического подхода состоит в том, что наиболее часто используемые термины должны быть отнесены к более высоким ярусам классификации, а более редкие термины следует помещать на нижних уровнях графа (Salton, 1968/1973, с. 77; Черный, 1968).

Как показывает опыт группы «Статистика речи», наибольший эффект можно получить, используя многоступенчатый прием статистической корректировки тезауруса. По выборке текстов, представляющих определенное СП (подъязык), автоматически составляется частотный словарь. Из ЧС извлекаются термины, каждый из которых снабжается научно-техническим определением. Оно может быть заимствовано из отраслевых толковых или энциклопедических словарей, учебников или составлено специалистом в данной области знаний. При этом предполагается, что дефиниции включают родовой термин, а также один или несколько видовых спецификаторов. Поэтому каждую дефиницию можно рассматривать как имплицитную ветку тезауруса, описывающую родо-видовые отношения.

Родовые термины встречаются в дефинициях чаще, чем видовые термины и отдельные спецификаторы. Поэтому автоматизированная статистическая обработка дефиниций дает ЧС, в начале которого оказываются часто употребляющиеся терминологические словоформы, имеющие обычно общие значения, а затем идут более редкие и одновременно более конкретные терминологические единицы.

На следующем этапе ЧС дефиниционных словоформ преобразуется в частотный словарь терминов (это достигается путем удаления служебной и общеупотребительной лексики, а также приведения словоформ к канонической форме — см. 61. Как показывают результаты этой операции (см. табл. 59), родовые понятия оказываются статистически отграниченными от видовых, а последние от подвидовых спецификаторов. Размещение всех этих понятий и обозначающих их терминов по вершинам тезауруса, а также сопоставление получаемых результатов с традиционной классификацией и ее корректировка осуществляется специалистом в данной области знаний (ср.: Бектаев, Пиотровский, Шабес, 1974, с. 28).

Таблица 59

Фрагмент дефиниционного частотного словаря терминов по нефтехимии (ср.: Овсяников, Пиотровский, Сидоров, Шабес, 1974)

i	Термин	F
1	атом	335
2	соединение	196
3	вещество	167
4	группа	155
5	молекула	151
. . .	. . .	. . .
31	кислород	42
32	ряд	42
33	свойство	42
34	электрон	42
35	бесцветный	39
. . .	. . .	. . .
111	катализатор	12
112	кипение	12
113	молекулярный	12
114	нитрогруппа	12
115	переход	12

Разрешающую силу только что описанного приема остранения можно было бы увеличить, во-первых, учитывая характер распределения отдельных лексических единиц в тексте (28—29), а во-вторых, используя численные оценки смысловой информации, которую несет каждый термин. Можно предполагать, что родовые термины несут меньше смысловой информации, чем термины конкретного видового значения (52).

2. Исследование совместной встречаемости терминов. Основная идея этого метода заключается в предположении, что термины, подобные или связанные по значению, встречаются в тексте на небольшом удалении друг от друга. Выявить эту совместную встречаемость можно либо путем автоматического составления частотно-алфавитных списков сегментов (21—23), обнаруживающих контактную встречаемость

здесь используется два подхода, опирающихся либо на индивидуальную, либо на коллективную интуицию. Во-первых, составитель может строить тезаурус, опираясь на собственное «виденье» объективной действительности без обращения к приемам остранения (ср.: Masterman, 1957/1964 и др. — см. также 13). Во-вторых, при составлении тезауруса используются научно-библиографические и терминологические иерархии типа УДК (ср.: Тезаурус. . ., 1972, с. 3; Словарь. . . , 1966, с. 4—5), опирающиеся на различные классификационные подходы, каждый из которых имеет свой принцип построения, отражающий научные позиции его создателей и специфику описываемого им материала.

Субъективизм этих подходов пытаются иногда корректировать методом экспертных оценок. Однако этот прием оказывается слишком громоздким и недостаточно эффективным (Salton, 1968/1973, с. 67). Поэтому более целесообразным является введение таких приемов остранения, которые опираются на массовые психолингвистические тесты, а также на методы лингвистики текста (ср.: Garvin, 1962, с. 138). Среди них наиболее перспективными являются следующие приемы.

1. Использование данных частотных словарей и информационных измерений семантики. Основная идея статистического подхода состоит в том, что наиболее часто используемые термины должны быть отнесены к более высоким ярусам классификации, а более редкие термины следует помещать на нижних уровнях графа (Salton, 1968/1973, с. 77; Черный, 1968).

Как показывает опыт группы «Статистика речи», наибольший эффект можно получить, используя многоступенчатый прием статистической корректировки тезауруса. По выборке текстов, представляющих определенное СП (подъязык), автоматически составляется частотный словарь. Из ЧС извлекаются термины, каждый из которых снабжается научно-техническим определением. Оно может быть заимствовано из отраслевых толковых или энциклопедических словарей, учебников или составлено специалистом в данной области знаний. При этом предполагается, что дефиниции включают родовой термин, а также один или несколько видовых спецификаторов. Поэтому каждую дефиницию можно рассматривать как имплицитную ветку тезауруса, описывающую родо-видовые отношения.

Родовые термины встречаются в дефинициях чаще, чем видовые термины и отдельные спецификаторы. Поэтому автоматизированная статистическая обработка дефиниций дает ЧС, в начале которого оказываются часто употребляющиеся терминологические словоформы, имеющие обычно общие значения, а затем идут более редкие и одновременно более конкретные терминологические единицы.

На следующем этапе ЧС дефиниционных словоформ преобразуется в частотный словарь терминов (это достигается путем удаления служебной и общеупотребительной лексики, а также приведения словоформ к канонической форме — см. 61. Как показывают результаты этой операции (см. табл. 59), родовые понятия оказываются статистически отграниченными от видовых, а последние от подвидовых спецификаторов. Размещение всех этих понятий и обозначающих их терминов по вершинам тезауруса, а также сопоставление получаемых результатов с традиционной классификацией и ее корректировка осуществляется специалистом в данной области знаний (ср.: Бектаев, Пиотровский, Шабес, 1974, с. 28).

Таблица 59

Фрагмент дефиниционного частотного словаря терминов по нефтехимии (ср.: Овсяников, Пиотровский, Сидоров, Шабес, 1974)

i	Термин	F
1	атом	335
2	соединение	196
3	вещество	167
4	группа	155
5	молекула	151
...	...	...
31	кислород	42
32	ряд	42
33	свойство	42
34	электрон	42
35	бесцветный	39
...	...	...
111	катализатор	12
112	кипение	12
113	молекулярный	12
114	нитрогруппа	12
115	переход	12

Разрешающую силу только что описанного приема остранения можно было бы увеличить, во-первых, учитывая характер распределения отдельных лексических единиц в тексте (28—29), а во-вторых, используя численные оценки смысловой информации, которую несет каждый термин. Можно предполагать, что родовые термины несут меньше смысловой информации, чем термины конкретного видового значения (52).

2. Исследование совместной встречаемости терминов. Основная идея этого метода заключается в предположении, что термины, подобные или связанные по значению, встречаются в тексте на небольшом удалении друг от друга. Выявить эту совместную встречаемость можно либо путем автоматического составления частотно-алфавитных списков сегментов (21—23), обнаруживающих контактную встречаемость

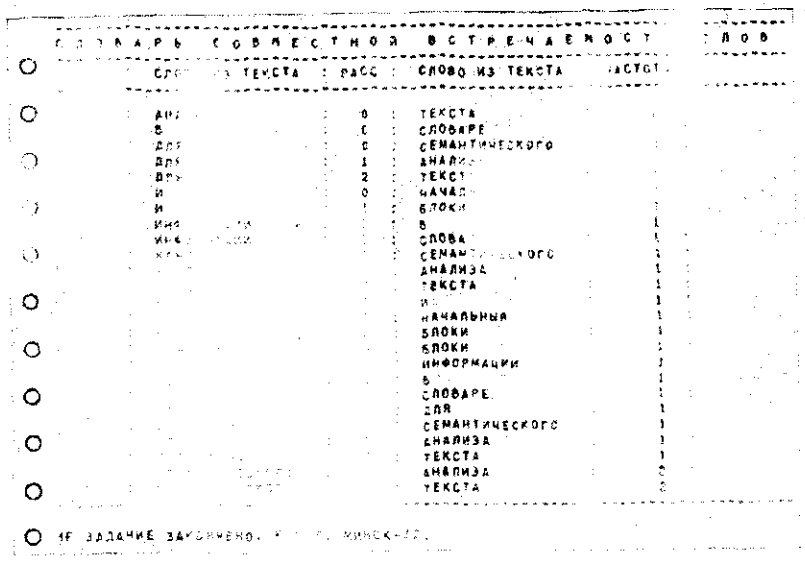


Рис. 52. Фрагмент машинного частотного списка совместной встречаемости словоформ (результат В. С. Крисевича).

терминов, либо с помощью специальных программ, выявляющих как контактную, так и дистантную совместную встречаемость (рис. 52; см. также: Джолос, 1965; Borko, Blankenship, 1968). Кроме того, степень подобия терминов оценивается в зависимости от частоты, с которой эти термины совместно встречаются в документах или тематических областях рассматриваемого СП. В качестве элементарного примера рассмотрим статистическую группировку французских сельскохозяйственных терминов. Как явствует из табл. 60, первые четыре термина совместно встречаются с близкими частотами в областях III и IV. Это дает право считать их значения связанными и отнести их к одному кусту тезауруса. Аналогичным образом можно поступить и с последними двумя терминами chargeur и hachoir-déchargeur. Эти статистические решения согласуются с семантическими связями, существующими внутри каждой из рассмотренных групп терминов. Что же касается пары boisseau 'буассо'—cellule 'секция силосной башни', то, входя в одни и те же области СП, эти термины дают слишком заметные расхождения частот и поэтому вряд ли могут быть включены в один куст тезауруса. Значения этих дескрипторов также не дают оснований для того, чтобы считать их подобными.

Наблюдения над совместной встречаемостью терминов широко использовались при построении американской АИПС SMART (Информационно-поисковая система. . ., 1966, с. 144 и сл.)

Таблица 60

Распределение частот французских сельскохозяйственных терминов по четырем областям СП ТСХМ (Рахубо*, 1974)

ПКС	Номера областей			
	III	IV	IX	X
ameneur 'лафет'	4	3	—	—
aligneur 'наклонная камера'	3	8	—	—
amasseur 'мотовило'	2	3	—	—
andainage 'образование валков'	2	15	—	—
boisseau 'буассо' (старинная мера сыпучих тел)	—	—	1	8
cellule 'секция силосной башни'	—	—	2	86
chargeur 'погрузчик'	2	—	—	4
hachoir-déchargeur 'силосноуборочная машина'	2	—	—	4

3. **Статистика семантико-синтаксических классов.** Эффективным средством коррекции интуитивных классификаций лексических единиц и понятий является текстовая статистика семантических и семантико-синтаксических классов. Идея этого подхода состоит снова в том, что наиболее часто встречающиеся в тексте семантические и синтаксические классы лексических единиц должны стоять в верхних ярусах древовидного тезауруса.

Эксперимент может быть реализован следующим образом. Каждому словоупотреблению обучающей выборки текста приписывается семантический или семантико-синтаксический код. Этот код берется либо из заранее построенного кодировочного древовидного графа или матрицы (Шабес*, 1973; Славина, 1974; Угодчикова*, 1975), либо вырабатывается путем психолингвистического текста (Salton, 1968/1973, с. 74 и сл.). Затем вся последовательность кодов обрабатывается на ЭВМ, которая выдает ЧС кодов или их комбинаций (кодовые диады, триады, тетрады и т. д.).

На основании полученного ЧС кодов строится тезаурусный граф семантических или семантико-синтаксических классов. Высокочастотные классы помещаются в верхние ярусы классификации, а низкочастотные ставятся в нижние ярусы иерархии. Полученный этим путем граф сравнивается с классификациями, полученными при помощи индивидуальной интроспекции или коллективной интуиции.

Описанные приемы интроспективного и формального построения тезауруса использовались в работах группы «Статистика речи» (Рахубо*, 1974; Славина, 1974; Шабес*, 1973).

64. Тезаурус СП «Сушка лакокрасочных поверхностей» и его машинная реализация

Основные идеи тезаурусного распознавания реализованы А. Н. Попеску и М. С. Хажинской (1973, с. 296—324) в машинном алгоритме аннотирования французского текста заглавной специальности.

Опираясь на имеющиеся классификации (УДК и др.), а также используя психолингвистический тест и дистрибутивно-статистические данные, М. С. Хажинская построила тезаурусный граф, отражающий родо-видовые и ассоциативные отношения между отдельными областями, СПо, КУЭ, и дескрипоторами исследуемого СП (рис. 53, 54).

Этот тезаурус, использующий некоторые приемы позиционного реферирования (ср.: Rathkirch-Trach, 1966) включает следующие три вспомогательных словаря-списка:

1) список русских названий (меток) областей и семантических полей тезауруса;

2) словарь дескрипторов, который включает списки французских словоформ и словосочетаний, составляющих классы условных эквивалентностей, а также русские метки этих КУЭ;

3) список русских словоформ и словосочетаний, выполняющих роль реляторов.

Выбор и функции меток и дескрипторов уже рассматривались в 56, поэтому остановимся на построении и функциях реляторов.

Если дескрипторы являются средством передачи терминологических значений, то реляторы выполняют в рассматриваемом МИПЯ семантико-синтаксическую функцию.

Различаются два типа реляторов: реляторы-индикаторы (r) и реляторы-шаблоны (rs). 33 релятора-индикатора (r, r_2,. . ., r_{33}) призваны выражать конкретные отношения между отдельными дескрипторами или группами дескрипторов. К этим отношениям принадлежат отношения результативности (ср. r_{24} — *вызывает такие физические явления, как*), пространственной или временной зависимости (ср. r_{11} — *производимая в*, r_8 — *уносимого из*, r_{33} — *в течение*) и т. п.

24 релятора-шаблона имеют более общее значение: они показывают, с какой точки зрения следует рассматривать вводимые ими дескрипторы. Каждый релятор-шаблон привязывается к определенному ярусу тезауруса и соответственно к определенному предметному классу, СПо, КУЭ и дескриптору. Эта привязка отражена в индексах реляторов-шаблонов. Так, например, шаблон *в общем плане рассматриваются вопросы. . .*, относящийся ко всему СП, получает обозначение rs_0, шаблон *описывается сушилка. . .*, связанный с предметным классом 2 — «Автоматическая сушка на конвейере», обозначается как rs_2, шаблон *приводятся и другие типы теплоносителей, как например. . .*, относящийся к предметному классу 3 — «Теплоноситель для сушки лакокра-

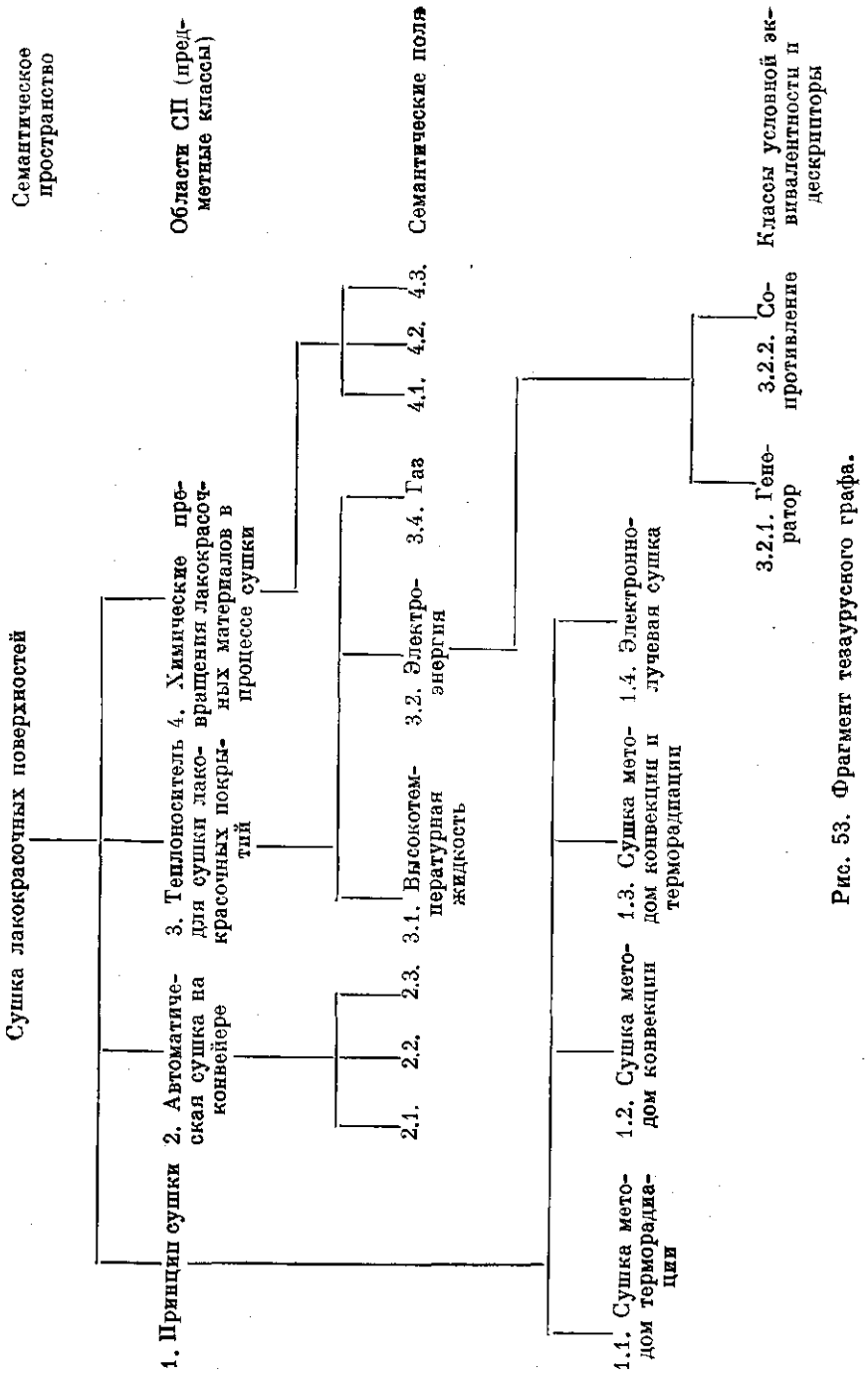


Рис. 53. Фрагмент тезаурусного графа.

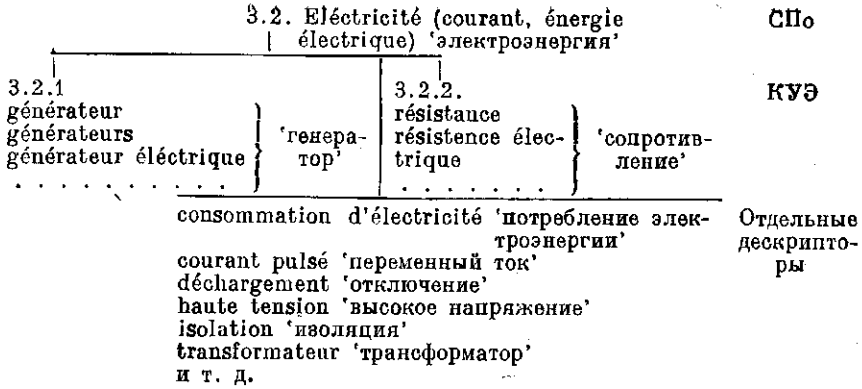


Рис. 54 Семантическое поле «3.2. Электроэнергия» из тезауруса СЛП, представленное во французских терминах.

сочных материалов», обозначается как rs_3, шаблон *подробно описываются химические вещества и их превращения*, ориентированный на одно из СПо четвертого предметного класса «Химические превращения лакокрасочных материалов в процессе сушки», обозначается символом rs_{42}, и т. д.

Реляторы-индикаторы и реляторы-шаблоны не только объединяют дескрипторы и метки, относящиеся к одному и тому же предметному классу, но способны также выступать в функции перекрестных ссылок, соединяя лексические единицы, относящиеся к разным кустам (классам, СПо) тезаурусного графа (подробнее см.: Попеску, Хажинская, 1972, с. 109—116).

* * *

Машинный алгоритм тезаурусного распознавания смысла, ориентированный на ЕС-ЭВМ, построен А. Н. Попеску. Схема записи информации к отдельным дескрипторам показана на рис. 55. Эти записи образуют документы D_1, которые объединяются в файл Ф-I.

Чтобы уяснить, каким образом записи отражают смысловые связи, существующие внутри тезауруса, рассмотрим три состояния байтов 25 и 26:

1) $I \neq 0$, $r_i \neq 0$ — дескриптор R_{ij} имеет связь со всеми дескрипторами, имеющими ключ I, связь воплощена в языковой форме релятора r_i;

2) $I = 0$, $r_i = 0$ — дескриптор R_{ij} не имеет связей с другими дескрипторами области A_j;

3) $I \neq 0$, $r_i = 0$ — дескриптор R_{ij} связан с дескрипторами, имеющими ключ I, языкового воплощения связь пока не имеет.

В записях русских переводов входных дескрипторов, меток и реляторов фиксируется их лингвистическая форма (т. е. записывается соответствующая словоформа или словосочетание) и указывается номер перевода метки (A_j) и релятора (r_i). При этом используются записи как постоянной (6 байт), так и переменной длины. Записи русских переводов дескрипторов оформляются как документы D_2, объединяющиеся в Ф-II. Записи меток и реляторов составляют документы D_3, попадающие в Ф-III.

Наконец, формируется файл Ф-IV, содержащий документы D_4 (записи меток A_j), упорядоченные по возрастанию j.

Вся эта информация, содержащаяся в тезаурусе, перфорируется и вводится в ЭВМ согласно блок-схеме, приведенной в работе: Попеску, Хажинская, 1973, с. 312—313.

Аннотирование текста с помощью только что описанного тезауруса осуществляется следующим образом.

Вводится текст, который проходит вероятностную атрибуцию относительно областей (A_j), заданных в тезаурусе (см. 56). Затем текст пропускается через фильтр антипризнаков и ядерных словоформ, в результате чего в нем выделяются потенциальные однословные и сложные дескрипторы вместе со стоящими рядом циф-

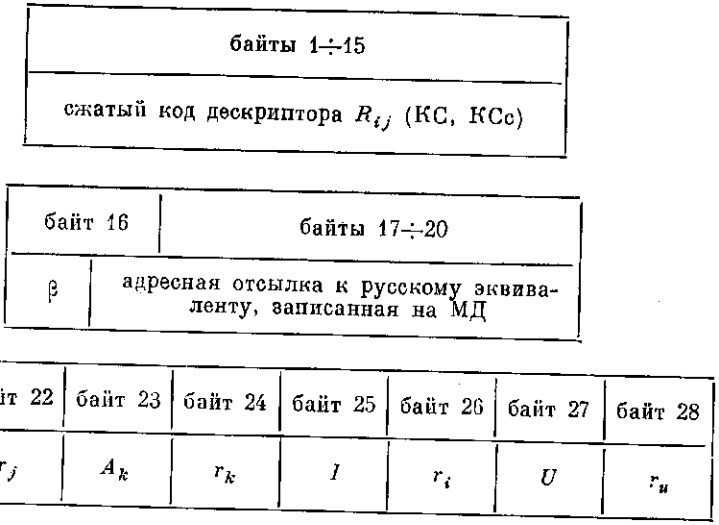


Рис. 55. Схема записи КС и КСс при тезаурусном распознавании смысла документа на ЕС-ЭВМ (документ типа D_1 из Ф-I).

Условные обозначения: A_j — область (предметный класс; СПо), к которой относится дескриптор R_{ij}; β — нулевой бит 16-го байта, принимающий значение 0, если R_{ij} имеет очень малую вероятность появления в области A_j, если эта вероятность не мала, то $\beta = 1$; r_j — релятор, связывающий R_{ij} с A_j; A_k — область, в которую дескриптор R_i не входит, но с которой он связан ассоциативной связью; r_k — релятор, воплощающий ассоциативную связь между R_{ij} и A_k; r_i — релятор, воплощающий связь между дескрипторами R_{nj} и R_{ij}, входящими в область A_j; I — ключ связанных дескрипторов R_{nj} и R_{ij}; r_u — релятор, выражающий связь между дескриптором R_{ij} из области A_j и дескриптором R_{kl}, относящимся к области A_l; U — ключ связанных дескрипторов R_{ij} и R_{kl}.

рами μ_1, μ_2. . . (если таковые имеются). Каждая из этих цепочек записывается в виде документа D_5 (ср. описание записей словосочетаний-заготовок в 60).

Происходит сравнение документов D_5 и D_1. При совпадении, кода словоформы (словосочетания) из документа D_5^i с аналогичным кодом из документа D_1^k формируется документ D_6^j, в который вносится вся информация из D_1^k и величины μ_1, μ_2. . . из D_5^i. Массив документов D_6 сортируется по возрастанию кодов меток областей A_j. Затем этот массив компрессируется путем удаления тех документов, у которых для всех одинаковых A_j $\beta=0$. Этим путем из дальнейшего рассмотрения устраняются дескрипторы, имеющие в области A_j очень малую вероятность. В результате этих операций создается файл Ф-VI. Затем с помощью адресных отсылок, перенесенных в D_6 из D_1 (ср. байты 16÷20, рис. 55) из соответствующих документов D_2, которые записаны на *MD*, выбираются русские эквиваленты входных дескрипторов. Документы D_6 преобразовываются в документы D_7, которые представляют собой цепочки следующего вида:

$$A_j,\ \mu_1,\ \mu_2,\ \mu_3,\ R_{ij},\ \beta,\ r_i,\ A_k,\ r_k,\ I,\ r_i,\ U,\ r_u.$$

Список Ф-VII, включающий документы D_7, разбивается на 6 подмассивов:

— первый подмассив содержит документы, у которых $I\neq0$, $U=0$, $r_u=0$, а r_i и r_j могут быть и равны и не равны нулю;

— второй подмассив включает документы со значениями $I=0$, $r_i=0$, $U\neq0$, r_j и r_u либо равны, либо не равны нулю;

— третий подмассив состоит из документов, у которых все указанные пять величин равны нулю;

— четвертый подмассив содержит документы, для которых все перечисленные величины, кроме r_j, равны нулю;

— пятый подмассив состоит из документов со следующими характеристиками: $I\neq0$, $r_i\neq0$, $U\neq0$, $r_u\neq0$, $r_j=0$ либо $r_j\neq0$;

— шестой подмассив включает документы, у которых $I\neq0$, $r_i=0$ или $r_i\neq0$, $U\neq0$, $r_u\neq0$ или $r_u=0$, $r_j=0$ (вариант $r_j\neq0$ не рассматривается).

Каждый из массивов характеризуется определенными схемами *связей, которые возникают между метками классов и СПо*, дескрипторами и соединяющими их реляторами.

В первом подмассиве существуют две схемы:

1-я схема $R_{hj}r_iR_{ij}$, если $r_i\neq0$, $r_j=0$ и

2-я схема $A_jr_jR_{hj}r_iR_{ij}$, если $r_i\neq0$, $r_j\neq0$, при $r_i=0$ схем связи нет.

Во втором подмассиве также имеется две схемы:

3-я схема $R_{ij}r_uR_{kl}$, если $r_u\neq0$, $r_j=0$ и

4-я схема $A_jr_jR_{ij}r_uR_{kl}$, если $r_u\neq0$, $r_j\neq0$, при $r_u=0$, связей с R_{kl} не возникает.

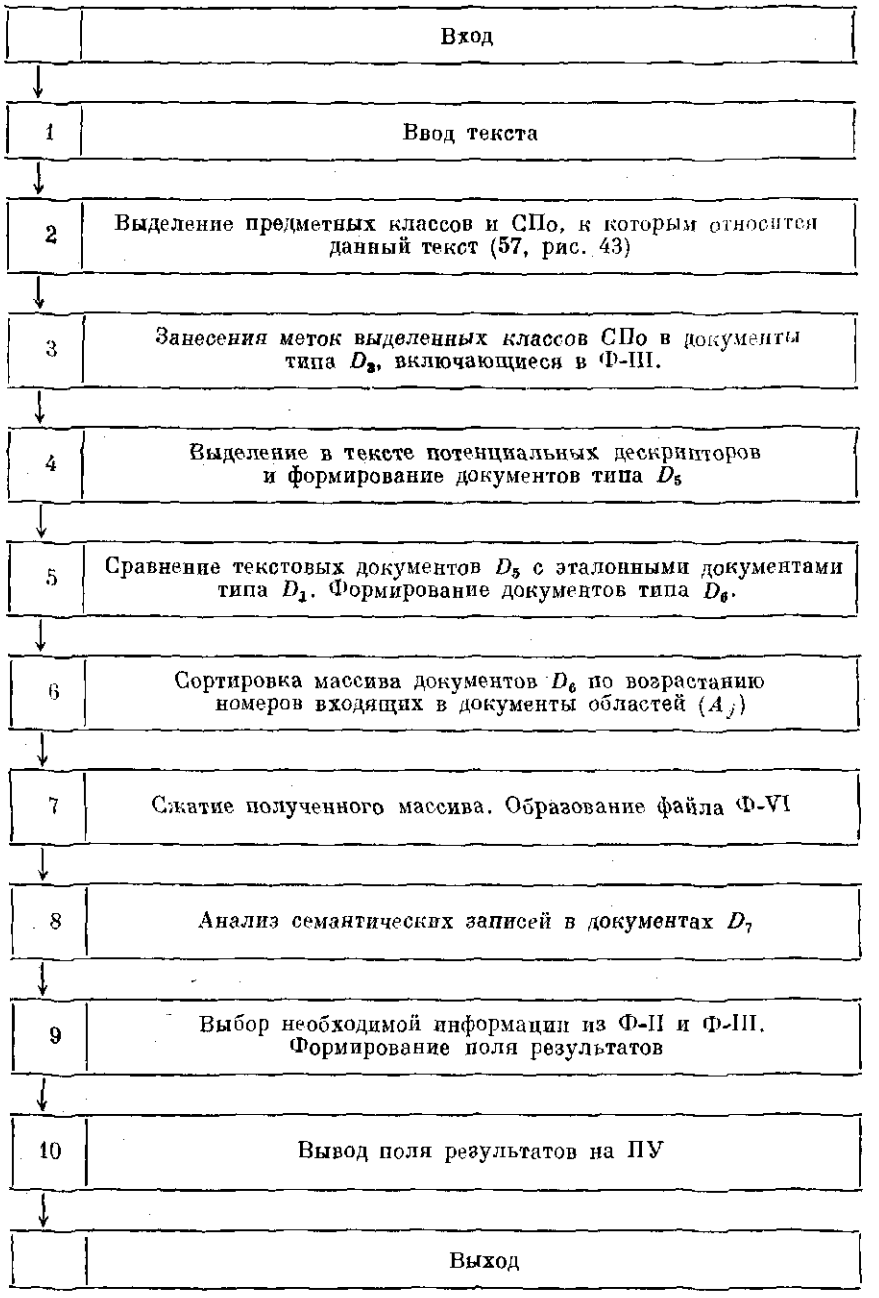


Рис. 56. Принципиальная блок-схема тезаурусного аннотирования текста на ЕС-ЭВМ.

В третьем подмассиве все русские эквиваленты R_{ij} в области A_j могут быть объединены релятором r_j (схема 5).

В четвертом подмассиве образуется единственная 6-я схема $A_j r_j R_{ij}$.

В пятом подмассиве снова возникают две схемы:

7-я схема $R_{kj} r_i R_{ij} r_u R_{kl}$, если $r_j = 0$;

8-я схема $A_j r_j R_{kj} r_i R_{ij} r_u R_{kl}$, если $r_j \neq 0$.

В шестом подмассиве также возможны две схемы:

9-я схема $R_{kl} r_u R_{ij} r_u R_{kk}$, при $r_i = 0$ и $r_u \neq 0$ и 1-я схема при $r_u = 0$.

Между разными предметными классами СПо также имеются свои семантические связи, поэтому 2-я схема может иметь расширенный вид:

$$A_1 r_j^1 A_2 r_j^2 \ldots r_j^m A_m r_i R_{ij} \ldots$$

Аналогичным образом могут быть расширены схемы 4, 6, 8. Русские эквиваленты из подмассивов I, II, V и VI, не вошедшие в схемы 1—4 и 6—10, объединяются в схему 5:

$$rs_j (R_{kj},\ R_{lj},\ R_{mj} \ldots),$$

в которой один релятор-шаблон может вводить целую группу дескрипторов.

На заключительном этапе работы алгоритма происходит формирование поля результатов, в которые заносятся русские эквиваленты дескрипторов из документов D_7, содержащиеся в Ф—II, а также русские обозначения меток и реляторов (Ф-III), адреса которых заложены в семантической записи документов D_7. Сформированное поле результатов выводится на печать (ср. рис. 56).

Ниже приводится результат тезаурусного аннотирования на ЭВМ французского текста по лакокрасочным покрытиям.

Задание машине (вопрос):

ДАТЬ АННОТАЦИЮ СТАТЬИ: «LE SE2CHAGE RAPIDE DE PEINTURE PAR BOMBARDEMENT E2LECTRONIQUE».

LES INSTALLATIONS DE COUCHAGE CLASSIQUE EMPLOYE2ES DANS L'INDUSTRIE SONT PRE2VUES POUR LE SE2CHAGE DE PEINTURE D'ORIGINE ORGANIQUE PAR LA CHALEUR, PARFOIS AVEC L'AIDE D'UN CATALYSEUR OR, CETTE ME2THODE COMPORTE UN CERTAIN NOMBRE D'INCONVE2NIENTS, DONT LES PRINCIPAUX SONT LES SUIVANTS: EMPLOI D'UN FOUR ASSEZ LONG, LEQUEL, AVEC LE POSTE DE REFROIDISSEMENT QUI EST SOUVENT NE2CESSAIRE, PEUT ATTEINDRE JUSDU'A8 60 M: FORTE CONSOMMATION D'E2NE2RGIE: DIFFICULTE2 D'E2VITER LA POUSSIE8RE AVANT SE2CHAGE COMPLET: RISQUE DE DE2TE2RIORATION DU MATE2RIAU DE BASE PAR LA CHALEUR, OU D'ADOUCISSEMENT DANS LE CAS DES ME2TAUX ET, TRE8S SOUVENT, PERTE DE SOLVANTS COU5TEUX OU FRAIS DE RE2CUPE2RATION DE CEUX-CI, ON SAIT, DEPUIS DE NOMBREUSES ANNE2ES, QUE L'IRRADIATION DE POLYME8RE CONVENABLEMENT CHOISIS PEUT DONNER, A8 L'AMBIANTE, DE COMPOSE2 RE2TICULE2 ET CE PROCE2DE2 A DE2JA8 E2TE2 APPLIQUE2, A8 L'E2CHELLE INDUSTRIELLE, AU SE2CHAGE DE MATIE8RES PLASTIQUES COMME LE POLYE2THYLE8NE (SOUS FORME DE PELLICULE OU AUTRE) AU MOYEN D'UN BOMBARDEMENT E2LECTRONIQUE SOUS UNE TENSION ASSEZ E2LEVE2 (FIG. 1). TOUTEFOIS, LE MATE2RIEL EXISTANT DANS CE DOMAINE NE POUVAIT E5TRE UTILISE2 POUR LE SE2CHAGE D'APPLICATION-DE FAIBLE E2PAISSEUR ET D'AUTRE PART, LES MATE2RIAUX HABITUELLEMENT EMPLOYE2S POUR CES APPLICATIONS ET DONT ON CONNAISSAIT BIEN LE COMPORTEMENT NE SE PRE5TAIENT PAS, D'UNE MANIE8RE GE2NE2RALE, A8 CE MODE DE SE2CHAGE. LES PREMIE8RES RECHERCHES, DE8S LES DERNIE8RES ANNE2ES 50, DES RECHERCHES E2TAIENT ENTREPRISES EN ANGLETERRE, AU TUBE INVESTMENT LABORATORIES (TIL), SUR LES POSSIBILITE2S DE SE2CHAGE DES PEINTURES PAR BOMBARDEMENT E2LECTRONIQUE. LES PREMIERS TRAVAUX, EFFECTUE2S AU MOYEN D'UN ACCE2LE2RATEUR DE VAN GRAAF, INDIQUAIENT QU'UN FAISCEAU D'E2LECTRONS D'E2NERGIE RELATIVEMENT MODE2RE2E (INFE2RIEURE A8 ENVIRON 250 KEV) POUVAIT PERMETTRE DE RE2ALISER DES E2CONOMIES SUBSTANTIELLES SI L'ON PARVENAIT A8 RE2SOUDRE LES PROBLE8MES LIE2S A8 LA MISE AU POINT MATE2RIEL APROPRIE2 ET A8 LA COMPOSITION DES PRODUITS D'APPLICATION NE2CESSAIRE. A CETTE E2POQUE, IL E2TAIT SURTOUT QUESTION D'APPLICATIONS SUR ME2TAUX ET LA FIRME DRYNAMEL FIT ALORS DE GROS TRAVAUX PRATIQUES DANS LE DOMAINE DES PRODUITS DE8S 1960, LES TRAVAUX E2TAIENT ASSEZ AVANCE2S POUR JUSTIFIER LE DE2PO5T D'UNE DEMANDE DE BREVET (LEQUEL BREVET A, PAR LA SUITE, E2TE2 HOMOLOGUE2 EN ANGLETERRE SOUS LE X) PORTANT SUR UN PROCE2DE2 DE SE2CHAGE OU DURCISSEMENT D'UN REVE5TEMENT EN RE2SINE SYNTHE2TIQUE ORGANIQUE CONSISTANT A8 SOUMETTRE LE DIT REVE5TEMENT A8 L'ACTION DE RADIATIONS IONISANTES SOUS FORME D'E2LECTRONS D'UNE ENERGIE EFFICACE NE DE2PASSANT PAS 250 KEV EN COLLABORATION AVEC LES BRITISH ALUMINIUM RESEARCH LABORATORIES QUI E2TAIENT, A8 L'E2POQUE, INTE2RESSE2S PAR UNE ME2THODE DE SE2CHAGE DES APPLICATIONS SUR BANDE D'ALUMINIUM, UNE LIGNE DE COUCHAGE ET DE SE2CHAGE E2TAIT E2TUDIE2E ET RE2ALISE2E, A8 TITRE D'INSTALLATION-PILOT, AUX TUBE INVESTMENTS RESEARCH LABORATORIES. CETTE INSTALLATION... DONNA DES RE2SULTATS QUI CONFIRMAIENT LES ESPE2RANCES, TOUT EN INDIQUANT QU'IL RESTAIT ENCORE BEAUCOUP A8 FAIRE, EN PARTICULIER POUR LA FABRICATION DE PRODUITS D'ORIGINE ORGANIQUE SUSCEPTIBLES DE SE2CHER ET DE DURCIR SOUS L'ACTION DE RADIATIONS PRODUITES DANS DES CONDITIONS RENTABLES ET D'ADHE2RER CONVENABLEMENT A8 LA SURFACE ME2TALLIQUE. PARALLE8LEMENT, DES PROGRE8S CONSIDE2RABLES E2TAIENT RE2ALISE2S DANS LE SENS DE LA SOLUTION DE PROBLE8ME LIE2 A8 LA MISE AU POINT D'UNE FENE5TRE PERMETTANT LE PASSAGE D'UN PUISSANT FAISCEAU D'E2LECTRONS EN PROVENANCE DE LA CHAMBRE A8 VIDE DE L'ACCE2LE2RATEUR DE PARTICULE VERS L'EXTE2RIEUR (C'EST-A8-DIRE VERS LA SURFACE TRAITE2E) TOUT EN RE2SISTANT A8 LA PRESSION DE L'AIR ATMOSPHE2RIQUE SUR UNE SUPERFICIE APPRE2CIABLE VERS LA FIN DE CETTE PE2RIODE, DE TRAVAUX E2TAIENT POURSUIVIS INDE2PENDAMMENT PAR LE VANTAGE RESEARCH LABORATORIES DE LA DIRECTION DE L'E2NERGIE ATOMIQUE DU ROYAUME-UNI, QUI AVAIENT E2GALEMENT CONSTRUIT UNE MACHINE PROTOTYPE, MAIS QUI SE SPE2CIALISAIENT DANS LE DOMAINE DE APPLICATION SUR BOIS ET AUTRES MATE2RIAUX NON ME2TALLIQUES, POUR LESQUELLES LES PRODUITS SENSIBLES AUX RADIATIONS QUI EXISTAIENT ALORS E2TAIENT PLUS FACILEMENT ADAPTABLES A8 CETTE TECHNIQUE, PARTANT DES RECHERCHES PRE2LIMINAIRES EFFECTUE2ES PAR LE WRL. UNE AUTRE FIRME ANGLAISE, PORTER PAINT, SE MIT ALORS EN RAPPORT AVEC TUBE INVESTMENT (TITULAIRES DU BREVET COUVRANT LE PROCE2DE2) ET LA MISE EN COMMUN DE RE2SULTAT TECHNIQUE OBTENUS DANS LE DOMAINE DE LA FABRICATION DE PRODUITS, D'UNE

PART, ET DANS CELUI DE LEUR MISE EN OEUVRE. INSTALATION-TE2MO EN VRAIE GRANDEUR... AU TERMES DE L'ACCORD CI-DESSUS, LE MATE2RIEL TIGER DEVAIT E5TRE INSTALLE2 DANS LES LABORATOIRES D'UNE AUTRE FIRME ENCORE, SISSIONS BROS, QUI FAIT PARTIE DU GROUPE PORTER PAINTS. LA TA5CHE A8 RE2ALISER E2TAIT DOUBLE: MISE AU POINT D'UNE GAMME DE PLUS EN PLUS E2TENDUE DE PRODUITS SPE2CIAUX ET EXPLOITATION DE L'INSTALLATION EN VUE DE FOURNIR LES PARAME8TRES D'INDUSTRIALISATION. CE DOUBLE OBJECTIF IMPLIQUAIT L'EXPLOITATION DU MATE2RIEL DANS DES CONDITIONS D'UTILISATION EXTRE5MEMENT VARIABLES. CE MATE2RIEL SE COMPOSAIT DE TROIS E2LE2MENTS PRINCIPAUX: LE POSTE D'APPLICATION (ICI, UNE COUCHEUSE AU VOILE), LE POSTE DE BOMBARDEMENT E2LECTRONIQUE ET UN DISPOSITIF TRANSPORTEUR ASSURANT LE CHEMINEMENT DE PIE8CE A8 TRAVERS L'ENSEMBLE A8 LA VITESSE APPROPRIE2E. LES PIE8CES SONT ICI DE PLAT D'AGGLOME2RE2 DONT LES DIMENSIONS MAXIMALES SONT DE 46 CM DE LONG PAR 15 DE LARGE ET QUI SONT FIXE2S, AU MOYEN DE RUBAN ADHE2SIF DOUBLE FACE, SUR DES TROLLEYS EN ALLIAGE DE MAGNE2SIUM SE DE2PLAC7ANT SUR DE PAIL SOUS L'ACTION D'UNE COMMANDE A8 COURROIE A8 FRICTION. CES TROLLEYS A8 FAIBLE INERTIE SONT RAPIDEMENT ACCE2LE2RE2S POUR ATTEINDRE LA VITESSE DE COUCHAGE AU VOILE, LAQUELLE PEUT VARIER DE 60 M/MN A8 600 M/MN, SELON LE CAS. LE COUCHAGE. LA MAJEURE PARTIE DE TRAVAUX PRATIQUE2S A E2TE2E BASE2E SUR L'EMPLOI D'UNE COUCHEUSE AU VOILE, PROCE2DE2 QUI SEMBLE TOUT PARTICULIE8REMENT ADAPTE2 A8 LA NOUVELLE ME2THODE DE SE2CHAGE. LA COUCHEUSE COMPORTE UNE FILIE8RE DONT L'OUVERTURE D'ENVIRON 15 CM DE LONGUEUR PEUT LIVRER PASSAGE A8 UN VOILE DE 0,5 MM D'E2PAISSEUR. LA PEINTURE TOMBE AINSI SUR LA PIE8CE SOUS FORME D'UN VOILE NET, A8 UNE VITESSE TRE8S PROCHE DE CELLE DE LA CHUTE LIBRE, LA LAIZE E2TANT MARGE2E PAR DEUX FILS ME2TALLIQUES VERTICAUX. L'EXCE2DENT DE PEINTURE EST RENVOYE2 EN CONTINU AU BAC CYLINDRIQUE DE LA FILIEBRE A8 TRAVERS UN FILTRE. LE PRINCIPAL AVANTAGE DU COUCHAGE AU VOILE, DANS CE CAS PARTICULIER, C'EST QU'IL S'ADAPTE A8 LA VITESSE E2LEVE2E DU SE2CHAGE. AVEC UN AUTRE SYSTE8ME D'APPLICATION, FAUDRAIT TENIR COMPTE DU TEMPS NE2CESSAIRE A8 LA FORMATION D'UN FILM LISSE ET UNIFORME AINSI QU'A8 L'E2VAPORATION DU SOLVANT EMPLOYE2 POUR FACILITER L'APPLICATION ET RE2DUIRE LE TEMPS D'E2COULEMENT POUR LE SE2CHAGE E2ELECTRONIQUE, IL N'EST PAS, NORMALEMENT, QUESTION DE SOLVANT ET, D'AUTRE PART, LE FILM OBTENU AVEC UNE COUCHEUSE AU VOILE EST UNIFORME ET HOMOGE8NE, MALGRE2 LA VISCOSITE2 E2LEVE2E DU PRODUIT.

Ответ:

ДАННАЯ СТАТЬЯ ОТНОСИТСЯ К ПОДТЕМЕ «СУШКА ЛАКОКРАСОЧНЫХ ПОКРЫТИЙ».

Аннотация статьи:

ПРИНЦИП СУШКИ. АВТОМАТИЧЕСКАЯ СУШКА НА КОНВЕЙЕРЕ. ПРЕДУСМАТРИВАЕТСЯ: СУШКА, ПОЛИМЕРИЗАЦИЯ, ИСПАРЕНИЕ, ОХЛАЖДЕНИЕ, ЭЛЕКТРОННОЛУЧЕВАЯ СУШКА

СПЕЦИАЛЬНОЕ ОБОРУДОВАНИЕ: АНОД, КАТОД, ЭЛЕКТРОННЫЙ УСКОРИТЕЛЬ, ПОСТ ОХЛАЖДЕНИЯ ЭЛЕКТРОННОЙ БОМБАРДИРОВКОЙ.

ЭЛЕКТРОННОЛУЧЕВАЯ СУШКА ВЫЗЫВАЕТ ТАКИЕ ФИЗИЧЕСКИЕ ЯВЛЕНИЯ, КАК БОМБАРДИРОВКА, ИРРАДИАЦИЯ ПОЛИМЕРОВ, РАДИАЦИЯ, ДЕЙСТВИЕ РАДИАЦИИ.

РАССМАТРИВАЕТСЯ ВОПРОС ОЦЕНКИ КАЧЕСТВА ПЛЕНКИ. ОПИСЫВАЮТСЯ ХИМИЧЕСКИЕ ВЕЩЕСТВА И ИХ ПРЕВРАЩЕНИЯ: ЛАКОКРАСОЧНЫЙ МАТЕРИАЛ, СОЛЬВЕНТЫ, СМОЛЫ.

ОСВЕЩАЮТСЯ ВОПРОСЫ, СВЯЗАННЫЕ С ИНТЕНСИВНОСТЬЮ СКОРОСТИ ОБМЕНА И ВЫБРОСА ЛЕТУЧИХ ВЕЩЕСТВ, РАСПРЕДЕЛЕНИЕМ ТЕМПЕРАТУРЫ ПО ДЛИНЕ И ПАРАМЕТРУ СУШИЛЬНОЙ ПЕЧИ, РАСТВОРИТЕЛЕМ, УНОСИМЫМ ВЕНТИЛЯЦИЕЙ.

ОБРАЩАЕТСЯ ВНИМАНИЕ НА ТАКИЕ ФАКТОРЫ, КАК ПОВЕРХНОСТЬ.

В ЗАКЛЮЧЕНИЕ ПРИВОДЯТСЯ СЛЕДУЮЩИЕ ДАННЫЕ: ВРЕМЯ ЭЛЕКТРОННОЙ СУШКИ ДЕТАЛИ МЕНЕЕ 1 СЕКУНДЫ.

Каждая из фраз аннотации (машинный оригинал см. в: ЛААТ, 1973, с. 324) представляет реализацию одной из приведенных выше формул. Например, по 2-й схеме построено предложение *электронно-лучевая сушка* ($A_{1.4}$) *вызывает такие физические явления, как* {r_{24}} *бомбардировка* ($R_{1.1.4}$), *иррадиация* полимеров ($R_{2.1.4}$), *радиация* ($R_{3.1.4}$}, *действие радиации* {$R_{4.1.4}$}, 5-я схема реализуется в предложении: *описываются химические вещества и их превращения* {r_{42}}; *лакокрасочный материал* {$R_{5.1.3}$}, *сольвенты* {$R_{5.1.5}$}, *смолы* {$R_{5.1.4}$}.

65. Многоотраслевой автоматический словарь русского языка

Многолетний опыт деятельности группы «Статистика речи» в области автоматической переработки русских и иноязычных текстов говорит о том, что задачи автоматического распознавания смысла, возникающие при индексации, аннотации, реферировании и машинном переводе документа должны решаться на основе многоотраслевого автоматического русского словаря (МАРС), охватывающего бо́льшую часть лексики (в идеале — всю лексику) современного русского языка, включая и специальную терминологию (Беляева, Крисевич, Липницкий, Пиотровский, 1975).

Необходимость построения такого словаря диктуется следующими соображениями.

1. Поскольку основная часть синтактической, сигматической и семантической информации текста концентрируется в лексике и фразеологии языка (ср. 21, 43), основным блоком автоматической переработки должен быть АС, в который следует заложить основную часть грамматической информации, необходимой для анализа и синтеза текста.

2. Нет необходимости строить автономные словари для вероятностного, детерминистского и тезаурусного распознавания, для свертывания текста и его машинного перевода. Во всех этих видах переработки текста используются однотипные операции, для осуществления которых часто необходима одна и та же лексико-

логическая и грамматическая информация. Эту информацию удобнее всего хранить в одном универсальном АС.

3. Если переработка текста на ЭВМ осуществляется в условиях двуязычной ситуации, то описание семантических пространств и их областей, а также некоторых операций по распознаванию целесообразно осуществлять средствами выходного (русского) языка. При таком подходе отпадает необходимость повторять в каждом из входных алгоритмов описания семантических пространств и их областей, а также строить в этих алгоритмах одни и те же семантические программы.

Как уже говорилось (12), весь корпус лексики русского языка оценивается в один миллион слов. Если, учитывая опыт организации АС (см.: Вертель, Вертель, Крисевич, Пиотровский, Трибис, 1971, с. 34), строить МАРС в виде автоматического словаря словоформ, то каждое словарное гнездо нужно будет развернуть примерно в десять словоформ (Nündel, Klimonow, Starke, Brand, 1969, с. 23) и весь словарь будет содержать около 10 млн. единиц.

Попробуем теперь оценить тот объем памяти, который понадобился бы для некомпрессированной записи всей информации, необходимой для функционирования МАРС.

Каждая словоформа снабжается набором кодов, дающих характеристику ее грамматических и семантических особенностей. Если словоформу идентифицировать не набором кодов-признаков, а только номерами этих наборов в некотором заранее составленном списке (ср.: Борисевич, Гончаренко и др., 1970), для размещения словоформы вместе с грамматическими характеристиками понадобится 30 байт (20 байт — на запись словоформы, 10 байт для информационного кода) и еще 5 байт придется отвести для семантического кода, характеризующего каждое словарное гнездо. При таком подходе для хранения словаря понадобится как минимум $30 \cdot 10^7 + 5 \cdot 10^6 = 305$ Мбайт.

Посмотрим теперь, как соотносится этот словарь с объемом памяти наиболее мощных отечественных машин третьего поколения. В ближайшее время ожидается серийное производство электронных вычислительных машин типа ЕС-1050. В дополнение к сведениям, приведенным в табл. 1, укажем на следующие характеристики этой машины. Объем ее оперативной памяти составляет, в зависимости от комплектации, от 512 до 1024 Кбайт. В минимальный комплект ее внешней памяти входит 8 накопителей на магнитной ленте (МЛ), работающих со сменными бобинами магнитной ленты, на каждую из которых можно записать 25 М байт информации. Кроме того, при рассмотрении возможностей ЕС-1050 следует особое внимание обратить на накопители на сменных магнитных дисках (МД). В минимальном комплекте их 5; в случае необходимости можно использовать дополнительные комплекты. Объем каждого сменного пакета дисков — 7.25 Мбайт. Среднее время доступа к информации, записанной на дисках,— 90 м/сек. (Шелихов, Селиванов, 1973, с. 23—27). Общая емкость внешней памяти ЭВМ ЕС-1050, таким образом, составляет:

8 МЛ:	200.0 Мбайт
5 МД:	37.5 Мбайт
Итого:	237.5 Мбайт

Это в полтора раза меньше того объема, который необходим для некомпрессированного хранения и функционирования МАРС.

Преодолеть это несоответствие между ограниченной емкостью памяти ЭВМ и объемом информации в МАРС можно либо за счет введения системы сменных носителей информации, либо путем сжатого представления в машине словарной информации. Первый путь, состоящий в увеличении емкости памяти ЭВМ за счет широкого использования сменных дисков и магнитных лент, привел бы к отказу от принципа универсальности, на котором должен строиться МАРС. Задачу размещения МАРС в памяти ЭВМ целесообразнее решать за счет компрессии информации, задаваемой словарем. Необходимо, сохранив число разных словарных гнезд, входящих в словарь русского языка, максимально сократить объем поля, необходимого для записи гнезда.

Наиболее распространенным способом такого сокращения является сжатие парадигмы каждого слова по методу Купера (Cooper, 1958; Курбаков, 1968, с. 32—33). Существо этой компрессии заключается в замене общей части двух словоформ специальным «числом Купера», показывающим, сколько букв от предыдущей словоформы сохраняются неизменными в следующей словоформе. За этим числом указываются буквы, которые следует добавить к буквенной цепочке, сохранившейся от предыдущей словоформы. Использование сжатия Купера в АС группы «Статистика речи» (Крисевич*, 1970) показало, что на запись только словоформы потребуется в среднем 5 байт вместо 20 байт при некомпрессированном побуквенном кодировании. Однако свертка Купера не касается лексико-грамматической и семантической информации, которая по-прежнему занимает 10 байт на словоформу и 5 байт на гнездо. Поэтому использование свертки Купера дает только двойное сокращение памяти, необходимое для хранения МАРС. При этом следует иметь в виду, что поиск в словаре, использующем свертку Купера, осуществляется медленно, так как для получения нужной словоформы приходится применять алгоритмическое развертывание всего гнезда.

Более эффективным является изменение структуры словаря и отказ от явного задания всех словоформ одной парадигмы и использования адресных отсылок к отдельным автономным подмножествам информации. При этом АС должен представлять собой сочетание словарей двух типов: словаря машинных основ и традиционного словаря словоформ.

В словаре основ все словоформы парадигмы порождаются от машинной основы, представляющей собой наибольшую начальную цепочку букв, встречающуюся во всех членах парадигмы. Машинная основа может совпадать, а может и не совпадать с традиционной основой.

Порождение каждой словоформы осуществляется путем прибавления к машинной основе машинных окончаний, которые также могут быть идентичными или неидентичными машинным окончаниям.

Рассмотрим в этой связи два существительных — *чугун* и *сварка*. Первое дает машинную парадигму, использующую традиционную основу и окончания *чугун-а*, *чугун*, *чугуну*. . . и т. д.; второе образует парадигму, использующую нетрадиционную основу и окончания: *свар-ка*, *свар-ки*, *свар-ке*, *свар-ку*, *свар-кой*, *свар-ки*, *свар-ок*, *свар-кам* и т. д.

При таком подходе вся парадигма записывается в памяти ЭВМ в виде словарного гнезда, в начале которого располагается машинная основа, а за ней, разделенные специальным признаком, даются адресные ссылки к типовым парадигмам машинных флексий. Флексии в этих парадигмах могут быть либо ранжированы по степени их употребительности, либо расположены в алфавитном порядке. Всего в МАРС используется около 180 типовых парадигм.

В МАРС входит некоторое количество лексем, которые при описанном выше подходе будут давать омонимию основ. Примером могут служить существительные *лев* и *лён*, имеющие одну машинную основу *л*; ср.: *л-ев*, *л-ьва*, *л-ьву* . . . и т. д. и *л-ён*, *л-ьна*, *л-ьну* . . . и т. д. Чтобы избежать ошибок при отождествлении этих омонимичных основ парадигмы слов указанного типа приходится задавать в виде последовательности словоформ, записанных по методу Купера.

В каждой словарной статье МАРС, кроме основы или исходной формы, а также указания на способ формообразования, должны быть заложены следующие сведения:

а) номер словарной статьи (ср.: Ястребова, Дегтярева, 1972; Берман, Мандрыка, Рабинович, 1974);

б) грамматическая информация о слове, задаваемая с помощью сжатого кода (СК), вторичного сжатого кода (ВСК) или сжатого кода парадигмы (СКП) — грамматический код, признак фразеологичности, лексико-семантический код, код подъязыка (или подъязыков), в котором используется данное слово, указание на вхождение слова в эталонные списки или списки граничных каналов и коды этих списков, код родо-видовых связей, код ассоциативных связей. Введение всех этих сведений в МАРС дает возможность использовать его в качестве универсального хранилища той информации, которая необходима для решения на ЭВМ различных задач переработки текста, начиная от простой индексации и кончая его семантико-синтаксическим анализом.

Всю указанную информацию предполагается разместить в словарной статье следующим образом.

Номер (адрес) статьи записывается в трех байтах. Машинные основы группируются в три класса: короткие основы (до 5 букв), основы средней длины (от 6 до 10 букв), длинные основы (от 11 до 35 букв). Для записи в машинной словарной статье коротких основ используется поле в 5, для средних — в 10, для длинных — в 35 байт. Для записи адреса и длины машинной парадигмы окончаний, прибавляемых к основе, отводится один байт. Сами машинные парадигмы окончаний (их в русском языке для всех частей речи будет, очевидно, около 200) записываются в отдельном поле памяти ЭВМ.

Если словоформы, образующие данное гнездо, не могут быть построены путем комбинации машинной основы с машинными окончаниями (ср. слова *лев* и *лён*), то поле основ и поле адреса парадигмы объединяются в одно поле, в которое заносится вся последовательность словоформ, сжатых по методу Купера.

Грамматическая информация, записываемая в одно полуслово, задается с помощью СК, ВСК или СКП — см.: Вертель, Вертель, Крисевич, Пиотровский, Трибис, 1971, с. 39.

Признак фразеологичности, занимающий поле переменной длины, указывает на способность слова выступать в качестве ядра оборота. Если слово не образует оборота, то нулевой бит первого байта принимает значение 0, если же слово выступает в качестве ядра оборота или оборотов, то нулевой бит получает значение 1. Далее указывается позиция ядра относительно других словоформ, входящих в оборот, а также даются адреса этих словоформ. Код, указывающий на принадлежность слова к определенному лексико-семантическому классу (ср.: Угодчикова *, 1975), занимает два байта.

Для записи кода подъязыка или сжатого кода группы подъязыков, к которым относится данное слово, отводится один байт. Одно полуслово отводится для записи номера дескрипторного или реляторного списка, в который входит данное слово или образуемый им оборот. Если слово выступает в качестве граничного сигнала, (ср. 23), то здесь вводится номер того списка границ, к которому относится данное слово.

Коды, отражающие тезаурусную организацию МАРС (родовидовые и ассоциативные связи), помещаются в поле, состоящее из двух слов (8 байт). Организация этого поля аналогична построению поля, в котором записывается информация при тезаурусном распознавании документа (см. рис. 55, байты 21—28).

Организация словарной статьи МАРС показывает, что этот словарь объединяет в себе тезаурусную структуру вместе с традиционными принципами построения АС, использующими компрессированное представление лингвистической информации.

Можно показать, что каждая статья МАРС записывается в среднем в 35 байт, а для записи всего словаря понадобится около 35 Мбайт, что вполне соответствует возможностям дисковой памяти наиболее мощных ЕС-ЭВМ (см. табл. 1).

Программное обеспечение МАРС предусматривает автоматическое пополнение словаря новыми словарными статьями, а также введение новой лингвистической информации в уже имеющиеся статьи.

Построение МАРС связано с объединением большого числа словарей, в том числе и отраслевых тезаурусов. Необходимость объединения отраслевых тезаурусов возникает и при построении ИПЯ для Единой автоматизированной системы научно-технической информации. В связи с тем что объединение словарей, и особенно словарей тезаурусного типа, представляет сложную и трудоемкую задачу, целесообразно проводить ее с помощью ЭВМ.

Вся процедура объединения строится на сочетании дедуктивного (классификационного) и индуктивного (текстового) подходов. Это значит, что объединение, с одной стороны, опирается на уже имеющиеся классификационные иерархии типа УДК или Тезауруса научно-технических терминов, 1972, а с другой — базируется на информации, получаемой с помощью приемов лингвистики текста (Мотылев, Пиотровский, Сова, Шабес, 1974, с. 10—12).

Последовательность операций при объединении тезаурусов должна выглядеть следующим образом.

1. Выбирается общая научно-техническая классификация, представляемая в виде древовидного графа, и аналогичным образом оформленные отраслевые тезаурусы.

2. На основании уже имеющихся лингвистических описаний тезаурусов окончаниям (веткам) классификационного дерева приписываются наборы КС и КСс, а к узлам привязываются списки ПКС и ПКСс, которые относятся ко всему кусту. Все эти лексические единицы вводятся в списки вместе с их вероятностными или информационными весами, если, разумеется, таковые имеются. Каждому КС, ПКС, КСс и ПКСс приписывается код той ветки или узла, к которым привязана данная лингвистическая единица.

3. Проверяется соответствие кодов КС, ПКС, КСс и ПКСс их значениям и функциям в отраслевых тезаурусах. На основании этой проверки осуществляется первая коррекция общей иерархической классификации.

4. Автоматически обрабатываются научно-технические документы по различным областям знаний, в результате чего для каждого документа формируется ПОД, представляющий собой список КС, ПКС, КСс и ПКСс с их статистическими весами (55).

5. Каждый ПОД последовательно сравнивается со списками, привязанными к узлам и веткам классификационного дерева (ср. п. 2). При сравнении учитываются веса лингвистических единиц.

6. В результате автоматического сравнения, осуществленного в п. 5, происходит пополнение и перераспределение списков КС, ПКС, КСс и ПКСс. Одновременно осуществляется вторичная классификация, объединение или разведение узлов и ветвей дерева с соответствующей коррекцией кодов.

7. Производится статистическая обработка дефиниций (63), по результатам которой происходит третья коррекция иерархической классификации. Одновременно достраивается верхняя (общепонятийная) часть дерева.

8. Путем исследования совместной встречаемости терминологических слов и словосочетаний в тексте (63) выявляются ассоциативные связи, которые затем вводятся в общее классификационное дерево.

Глава 11

МАШИННЫЙ ПЕРЕВОД

66. Лингвистика текста и машинный перевод

Машинный перевод является высшей формой автоматической переработки текста. С одной стороны, если машинная атрибуция, аннотирование и реферирование, имеющие целью извлечь из текста только общий смысловой инвариант (обычно тему текста), ограничиваются простыми приемами распознавания смысла у отдельных элементов текста, то МП, ориентированный на полную смысловую переработку текста, использует более сложные формы распознавания, охватывающие в идеале все единицы текста. Поэтому, с одной стороны, в алгоритмах МП оказываются представленными все приемы отбора, сортировки, сегментации, сравнения и отождествления ЛЕ, использующиеся в низших формах переработки текста. С другой стороны, успешное создание действующего МП в гораздо большей степени, чем построение простых форм переработки зависит от выбора правильной лингвистической теории.

Как уже говорилось (13—16), разработка систем МП может осуществляться либо на основе теории порождающих грамматик, либо опираясь на идеи лингвистики текста.

Порождающая грамматика обещает заманчивую перспективу полного формального распознавания и порождения любого правильного текста при условии, что будет определен словарь исходных единиц и глубинных структур, а также задан алгоритм порождения. Однако реализация этой перспективы наталкивается на барьеры антиномии индивидуального и коллективного восприятия языка, а также антиномии человека и робота (10—11).

Выделение исходных единиц и глубинных структур, осуществляющееся на основе индивидуальных интроспекций без учета

коллективного опыта носителей языка и информационной многоуровневой иерархии текста, дает обычно неоднозначные результаты. Приведение этих результатов к инварианту, описываемое в теории нечетких множеств и алгоритмов (гл. 8), упирается в барьер антиномии человека и робота, сущность которой состоит в том, что мышление человека, хронологически находясь впереди создаваемого им математического аппарата, всегда оказывается в данный конкретный момент богаче той дедуктивной порождающей процедуры, которая должна быть заложена в ЭВМ для 100-процентного распознавания и порождения текста (9). Если учесть к тому же, что порождающая грамматика не использует в своих лингвистических построениях правила остранения (13), станет ясным, почему эта теория не смогла до сих пор стать лингвистическим фундаментом для построения работающих систем МП.

Что же предлагает машинному переводу лингвистика текста?

Во-первых, ЛТ рассматривает текст как избыточную многоуровневую систему, на верхних ярусах которой находятся наиболее информативные в среднем ЛЕ, а на нижних — малоинформативные элементы.

Во-вторых, ЛТ ориентируется только на бинарный перевод, при построении которого структура входного и выходного языков объединяются в суперструктуру, которую мы будем называть двуязычной ситуацией (Пиотровский, Чижаковский, 1971, с. 444—451; см. также ниже, с. 276).

В-третьих, все алгоритмы МП предполагается строить на основе идей формального распознавания смыслового образа входного текста (55).

В-четвертых, отказываясь от 100-процентной переработки входного текста, ЛТ предлагает последовательное развертывание системы МП, от простых алгоритмов, работающих на верхних ярусах двуязычной ситуации и обеспечивающих извлечение из входного текста максимума основной информации текста, к более сложным алгоритмам, которые ориентированы на глубинные уровни этой ситуации и с помощью которых из входного текста извлекаются малые дозы более тонких и детальных сведений.

Решить вопрос о том, какие блоки МП являются с информационной точки зрения и с точки зрения теории распознавания наиболее эффективными и поэтому должны разрабатываться в первую очередь, а какие должны строиться на последующих этапах формирования системы МП, позволяют информационно-статистические измерения текста, психолингвистические эксперименты, а также опыт оптимизации преподавания языков.

В частности, информационно-статистические исследования текстов по основным европейским языкам показали, что львиная доля синтактической (структурной) и, очевидно, смысловой информации заложена в лексике и фразеологии текста, в то время

как морфология дает сравнительно скромный процент этой информации (ср. табл. 61).

Таблица 61

Процент синтактической информации, передаваемый различными лингвистическими единицами

Языки	Неактуализованное слово	Словосочетание	Морфологические аффиксы
Английский	65.0	81.1	3.5
Немецкий	70.8	85.9	—
Французский	76.5	86.1	16.0

Примечание. При составлении настоящей таблицы использованы данные табл. 33, 45, 47; процент синтактической информации, заложенной в отдельных ЛЕ, определяется из отношения избыточности данной ЛЕ к общей избыточности текста.

Эксперименты по восстановлению пропущенных букв в связном тексте показывают, что при отсечении конечных букв слова, т. е. тех букв, которые оформляют грамматические морфемы, носители языка восстанавливают текст на 95—99%. При разрушении тех частей слова, которые воплощают его лексическое значение, процент восстановления текста заметно падает (45).

В ходе статистической оптимизации преподавания иностранных языков выяснилось, что адресату, воспринимающему в тексте около 80% информации, обычно удается распознать содержание текста, опираясь на догадку и избыточность письменной речи (Алексеев, Герман-Прозорова, Пиотровский, Щепетова, 1974, с. 220—221).

Опираясь на эти результаты, можно было ожидать, что правильное распознавание и переработка на ЭВМ актуальных значений ЛЕ входного текста даст возможность выдавать такой машинный результат, который не только воспринимался бы потребителем, но и сохранял основную смысловую информацию входа.

Отсюда следует, что фундаментом всякой системы МП должен быть лексический перевод, опирающийся на автоматические словари (АС и АСО).

На следующем этапе следует строить семантический перевод, который опирается на алгоритмы устранения многозначности, позволяющие автоматически распознавать актуальное значение ЛЕ в перерабатываемом тексте.

Составление грамматических алгоритмов целесообразно отнести на заключительный этап построения всей системы МП.

67. Лексический перевод

Технологические вопросы построения и программирования автоматических словарей обсуждались уже в десятках работ (Мельчук, Равич, 1967, с. 203—224; Perschke, 1970; Вертель, Вертель, Крисевич и др., 1971; Гончаренко*, 1972а; Добрускина*, 1973; Садчикова*, 1975, и др.). Поэтому сосредоточимся на рассмотрении только теоретических аспектов лексического МП.

Начнем с автоматического словаря слов и словоформ.

Каждый бинарный словарь, описывающий двуязычную лексическую ситуацию, представляет собой суперструктуру, включающую два множества, первое из которых включает входную, а второе — выходную лексику. При построении и функционировании АС интерес представляют однонаправленные лексико-семантические отношения вход—выход, а отношения между выходными и входными единицами в расчет могут не приниматься. При таком подходе входные и выходные ЛЕ оказываются связанными обычно одно-многозначными отношениями, причем одна входная ЛЕ может иметь несколько десятков выходных эквивалентов (ср.: Борисевич, Гончаренко и др., 1970).

Предположим теперь, что иностранный текст перерабатывается на выходной язык с помощью одного лишь АС. В этом случае схема распознавания выглядит следующим образом.

Имеется СП, либо охватывающее язык в целом, либо отражающее определенную тематику. СП распадается на области (классы), в качестве которых выступают СПо входных лексических единиц, заложенных в АС. ЛЕ выполняют одновременно функции признаков СП и эталонов областей. В качестве метки класса выступает набор переводных эквивалентов.

Организация алгоритма распознавания зависит от того, из каких ЛЕ строится входной или выходной словник. Если на входе АС используются исходные формы слова, а также традиционные или машинные основы (Perschke, 1970; Зореф,* 1972; Марчук, 1973), то распознающая процедура должна включать морфологический алгоритм приведения текстовой словоформы к основе или исходной форме (ср. 61). Если же входной словник АС строится на основе словоформ (Вертель, Вертель, Крисевич и др., 1971), то процесс распознавания сводится к простому сравнению текстовых и словарных ЛЕ с последующим выводом на печать в случае их совпадения меток классов, т. е. набора переводных эквивалентов (ср. рис. 57).[1]

[1] Независимо от структуры АС каждая ЛЕ входного словника и соответственно каждая лексема в МАРС кодируется по системе, описанной в работе: Борисевич, Гончаренко и др., 1970, с. 331—614. Фрагмент кодировочной матрицы, с помощью которой происходит кодирование отдельных лексических единиц во входном и выходном словниках АС, показан в табл. 67.

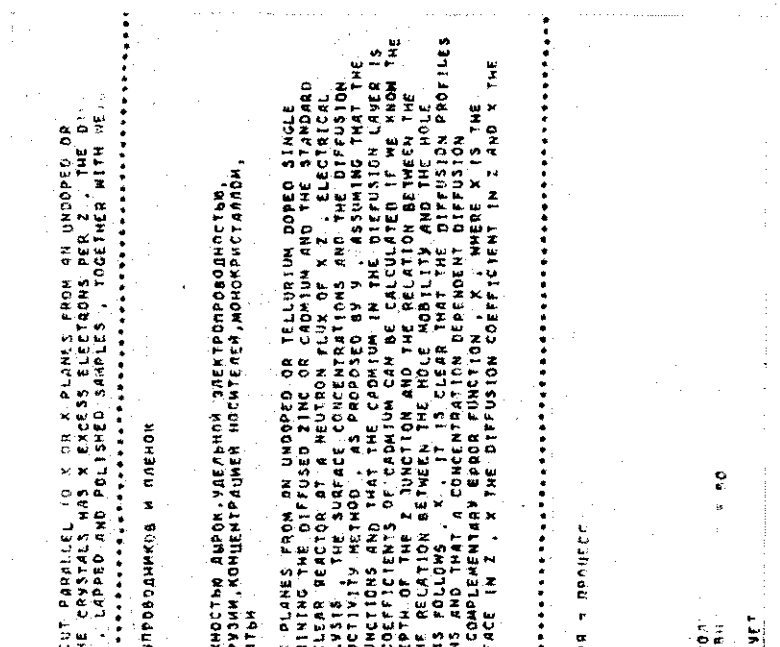


Рис. 57. Работа автоматизированной системы «Информационное обслуживание специалиста: индексирование, аннотирование, реферирование и пословный перевод английского текста по полупроводниковой технике» (результат В. В. Гончаренко, В. С. Крисевича, А. Н. Попеску, Е. С. Тарасовой).

Итак, осуществленный с помощью АС пословный перевод представляет собой последовательность наборов русских синонимов — наборов, описывающих не текстовое (актуальное), но словарное (виртуальное) значение входных словоформ. Переход к актуальным эквивалентам входных ЛЕ и тем самым распознание смысла исходного текста может быть осуществлен путем просмотра всех синтагматических комбинаций из предлагаемых машиной эквивалентов и выбора среди этих комбинаций такой последовательности русских словоформ, которая передавала бы содержание исходного текста.

Опыт использования пословного МП (Masterman, 1967) показывает, что рядовой потребитель, не владеющий или плохо владеющий входным языком, с этой задачей не справляется. Число выходных цепочек, правдоподобных с точки зрения профессиональной осмысленности, настолько велико, что безошибочно выбрать среди них ту последовательность, которая передает содержание входного текста, не удается.

Эту конфликтную ситуацию, возникающую во второй «горячей точке» коммуникационной системы человек—машина—человек (ср. 14), можно преодолеть путем создания системы фильтров, последовательно сокращающих информационный шум в выдаваемом компьютером переводе.

Выделяется два типа таких фильтров — доалгоритмические и алгоритмические.

При построении лексических фильтров первого типа, закладываемых на лингвистическом уровне построения МП, чаще всего применяются следующие приемы.

Во-первых, для перевода входных общеупотребительных ЛЕ из многоцелевого автоматического русского словаря выбирается ограниченное количество достаточно часто употребляющихся эквивалентов, которые покрывают своими значениями СПо входного слова (ср.: Ястребова, Мандрыка, Моисеева и др., 1974; Садчикова,* 1975).

Во-вторых, используется прием статистического отбора отраслевых эквивалентов. Алфавитно-частотные и частотные словники, составленные по русским текстам определенной специальности, позволяют отличать эквиваленты, типичные для данных текстов, от редко использующихся и вообще не употребляющихся в текстах данной тематики слов. Например, «Англо-русский словарь по радиоэлектронике и связи» (М., 1959) указывает в качестве переводов английского существительного mark эквиваленты *метка*, *пометка*, *отметка*, *след*, *знак*. Между тем ЧС русских текстов по радиоэлектронике Е. А. Калининой (1968б) и А. Б. Межлумовой (1973) показывают, что в современных текстах названной тематики используются только три из указанных ЛЕ — *отметка*, *знак*, *след*.

С помощью доалгоритмических фильтров удается несколько сократить число выходных элементов. Однако оно все же остается

слишком высоким для того, чтобы полностью погасить информационный шум, мешающий пониманию пословного МП.[1] Снять информационный шум и превратить пословный МП в связный, легко воспринимаемый адресатом текст можно либо путем постредактирования или интерредактирования этого перевода с помощью дисплея, либо путем последовательного использования все более сильных алгоритмических фильтров, либо путем комбинированного применения фильтров и интерредактирования.

В качестве простейшего алгоритмического фильтра, заметно приглушающего информационный шум в пословном МП, может выступать система машинного индексирования, аннотирования и реферирования, образующая вместе с автоматическим словарем комплекс информационного обслуживания специалиста (рис. 57). Выданный компьютером индекс и аннотация статьи определяет область рассуждения и тем самым заметно сокращает число возможных вариантов перевода, которые можно получить с помощью пословного МП.

Следующим по эффективности фильтром является автоматический словарь оборотов (АСО), с помощью которого должно быть осуществлено распознавание и перевод идиоматических выражений входного текста.

Прежде чем говорить об организации АСО, остановимся на самом термине «идиоматичность», который в условиях двуязычной ситуации расшифровывается иначе, чем в традиционной лингвистике, ориентирующейся на одноязычную ситуацию. Идиоматичными в двуязычной ситуации следует считать лишь такие единицы или их сочетания во входном языке, для которых в результате обращения к двуязычному словарю и к грамматическим правилам анализа и синтеза невозможно найти правильный перевод на выходной язык. С этой точки зрения эквивалентные устойчивые словосочетания Adam's apple, Adamsapfel, *адамово яблоко* не являются взаимоидиоматичными. Напротив, сложный английский термин Morse recorder (receiver) или немецкий Morseschreiber могут рассматриваться как идиомы по отношению к русскому *аппарат Морзе* при условии, что для словоформ recorder (receiver), Schreiber в англо-русском и немецко-русском словарях не задан эквивалент *аппарат*. Одновременно французский appareil de Morse можно считать неидиоматичным по отношению к русскому *аппарат Морзе*.

В условиях двуязычной ситуации идиоматичность можно рассматривать как форму неоднозначности. Действительно, если каждое из значений полисемантического слова или морфемы

[1] В результате применения доалгоритмических фильтров в ходе построения англо-русского АС оказалось, что на три тысячи словоформ в английской части СОЛ приходится около 180 тыс. русских словоформ, а на 9.5 тыс. английских словоформ в ОтС по стройматериалам падает 120 тыс. выходных единиц.

выявляется исходя из его текстового контакта с целым классом лексических или морфологических единиц, то идиоматическое значение или употребление лингвистического элемента обнаруживается при его соединении только с одной или несколькими строго фиксированными ЛЕ. Например, англ. recorder *самопишущий прибор, фототелеграфный приемник* и т. д., receiver *получатель, радиоприемник, телефонная трубка, резервуар* и т. д., нем. Schreiber *писарь, переписчик, самописец* становятся эквивалентами русск. *аппарат* только в соседстве с именем собственным Morse.

При бинарном МП почти в каждом сегменте входного текста можно найти неоднозначность и идиоматичность составляющих его словоформ по отношению к словоформам выходного сегмента.[1] Эта неоднозначность и идиоматичность устраняется либо с помощью стандартных грамматических и семантических алгоритмов анализа, работающих во взаимодействии с АС, либо путем прямого соотнесения входного и выходного сегментов. В последнем случае и тот, и другой рассматриваются как неделимые обороты. Список, в котором каждому входному обороту поставлен в соответствие один или несколько входных оборотов или словоформ, называется словарем оборотов (его входную часть будем называть списком оборотов).

Вопрос о том, какой сегмент считать оборотом, а какой нет, решается прагматически. Если данный сегмент текста принадлежит к достаточно обширному классу однотипных сегментов, в которых многозначность и идиоматичность снимаются в ходе машинного перевода с помощью АС и стандартных алгоритмов, то данный сегмент следует считать не оборотом, но свободным словосочетанием. Если же с помощью имеющихся стандартных алгоритмов и АС данный сегмент перевести нельзя, то целесообразней будет не составлять специальную программу его анализа, а поместить указанный сегмент в АСО. Само собой понятно, что в словарь оборотов помещаются только достаточно часто встречающиеся сегменты: включить в этот словарь все комбинации словоформ входного текста, разумеется, невозможно.

Организация списка машинных оборотов во многом определяется построением словников словоформ. В этом смысле можно указать на три основных ограничения, которые приходится учитывать при построении списка оборотов в системе МП группы «Статистика речи». Во-первых, все иностранные словоформы, из которых составлены обороты, должны входить во входной словник словоформ. Если оборот входит в общую часть словника, то все составляющие его словоформы должны быть представлены в СОЛ. Если оборот является отраслевым, составляющие его словоформы могут находиться как в СОЛ, так и в соответствующем

[1] Случаи полной изоморфности (калькируемости) входного и выходного сегментов типа англ., фр. fig. 1=рис. 1 встречаются в виде исключения.

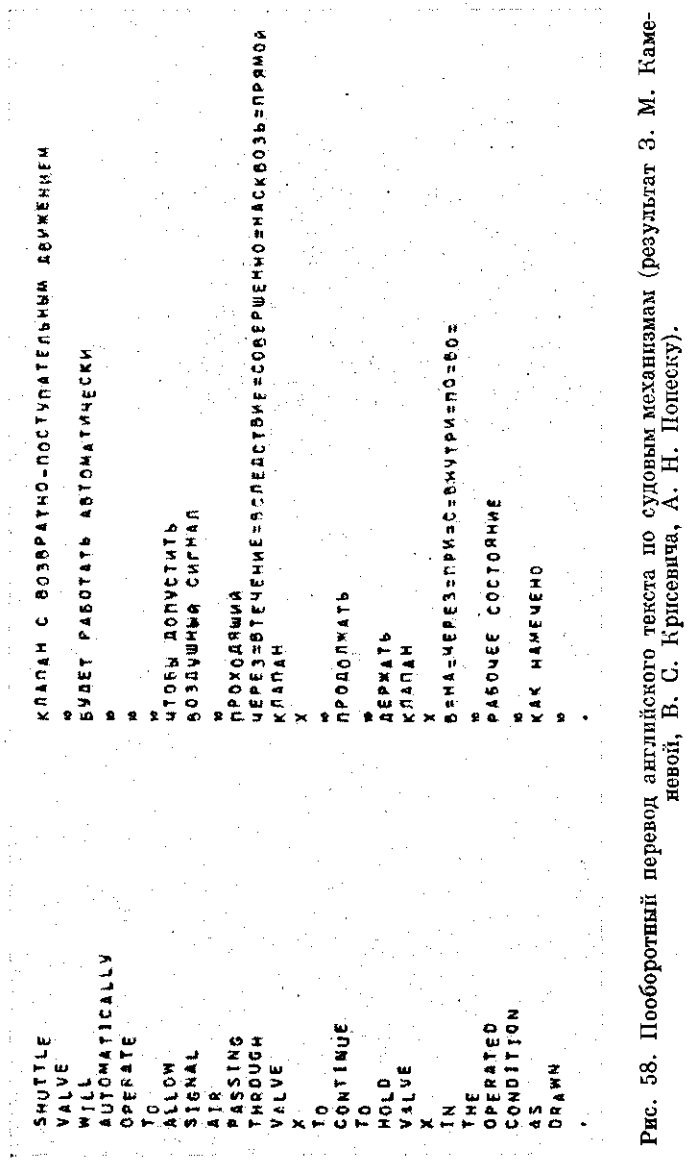


Рис. 58. Пооборотный перевод английского текста по судовым механизмам (результат З. М. Каменевой, В. С. Крисевича, А. Н. Попеску).

ОтС. Во-вторых, русские словоформы, составляющие эквиваленты выражений, должны быть зафиксированы с соответствующими кодами в МАРС (65). В-третьих, каждое выражение может состоять не более чем из шести словоформ.

Технология построения и функционирования АСО, а также связанные с ним теоретические проблемы распознавания были подробно рассмотрены выше (см.: 22—23, 60, 63—64). Второй раз освещать эти вопросы нет необходимости. Поэтому перейдем к оценке эффективности АСО.

Опытная эксплуатация англо-, немецко- и французско-русских АСО, созданных в группе «Статистика речи» (Ионицэ*, 1971; Чижаковский*, 1971; Борисевич*, 1972; Каменева*, 1973; Садчикова*, 1975), а также результаты, полученные в американском коллективе Logos Development Corporation (Арсентьева, Кулагина, 1974, с. 200), показали, что использование словаря оборотов дает возможность ЭВМ распознавать не только виртуальные (словарные), но и актуальные (текстовые) значения у большого числа ЛЕ. Эта актуализация приводит к заметному ослаблению информационного шума, в результате чего выдаваемая с помощью АС и АСО машинная переработка текста дает возможность потребителю извлекать из пословно-пооборотного перевода — хотя и с некоторым напряжением — ту информацию, которая была заложена во входном тексте (ср. рис. 58). Поэтому пословно-пооборотный МП можно рассматривать как первое приближение к промышленной переработке иностранного текста на ЭВМ.

68. Семантический перевод

Хотя работающий лексический перевод и является первоначальным необходимым условием для организации промышленного МП, качественная переработка иностранного текста на ЭВМ не может быть достигнута без широкой актуализации, т. е. без устранения многозначности у подавляющего числа ЛЕ.

Как уже говорилось, словарная многозначность может устраняться иконическим путем, т. е. с помощью АСО. Однако сплошное иконическое снятие многозначности может быть осуществлено с помощью только такого АСО, который включал бы сотни тысяч входных оборотов. Реализация этого подхода потребовала бы десятки лет подготовительной работы и привела бы к построению громоздких и неэффективных алгоритмов.

Более экономным является использование семантических фильтров, применение которых позволяет разрешать типовые схемы многозначности. Такой подход позволяет, с одной стороны, актуализовать лексические единицы, не охваченные словарем оборотов, а с другой — сокращает объем самого АСО.

В настоящее время в действующих системах МП используются два типа семантических фильтров. Фильтры первого типа используют в качестве снимающих многозначность признаков словоформы и морфемы, которые задаются в виде специальных списков. Хотя применение этой методики дает хорошие результаты с точки зрения устранения многозначности (ср. Зубов*, 1969; Марчук, 1973; Трибис, 1973), оно оказывается громоздким с точки зрения программирования. Дело в том, что для небольших групп ЛЕ, а иногда даже для одиночных слов приходится составлять специальные актуализирующие алгоритмы, число которых в работающих системах МП достигает таким образом нескольких тысяч. Чтобы избежать перегрузки памяти ЭВМ, приходится строить более экономные фильтры второго типа, опирающиеся на парадигматическое описание элементарных смыслов и их синтагматику в тексте (ср.: Шабес*, 1973).

Парадигматическое описание семантики осуществляется следующим образом. Выбирается СП, определяемое набором дифференциальных семантических признаков (элементарных смыслов), которые могут быть упорядочены в виде дихотомического графа (ср. рис. 59) или каким-то другим образом (ср. Nilsson, 1971/1973). В работах группы «Статистика речи» в качестве такой классификационной иерархии чаще всего применяется комбинация из двух дихотомических древовидных графов, входом в которые служит грамматический код части речи, применяемый в стандартной матрице кодирования лексико-грамматической информации (см.: Борисевич, Гончаренко и др., 1970, с. 341, 355 и сл.).

Первый граф охватывает признаки, являющиеся общими для большинства рассматриваемых СП (табл. 62). Второй граф описывает соотношение признаков, характеризующих конкретное СП (табл. 63, 64).

Первый граф является постоянным для данного грамматического класса. Графы второго типа являются сменными: подключение того или иного графа определяется тематикой перерабатываемых на ЭВМ текстов. На рис. 59 показан общий граф, использующийся для кодирования имен существительных, а также части двух сменных графов, один из которых описывает СП «общественно-политическая тематика», а другой моделирует СП «радиоэлектроника».

Только что сформированный классификационный граф позволяет описывать с большей или меньшей степенью детализации семантику именных ЛЕ. Рассмотрим в качестве примера семантическое кодирование французских терминов alimentation ‘питание’, ‘источник питания’ и commande ‘управление’ и ‘механизм управления’ (см.: Угодчикова*, 1975, с. 7—8).

Первые значения этих терминов, которые могут быть отнесены к семантическому классу «процесс», описываются двоичным кодом 10-01 (десятичный код 2-01), соответствующим цепочке: имя существительное—неконкретное—процесс. Вторые значения, относящиеся к классу «устройство», обозначаются кодом 10-111111 (десятичный код 2-63), соответствующим последовательности: имя

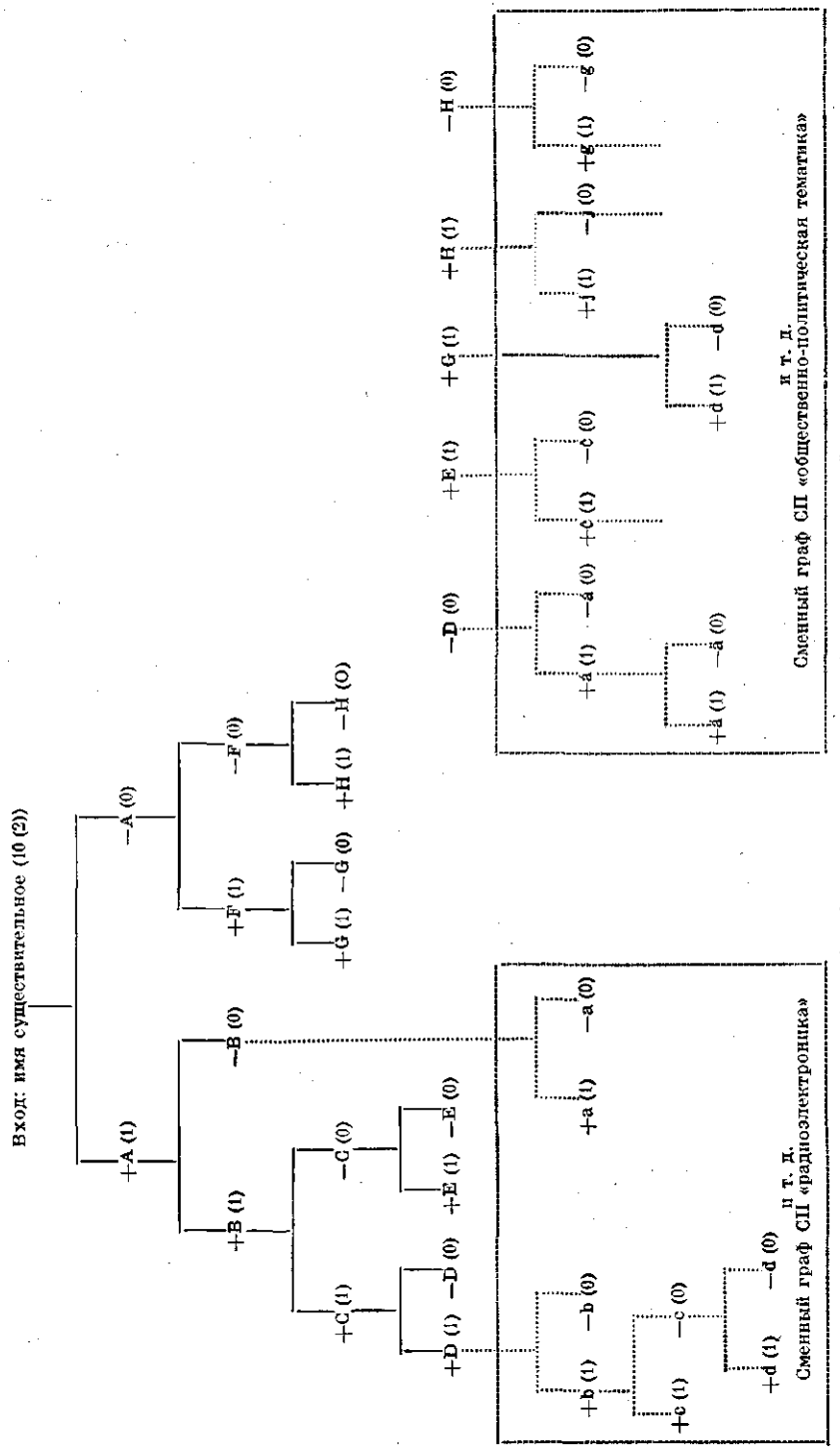


Рис. 59. Семантические классификационные графы.

Таблица 62

Общие дифференциальные семантические признаки

№ пп.	Название дифференциальных семантических признаков	Коды		
		буквенный	цифровой двоичный	цифровой десятичный
1	Конкретное/неконкретное	+A/—A	1/0	1/0
2	Дискретное/недискретное	+B/—B	11/10	3/2
3	Тело/нетело	+C/—C	111/110	7/6
4	Артефакт/неартефакт	+D/—D	1111/1110	15/14
5	Место/неместо	+E/—E	1101/1100	13/12
6	Процесс/непроцесс	+F/—F	01/00	01/00
7	Спец. процесс/неспец. процесс	+G/—G	011/010	03/02
8	Признак/непризнак	+H/—H	001/000	001/000

Таблица 63

Отраслевые ДСП для СП «радиоэлектроника»

№ пп.	Дифференциальные семантические признаки	Коды		
		буквенный	цифровой двоичный	цифровой десятичный
1	Материал/нематериал	+a/—a	10—101/10—100	2—5/2—4
2	Целое/нецелое	+b/—b	10—11111/10—11110	2—31/2—30
3	Устройство/неустройство	+c/—c	10—111111/10—111110	2—63/2—62
4	Сооружение/несооружение и т. д.	+d/—d	10—1111101/10—1111100	2—125/2—124

существительное—конкретное—дискретное—тело—артефакт—целое—устройство. Аналогичным образом слова oxygene 'кислород', fer 'железо', plomb 'свинец', получат семантический код 10-101 (десятичный код 2-5) и будут отнесены к классу «материал».

Всего в СП «радиоэлектроника» Н. Ф. Угодчикова* (1975, с. 59) выделила 55 типов семантических классов. На следующем шаге построения нашей распознающей системы происходит объединение ЛЕ, обнаруживающих однотипную многозначность в типовые множества смыслов (ТМС). Так, например, термины alimentation и commande объединяются в ТМС процесс/устройство, которому в классификации Н. Ф. Угодчиковой присвоен номер 1. Всего Н. Ф. Угодчикова обнаружила во французских текстах по радиоэлектронике 21 ТМС. Термины, не вошедшие в эти ТМС и дающие нерегулярную многозначность, должны обрабатываться с помощью АСО.

Описанная процедура позволяет устранять в ходе автоматической переработки текста регулярную многозначность отдельных

Таблица 64

Отраслевые дифференциальные признаки для СП «общественно-политическая тематика»

№ пп.	Название дифференциальных семантических признаков	Буквенные коды
1	Одушевленное/неодушевленное	+a′/—a′
2	Отдельное лицо/неотдельное лицо	+ä/—ä
3	Группа лиц/негруппа лиц	+b′/—b′
4	Собственное/несобственное	+c′/—c′
5	Политический процесс/неполитический процесс	+d′/—d′
6	Событие/несобытие	+e/—e
7	Экономический процесс/неэкономический процесс	+f/—f
8	Торгово-промышленная операция/неторгово-промышленная операция	+g/—g
9	Финансовая операция/нефинансовая операция	+h/—h
10	Способность/неспособность	+k/—k
11	Возможность/невозможность	+l/—l
12	Параметр/непараметр	+m/—m
13	Институт/неинститут	+t/—t
14	Внутригосударственный институт/невнутригосударственный институт	+v/—v
15	Предприятие (объединение предприятий)/учреждение	+y/—y
16	Финансово-экономическая категория/нефинансово-экономическая категория	+β/—β
17	Денежный документ/неденежный документ и т. д.	+γ/—γ

лексических единиц, не затрагивая их синтаксических функций (см. рис. 60). Однако при построении действующих алгоритмов МП редко удается отделить лексико-семантические операции от синтаксических задач. Поэтому следующим шагом в разработке семантического МП является создание процедур устранения семантико-синтаксической неоднозначности текста. Рассмотрим эту задачу на примере алгоритма переработки английских двухсловных сочетаний типа stone wall, составленного и реализованного на ЕС-ЭВМ В. Н. Билан* (1975) и И. Г. Низамутдиновой.

Ход процедуры обучения ЭВМ, а также сам процесс синтаксико-семантического распознавания проследим на примере кодирования английских существительных exchange и labour, а также перевода двухсловного сочетания labour exchange.

На первом этапе обучения все выходные существительные распределяются по семантическим классам и кодируются с помощью классификационного дерева, показанного на рис. 59. При этом русские словарные эквиваленты указанных выше английских существительных получают следующие коды:

exchange *обмен* (2-027)
биржа (2-00093)
биржевой (3-00093)
мена (2-00017)
меновой (3-00017)

labour *труд* (2-02)
лейборист (2-59)
трудовой (3-02)
лейбористский (3-59)

Затем выявляются типы интеркомпонентных смысловых отношений (ТИСО), в которые вступают входные существительные схемы stone wall. Выделяются следующие ТИСО с их условными обозначениями, показанными в скобках:

1) отношения принадлежности (P);
2) уточняюще-определительные (D);
3) определительно-субъективные (S);
4) определительно-объектные (O);
5) определительно-временные (T);
6) определительно-пространственные (L).

Каждая входная (английская) ЛЕ снабжается валентностными характеристиками, которые указывают, в каких ТИСО может участвовать данная словоформа. Эта информация закладывается в словарную статью каждой словоформы, находящейся в АС. Образцы записи словоформ exchange и labour показаны в табл. 65.

Таблица 65

Валентностные характеристики двух английских существительных
(1 — присутствие данного ТИСО, 0 — отсутствие ТИСО)

	I позиция	II позиция
Лексические единицы	P D S O T L	P D S O T L
exchange	0 1 0 0 0 0	0 1 0 0 0 0
labour	0 1 0 1 0 0	0 1 0 0 0 0

Оба существительные имеют единицы только в столбцах D, в остальных столбцах для них показаны либо нули, либо сочетание нуля и единицы (столбец O). Это значит, что двусловное сочетание составленное из ЛЕ exchange и labour может иметь только уточняюще-определительное значение.

Каждому ТИСО соответствует множество распознающих эталонных наборов (РЭН), представляющих собой разрешенные бинарные комбинации семантических кодов выходных (русских) ЛЕ. В нашем случае, поскольку английские словосочетания типа stone wall переводятся на русский язык либо комбинацией двух существительных — *стена* (*из*) *камня*, либо словосочетанием при-

Рис. 60. Автоматическое снятие многозначности у французских словоформ (*а*) и английских именных словосочетаний (*б*, *в*).

а — результат Г. Е. Кузнецовой и Н. Ф. Угодчиковой; *б* — результат В. Н. Билан и И. Г. Назамутдиновой; *в* — результат Л. А. Славиной.

лагательное+существительное — *каменная стена*, то для каждого ТИСО задается две группы РЭН. Одна включает комбинации кодов, имеющих вход 2(10), — «существительное», другая содержит пары кодов, первый из которых имеет вход 2 (10) (существительное), а второй — 3 (11) (прилагательное) — см. табл. 66. Первая группа РЭН выдает переводы типа *стена (из) камня* с инверсией эквивалентов, вторая порождает переводы *каменная стена*.

Таблица 66

Фрагмент списка РЭН

ТИСО	I схема перевода (существительное + + существительное)	II схема перехода (прилагательное + + существительное)
D	2-02/2-00093 2-00001/2-015 2-119/2-00047 и т. д.	3-00017/2-053 3-02/2-03 3-015/2-00045 и т. д.
O	2-00071/2-02 2-00019/2-013 2-15/2-02 и т. д.	3-00017/2-15 3-00017/2-02 3-117/2-07 и т. д.

В качестве семантического пространства в нашем случае выступает вся совокупность значений комбинаций типа stone wall в данном подъязыке. Областями этого СП являются значения конкретного английского двусловного сочетания, меткой области — его русский перевод. В качестве признаков СП и областей выступают семантические коды и коды ТИСО. Эталонами являются ТИСО и РЭН. Смысловой образ считается распознанным тогда, когда ЭВМ выдает перевод входного словосочетания.

Теперь на примере перевода английского словосочетания labour exchange проследим всю процедуру распознавания.

Указанное словосочетание, выделенное из текста в результате сегментации или поиска по опорным словам, в АСО не обнаружено. Поэтому его машинная обработка идет по семантическому алгоритму в следующем порядке.

Сначала с помощью табл. 65 определяется тот ТИСО, к которому должно принадлежать словосочетание labour exchange. Им оказывается D-ТИСО (уточняюще-определительные отношения), в этой зоне и будет осуществляться поиск нужного РЭН (см. табл. 66). Затем из МАРС выбираются русские эквиваленты с их семантическими и грамматическими кодами. Эти эквиваленты образуют $4\times5=20$ двухсловных комбинаций.

Чтобы определить, какая из этих пар может быть использована при построении русского перевода англ. labour exchange, комбинации семантических кодов русских ЛЕ сопоставляются

с РЭН, находящимися в зоне D табл. 66. Совпадение с РЭН дает только одна комбинация 2-02/2-00093, соответствующая паре *труд—биржа*. Эта пара и выбирается для формирования переводного эквивалента.

Совпадение кодовой комбинации и РЭН обнаружено в зоне действия переводной схемы существительное+существительное, в которой порядок выходных словоформ является обратным по отношению к расположению входных ЛЕ *стена* (*из*) *камня* < stone wall. Поэтому производится инверсия русских словоформ с одновременным выбором из парадигмы MAPC родительного падежа второго существительного:

labour exchange *труд—биржа* > *биржа труда*

На этом весь процесс семантико-синтаксического распознавания двухкомпонентного английского словосочетания заканчивается. Аналогичная процедура распознавания для *n*-членных именных словосочетаний разработана также Л. А. Славиной (1974).

Оба алгоритма запрограммированы на языке КОБОЛ и реализованы на ЭВМ ЕС-1020 (см. рис. 60, *б*, *в*).

Дальнейшее развитие семантического МП идет по пути формализации новых СП, подключения описывающих их графов к основному дереву смыслов и создания на этой основе новых распознающих алгоритмов.

69. Грамматический перевод

Описание грамматики одного и того же языка, как, впрочем, и других его аспектов может быть осуществлено на основе различных моделей. Что касается морфологических алгоритмов, то они строятся либо на базе традиционных грамматик, либо с применением особых «машинных» морфологических моделей. Характерный для ранних этапов развития МП традиционный подход (Мельчук, Равич, 1967, с. 247—263) в настоящее время используется редко. Дело в том, что множество морфологических форм языка, в особенности языка, пользующегося флективной техникой, представляет собой переплетение регулярных моделей, порожденных современной системой, с архаическими парадигмами и нормализовавшимися диалектизмами. В результате традиционное описание морфологии образует довольно непоследовательную и поэтому трудно формализуемую систему правил.

Чтобы преодолеть эти трудности, приходится обращаться к построению так называемой машинной морфологической модели. Выбор той или иной морфологической модели определяется здесь строем языка, возможностями ЭВМ, а также архитектурой данной системы МП. В действующих системах группы «Статистика речи» используется два подхода.

Идея первого подхода заключается в том, что фузионная техника индоевропейских языков заменяется при их машинном моделировании квазиагглютинативными схемами. Согласно этим схемам, словоформы, входящие в парадигму определенной ЛЕ, расщепляются на неизменяемые основы и приклеивающиеся к ним аффиксы. Чтобы обойти случаи чередования основы и ассимиляционные процессы, возникающие на границах основы и флексий, приходится выделять иногда у одной ЛЕ несколько основ и создавать нетрадиционные парадигмы аффиксов. Так, например, немецкое существительное Tag имеет в машинной грамматике М. Г. Зорефа* (1972) две основы — Tag (ед. ч.) и Tage (мн. ч.), причем первая в род. пад. получает аффикс s, а вторая снабжается в дат. пад. окончанием n, в остальных падежах Tag и Tage окончаний не имеют (см.: СтРААТ, 1973, с. 255). Аналогичные примеры перестройки русской морфологии см. в 65 и работах: Пиотровская, 1973, 1974а, 1974б. «Агглютинизация» фузионной морфологии вполне оправдывает себя при построении больших АС для языков, имеющих богатую флексию. Использование этого приема заметно упрощает морфологический анализ, а вместе с тем не слишком увеличивает объем АС. Это особенно важно в тех случаях, когда на построение системы МП накладываются ограничения, связанные с объемом машинной памяти. Кроме того, агглютинативное моделирование лежит в основе независимого морфологического анализа и синтеза, использующегося в алгоритмах приведения текстового словоупотребления к канонической форме (61), а также в алгоритмах пополнения АС.

Идея второго подхода состоит в оформлении индоевропейской лексики с помощью техники аморфных языков (вьетнамского, китайского). При этом каждая словоформа рассматривается в качестве самостоятельной морфологически неразложимой ЛЕ (ср.: Вертель, Вертель, Крисевич, Пиотровский, Трибис, 1971). Хотя этот подход приводит к предельному упрощению морфологического алгоритма, он вызывает одновременно увеличение объема АС. Исходя из этих соображений аморфную технику целесообразно применять в системах МП, ориентированных на ЭВМ с практически неограниченным объемом внешней памяти и малым временем обращения.

Если же речь идет о средних и малых машинах, то использование аморфной техники оправдывает себя в двух случаях.

Во-первых, ее можно применить к таким разновидностям флективных языков, которые используют более или менее фиксированный и ограниченный по объему набор шаблонных словоформ и оборотов. Примером может служить МП переговоров в коммуникативных системах «земля—воздух», «земля—вода» или лингвистические алгоритмы программированного обучения иностранным языкам.

Во-вторых, с помощью аморфной техники следует моделировать словарь и грамматику таких языков, которые, располагая слабыми формообразующими возможностями и высоким коэффициентом аналитизма (49, табл. 46), дают при переходе от лексем

к словоформам не более чем двойное увеличение АС. Исходя из этих соображений в системах англо-, французско-, испанско-, итальянско- и румынско-русского МП, разрабатываемых в группе «Статистика речи», используются входные словники словоформ и не применяются морфологические трансформации и отсечения (Ястребова, Мандрыка, Моисеева и др., 1974).

В больших автоматизированных системах переработки текстовой информации обычно приходится применять комбинированное аморфно-агглютинативное моделирование морфологии и лексики. Этот подход использован в частности при разработке МАРС (65), а также при построении систем индексирования и аннотирования английских и французских текстов (56—60).

Независимо от того, какая морфологическая модель использована при построении АС и грамматического алгоритма, каждая ЛЕ машинного словаря снабжается морфолого-синтаксическими кодами. Система кодирования, разработанная в группе «Статистика речи» (см.: Борисевич, Гончаренко и др., 1970), использует специальную кодировочную матрицу (табл. 67), с помощью которой для каждой ЛЕ вырабатываются сжатые коды. Сжатые коды омонимичных форм укрупняются во вторичные сжатые коды. Одновременно используются сжатые коды парадигмы. Между кодами входного и выходного языка устанавливаются соответствия, задаваемые специальными таблицами (Гончаренко, Гречишкина, Добрускина, 1972, с. 43—64).

Если говорить о грамматическом МП в целом, то его также можно рассматривать в русле теории распознавания. В этом случае в качестве СП выступает все множество синтаксических структур выходного языка, а областями (классами) этого СП являются сами синтаксические структуры. Формальное представление каждой выходной структуры, будь то дерево зависимости или запись линейной последовательности членов предложения, можно считать эталоном и одновременно меткой.[1]

Весь алгоритм распознавания «грамматического образа» входного предложения складывается из нескольких этапов.

На первом этапе путем использования граничных сигналов, в качестве которых могут выступать либо словоформы и их части (23), либо коды этих словоформ, производится сегментация текста. Затем с помощью кодов словоформ, образующих сегмент, распознается сначала его морфологическая природа, а затем и синтаксическая роль в предложении (Коверин, Скитневский, 1973).

На втором этапе путем комбинирования синтаксических кодов, выработанных на первом этапе распознавания, а также путем использования различных формальных признаков текста, напри-

[1] Иногда в качестве метки используется словесное описание входной грамматической структуры на выходном (русском) языке (23; ср. также: Афаунова, 1973, стр. 327—335). Эта форма распознавания может быть использована в обучающих, репетирующих и экзаменующих алгоритмах.

Таблица 67

Фрагмент матрицы кодов для словоформ англо-русского АС

№ пп.	Лексический код (номер первоначального экземпляра)	Словоформы	I Грамматический класс (часть речи)	II Управление	III Число	IV Род	V Падеж	VI А Степень сравнения	VI Б Полнота (краткость)	VIII А Время	XI Лицо	X Залог	IX Вид	VII Б Видо-временные отношения	VII В Наклонение	XIX Репрезентация	XX Переходность/непереходность	XIX Дополнительная информация	СК	ВСК	СКП
1		American (прил.)	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7001	10007	
2		American (сущ.)	2	0	2	0	8	0	0	0	0	0	0	0	0	0	0	2	7501		
3	2130	американский (имен.)	3	0	2	2	2	2	2	0	0	0	0	0	0	0	0	4	2801	9100	12035
4		американского (род.)	3	0	2	2	3	2	2	0	0	0	0	0	0	0	0	4	2802	9101	
5		американскому (дат.)	3	0	2	2	4	2	2	0	0	0	0	0	0	0	0	4	2803	9102	
6		американский (вин.)	3	0	2	2	5	2	2	0	0	0	0	0	0	0	0	3	2804	9100	
7		американского (вин.)	3	0	2	2	5	2	2	0	0	0	0	0	0	0	0	2	2805	9101	
		и т. д.																			
8	2500	американец (имен.)	2	0	2	1	1	2	2	0	0	0	0	0	0	0	0	2	3013	9001	12011
9		американца (вин.)	2	0	2	1	5	2	2	0	0	0	0	0	0	0	0	2	3016	9003	
		и т. д.																			

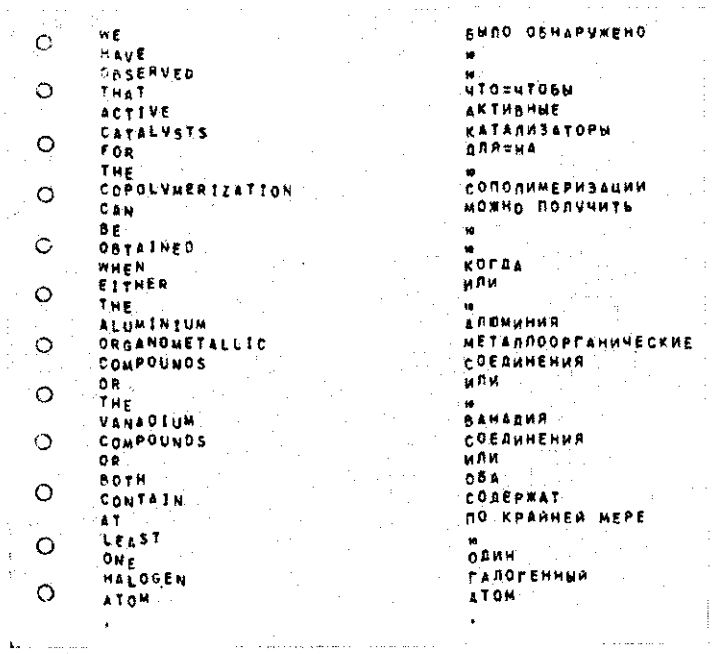


Рис. 61. Лексико-грамматический МП с частичным устранением многозначности английского текста по химии полимеров (Садчикова, 1975).

мер знаков препинания (Мануков, Зубов, 1972), автоматически определяется структурная формула входного предложения. В этой формуле должны быть, в первую очередь, отражены предикативная связь предложения, а затем указаны отношения между главными и второстепенными членами.

Структурная формула последовательно сопоставляется с входными эталонными структурами, заданными в машинном двуязычном синтаксическом словаре (табл. 68). В случае совпадения текстовой и эталонной структуры выдается ее выходной эквивалент, выступающий в качестве метки распознанного грамматического образа предложения. Полученная этим путем выходная синтаксическая структура получает лексическое наполнение из АС (67) и семантических алгоритмов (68). Идя по этому пути, группа «Статистика речи» получила первые экспериментальные образцы англо-русского лексико-грамматического перевода научно-технических текстов (рис. 61). Аналогичные результаты имеются в группе Ю. Н. Марчука (см.: СтРААТ, 1973, с. 226—229).

Само собой разумеется, что описанный только что подход дает на выходе еще достаточно большую синтаксическую омонимию. Устранить эту омонимию можно только с помощью сильных семантико-синтаксических алгоритмов. Лингвистические иссле-

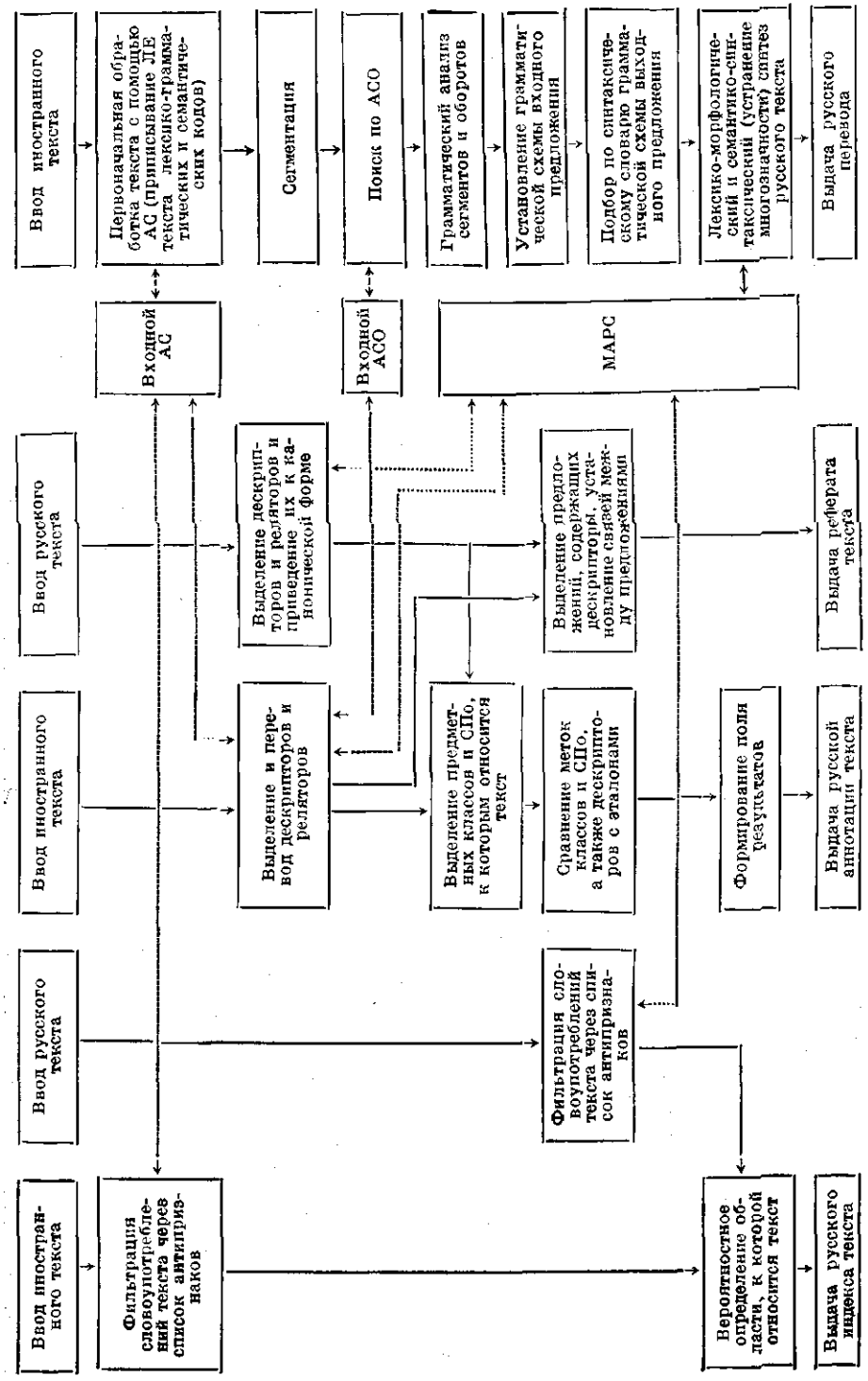


Рис. 62. Построение лингвистического обеспечения АСНТИ, разрабатываемого в группе «Статистика речи» (ср. рис. 43, 47—49, 56).

Таблица 68

Фрагмент англо-русского двуязычного словаря синтаксических структур сказуемого (по данным И. Н. Сафоновой)

№ п.п.	Английская структура	*f*	*Русские эквиваленты*	*f*
1	S (*is*) S (Pr-n)	0.085	а) S (Vcs+*ся*)	0.079
			б) S (Pr кр. ед. ч.)	0.006
2	S (are) S (Pr-n)	0.060	а) S (Vcp+*ся*)	0.047
			б) S (Pr кр. мн. ч.)	0.013
3	S (*is*) S (U)	0.056	а) S (U кр.)	0.016
			б) S (является) S (U тв.)	0.011
			в) S (U полн.)	0.029

Примечание. В таблице используются следующие сокращения: *S* — сказуемое; кр. — краткая форма; полн. — полная форма. Остальные сокращения см. в табл. 14.

дования, ориентированные на создание этих алгоритмов, ведутся уже давно (Bobrov, 1963; Vauquois, Veillon, Veyrunes, 1966; Weizenbaum, 1968/1970; Schank, 1972а, 1972б; Walker, 1973; Мельчук, 1974, и др.). Ближайшей задачей инженерной лингвистики является доведение этих разработок до уровня машинной реализации.

Объединение АС, АСО, семантических, морфолого-синтаксических, семантико-синтаксических алгоритмов вместе с алгоритмами индексирования, аннотирования и реферирования документа даст в перспективе единую систему лингвистического обеспечения АСНТИ, один из вариантов которой показан на рис. 62.

ЗАКЛЮЧЕНИЕ

Содержащиеся в работе результаты могут быть оценены и интерпретированы с двух точек зрения.

Во-первых, их можно оценивать с языковедческой точки зрения, точнее — с точки зрения перспектив развития лингвистики текста, стремящейся к преодолению противоречия между индивидуальной (интроспективной) оценкой языка лингвистами и коллективной (социальной) его природой путем введения правил остранения.

В этом случае центральное место в работе займет вторая часть (главы 4—8) и начало третьей части книги. В главах 4—7 сделана попытка вскрыть с помощью статистико-дистрибутивных и информационных приемов текстообразовательную иерархию лексико-фразеологического, морфологического и синтаксического уровней. Эти приемы дают также возможность раскрыть вероятностную природу нормы языка. Наличие приемов, позволяющих получать из текста вероятностные характеристики и количественные измерения смысла отдельных ЛЕ, еще не решает всей проблемы создания общей процедуры для адекватного описания системы языка, в том числе для экспликации плана содержания. Поэтому вероятностно-информационная методика должна сочетаться с приемами неколичественного описания, в частности с методами, предлагаемыми теорией нечетких множеств и алгоритмов (глава 8) и теорией распознавания смыслового образа текста (глава 9). Пока созданы лишь общие контуры этих последних теорий.

Таким образом, ЛТ представляет собой композицию различных лингвистических и математико-лингвистических идей. Уточнение, развитие и консолидация этой композиции будет стимулироваться, очевидно, прикладными задачами.

Если оценивать книгу только с точки зрения лингвистики текста, то третью часть следует рассматривать как вспомогательный раздел, в котором осуществляется формальная проверка объяснительной силы тех гипотез и моделей, которые были заданы во второй части работы.

Во-вторых, книгу можно рассматривать и в инженерно-лингвистическом ключе. В этом случае первая часть и большинство параграфов второй части выступают в качестве вспомогательных разделов, задающих кибернетические и языковедческие предпосылки

для построения здания инженерной лингвистики. Напротив, в третьей части и отдельных разделах первой и второй частей (5, 7, 9, 12—16, 19, 23) излагаются основные идеи и технология инженерного языкознания. Здесь автору хотелось показать, что пафос инженерной лингвистики заключается не в постулировании и программистском воплощении очередной «единственно правильной» лингвистической концепции, но в создании разумной композиции разных идей — такой теоретической композиции, которая обеспечивала бы поэтапное и все более эффективное моделирование на ЭВМ процессов текстообразования и приводила бы к воспроизведению таких сложных лингвистических объектов, как индекс, аннотация, реферат и, наконец, перевод текста.

Каждая научно-техническая задача, будь то создание самодвижущегося экипажа — автомобиля, или построение летательного аппарата тяжелее воздуха — самолета, или, наконец, создание современного кинематографа не могла сразу найти своего оптимального решения. Более того, прежде чем приступить к комплексному решению каждой из этих задач, человечество должно было создать и отработать основные блоки этих систем (колесо, двигатель внутреннего сгорания, искусственный источник света, фото- и проекционный аппарат и т. п.).

Аналогичная ситуация складывается при построении систем лингвистического обеспечения АСНТИ. Отправными точками этого построения является, с одной стороны, автоматический словарь, дающий пословный перевод, а с другой—машинная атрибуция, выдающая элементарный смысловой индекс текста (ср. рис. 57). Можно ожидать, что оптимальное решение задачи автоматической переработки текста, предусматривающее выдачу не только темы, но и ремы (новизны) текста, будет найдено на скрещении встречных путей формирования и компановки, с одной стороны, блоков МП, а с другой — автоматического индексирования, аннотирования и реферирования.

И тот и другой путь развития лингвистического обеспечения предусматривает составление, освоение на ЭВМ и поэтапное объединение все более сложных блоков.

При этом блоки МП строятся в виде последовательности фильтров, задерживающих проникновение в текст перевода информационного шума. Так, сформированный на основе ЧС (19) двуязычный АС дает в пословном переводе большое количество лишней информации. Первым фильтром, ограничивающим это разнообразие, является объединяемый с АС словарь оборотов (67), затем строятся семантические фильтры (68), подключаемые к АС и АСО и т. д.

Напротив, система «индекс—аннотация—реферат» (56—60, 63—64) строится по принципу все более развернутого распознавания смысла в каждом последующем блоке.

Следует иметь в виду, что эти блоки не обязательно повторяют прототипы естественного языка. Как было показано в 69, описание того или иного уровня языка может при необходимости осуществляться с помощью специальных машинных моделей, имеющих мало общего с привычными лингвистическими представлениями.

Необходимость построения, а затем и объединения отдельных блоков лингвистического обеспечения АСНТИ требует каждый раз привлечения новых языковедческих и лингво-математических идей и концепций. Так, например, построение ЧС и двуязычных АС осуществлялось на основе достижений лингвостатистики. Создание АСО стимулировало лингвостатистические исследования в области фразеологии и малого синтаксиса, а также потребовало более глубокого сопоставительного анализа языков и исследования двуязычной ситуации. Построение семантического МП, автоматического индексирования и аннотирования опирается на теорию формального распознавания смыслового образа текста. Семантико-синтаксический МП и тезаурусное аннотирование заставляют ввести в инженерно-лингвистический обиход тезаурусные графы, а также теорию нечетких множеств и алгоритмов. Так, возникновение и решение прикладных задач инженерной лингвистики стимулирует развитие своего теоретического субстрата — лингвистики текста.

Приложение

СПИСКИ ГРАНИЧНЫХ СИГНАЛОВ К АЛГОРИТМУ СЕГМЕНТАЦИИ АНГЛИЙСКОГО ТЕКСТА

(Волосевич, 1973, с. 318—322)

Список 1. Левые инклюзивные границы именной группы, правые эксклюзивные границы именной и глагольной групп:

1) a
2) an
3) my
4) its
5) our
6) the
7) your
8) their

Список 2. Промежуточные границы именной и глагольной групп:

1) all
2) any
3) few
4) her
5) his
6) both
7) each
8) many
9) most
10) *much*
11) none
12) same
13) some
14) such
15) that
16) this
17) every
18) other
19) these
20) those
21) anyone
22) nobody
23) anybody
24) nothing
25) someone
26) another
27) anything
28) somebody
29) something
30) *everything*

Список 3. Эксклюзивные границы именной группы:

1) be
2) do
3) is
4) go
5) add
6) are
7) buy
8) did
9) fix
10) get
11) had
12) has
13) let
14) lie
15) mix
16) put
17) say
18) see
19) adds
20) buys
21) came
22) come
23) deal
24) does
25) fall
26) feel
27) fell
28) felt
29) find
30) gave
31) give
32) goes
33) have
34) know
35) knew
36) keep
37) lies
38) live
39) make
40) meet
41) move
42) obey
43) puts
44) says
45) said
46) seem
47) show
48) take
49) tend
50) took
51) want
52) went
53) were
54) vary
55) uses
56) gets
57) sees
58) admit
59) adopt
60) agree
61) allow
62) apply
63) arise
64) avoid
65) began
66) begin
67) bring
68) carry
69) cause
70) cease
71) cover
72) drive
73) enter
74) erase
75) exist
76) holds
77) imply
78) let's
79) leads
80) leave
81) occur
82) reach
83) refer
84) relax
85) serve
86) spend
87) think
88) write
89) absorb
90) accept
91) adjust
92) affect
93) anneal
94) appear
95) assume
96) attain
97) became
98) become
99) behave
100) cancel
101) denote
102) depend
103) detect
104) define
105) differ
106) employ
107) enable
108) ensure
109) equals
110) exceed
111) expect
120) extend
113) follow
114) happen
115) ionize
116) obtain
117) occupy
118) passes
119) permit
112) *remain*
121) remove
122) render
123) reduce
124) rotate
125) modify
126) starts
127) suffer
128) vanish
129) varries
130) achieve
131) acquire
132) analyze
133) applies
134) believe
135) carries
136) collect
137) compute
138) compare
139) conduct
140) consist
141) diffuse
142) discuss
143) examine
144) exhibit
145) explain
146) express
147) furnish
148) imagine
149) implies
150) improve
151) include
152) involve
153) observe
154) operate
155) overlap
156) perform
157) portray
158) prepare
159) prevail
160) prevent
161) proceed
162) produce
163) protect
164) provide
165) receive
166) release
167) replace
168) require
169) suggest
170) suppose
171) sustain
172) undergo
173) utilize
174) compress
175) concerns
176) consider
177) *continue*
178) decrease
179) describe
180) discover
181) endanger
182) estimate
183) evaluate
184) generate
185) increase
186) indicate
187) maintain
188) minimize
189) modifies
190) occupies
191) overcome
192) preclude
193) reappear
194) remember
195) restrict
196) solidify
197) alternate
198) calculate
199) determine
200) eliminate
201) encounter
202) establish
203) intersect
204) introduce
205) *oscillate*
206) represent
207) transform
208) withstand
209) accelerate
210) accomplish
211) contribute
212) coordinate
213) correspond
214) degenerate
215) distribute
216) illustrate
217) neutralize
218) solidifies
219) strengthen
220) understand
221) understood
222) communicate
223) concentrate
224) familiarize
225) incorporate

Список 4. Левые инклюзивные и правые эксклюзивные границы именной группы:

1) at
2) by
3) in
4) of
5) on
6) up
7) for
8) off
9) out
10) via
11) down
12) from
13) into
14) onto
15) over
16) upon
17) with
18) about
19) above
20) after
21) along
22) among
23) below
24) under
25) around
26) before
27) behind
28) beyond
29) during
30) except
31) inside
32) toward
33) versus
34) within
35) against
36) between
37) outside
38) through
39) towards
40) without
41) according
42) throughout

Список 5. Начала придаточных предложений, правые эксклюзивные границы именной и глагольной групп:

1) as
2) if
3) how
4) who
5) why
6) when
7) whom
8) until
9) which
10) where
11) while
12) whose
13) unless
14) because

Список 6. Левые инклюзивные границы глагольной группы, правые эксклюзивные границы именной и глагольной групп:

1) I
2) it
3) he
4) we
5) she
6) you
7) they

Список 7. Промежуточные инклюзивные границы:

а. 1) me
2) us
3) him
4) them

б. 1) no
2) so
3) due
4) not
5) now
6) too
7) yet
8) able
9) also
10) away
11) back
12) best
13) dead
14) easy
15) even
16) fast
17) full
18) here
19) less
20) long
21) just
22) main
23) once
24) only
25) poor
26) pure
27) soon
28) then
29) thin
30) thus
31) true
32) very
33) well
34) what
35) again
36) ahead
37) apart
38) hence
39) never
40) often
41) rapid
42) ready
43) right
44) quite
45) still
46) there
47) twice
48) abrupt
49) almost
50) always
51) across
52) common
53) direct
54) enough
55) indeed
56) itself
57) present
58) rather
59) seldom
60) slight
61) though
62) unique
63) within
64) already
65) complex
66) however
67) perhaps
68) thereby
69) somewhat
70) straight
71) together
72) whenever
73) according
74) otherwise
75) sometimes
76) therefore
77) nevertheless

в. 1) or
2) and
3) but
4) since
5) either
6) whether
7) whereas
8) although

Список 8. Морфемы, выполняющие функции промежуточных инклюзивных границ:

1) -al
2) -ar
3) -el
4) -ic
5) -le
6) -ly
7) -se
8) -age
9) -ant
10) -ary
11) -ate
12) -ent
13) -ete
14) -ful
15) -ine
16) -ire
17) -ite
18) -ive
19) -ous
20) -ult
21) -ute
22) -able
23) -ible
24) -less
25) -ed
26) -ing

Словоформы с выделенными морфемами после соответствующего анализа выводятся на печать в составе или именной, или глагольной группы.

Список 9. Промежуточные границы именной и глагольной групп:

1) got
2) fed
3) led
4) met
5) won
6) been
7) made
8) seen
9) sent
10) done
11) left
12) lost
13) held
14) kept
15) shown
16) given
17) found
18) heard
19) taken
20) known
21) drawn
22) built
23) bound
24) grown
25) chosen
26) driven
27) written
28) brought
29) forbidden

В алгоритме указанные словоформы могут явиться левым эксклюзивным граничным сигналом именной группы. В этом случае словоформа выводится на печать с пометой: «начало причастного оборота».

Список 10. Эксклюзивные границы именной группы:

1) can
2) may
3) will
4) must
5) would
6) could
7) shall
8) should
9) cannot

В составе сложной глагольной группы перечисленные словоформы являются компонентом этой группы.

Список 11. Левые инклюзивные границы глагольной группы, эксклюзивные границы глагольной и именной групп:

1) I'm
2) I'll
3) I've
4) he's
5) it's
6) he'll
7) it'll
8) she's
9) we're
10) we've
11) she'll
12) you'll
13) you're
14) you've
15) they'll
16) they're
17) they've

Являясь граничным сигналом как именной, так и глагольной группы, указанные словоформы выделены в отдельный список для восстановления полной формы вспомогательного глагола.

Окончание -ed может выступать в качестве диагносцирующего признака границы причастного, а -ing — деепричастного оборотов. В качестве граничных сигналов используются также знаки препинания.

СОКРАЩЕНИЯ И УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

АДД — автореферат докторской диссертации.
АИПС — автоматизированная информационно-поисковая система.
АИС — автоматизированная информационная система (см. АИПС).
акад. — академик.
АКД — автореферат кандидатской диссертации.
АН — Академия наук.
АП — Автоматический перевод. Сборник статей. Перевод с английского, итальянского, немецкого и французского языков. Под редакцией, с предисловием О. С. Кулагиной и И. А. Мельчука. М., 1971.
АПТ — сб.: Автоматическая переработка текста. Кишинев, 1972.
АПТМПЛ — сб.: Автоматическая переработка текста методами прикладной лингвистики. Материалы Всесоюзной конференции 6—8 октября 1971 г. Кишинев, 1971.
АС — автоматический словарь.
АСНТИ — автоматизированная система научно-технической информации.
АСО — автоматический словарь оборотов.
АСУ — автоматизированная система управления.
АЦПУ — алфавитно-цифровое печатающее устройство.
БСЭ-II — Большая Советская Энциклопедия. Второе издание, тт. 1—51. М., 1949—1968.
БСЭ-III — Большая Советская Энциклопедия. Третье издание, тт. 1—15. М., 1970—1974.
БЯ — базовый язык.
ВВ — виноградарство и виноделие.
вин. — винительный падеж.
ВИНИТИ — Всесоюзный институт научно-технической информации.
ВММ — сб.: Вычислительные машины и мышление. Перевод с английского. М., 1967.
ВНТС'ЛО — сб.: Всесоюзный научно-технический симпозиум «Лингвистическое обеспечение автоматизированных систем управления и информационно-поисковых систем». Махачкала, 1974.
ВОПЛ — сб.: Вопросы общей и прикладной лингвистики. Минск, 1975.
ВП — журн.: Вопросы психологии. М.
ВСК — вторичный сжатый код.
ВФ — журн.: Вопросы философии. М.
ВЯ — журн.: Вопросы языкознания. М.
ГНГ — геология нефти и газа.
ГПИ — Государственный педагогический институт.
ГПИИЯ — Государственный педагогический институт иностранных языков.
ГС — граничный сигнал.
дат. — дательный падеж.
дв. ед. — двоичная единица.

ДСАТ — сб.: Дистрибутивно-статистическое описание текстов (для нужд машинной переработки текстов в АСУ и АИС). Иркутск, 1973.
ед. ч. — единственное число.
ЕС — единая система.
ЗУ — запоминающее устройство.
ИЗ — искусственный знак.
ИЛ — сб.: Инженерная лингвистика. Уч. зап. Ленинградского ордена Трудового Красного Знамени ГПИ им. А. И. Герцена, т. 458, ч. I—II, Л., 1971.
им. ф. — форма имени существительного, прилагательного, местоимения, причастия.
имен. — именительный падеж.
инф. — инфинитив.
ИП — информационный процесс.
ИПС — информационно-поисковая система.
ИПЯ — информационно-поисковый язык.
ИЯШ — журн.: Иностранные языки в школе. М.
Кбайт — килобайт (1 тыс. байт).
КД — кандидатская диссертация (рукопись).
Кн. — книга.
КОКН — сб.: Кибернетика ожидаемая и кибернетика неожиданная. М., 1968.
КС — ключевое слово.
КСб — сб.: Кибернетический сборник, вып. 1—7, 1960—1963, Новая серия вып. 1— , 1965.
КСс — ключевое словосочетание.
КУЭ — класс условной эквивалентности.
КФ — каноническая (исходная) форма слова.
Л. — Ленинград.
ЛААТ — сб.: Лингвостатистика и автоматический анализ текста. Минск, 1973.
ЛГПИ — Ленинградский государственный педагогический институт.
ЛГУ — Ленинградский государственный университет.
ЛЕ — лексическая единица
ЛО — сб.: Лингвистическое обеспечение автоматизированных систем управления и информационно-поисковых систем. (Краткое содержание докладов научно-технической конференции 29—31 мая 1974 г.). Тюмень, 1974.
ЛОТКЗМИ — Ленинградский ордена Трудового Красного Знамени медицинский институт.
ЛТ — лингвистика текста.
М. — Москва.
МАРС — многоотраслевой автоматический русский словарь.
МБ — магнитный барабан.
Мбайт — мегабайт (1 млн. байт.).
МБЯ — машинный базовый язык.
МГУ — Московский государственный университет.
МД — магнитный диск.
МДо — машинный документ.
МИПЯ — машинный информационно-поисковый язык.
МК — магнитная карта.
МЛ — магнитная лента.
ММЯ — сб.: Математические методы в языкознании. Рига, 1969.
мн. ч. — множественное число.
МНК — Материалы научной конференции профессорско-преподавательского состава Чимкентского педагогического института. Чимкент, 1970.
МП — машинный перевод (реализуемый на ЭВМ).

МПИЯР — сб.: Методика преподавания иностранных языков за рубежом. М., 1967.
МСС — сб.: Морфологическая структура слова в языках различных типов. М.— Л., 1963.
МТ — морские тексты.
НДВШ — журн.: Научные доклады высшей школы. Филологические науки.
НЛ — сб.: Новое в лингвистике. I—VII. М., 1960—1975.
НТИ — журн.: Научно-техническая информация. М.
ОЗУ — оперативное запоминающее устройство (ср. ОП.).
ОП — оперативная память.
ОС — Обратный словарь русского языка. М., 1974.
ОтС — отраслевой словарь.
п. — пункт.
пад. — падеж.
ПИП — сб.: Проблемы инженерной психологии. Выпуск 3. Психология памяти. Л., 1965.
ПКС — полиключевое слово.
ПКСс — полиключевое словосочетание.
ПО — поисковый образ.
ПОД — поисковый образ документа.
ПОЗ — поисковый образ запроса.
ПП — поисковое предписание (см. ПОЗ).
предл. — предложный падеж.
прил. — прилагательное.
ПТ — полупроводниковая техника.
ПУ — печатающее устройство.
РЖ — реферативный журнал.
РО — сб.: Распознавание образов. Исследование живых и автоматических распознающих систем. М., 1970.
род. — родительный падеж.
РТИК — К. Шеннон. Работы по теории информации и кибернетике. Перевод с английского. М., 1963.
РЭН — распознающий эталонный набор семантических кодов.
сб. — сборник.
СК — сжатый код.
СКП — сжатый код парадигмы.
СКТ — сб.: Статистика казахского текста. I. Труды группы «Статистико-лингвистическое исследование и автоматизация». Выпуск III. Алма-Ата, 1973.
СЛП — сушка лакокрасочных поверхностей.
см. — смотри.
СОЛ — словарь общеупотребительной лексики.
СП — семантическое пространство.
СПо — семантическое поле.
ср. — сравни.
СТ I — сб.: Статистика текста I. Лингвостатистические исследования. Минск, 1969.
СТ II — сб.: Статистика текста II. Автоматическая переработка текста. Минск, 1970.
Ст — стоматология.
СтР — сб.: Статистика речи. Л., 1968 — итальянский перевод: Statistica linguistica (con l'aggiunta di due appendici). Linguistica 3. Bologna; немецкий перевод:Sprachstatistik. Sammlung Akademie—Verlag 22. Sprache. Berlin. 1973.
СтРААТ — сборники: Статистика речи и автоматический анализ текста. Л., 1971; 1973; 1974.
сущ. — существительное.
табл. — таблица.

тв. — творительный падеж.
ТИСО — типы интеркомпонентных смысловых отношений.
ТМС — типовые множества смыслов.
ТСХМ — тракторное и сельскохозяйственное машиностроение.
ТЯИЛ — сб.: Теория языка и инженерная лингвистика. Л., 1973.
УДК — Универсальная десятичная классификация.
УМН — журн.: Успехи математических наук, М.
УС — устойчивое словосочетание.
УСиМ — журн.: Управляющие системы и машины.
Уч. зап. — ученые записки.
Ф — файл.
Фл — флексия.
ФН — журн.: Филологические науки (см. НДВШ).
ЦНИИПИ — Центральный научно-исследовательский институт патентной информации.
ЧВААТ — сб.: Частные вопросы автоматического анализа текста. Минск, 1972
ЧС — частотный словарь.
ЭВМ — электронно-вычислительная машина.
ЭСП — сб.: Экспериментальная система англо-русского автоматического перевода патентной документации. М., 1970.
ЭЯСР — сб.: Энтропия языка и статистика речи. Минск, 1966.

a — adjectif, adjective.
adv. — adverbe, adverb.
AFIPS — American Federation of Information Processing Societies.
art — article.
BSLP — журн.: Bulletin de la Société de Linguistique de Paris. P., 1900.
BSTJ — журн.: Bell System Technical Journal. Devoted to the Scientific and Engineering Aspects of Electrical Communication. N. Y., 1922.
BZB — Burghagens Zeitschrift für Bürotechnik und Informatik.
C — журн.: Cybernetica.
C — conjonction, conjunction.
CTh — Computers and Thought. A Collection of Articles. Eds: E. A. Feigenbaum and J. Feldman. N. Y.—S.-Francisco—Toronto—London, 1963.
dem. — demonstrative.
F — абсолютная частота.
*F** — абсолютная накопленная частота.
HMP — журн.: Handbook of Mathematical Psychology. ed. R. D. Luce, R. R. Bush, E. Galanter. N. Y., 1963—1965.
i — номер лингвистической единицы в ЧС.
IC — журн.: Information and Control.
IEEE — Institute of Electrical and Electronics Engineers.
IEEE TEC — журн.: IEEE Transactions on Electronic Computers. N. Y.
IEEE TSMC — журн.: IEEE Transactions on Systems, Man and Cybernetics. N. Y.
imp. — imperative.
inf. — infinitive.
IRE — Institute of Radio Engineers.
ISIT — сб.: 2nd International Symposium of the Information Theory, Tsahkadsor, 1971. Budapest, 1973.
JVLVB — Journal of Verbal Learning and Verbal Behavior.
L — Linguistics. An International Review.
M, m — количество лингвистических единиц, обладающих данной частотой.
McT — журн.: Mechanical Translation and Computational Linguistics. An International Journal. Chicago.

MIT — Massachusetts Institute of Technology.
mod. — modal (verb).
MS — журн.: Management Science. Journal of Management Sciences. Providence, Rhode Island.
MT — сб.: Machine Translation. Ed. A. D. Booth, Amsterdam.
N — объем выборки.
*N** — накопленное количество словоупотреблений.
n — noun.
ND — журн.: Nachrichten für Dokumentation.
num. — numeral.
N. Y. — New York.
OR — журн.: Operations Research. The Journal of the Operations Research Society of America. Baltimore.
part. — participle.
pr. — pronom, pronoun.
pref. — prefix, Präfix.
prp. — preposition.
prtc. — Partikel.
RP — сб.: Recognizing Patterns. Studies in Living and Automatic Systems. Ed. P. A. Kolers and M. Eden. Cambridge, Mass.— London, 1968.
RRL — журн.: Revue (Roumaine) de Linguistique. Bucureşti.
RTCL — сб.: Research Trends on the Computational Linguistics. Arlington, Va, 1972.
s — substantif, substantiv.
SCL — журн.: Studii şi cercetări lingvistice. Bucureşti.
sing. — singular.
SL — сб.: Statistica linguistica (con l'agguinta di due appendici). «Linguistica». Collezione di monografie originali o tradotte di linguistica generale, speciale ed applicata, diretta da C. Tagliavini. 3. Bologna, 1971.
SLMA — сб.: Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics, vol. XII, Providence, Rhode Island, 1961.
TL — сб.: Textlinguistik. B. I. Dresden, 1970.
v — Verb, verbe.
y — формула.
X — цифра.
Z — символ и формула.
ZD — журн.: Zeitschrift für Datenverarbeitung.
ZPhSK — журн.: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung. Berlin.
* — диссертационная работа, выполненная под руководством Р. Г. Пиотровского.

ЛИТЕРАТУРА

Маркс К., Энгельс Ф. Сочинения. Изд. 2. М.
Ленин В. И. Философские тетради. М., 1973.
Авербух О. З.* Формально-смысловые структуры английского языка. — *АКД. Калинин, 1970.*
Алексеев П. М. Частотный словарь английского подъязыка электроники. — СтР, 1968.
Алексеев П. М. Некоторые вопросы теории и практики статистической лексикографии. — СТ I, 1969.
Алексеев П. М. Частотные словари английского языка и их практическое применение. — СтРААТ, 1971а.
Алексеев П. М. Частотный англо-русский словарь-минимум по электронике. М., 1971б.
Алексеев П. М., Герман-Прозорова Л. П., Пиотровский Р. Г., Щепетова О. П. Основы статистической оптимизации преподавания иностранных языков. — СтРААТ, 1974.
Алексеев П. М., Навальна С. Т. Про графічний опис залежності «ранг—частота» лінгвистичних одиниць. — Вісник Харьківського університету, № 64, філологія, вип. 8, Харків, 1971.
Алексеев П. М., Турыгина Л. А. Частотный англо-русский словарь-минимум газетной лексики. М., 1974.
Амосов Н. М., Касаткин А. М., Касаткина Л. М., Талаев С. А. Автоматы и разумное поведение. Опыт моделирования. Киев, 1973.
Анохин П. К. Философский смысл проблемы естественного и искусственного интеллекта. — ВФ, 1973, № 6.
Апресян Ю. Л. Лексическая семантика. Синонимические средства языка. М., 1974.
Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А. О смысле ранговых распределений. — НТИ, серия 2, 1975, № 1.
Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А. Ранговые распределения в тексте и языке. — НТИ, серия 2, 1975, № 2.
Арсентьева Н. Г., Кулагина О. С. Международная конференция по вычислительной лингвистике. Машинный перевод и прикладная лингвистика. Вып. 17, М., 1974.
Афаунова О. А. Автоматическое устранение грамматической многозначности английских ing-овых форм. — СтРААТ, 1973.
Баженов Л. Б. О некоторых философских аспектах проблемы моделирования мышления кибернетическими устройствами. — В кн.: Кибернетика, мышление, жизнь. М., 1964.
Байтанаева Д. А., Бектаев К. Б. Энтропия казахского текста. — СКТ, 1973.

Байтанаева Д. А., Бектаев К. Б., Чалабаева М. Ч. Информационная структура казахской и узбекской словоформы. — Труды VII Всесоюзной школы-семинара. Автоматическое распознавание слуховых образов. Алма-Ата, 1973.
Бальмонт К. Д. Полное собрание стихов. Том первый. М., 1909.
Башарин Г. П. О статистической оценке энтропии последовательности независимых случайных величин. — Теория вероятностей и ее применение, т. 4, № 3, 1959.
Бектаев К. Б.* Статистико-информационная типология тюркского текста. АДД. Ленинград, 1975.
Бектаев К. Б., Белоцерковская Л. И., Борисевич А. Д., Букович В. А., Булашева Н. С., Гончаренко В. В., Данейко М. В., Зуева Т. Р., Исабекова Н. И., Нехай О. А., Пиотровский Р. Г., Соркина В. А. Статистическое выделение русских машинных оборотов и построение двуязычных автоматических словарей. — СтРААТ, 1973.
Бектаев К. Б., Зубов А. В., Ковалевич Е. Ф., Машкина Л. Е., Нехай О. А. К исследованию законов распределения лингвистических единиц. — СТ I, 1969.
Бектаев К. Б., Лукьяненков К. Ф. О законах распределения единиц письменной речи. — СтРААТ, 1971.
Бектаев К. Б., Машкина Л. Е., Микерина Т. А., Ротарь А. С. Энтропия словоформы и словосочетания в английских и немецких текстах. — ЭЯСР, 1966.
Бектаев К. Б., Пиотровский Р. Г. Математические методы в языкознании. Часть I. Теория вероятностей и моделирование нормы языка. Алма-Ата, 1973; часть II. Математическая лингвистика и моделирование текста. Алма-Ата, 1974.
Бектаев К. Б., Пиотровский Р. Г., Шабес В. Я. Тезаурусное распознавание смысла документа в АСУ и АИПС—ЛО, 1974.
Белоногов Г. Г., Богатырев В. И. Автоматизированные информационные системы. М., 1973.
Белоногов Г. Г., Котов Р. Г. Автоматизированные информационно-поисковые системы. М., 1968.
Белоногов Г. Г., Шемакин Ю. И., Новоселов А. П., Чиркин В. А., Рыбаков Б. П. Автоматическое индексирование документов и запросов. НТИ, серия 1, 1973, № 7.
Белоцерковская Л. И. Исследование сегментов текста в аспекте нормы и узуса. АКД. Алма-Ата, 1973.
Беляева Л. Н.* Статистическая структура словообразовательных гнезд и машинный словарь русских основ. АКД. Л., 1974.
Беляева Л. Н., Крисевич В. С., Липницкий С. Ф., Пиотровский Р. Г. О многоцелевом автоматическом словаре русского языка (МАРС). — ВОПЛ, 1975.
Берзон В. Е.* Исследование связности текста при разработке автоматических методов его свертывания. — АКД. Л., 1972.
Берзон В. Е. Автоматическое смысловое свертывание на основе анализа сверхфразовых связей текста. — СтРААТ, 1974.
Берман Р. Л., Мандрыка О. А., Рабинович Е. Е. Отбор французской и румынской общеупотребительной лексики. — СтРААТ, 1974.
Бернштейн Н. А. О построении движений. М., 1947.
Бехтерева Н. П., Бундзен П. В., Матвеев Ю. К., Каплуновский А. С. Функциональная реорганизация активности нейронных популяций мозга человека при кратковременной вербальной памяти. — Физиологический журнал СССР им. И. М. Сеченова, т. VII, № 12, 1971.
Билан В. Н.* Семантико-синтаксическое распознавание английских двухкомпонентных именных словосочетаний. — АКД. Минск, 1975.

Богодист В. И.* Измерение смысловой информации лингвистических единиц французского текста. — АКД. Л., 1974.
Богодист В. И., Георгиев Х. Ц., Пестунова В. Н., Пиотровский Р. Г., Райтар С. В. Как измерить смысловую информацию. — ФН, 1975, № 4.
Богуславская Г. П. Информационно-статистическая оценка аналитизма в английском языке. — СТI, 1969.
Богуславская Г. П.* Энтропия английского печатного текста. — АКД. Минск, 1966.
Богуславская Г. П., Новак Л. А. Энтропия английского и румынского языков. Серия: Статистика речи II. Минск, 1968.
Большаков И. А. Статистические порождающие грамматики для синтетических языков и основанные на них алгоритмы анализа и сжатия предложений. Известия АН СССР. Техническая кибернетика, 1970, № 1, 4; 1971, № 1.
Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., 1965.
Бонгард М. М. О понятии «полезная информация». — ПК, вып. 9, 1963.
Борисевич А. Д. Словарь трехсловных сочетаний, полученных с помощью ЭВМ (на материале английских строительных текстов). — СТ I, 1969.
Борисевич А. Д.* Англо-русский автоматический словарь оборотов. (К проблеме идиоматичности при обращении текста в системе «человек—машина—человек») — АКД. Минск, 1972.
Борисевич А. Д., Гончаренко В. В., Гречишкина В. И., Добрускина Э. М., Кочеткова В. К., Крисевич В. С., Нехай О. А., Пиотровский Р. Г., Штирбу Т. А., Ястребова С. В. Кодирование грамматической информации в машинном словаре. — СТ II, 1970.
Боркун М. Н. Частотный словарь трехчленных словосочетаний в подъязыке английской публицистики. — СТ I, 1969.
Братко А. А. Моделирование психики. М., 1969.
Будагов Р. А. Человек и его язык. Изд. МГУ, 1974.
Булашева Н. С. Статистика трехсловных сочетаний с опорным частотным словом из текстов по русской радиоэлектронике. — СТ I, 1969.
Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. Статистические проблемы обучения. М., 1974.
Вертель В. А. Автоматический немецко-русский словарь для перевода научно-технических текстов (подъязык экономической литературы ГДР). — ЧВААТ, 1972.
Вертель В. А.* Теоретическое и экспериментальное исследование одной модели немецко-русского автоматического словаря. — АКД. Л., 1974.
Вертель В. А., Вертель Е. В. Алгоритм получения частотного словаря с учетом длины словоформ. — СТ II, 1970.
Вертель В. А., Вертель Е. В., Крисевич В. С., Пиотровский Р. Г., Трибис Л. И. Автоматические словари в системе бинарного вероятностного МП. — Сб. «Инженерная лингвистика». Уч. зап. ЛГПИ им. Герцена, т. 458, Л., 1971.
Ветров А. А. Семиотика и ее основные проблемы. М., 1968.
Волосевич И. И.* Опыт формальной сегментации английского предложения. — АКД. Л., 1971.
Волосевич И. И. Машинная сегментация английского предложения. — СтРААТ, 1973.
Волоцкая З. М., Молошная Т. Н., Николаева Т. М. Опыт описания русского языка в его письменной форме. М., 1964.
Гаврин М. К. О ценности информации. — Вестник ЛГУ. Серия математики, механики и астрономии, № 19, вып. 4, 1963.

Гак В. Г. Беседы о французском слове. М., 1966.
Гансовский С. Машина как личность. — КОКН, 1968.
Гачечиладзе Т. Г., Цилосани Т. П. Об одном методе изучения статистической структуры текста. — СтРААТ, 1971.
Георгиев Х. Ц.* Информационные измерения болгарского языка. — АКД, Л., 1973.
Георгиев Х. Ц., Богодист В. И., Коженец Т. Ф., Пестунова В. Н., Пиотровский Р. Г., Райтар С. В. Измерение смысловой информации. — ЧВААТ, 1972.
Герман-Прозорова Л. П.* Опыт вероятностно-статистической оптимизации обучения иностранному языку в неязыковом вузе (на материале английского подъязыка электроники) — АКД. Л., 1973.
Гладкий А. В., Мельчук И. А. Элементы математической лингвистики. М., 1969.
Головин Б. Н. Из курса лекций по лингвистической статистике. Горький, 1966.
Гончаренко В. В.* Лексикографические, лингвостатистические и инженерно-лингвистические вопросы построения автоматического словаря (англо-русский автоматический словарь по полупроводникам). — АКД. Кишинев, 1972а.
Гончаренко В. В.* Лексикографические, лингвостатистические и инженерно-лингвистические вопросы построения автоматического словаря (англо-русский автоматический словарь по полупроводникам). — КД. Ленинград, 1972б.
Гончаренко В. В., Гречишкина В. А., Добрускина Э. М. Англо-русская морфологическая неадекватность и кодирование грамматической информации в автоматическом словаре (АС). — АПТ, 1972.
Гончаренко В. В., Добрускина Э. М. Лексический перевод английских научно-технических текстов с помощью ЭВМ. — АПТ, 1972.
Граудина Л. К. Развитие нулевой формы родительного множественного у существительных — единиц измерения. — В кн.: Развитие грамматики и лексики современного русского языка. М., 1964.
Григорьев В. П. Предварительные итоги статистического исследования поэтики испанского народного романса. — СтРААТ, 1971.
Грубов В. И., Кирдан В. С. Электронные вычислительные машины и моделирующие устройства. Киев, 1969.
Гурова Н. В. Базовая лексика и морфология английского подъязыка металлургии (прокатное производство). — АКД. Л., 1974.
Данейко М. В. Статистика глагольных словосочетаний в английских текстах по радиоэлектронике. — СТ I, 1969.
Данейко М. В., Петровская В. М. Алгоритм автоматической сегментации английского предложения. — СТ II, 1970.
Джолос Е. М. Алгоритмы определения расстояний между словами. — ПТИ, 1965, № 8.
Джубанов А. Х. Статистическое исследование казахского текста с применением ЭВМ (на материале романа М. Ауэзова «Абай жолы»). — АКД. Алма-Ата, 1973.
Добрускина Э. М.* Проблемы фразеологии и идиоматичности текста в электронно-вычислительной машине (англо-русский автоматический словарь оборотов). — АКД. Л., 1973.
Дунаевский В. А., Перцовая Г. М. Автоматическое реферирование и перевод-аннотация английских текстов по стоматологии. — Материалы I научной сессии стоматологического факультета I ЛОТКЗМИ им. акад. И. П. Павлова. Л., 1972.
Ешан Л. И.* Опыт статистического описания научно-технического стиля румынского языка (на материале текстов по радиоэлектронике). — АКД. Л., 1966.
Жинкин Н. И. Звуковая коммуникативная система обезьян. — Известия Академии педагогических наук РСФСР, вып. 113, 1960.

Загоруйко Н. Г. Методы распознавания и их применение. М., 1972.
Звонкин А. К., Левин Л. А. Сложность конечных объектов и обоснование понятий информации и случайности с помощью теории алгоритмов. — УМН, т. 25, № 6, 1970.
Зиновьев А. А. Очерк многозначной логики. — В кн.: Проблемы логики и теории познания. М., 1968.
Зореф М. Г.* Машинные основы и машинная морфология в немецко-русском автоматическом словаре. — АКД. Кишинев, 1972.
Зубов А. В.* Переработка текста естественного языка в системе «человек—машина». — АКД. Л., 1969.
Зубов А. В., Лукьяненков К. Ф., Пиотровский Р. Г., Хотяшов Э. Н. Лексико-статистическое описание текста на электронно-вычислительных машинах. — СтР, 1968.
Зуева Т. Р. Статистическая характеристика четырехсловных сочетаний русского текста по электронике. — В кн.: Материалы научной конференции профессорско-преподавательского состава Чимкентского педагогического института, посвященной 100-летию со дня рождения В. И. Ленина и 50-летию Казахстана. Чимкент, 1970.
Иванкин В. И. Семантико-дистрибутивный метод составления тематических списков ключевых слов, рекомендуемых для ввода в поисковые образы документа. — НТИ, серия 2, 1974, № 10.
Иванкин В. И. Опыт анализа семантических связей между текстами запросов и релевантных документов. — НТИ, серия 2, 1975, № 1.
Ивахненко А. В природе запрета нет. — КОКН, 1968.
Информационно-поисковая система «БИТ». Киев, 1968.
Информационно-поисковая система SMART. Сборник переводов № 4, ВИНИТИ, М., 1966.
Ионицэ М. П.* Статистика и вероятностный машинный перевод французского глагола на русский и молдавский языки (на материале французских текстов по публицистике). — АКД. Л., 1971.
Исабекова Н. Именные словосочетания в русских текстах по радиоэлектронике. — СТ I, 1969.
Исаков В. П., Лоскутов Н. Г. О выборе алгоритмического языка для программирования задач обработки смысловой информации. — СтРААТ, 1971.
Калинин В. М. О статистике литературного текста. — ВЯ, 1964а, № 1.
Калинин В. М.* Развитие схемы Пуассона и ее применение для описания статистических свойств речи. АКД. Л., 1964б.
Калинина Е. А. Изучение лексико-статистических закономерностей на основе вероятностной модели. — СтР, 1968а.
Калинина Е. А. Частотный словарь русского подъязыка электроники. — СтР, 1968б.
Калугина В. В. Словарь ложной омонимии флексий русского языка. Кишинев, 1965.
Каменева З. М.* О межъязыковой идиоматичности (на материале формального анализа перевода английских составных терминов). — АКД. Минск, 1973.
Карпилович Т. П. Программа построения частотно-алфавитного словаря на машине с байтовой структурой. — ЛААТ, 1973.
Каушанская М. В., Лукьяненков К. Ф. Статистический отбор ключевых и терминологических единиц для атрибуции и реферирования французского научно-технического текста. — АПТ, 1972.
Кацнельсон С. Д. Типология языка и речевое мышление. Л., 1972.
Каширина М. Е. О типах распределения лексических единиц в тексте. — СтРААТ, 1974.
Киссен И. А. Словарь наиболее употребительных слов современного узбекского литературного языка (высокочастотная лексика подъязыка художественной прозы). Ташкент, 1972.
Клименко А. П. Лексическая системность и ее психолингвистическое изучение. Минск, 1974.

Кобозев Н. И. Исследования в области термодинамики процессов информации и мышления. МГУ, 1971.
Коверин А. А.* Грамматический анализ на ЭВМ французских научно-технических текстов. — АКД. Л., 1972.
Коверин А. А., Скитневский Д. М. Опыт грамматического анализа на ЭВМ французских научно-технических текстов. — ДСАТ, 1973.
Кодухов В. И. Общее языкознание. М., 1974.
Коженец Т. Ф.* Энтропия немецкого печатного текста. — АКД. Минск, 1973.
Колева Т. Р. Частотный словарь немецких текстов по виноградарству и виноделию. — АПТ, 1972.
Колесникова В. В.* Лингвистические вопросы автоматического распознавания смысла английских научно-технических текстов. — АКД. Л., 1974.
Колмогоров А. Н. Три подхода к определению понятия «количество информации». — В кн.: Проблемы передачи информации, т. I, вып. 1. М., 1965.
Кондаков Н. И. Логический словарь. М., 1971.
Копачева Н. А. Статистическое исследование синтаксиса английских текстов по радиоэлектронике. — ЭЯСР, 1966.
Копылова М. Е., Кравцова И. С., Кулешова М. В. Вероятностное определение коэффициента распространенности и меры употребительности. ВНТС ЛО, 1974.
Кочеткова В. К.* Вероятностно-статистическое построение автоматического словаря. (На материале французских текстов по электронике). — АКД. Л., 1969.
Кочеткова В. К., Скрелина Л. М. Частотный словарь французского подъязыка электроники. — СтР, 1968.
Крайзмер Л. П., Матюхин С. А., Майоркин С. Г. Память кибернетических систем (основы мнемологии). М., 1971.
Крисевич В. С.* Серийный автоматический поиск при обработке информации с помощью ЭВМ. — СТ II, 1970.
Кунин А. В. (сост.). Англо-русский фразеологический словарь. Изд. 3-е, исправленное, в двух книгах. I—II. М., 1967.
Курбаков К. И. Кодирование и поиск информации в автоматическом словаре. М., 1968.
Кушнерев Н. Т., Неменман М. Е., Цагельский В. И. Программирование для ЭВМ «Минск-32». М., 1973.
Ларионов А. М., Левин В. К., Пржиялковский В. В., Фатеев А. Е. Основные принципы построения и технико-экономические характеристики Единой системы ЭВМ (ЕС ЭВМ). — УСиМ, 1973, № 2 (4).
Лебедев Д. С., Гармаш В. А. Статистический анализ трехбуквенных сочетаний русского текста. — В кн.: Проблемы передачи информации, вып. 2. М., 1959.
Левит З. Н. К проблеме аналитического слова в современном французском языке. Минск, 1968.
Лейбниц Г. В. Избранные философские сочинения. М., 1908.
Лейтес А. Хлебников — каким он был. . . — Новый мир, 1973, № 1.
Лукьяненков К. Ф.* Лексико-статистическое описание английского научно-технического текста с помощью электронно-вычислительной машины. — АКД. Минск, 1969.
Лукьянова Е. М. О соотношении лингвистического и математического обеспечения АСУ и АИПС на базе ЭВМ третьего поколения. — ВОПЛ, 1975.
Лурия А. Р. Основы нейропсихологии. Изд. МГУ, М., 1973.
Лурия А. Р. Нейрофизиология памяти. М., 1974.
Лурия А. Р. Научные горизонты и философские тупики в современной лингвистике. — ВФ, 1975, № 4.

Манукова З. Г.* Опыт автоматического выделения и перевода немецких глагольных групп (на материале немецких публицистических текстов). — АКД. Калинин, 1972.
Манукова З. Г., Зубов А. В. Автоматический синтаксический анализ предложения при переводе немецких публицистических текстов. — ЧВААТ, 1972.
Марчук Ю. Н. Опыт машинной реализации дистрибутивной методики определения лексических значений. — СтРААТ, 1973.
Маслов Ю. С. Знаковая теория языка. — В кн.: Вопросы общего языкознания. Материалы республиканского семинара преподавателей общего языкознания, Л., 1967.
Маткаш Н. Г.* Лексика и морфология молдавской публицистики в сравнении с лексикой и морфологией других дако-романских функциональных стилей. — АКД. Л., 1967.
Машкина Л. Е.* О статистических методах исследования лексико-грамматической дистрибуции (на материале публицистических текстов политической тематики современного немецкого языка). — АКД. Минск, 1968.
Маяковский В. Как делать стихи? М., 1952.
Межлумова А. Б. Статистическая характеристика лексики и морфологии русских текстов по радиотехнике. — АКД. Минск, 1973.
Мельчук И. А. Опыт теории лингвистических моделей «смысл—текст». Семантика, синтаксис. М., 1974.
Мельчук И. А., Равич Р. Д. Автоматический перевод. 1949—1963. М., 1967.
Микерина Т. А.* Некоторые статистические приемы лексико-морфологического описания функционального стиля (на материале английских текстов по судостроению). — АКД. Л., 1967.
Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. М., 1968.
Михайлова И. В.* Основы автоматического сегментирования испанского текста. — АКД. Л., 1972.
Михлин Г. З., Пиотровский Р. Г., Фрумкин В. А. Свертывание кодов слов при автоматической переработке текстов. — НТИ, серия 2, 1974, № 9.
Мотылев В. М., Пиотровский Р. Г., Сова Л. З., Шабес В. Я. Объединение отраслевых тезаурусов с помощью ЭВМ. — ВНТС ЛО, 1974.
Налимов В. В. Вероятностная модель языка. О соотношении естественных и искусственных языков. М., 1974.
Невельский П. Б. Объем памяти и количество информации. — ПИП, 1965.
Невельский П. Б., Розенбаум М. Д. Угадывание профессионального текста специалистами и неспециалистами. — СтРААТ, 1971.
Нехай О. А.* Статистика и автоматический анализ текста (на материале английских текстов по электронике). — АКД. Минск, 1968.
Никитина Л. С. Именные трехсловные сочетания в русских публицистических текстах. — СтРААТ, 1971.
Новак Л. А.* Частотный словарь молдавского и румынского языков. — АКД. Л., 1962.
Ноздрина В. А. Частотный список атрибутивных сочетаний в английских текстах по радиоэлектронике. — СТ I, 1969.
Овсянников Г. И., Пиотровский Р. Г., Сидоров В. А., Шабес В. Я. К вопросу о лингвистическом обеспечении АИПС по нефтехимии. В кн.: Вопросы прогнозирования, экономики, информатики и применения математических методов и ЭВМ в нефтехимии. Л., 1974.
Ольшанский И. Г. Текст как единство элементов и отношений. Лингвистика текста. — Материалы научной конференции. Часть I. М., 1974.

Орлов Ю. К. О статистической структуре сообщений, оптимальных для человеческого восприятия (К постановке вопроса). — НТИ, серия 2, 1970, № 8.
Орлова А. А., Буравцева Н. М. К вопросу о машинном переводе французских публицистических текстов. — В кн.: Вопросы современного французского языка. Калинин, 1973.
Палибина И. В. Автоматическое реферирование текстов по морской технике. — АПТМПЛ, 1971.
Панова Н. С., Шрейдер Ю. А. О знаковой природе классификации. — НТИ, серия 2, 1974. № 12.
Пашковский В. Э., Сребрянская И. И. Статистические оценки письменной речи больных шизофренией. — ИЛ, 1971.
Перцовая Г. М.* Автоматический анализ содержания текста (на материале английских текстов по стоматологии. — АКД. Л., 1975.
Петренко Б. В. Исследование документального информационного потока на основе анализа запросов. — НТИ, серия 2, 1974, № 10.
Петрова Н. В.* Энтропия французского печатного текста. — АКД. Л., 1968.
Петрова Н. В., Пиотровский Р. Г. Слово, контекст, морфология. — ВЯ, 1966, № 2.
Пешковский А. М. Русский синтаксис в научном освещении. Изд. 7-е, М., 1956.
Пиотровская А. А. Машинная морфология русского глагола. — СтРААТ, 1973.
Пиотровская А. А. Алгоритм приведения именных словоформ документов, перерабатываемых в АСУ и АИПС. — ВНТС ЛО, 1974а.
Пиотровская А. А. Машинная грамматика для АСУ и АИПС. — ЛО, 1974б.
Пиотровская А. А., Пиотровский Р. Г. Математические модели диахронии и текстообразования. — СтРААТ, 1974.
Пиотровская А. А., Рихтер Н. А. Синтез видовых форм русского глагола, — СТ II, 1970.
Пиотровский Р. Г. Очерки по стилистике французского языка. Морфология и синтаксис. Л., 1960.
Пиотровский Р. Г. О теоретико-информационных параметрах устной и письменной форм языка. — В кн.: Проблемы структурной лингвистики. М., 1962.
Пиотровский Р. Г. Моделирование фонологических систем и методы их сравнения, Л., 1966.
Пиотровский Р. Г. Информационные измерения языка. Л., 1968.
Пиотровский Р. Г., Байтанаева Д. А., Бектаев К. Б., Богодист В. И., Георгиев Х. Ц., Пестунова В. Н., Райтар С. В. Можно ли измерять смысловую информацию слова? — СКТ, 1973.
Пиотровский Р. Г., Турыгина Л. А. Антиномия «язык—речь» и статистическая интерпретация нормы языка. — СтРААТ, 1971.
Пиотровский Р. Г., Чижаковский В. А. О двуязычной ситуации. — СтРААТ, 1971.
Полетаев И. А. Вступительная статья к сб. «Человеческие способности машин». М., 1971.
Поляков Г. И. О принципах нейронной организации мозга. Изд. МГУ, М., 1965.
Попеску А. Н. Автоматическое индексирование и аннотирование научно-технических текстов. — В кн.: Всесоюзный семинар по информационным языкам. Научный совет по комплексной проблеме «Кибернетика» АН СССР. Секция семиотики. Предварительные публикации. Вып. 4. М., 1972.
Попеску А. Н., Колесникова В. В., Палибина И. В., Рахубо Н. П., Тарасова Е. С., Чайковская И. И. Вероятностное и детерминистское распознавание смыслового образа документа. — ЛО, 1974.
Попеску А. Н., Хажинская М. С. Тезаурусный метод составления алгоритмов для автоматического реферирования французского научно-технического текста. — АПТ, 1972.
Попеску А. Н., Хажинская М. С. Тезаурусный метод автоматического распознавания смыслового образа научно-технического текста. — ЛААТ, 1973.
Попескул А. Н.* Анализ методов и построение алгоритмов вероятностно-детерминистского распознавания смысла связного текста. — АКД. Киев, 1975.
Приймов Ю. Г., Соркина В. А. Алгоритм автоматического выделения именной группы в техническом тексте на английском языке. — СТ II, 1970.
Прогресс биологической и медицинской кибернетики. М., 1974.
Радченко А. Н. Моделирование основных механизмов мозга. М., 1968.
Ратцева И. И., Строганов В. А. Семантика текста, V. Алгоритмизация. — В кн.: Лингвистика текста. Материалы научной конференции, ч. II. М., 1974.
Рахубо Н. П.* Автоматическое распознавание смысла французского научно-технического текста. — АКД. Л., 1974.
Репкина Г. В. Исследование оперативной памяти. — ПИП, 1965.
Рецкер Я. И. Теория перевода и переводческая практика. М., 1974.
Ротарь А. С.* Опыт статистического описания современных немецких публицистических текстов (к вопросу о некоторых дивергентных явлениях в газетных текстах ГДР и ФРГ). — АКД. Минск, 1971.
Рохваргер А. Е., Кирюшина Е. Н., Кудрин Д. Д. Статистическое моделирование работы информационно-поисковой системы. — НТИ, серия 2, 1974, № 10.
Садчикова П. В.* Опыт построения межъязыковой лексико-морфологической модели (англо-русский автоматический словарь по химии полимеров). АКД. Алма-Ата, 1975.
Сафонова И. Н. Дистрибутивно-статистический анализ английских микроконтекстов с опорным подлежащим и их эквиваленты в русском языке. — ДСАТ, 1973.
Севбо И. П. Структура связного текста и автоматизация реферирования. М., 1969.
Сидорченко В. Д. Типологическая классификация семантических структур тезаурусов. НТИ, серия 2, 1975, № 5.
Скитпевский Д. М. Система программирования лингвистических алгоритмов. — СтРААТ, 1974.
Скороходько Э. Ф. Моделирование языка в связи с задачами информационного поиска. Vol. IV. Recueil linguistique de Bratislava. Bratislava, 1973.
Скороходько Э. Ф. Определение значимости элементов текста на основе сетевой модели. — ВНТС ЛО, 1974.
Славина Л. А. Матрично-табличная переработка информации при машинном переводе словосочетаний. — СтРААТ, 1974.
Словарь дескрипторов по химии и химической промышленности. М., 1966.
Смолян Г. Л. Человек и ЭВМ. — ВФ, 1973, № 3.
Соркина В. А. Статистика именных словосочетаний в английских текстах по радиоэлектронике. — СТ I, 1969.
Соркина Д. А.* Дистрибутивно-статистическое выделение именной группы в английском языке. — АКД. Минск, 1972.
Степанов Г. В. О двух аспектах понятия языковой нормы. Методы сравнительно-сопоставительного изучения современных романских языков. М., 1966.
Степанов Ю. С. Семиотика. М., 1971.

Степанов Ю. С. Методы и принципы современной лингвистики. М., 1975.
Тарасова Е. С.* Распознавание смыслового образа английского научно-технического текста. — АКД. Кишинев, 1974.
Тезаурус научно-технических терминов. Под общей редакцией Ю. И. Шемакина. М., 1972.
Тихомиров О. К. Философско-психологические проблемы «искусственного интеллекта». — ВФ, 1975, № 1.
Толстой Л. Н. Детство. Отрочество. Юность. Л., 1966.
Трахтенброт Б. А. Алгоритмы и вычислительные автоматы. М., 1974.
Трибис Л. И.* Лексическая неоднозначность и ее автоматическое устранение в английском научно-техническом тексте. — АКД. Минск, 1972.
Трибис Л. И. Об одной модели распознавания лексических значений неоднозначных слов. — СтРААТ, 1973.
Тулдава Ю. А. О некоторых статистилических характеристиках текста. — В кн.: Материалы Всесоюзного симпозиума по вторичным моделирующим системам. 1 (5). Тарту, 1974.
Турыгина Л. А.* Статистическая интерпретация антиномии «язык и речь». — АКД. Л., 1970.
Угодчикова Н. Ф.* Автоматическое устранение регулярной лексической многозначности (на материале французских существительных подъязыка радиоэлектроники). — АКД. Горький, 1975.
Урбах В. Ю. К учету корреляции между буквами алфавита при вычислении количества информации в сообщении. —ПК, вып. 10, 1963.
Урсул А. Д. Информация. Методологические аспекты. М., 1971.
Фаддеев Д. К. К понятию энтропии конечной вероятностной схемы. — УМН, т. 11, № 1, 1956.
Фатовская М. А.* Статистическое описание испанской экономико-статистической литературы (алгоритмы автоматического морфологического анализа) — АКД. Л., 1974.
Филин Ф. П. Несколько слов о языковой норме и культуре речи. — В кн.: Вопросы культуры речи, вып. 7. М., 1966.
Харкевич А. А. О ценности информации. — ПК, вып. 4, 1960.
Целиковская И. П. Определение уровня языковых умений с помощью информационно-статистических методов. — В кн.: Проблемы прикладной лингвистики. Тезисы межвузовской конференции 16—19 декабря 1969 г. М., 1969.
Чапля А. И.* Опыт статистического описания лексической комбинаторики (на материале румынских публицистических текстов). — АКД. Л., 1967.
Чапля А. И., Чапля С. Г., Зубов А. В. Автоматический отбор ключевых и полиключевых слов. Сборник научных сообщений факультета иностранных языков. Дагестанский государственный университет им. В. И. Ленина. Махачкала, 1973.
Черный А. И. Общая методика построения тезаурусов. — НТИ, серия 2, 1968, № 5.
Чижаковский В. А.* Фразеология и машинный перевод (опыт составления и работы немецко-русского автоматического словаря оборотов для публицистических и научных текстов). — АКД. Кишинев, 1971.
Чуковский К. Высокое искусство. М., 1968.
Шабес В. Я.* Распознавание смысла с помощью семантико-синтаксического тезауруса (на материале предложно-падежной многозначности). — АКД. Л., 1973.
Шаранда А. Н.* Статистическое выделение типовых контекстов и автоматический перевод (на материале немецких научно-технических текстов). — АКД. Л., 1968.
Швейцер А. Д. Перевод и лингвистика. М., 1973.

Шелихов А. А., Селиванов Ю. П. Вычислительные машины. Справочник. М., 1973.
Шемакин Ю. И. Тезаурус в автоматизированных системах управления и обработки информации. М., 1974.
Шкловский В. Б. Искусство как прием. Пгр., 1919.
Шрейдер Ю. А. О семантических аспектах теории информации. — В кн.: Информация и кибернетика. М., 1967.
Шрейдер Ю. А. Информация и метаинформация. — НТИ, серия 2, 1974, № 4.
Штофф В. А. Моделирование и философия. М.—Л., 1966.
Эшби У. Р. Что такое разумная машина? — КОКН, 1968.
Яглом А. М., Яглом И. М. Вероятность и информация. Изд. 3-е, переработанное и дополненное. М., 1973.
Ярцева В. Н., Колшанский Г. В., Степанов Ю. С., Уфимцева А. А. Основные проблемы марксистского языкознания (доклад на Всесоюзной конференции по теоретическим вопросам языкознания). М., 1974.
Ястребова С. В., Дегтярева Л. И. Русская часть машинного словаря общеупотребительной лексики в системе англо-русского машинного перевода. — ЧВААТ, 1972.
Ястребова С. В., Мандрыка О. А., Моисеева С. А., Окулич Н. Э., Попова М. Ю., Рабинович Е. Е., Райтаровский В. В., Рояк Р. Л. Построение автоматических словарей общеупотребительной лексики. — ВНТС ЛО, 1974.

Abelson R. P., Carrol J. C. Computer Simulation of Individual Belief Systems. — American Behavioral Science. 9, May 1965.
Ackoff R. L. Towards a Behavioral Theory of Communication. — MS, vol. 4, 1958.
Ackoff R. L. Scientific Method. Optimizing Applied Research Decisions. N. Y., 1962.
Ackoff R. L., Emery F. E. On Purposeful Systems. Chicago—N. Y., 1972. (цит. по русск. переводу: Акофф Р., Эмери Ф. О целеустремленных системах. М., 1974).
Aitken A. J., Bailey R. W., Hamilton-Smith N. The Computer and Literary Studies. Edinburgh, 1973.
Allén S. Nusvensk frekvensordbok baserad på tidningstext. 2. Lemman. Stockholm, 1971.
Alouche F., Bely N., Cros R. C., Gardin J. C.,, Lévy F., Perridult J. Economie générale d'une chaine documentaire mécanisée. Paris, 1967.
Altmann G. Рецензия на: Пиотровский Р. Г. Информационные измерения языка. М., 1968. — L., 127, 1974.
Arbib M. A. Brains, Machines and Mathematics. N. Y.—S. Francisco—Toronto—London, 1964 (цит. по русск. переводу: Арбиб М. Мозг, машина и математика. М., 1968).
Ashby W. R. Design for a Brain. 2nd ed. London, 1960 (цит. по русск. переводу: Эшби У. Р. Конструкция мозга. Происхождение адаптивного поведения. М., 1964).
Attneave F. Psychological Probability as a Function of Experienced Frequency. — Journal of Experimental Psychology, XLVI, 2, 1953.
Bar-Hillel Y. Language and Information. Selected Essays on their Theory and Application. Jerusalem, 1964.
Bar-Hillel Y., Carnap R. Semantic Information. In: Introduction to Information Science. Compl. and ed. T. Saracevic. N. Y. 1970.
Beauclair W., de. Datenerfassung mit Seitenleser. — ZD., B. 11, H. 4, 1973.
Bellman R. E., Zadeh L. A. Decision-making in a Fuzzy Environment. — MS, vol. 17, 1970.

Benešová E., Sgall P. Remarks on the Topic Comment Articulation. Part I. — Prague Bulletin in Mathematical Linguistics, 1973, № 19.
Bense M. Theorie der Texte. Berlin, 1962.
Berztiss A. T. Data Structures. Theory and Practice. N. Y.—London, 1971 (цит. по русск. переводу: Берзтисс А. Т. Структуры данных. М., 1974).
Beth E. W. The Relationship between Formalised Languages and Natural Language. — Synthese, vol. 15, no 1, 1963.
Biology of Memory, eds.: Pribram K. H. and Broadbent D. E. N. Y., 1970.
Bobrov D. G. Syntactic Analysis of English by Computer. A Survey. — AFIPS Conference Proceedings, vol. 24, 1963.
Bogodist V., Guéorguiev Chr., Pestunova V., Piotrowski R., Raitar S. Une nouvelle méthode d'évaluation de l'information sémasiologique. — Linguistica, V, Tartu, 1974.
Boguslavskaja G., Koženec T., Piotrowski R. Informational Estimates of Text. — ZPhSK, B. 24, H, 1, 1971.
Bogusławskaja H., Korzeniec T., Piotrowski R. Language and Information. — Biuletyn fonograficzny, XIII, Poznań, 1972.
Booth A. D. A Law of Occurences for Words of Low Frequency. Introduction to Informational Science. N. Y.—London, 1970.
Borko H., Blankenship D. A. One-line Information Retrieval Using Associative Indexing. — In: Final Report System Development Corporation. Santa Monica (Calif), May 1968.
Brillouin L. Science and Information Theory. N. Y. 1956 (цит. по русск. переводу: Бриллюэн Л. Наука и теория информации. М., 1960).
Brinkmann K. H. Überlegungen zum Aufbau und Betrieb von Terminologie-Datenbanken als Voraussetzung der maschinenunterstüzten Übersetzung. — ND, XXV, 3, 1974.
Bühler K. Tatsachen und Probleme zu einer Psychologie der Denkvorgänge. — Archiv für die gesamte Psychologie, IX, 1907; XII, 1908.
Carroll J. B. Word-frequency Studies and the Lognormal Distribution. Proceedings of the Conference on Language and Language Behavior, ed.: Zale E. M. N. Y., 1968.
Carroll J. B. The American Heritage Frequency Word Book. N. Y.—Toronto, 1971.
Chafe N. L. Idiomaticity as an Anomaly in the Chomskyan Paradigm. — Foundations of Language, vol. 4, no 2, 1968.
Chang H. T. The Repetitive Discharges of Corticothlamic Reverberating Circuits. — Journal of the Neurophysiology, XIII, 1950.
Chang Sh. K. A Interactive System for Chinese Character Generation and Retrieval.—IEEE TSMC, vol. 3, no 3, 1973.
Cherry C. On Human Communication. A Review, a Survey, and a Criticism, 2-th ed., Cambridge Mass. and London, 1965 [цит. по русск. переводу: Черри К. Человек и информация (критика и обзор). М., 1972].
Chomsky N. Language and Mind. 1968 (цит. по русск. переводу: Хомский Н. Язык и мышление. Изд. МГУ, М., 1972).
Chomsky N., Miller G. A. Finitary Models of Language Users. — HMP, vol. 2, 1963 (цит. по русск. переводу: Хомский Н., Миллер Дж. Конечные модели использования языка. — КСб, новая серия, вып. 4, 1967).
Chomsky N., Miller G. A. Introduction to the Formal Analisis of Natural Languages. — HMP, vol. 2, 1963 (цит. по частичному русск. переводу: Хомский Н., Миллер Дж. Введение в формальный анализ естественных языков. — КСб, новая серия, вып. 1, 1965).
Church A. Introduction to Mathematical Logic, vol. I. Prinston, New Jersey, 1956 (цит. по русск. переводу: Черч А. Введение в математическую логику. М., 1960).
Climenson W. D., Hardwick N. H., Jacobson S. N. Automatic Syntax Analysis in Machine Indexing and Abstracting. — American Documentation, vol. 12, no 3, 1961.
Cooper W. The Storage Probleme. — McT, vol. 5, no 2, 1958.
Coseriu E. Determinación y Entorno. Dos problemas de una lingüística del hablar. — Romanistisches Jahrbuch, VII, 1955—1956.
Crosfield 7000-Schrift-Gelehrter. — BZB, 76, 5, 1973.
Danielsen N. Das Generative Abenteuer. — ZPhSK, B. 25, H. 4/5, 1972.
De Luca A., Termini S. A. Definition of Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. — IC, vol. 20, 1972.
Digby D. W. A Search Memory for Many-to-many Comparisons. — IEEE Transactions on Electronic Computers, XXII, 8, 1973.
Earl L. L. Experiments in Automatic Extracting and Indexing. — Information Storage and Retrieval, vol. 6, no 4, 1970.
Edmundson H. P. Mathematical and Computational Linguistics. Concepts of Communication: Interpersonal, Intrapersonal and Mathematical. N. Y.—London—Sydney—Toronto, 1971.
Endres W. A Comparision of Redundancy in Written and Spoken Language. — ISIT, 1973.
Feigenbaum E. A., Feldman J. (editors). Computers and Thought. A Collections of Articles. N. Y.—S. Francisco—Toronto—London, 1963 (цит. по русск. переводу: Вычислительные машины и мышление. М., 1967).
Feinstein A. Foundations of Information Theory. N. Y.—Toronto—London, 1958 (цит. по русск. переводу: Файнстейн А. Основы теории информации. М., 1960).
Findler N. Y. On a Computer Language Which Simulates Associative Memory and Parallel Processing. — C, X, 4, 1968.
Florès C. Mémoire. «Traité de psychologie expérimentale. IV. Apprentissage et *mémoire*», sous la direction de P. Fraisse et J. Piaget. Presses univ. de France, s. d. (цит. по русск. переводу: Экспериментальная психология. Редакторы-составители Фресс П. и Пиаже Ж. Вып. IV. М., 1973).
Foster J. M. List Processing (Ser. Computer Monographs). London, 1968 [цит. по русск. переводу: Фостер Дж. Обработка списков. (Серия «Математическое обеспечение ЭВМ»). М., 1974].
Frick F. C., Sumby W. H. Control Tower Language.—Journal of the Acoustical Society of America, XXIV, 6, 1952.
Fucks W. Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. — In: Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen. H. 34a. Köln und Opladen, 1955.
Garner W. R., Hake H. W. The Amount of Informationin Absolute Judgements. — Psychological Review, vol. 58, no 6, 1951.
Garvin P. Some Linguistic Aspects of Information Retrieval. Machine Indexing. N. Y., 1962.
Gellner E. Words and Things. A Critical Account of Linguistic Philosophy and a Study in Ideology. London, 1959 (цит. по русск. переводу: Гелднер Э. Слова и вещи. М., 1962).
George F. H. The Brain as a Computer. Oxford—London—N. Y.—Paris, 1961 (цит. по русск. переводу: Джордж Ф. Мозг как вычислительная машина. М., 1963).
Georgiev H., Piotrowski R. Meaning Information and Its Measures. — RRL, XIX, 2, 1974.
Germain C. Origine et évolution de la notion de «situation» de l'«Ecole linguistique de Londres»: de Malinowski à Lyons. — Linguistique, vol. 8, fasc. 2, 1972.

Germain C. B. Programming the IBM/360. Engelwood Clifts (New Jersey), 1967 (цит. по русск. переводу: Джермейн К. Программирование на IBM/360. М., 1971).
Ginsburg S. The Mathematical Theory of Context-free Languages. N. Y., 1966 (цит. по русск. переводу: Гинзбург С. Математическая теория контекстно-свободных языков. М., 1970).
Gödel K. Ueber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. I. — Monatshefte für Mathematik und Physik, XXXVIII, 1931.
Guiaşu S. Weighted Entropy. — Reports on Mathematics and Physics, vol. 2, no 3, 1971.
Hanley M. L. Word Index to James Joyce's Ulysses. Madison, 1965.
Hartley R. V. L. Transmission of Information. BSTJ, vol. 7, no 3, 1928 (цит. по русск. переводу: Хартли Р. Передача информации. В кн.: Теория информации и ее приложения. М., 1959).
Hartmanis J., Hopcroft J. E. What Makes Some Language Theory Problems Indecidable. — Journal of Computer and System Sciences, IV, 4, 1970 (цит. по русск. переводу: Хартманис Дж., Хопкрофт Дж. О причинах неразрешимости некоторых проблем теории языка. — КСб, 8, 1971).
Hartung W. Die gesellschaftliche Determiniertheit von Sprache und Kommunikation in der Sicht des Strukturalismus und der generativen Grammatik, I—II. — ZPhSK, B. 26, H. 3—4, 1973; III. — ZPhSK, B 27, H. 4, 1974.
Hayakawa S. I. Language in Thought and Action. 2-th ed. N. Y., 1964.
Heaps D. M., Ingram W. D. Computer Recognition and Graphical Reproduction of Patterns in Scientific and Technical Stile. — Proceedings of the American Society for Information Science, vol. 8, Westport Conn., 1971.
Hebb D. O. The Organization of Behavior. N. Y., 1949.
Herdan G. The Advanced Theory of Language as Choice and Chance. Berlin—Heidelberg—N. Y., 1966.
Higman B. A Comparative Study of Programming Languages. London—N. Y., 1969 (цит. по русск. переводу: Хигман Б. Сравнительное изучение языков программирования. М., 1974).
Hilpinen R. On the Information Provided by Observations. — In: Information and Inference. Dordrecht, 1970.
Hintikka J. The Varieties of Information and Scientific Explanation. — In: Logic, Methodology and Philosophy of Science, Vol. III. Amsterdam, 1968.
Hintikka J. On Semantic Information. — In: Information and Inference. Dordrecht, 1970.
Hjelmslev L. Prolegomena to a Theory of Language. Translated by Whitfield F. J. — Supplement to International Journal of American Linguistics, vol. 19, 1, 1953. Indiana University Publications in Antropology and Linguistics (цит. по русск. переводу: Ельмслев Л. Пролегомены к теории языка. — НЛ, вып. 1, М., 1960).
Hockett Ch. F. Grammar for the Hearers. — SLMA, 1961 (цит. по русск. переводу: Хоккетт Ч. Грамматика для слушающего. — НЛ, IV, 1965).
Holt E. B. Animal Drive and the Learning Processes. N. Y., 1931.
Horn E. A. A Basic Writing Vocabulary. 10 000 words Most Commonly Used in Writing. Iowa City, 1926.
Howes D. A Word Count of Spoken English. — JVLVB, vol. 5, no 6, 1966.
Ingarden R. S., Urbanik K. Information without Probability.— Colloquium mathematicum, vol. IX, fasc. 1, 1962.
Jacobson H. On Models of Reproduction. — American Scientist, vol. 46, no 3, 1958 (цит. по русск. переводу: Джекобсон Г. О моделях воспроизведения. — КСб. 7, 1963)

Juilland A., Roceric A. The Linguistic Concept of Word. The Hague—Paris, 1972.
Kaplan A. An Experimental Study of Ambiguity and Context. — McT, vol. 2, no 2, 1955.
Kappers C. U. A., Huber G. C., Crosby D. C. The Comparative Anatomy of the Nervous System of Vertebrates, including Man. N. Y., 1936.
Katzan H. Computer Organization and the System/370. N. Y.—Cincinnati—Toronto—London—Melbourne, 1971 (цит. по русск. переводу: Катцан Г. Вычислительные машины системы 370. М., 1974).
Kirschenmann P. Kybernetik, Information, Widerspiegelung. München—Salzburg, 1969.
Klaus G. Die Macht des Wortes. Ein erkenntnistheoretisch-pragmatisches Traktat. Berlin, 1965 (цит. по русск. переводу: Клаус Г. Сила слова. Гносеологический и прагматический анализ языка. М., 1967).
Kleene S. C. Introduction to Metamathematics. Amsterdam—Gröningen—N. Y.—Toronto, 1952 (цит. по русск. переводу: Клини С. К. Введение в метаматематику. М., 1957).
Klein Sh., Kuppin M. An Interactive Heuristic Program for Learning Transformational Grammars. — Computer Studies in the Humanities and Verbal Behaviour, vol. 1, no 3, 1970.
Körner E. Ferdinand de Saussure. Origin and Development of his Linguistic Thought in Western Studies of Language. Braunschweig, 1973.
Krämer O. Automat und Mensch.—Universitas, B. 13, H. 5, 1958.
Kučera H., Francis W. Computational Analysis of Present-day American English. Providence (Rhode Island), 1967.
Kuno S. The Predictive Analyzer and a Path Elimination Technique. Readings in Automatic Language Processing. Ed.: Hays D. G. N. Y., 1966.
Küpfmüller K. Die Entropie der deutschen Sprache. — Fermeldetechnische Zeitschrift, VII, 6. 1954.
Küpfmüller K. Informationsverarbeitung durch den Menschen.— Nachrichtentechnische Zeitschrift, XII, 2, 1959.
Laffitte J. Nous retournerons cueillir les jonquilles. Moscou, 1951.
Lakoff G. On Generative Semantics. Semantics, ed. by Steinberg D. D. and Jakobovits L. A. Cambridge University Press, 1971.
Lakoff G. Hedges: a Study in Meaning Criteria and the Logic of Fuzzy Concepts. Proceedings of the 8th Regional Meeting of the Chicago Linguistic Society. Chicago, 1972.
Langer S. K. Philosophy in a New Key. 2-nd ed. Harvard, 1958.
Lashley K. S. The Problem of Serial Order in Behavior. Cerebral Mechanisms in Behavior, ed.: L. Jeffres. N. Y.—London, 1951.
Lecomte J. Exploitation d'un corpus d'anglais scientifique écrit. 2. Traitement automatique des formes en -ING et -ED. — Publication Linguistique du Groupe de Traduction Automatique, no 14, Nancy, 1974.
Lenneberg E. H. Biological Foundation of Language. N. Y., 1967.
Levelt W. J. M. Formal Grammars in Linguistics and Psycholinguistics (vol. I). An Introduction to the Theory of Formal Languages and Automata. The Hague—Paris, 1974.
Lewis C. I. The Modes of Meaning. Semantics and the Philosophy of Language. Ed.: Linsky L. Urbana, 1952.
Linguistic and Engineering Studies in the Automatic Language Translation of Scientific Russian into English. Technical Report. Phase II. University of Washington Press. Seattle, 1960.
Lofgren L. Self-repair as the Limit for Automatic Error-correction. Proceeding Principles of Self-organisation. Transactions of the University of Illinois. Symposium on Self-organisation, eds.: Forster H. von and Zopf G. Oxford—London—N. Y. 1962 (цит. по русск. переводу: Лефф-

грен Л. Самовосстановление как предел для автоматической коррекции ошибок. — В кн.: Принципы самоорганизации. М., 1966).
Lorin H. A Guided Bibliography to Sorting. — IBM Systems Journal, vol. 10, no 3, 1971.
Mac Culloch W. S. In: Symposium. The Design of Machines to Simulate the Behavior of the Human Brain. — Transactions of IRE, EC-5, 4, 1956 (цит. по русск. переводу: Маккаллок У. С. «Мозг», вычислительное устройство с отрицательной обратной связью. — КСб, 1, М., 1960).
Machine Translation of Language, eds: Locke W. N., Booth A. D. N. Y. 1955 (цит. по русск. переводу: Машинный перевод, М., 1957).
Majernik V., Krútel' J. Prädiktionsentropie der slowakischen Schrift. — Kybernetika, R. 8, č. 4, 1972.
Mandelbrot B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse. — SLMA, 1961.
Marcus S. Gramatici şi automate finite. Bucureşti, 1964.
Marcus S. Algebraic Linguistics; Analytical Models. N. Y.—London, 1967 (цит. по русск. переводу: Маркус С. Теоретико-множественные модели языков. М., 1970).
Masterman M. The Thesaurus in Syntax and Semantics. — McT, vol. 4, No. 4, 1957 (цит. по русск. переводу: Мастерман М. Тезаурус в синтаксисе и семантике. — В кн.: Математическая лингвистика. М., 1964).
Masterman M. Mechanical Pidgin Translation. — MT, 1967.
Mazur M. Jakościowa teoria informacji. Warszawa, 1970 (цит. по русск. переводу: Мазур М. Качественная теория информации. М., 1974).
Meier H. Deutsche Sprachstatistik I/II. Hildesheim, 1964.
Miller G. A. The Magical Number Seven Plus or Minus Two. Some Limits on Our Capacity for Processing Information. — Psychological Review, 3, 1956 (цит. по русск. переводу: Миллер Дж. Магическое число семь, плюс или минус два. О некоторых пределах нашей способности перерабатывать информацию. — В кн.: Инженерная психология. М., 1964).
Miller G. A. The Organisation of Lexical Memory: Are Word Associations Sufficient? The Pathology of Memory, eds: Talland G. and Waugh N. N. Y., 1969.
Miller G. A., Friedman E. A. The Reconstruction of Mutilated English Texts.—IC, vol. 1, 1957.
Miller G. A., Galanter E., Pribram K. H. Plans and the Structure of Behavior. N. Y., 1960 (цит. по русск. переводу: Миллер Дж., Галантер Е., Прибрам К. Планы и структуры поведения. М., 1965).
Minsky M. Steps toward Artificial Intelligence. — CTh, 1963 (цит. по русск. переводу: Минский М. На пути к созданию искусственного разума. — ВММ, 1967).
Models of Human-memory, ed.: Norman D. A. N. Y., 1970.
Moles A. Théorie de l'information et perception esthétique. Paris, 1958 (цит. по русск. переводу: Моль А. Теория информации и эстетическое восприятие. М., 1966).
Moles A. Sociodynamique de la culture. Paris—La Haye, 1967 (цит. по русск. переводу: Моль А. Социодинамика культуры. М., 1973).
Morin Y. Ch. A Computer Tested Transformational Grammar of French. — L., 116, 1973.
Morris Ch. W. Foundations of the Theory of Signs. International Encyclopedia of Unified Science, vol. 1, no 2. Chicago, 1938.
Morris Ch. Signs, Language and Behavior. 2-nd ed. N. Y. 1955.
Morris Ch. W. Signification and Significance, a Study of the Relations of Sign and Values. Cambridge, Mass. 1964.

Nagel E., Newman J. R. Gödel's Proof. New York University Press, 1958.
Nauta D. The Meaning of Information. The Hague—Paris, 1972.
Neubert A. Elemente einer allgemeinen Theorie der Translation. Actes du X-ème Congrès international des Linguistes. Bucarest, II, 1970.
Neumann J., von. The Computer and the Brain. New Haven, 1958 (цит. по русск. переводу: Нейман Дж., фон. Вычислительная машина и мозг. — КСб, 1, 1960).
Newell A., Simon H. A. Heuristic Problem-solving: The Next Advance in Operation Research. — OR, vol. 8, no 1, 1958.
Nida E., Taber Ch. Theory and Practice of Translation. Leiden, 1969.
Nilsson N. J. Problem-solving Methods in Artificial Intelligence. N. Y., 1971 (цит. по русск. переводу: Нильсон Н. Искусственный интеллект. Методы поиска решений. М., 1973).
Norman D. A. Memory and Attention. An Introduction to Human Information Processing. N. Y., 1969.
Novak L. A., Piotrowski R. Experimentul de predicţie şi entropia limbii române. — SCL, XIX, 3, 1968 (см. также SL, с. 325—366).
Nündel S., Klimonow G., Starke I., Brand I. Automatische Sprachübersetzung Russisch-Deutsch. Berlin, 1969.
Olf F. K. Lesende Computer-Alternative der Datenerfassung. — ZD, b. 11, H. 4, 1973.
O'Malley M. H. What do Standard Transformational Grammars Produce? — A Computational Study. — International Journal of Man-Machine Studies, vol. 5, no 2, April 1973.
Paisley W. J. The Effects of Authorship, Topic Structure and Time of Composition on Letter Redundancy in English Texts. — IVLVB, vol. 5, no 1, 1966.
Penfield W., Roberts L. Speech and Brain-mechanisms. Princeton University Press, 1959 (цит. по русск. переводу: Пенфилд В., Робертс Л. Речь и мозговые механизмы. М., 1964).
Perception Mechanisms and Models. S. Francisco, 1972 (цит. по русск. переводу: Восприятие. Механизмы и модели. М., 1974).
Perschke S. Possibilities of Further Development of Automatic Language Translation. — Pensiero e linguaggio in operazioni. Milano, I, 1, 1970.
Perschke S., Fangmeyer H., Fassone G., Geoffrion G. Les travaux du CETIS dans le domaine de l'informatique documentaire. — Documentaliste, 1974, numéro spécial.
Petöfi J. S. Transformationsgrammatiken und eine kontextuelle Texttheorie. Frankfurt am Main, 1971.
Petrova N., Piotrowski R., Giraud R. L'entropie du français écrit. — BSLP, t. 59, fasc. 1, 1964.
Petrova N., Piotrowski R., Giraud R. Caractéristiques informationnelles du mot français. — BSLP, t. 65, fasc. 1, 1970.
Piotrowski R. Modèles structuraux du dialecte. — In: Communications et rapports du Premier Congrès International de Dialectologie générale (Louvain du 21 au 25 août, Bruxelles les 26 et 27 août 1960). Louvain, 1964.
Piotrowski R. Entropy and Redundancy in Four European Languages. — Statistical Methods in Linguistics, 5, Stockholm, 1969.
Piotrowski R. The Antinomies of Linguistics and Automatic Interpretation of the Text. — In: 3-rd International Congress of Applied Linguistics. Copenhagen, 1972.
Piotrowski R. El cuestionario fonológico en el estudio de los dialectos latinoamericanos. — In: Actas de la Primera Reunión Latinoamericana de Lingüística y Filología. Viña del Mar (Chile). Enero 1964. Bogotá, 1973.
Piotrowski R. Le vocabulaire de base des langues romanes. — In: XIV Congresso internazionale di linguistica e filologia romanza. Napoli, 1974.

Piotrowski R. G., Georgiev H. C. La traduction automatique en URSS. — RRL, XIX, 1, 1974.
Piotrowski R., Palibina I. Automatic Pattern Recognition Applied to Semantic Problems. Computational and Mathematical Linguistics. — In: Proceedings of the 1973 International Conference on Computational Linguistics. Pisa, 1974.
Pratt V. R. A Linguistics Oriented Programming Language. — MIT AI Memo no 277. Massachusetts Institute of Technology, Cambridge, Mass. February 1973.
Pribram K. H. Languages of the Brain. Englewood Cliffs, 1974 (цит. по русск. переводу: Прибрам К. Языки мозга. М., 1975).
Prütze M. Grundgedanken zu einer funktionalen Textlinguistik.—TL, I, 1970.
Ramsay A. Time, Space and Hierarchy in Zoosemiotics. Approaches to Animal Communication, eds: Sebeok Th. A. and Ramsay A. The Hague—Paris, 1966.
Ranson S. W., Clarc S. L. The Anatomy of the Nervous System, its Development and Function. 10th ed. Philadelphia—London, 1959.
Raphael B., Duda R., Fikes R. E., Hart P. E., Nilson N., Thorndyke P. W., Wilbur B. M. Research and Applications — Artificial Intelligence. Stanforod Research Institute. Menlo Park Calif., Final Report, Oct. 1971.
Rathkirch-Trach K. Methodische Textanalyse-Referierungtechnik im der automatischen Documentation. — ND, B. 17, N 5, 1966.
Reid I. R. French-frequency Distribution Curve. — Language, vol. 20. Language Monograph: The Psycho-biology of Language. Baltimore, 1944.
Reitman W. R. Cognition and Thought. An Information-Processing Approach. N. Y.—London—Sydney, 1965 (цит. по русск. переводу: Рейтман У. Р. Познание и мышление. Моделирование на уровне информационных процессов. М. 1968).
Rhodes I., Alt F. L. Hindsight Technique in Machine Translation of Natural Languages. — Journal of Research of National Bureau of Standards. B. Mathematics and Mathematical Physics, 66B, N 2, 1962 (цит. по русск. переводу: Родс И., Альт Ф. Л. Использование метода «возвратов» при автоматическом переводе. — АП, 1971).
Rinsland H. D. A Basic Writing Vocabulary of Elementary School Children. N. Y., 1945.
Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, 1962 (цит. по русск. переводу: Розенблатт Ф. Принципы нейродинамики. М., 1965).
Russell S. W. Semantic Categories of Nominals for Conceptual Dependency Analysis of Natural Language. — Stanford Artificial Intelligence Project Memory, N 172, 1972.
Salton G. Automatic Information Organisation and Retrieval. N. Y.—St. Louis etc. 1968 (цит. по русск. переводу: Сэлтон Г. Автоматическая обработка, хранение и поиск информации. М., 1973).
Santos E. S. Fuzzy Algorithms. — IC, vol. 17, no 4, 1970.
Saterland N. S. Machines Like Men. Science Journal, oct. 1968 (Special issue) (цит. по русск. переводу: Сатерленд Н. С. Человекоподобные машины. Сб.: Человеческие способности машин. М., 1971).
Saussure F., de. Cours de linguistique générale. 2-ème ed. Paris, 1922 (цит. по русск. переводу: Соссюр Ф., де. Курс общей лингвистики. М., 1933).
Schaff A. Wstęp do semantyki. Warszawa, 1960 (цит. по русск. переводу: Шафф А. Введение в семантику. М., 1963).
Schank R. C. Conceptual Dependency: A Theory of Natural Language Understanding. — Cognitive Psychology, vol. 3, no 4, 1972a.
Schank R. «Semantics» in Conceptual Analysis. — Lingua, 30, N 3—4, 1972b.

Schonell F. J., Middelton I. C., Shaw B. A. A Study of the Oral Vocabulary of Adults. Brisbane, 1956.
Schuhmacher W. W. Cybernetic Aspects of Language. The Hague, 1973.
Schveiger P. The Functionning of the Human Brain and of the Computer as Language Generating Devices. — RRL, XVII, 6, 1972.
Scriven M. The Complete Robot: A Guide to Androidology. Dimensions of Mind. N. Y., 1961.
Sgall P. K programu lingvistiky textu. — Slovo a Slovesnost, XXXIV, 1, 1973.
Shank R. — см. Schank R.
Shannon C. A Mathematical Theory of Communication. — BSTJ, vol. 27, no 3—4, 1948 (цит. по русск. переводу: Шеннон К. Математическая теория связи. РТИК, 1963).
Shannon C. Prediction and Entropy of Printed English. — BSTJ, vol. 30, no 1, 1951 (цит. по русск. переводу: Шеннон К. Предсказание и энтропия печатного английского текста. — РТИК, 1963).
Siebert W. M. Stimulus Transformations in the Peripheral Auditory System. — RP, 1968 (цит. по русск. переводу: Сиберт У. Преобразование стимула в периферической слуховой системе. — РО, 1970).
Simmons R. F. Integrated Computer Systems for Language. — RTCL, 1972.
Slagle J. R. Artificial Intelligence. The Heuristic Programming Approach. N. Y.—St. Louis—S. Francisco—Düsseldorf, 1971 (цит. по русск. переводу: Слэйгл Дж. Искусственный интеллект. Подход на основе эвристического программирования. М., 1973).
Sommers H. H. The Measurement of Grammatical Constraints. — Language and Speech, vol. 4, p. 3, 1961.
Stably D. Logical Programming with System/360. N. Y., 1970 (цит. по русск. переводу: Стэблн Д. Логическое программирование в системе/360. М., 1974).
Steinbuch K. Automat und Mensch. Kybernetische Tatsachen und Hypothesen. 3 Aufl. Berlin—Heidelberg—N. Y., 1965 (цит. по русск. переводу: Штейнбух К. Автомат и человек. Кибернетические факты и гипотезы. М., 1967).
Suppes P. Probabilistic Grammar for Natural Languages. — Synthese, 22, 1970.
Tarski A. O pojęciu prawdy w odniesieniu do sformalizowanych nauk. dedukcyjnych. — Ruch filozoficzny, LII, 1, 1930.
Taube M. Computers and Common Sense. The Myth of Thinking Machines. N. Y.—London, 1961 (цит. по русск. переводу: Таубе М. Вычислительные машины и здравый смысл. Миф о думающих машинах. М., 1964).
Thomason M. G., Marinos P. N. Deterministic Acceptors of Regular Fuzzy Languages. — IEEE TSMC, vol. 4, no 2, 1974.
Thorndike E. L., Lorge I. The Teacher's Word Book of 30 000 Words. N. Y. 1944.
Thorpe W. H. Learning and Instinct in Animals. London, 1956.
Thurber K. J., Patton P. C. The Future of Parallel Processing. — IEEE TEC, vol. 22, No. 12, 1973.
Tillich P. The Dynamics of Faith. N. Y., 1957.
Tinbergen N. The Study of Instinct, 4-th ed. Oxford, 1958.
Turing A. M. Can the Machine Think? — In: The World of Mathematics, vol. 4, N. Y., 1956 (цит. по русск. переводу: Тьюринг А. Может ли машина мыслить? М., 1960).
Ty Pak. Contradictions in Chomskian Semantics. — Studia Linguistica. Revue de linguistique générale et comparée, XXVIII, 1, 1974.
Uexkuell J., von, Kriszat G. Streifzüge durch die Umwelten von Tieren und Menschen. 2-te Aufl. Hamburg, 1956.

Vauquois B., Veillon G., Veyrunes J. Syntax and Interpretation. — McT, vol 9, no 2, 1966.
Vitek A. J. Russian-English Idiom Dictionary. Ed. by Josselson H. H. Wayne State University Press. Detroit, 1973.
Wald A. Sequential Analysis. N. Y., 1947 (цит. по русск. переводу: Вальд А. Последовательный анализ. М., 1960).
Walker D. E. Social Implications of Automatic Language Processing. — RTCL, 1972.
Walker D. E. Automated Language Processing — Stanford Research Institute. The Artificial Intelligence Center Technical Notes, 1973, N 77.
Weizenbaum J. Contextual Understanding by Computers. — RP, 1968 (цит. по русск. переводу: Вейценбаум Дж. Понимание связного текста вычислительной машиной. — РО, 1970).
Wells R. A Measure of Subjective Information. — SLMA, 1961 (цит. по русск. переводу: Уэлз Р. Мера субъективной информации. — НЛ, вып. IV, 1965).
Wenzel F. SPLIT. Ein Verfahren zu maschinellen morphologischen Segmentierung russischer Wörter. München, 1973.
Wickelgren W. A. Multitrace Strength Theory. Models of Human Memory, ed.: Norman D. A., N. Y., 1971.
Wilkes M. V. Time-sharing Computer Systems. London—N. Y., 1969 (цит. по русск. переводу: Уилкс М. Системы с разделением времени. М., 1972).
Wilks Y. A. Grammar, Meaning and the Machine Analysis of Language. London, 1972.
Winograd S., Cowan J. D. Reliable Computation in the Presence of Noise. Cambridge, Mass., 1963 (цит. по русск. переводу: Виноград С., Кован Дж. Д. Надежные вычисления при наличии шумов. М., 1968).
Wittgenstein L. Tractatus Logico-Philosophicus. London, 1933. (цит. по русск. переводу: Витгенштейн Л. Логико-философский трактат. М., 1958).
Wittmers E. Über das Modell einer allgemeinen Grundstruktur des Textes als Mittel zur Erfassung spezifischer Textgesetzmäsigkeiten. — TL, I, 1970.
Wójcik T. Prakseosemiotika. Zarys teorii optimalnego znaku. Warszawa, 1969.
Wooldridge D. E. The Machinery of the Brain. N. Y., 1963 (цит. по русск. переводу: Вулдридж Д. Механизмы мозга. М., 1965).
Yngve V. N. A Programming Language for Mechanical Translation. — McT. vol. 5, No. 1, 1958 (цит. по русск. переводу: Ингве В. Х. Язык для программирования задач машинного перевода. — КСб, 6, М., 1963).
Zadeh L. A. Quantitative Fuzzy Semantics. — Information Sciences, vol. 3, 1971.
Zadeh L. A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. — IEEE TSMC, vol. SMC—3. Jan. 1973 (цит. по русск. переводу: Заде Л. А. Основы нового подхода к анализу сложных систем и процессов принятия решений. — Математика сегодня, 7, 1974).
Zipf G. K. Human Behavior and the Principle of Least Effort. Cambridge, Mass., 1949.

ОГЛАВЛЕНИЕ

Раймонд Генрихович Пиотровский

ТЕКСТ, МАШИНА, ЧЕЛОВЕК

Утверждено к печати
Институтом языкознания
АН СССР

Редактор издательства *К. Н. Феноменов*
Художник *М. И. Разумевич*
Технический редактор *Г. А. Смирнова*
Корректоры *З. В. Гришина, Л. Б. Жукоборская* и *Н. П. Килим*

Сдано в набор 23/IX 1975 г. Подписано к печати 9/XII 1975 г. Формат 60×90¹/₁₆. Бумага № 2. Печ. л. 20½+1 вкл. (¼ печ. л.)= 20.75 усл. печ. л. Уч.-изд. л. 24.09. Изд. № 5911. Тип. зак. № 624. М-26914. Тираж 8300. *Цена 1 р. 45 к.*

Ленинградское отделение издательства «Наука»
199164, Ленинград, В-164, Менделеевская линия, д. 1

1-я тип. издательства «Наука»
199034, Ленинград, В-34, 9 линия, д. 12

ИСПРАВЛЕНИЯ И ОПЕЧАТКИ

Страница	*Строка*	*Напечатано*	*Должно быть*
78	6—8 снизу	развертывание начальной триады: Путем	развертывание начальной триады путем
167	13 сверху	Начало слов	Начала слов
186	6 снизу	оценки C	оценки B
231	5 сверху	$M_3 = b + kN$	$M_2 = b + kN$
248	13 »	Если не рассматривать	Если же рассматривать
Табл. 40	столбцы 27—28	приходящееся на словоформу	приходящееся на букву

Р. Г. Пиотровский

1 р. 45 к.

Издательство
«НАУКА»
Ленинградское
отделение